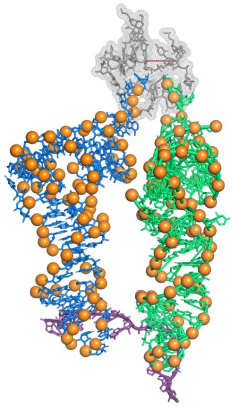
LIÈGE université
Sciences Appliquées

AEROSPACE & MECHANICS
DOCTORAL SCHOOL
Faculty of Engineering Science



Protein synthesis by ribosomes

Agent-based modeling of mRNA translation rates incorporating tRNA modification effects

Marc Joiret

Supervisors:
Prof. dr. ir. Liesbet Geris
(supervisor, Liège University)
Prof. dr. Pierre Close
(co-supervisor, Liège University)

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (PhD) in Engineering Science: Aerospace and Mechanics

April 2025

Protein synthesis by ribosomes

Agent-based modeling of mRNA translation rates incorporating tRNA modification effects

Marc JOIRET

Examination committee:
Prof. dr. ir. Davide Ruffoni, chair
(Liège University)
Prof. dr. ir. Liesbet Geris
(supervisor, Liège University)
Prof. dr. Pierre Close
(co-supervisor, Liège University)
Prof. dr. Andre Matagne (Liège University)
Prof. dr. Eveline Lescrinier
(KU-Leuven University)
Prof. dr. ir. Dick de Ridder
(Wageningen University)
Prof. dr. ir. Gerben Menschaert
(Ghent University)

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (PhD) in Engineering Science: Aerospace and Mechanics

April 2025

© 2025 ULiege – Faculty of Engineering Science
Edité au nom de Marc Joiret, Avenue de l'Hôpital, 1 (B34)+5CHU, B-4000 Liège (Belgium)

Tous droits réservés. Aucune reproduction de cet ouvrage, même partielle, quelque soit le procédé, impression, photocopie, microfilm, électronique ou autre, n'est autorisée sans la permission écrite de l'éditeur.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

Finding the need

The central research question of this PhD dissertation arose from discussions with leading cancer researchers in 2018, highlighting the challenge of reconciling transcriptomics (RNA-seq), proteomics (mass spectrometry), and ribosome profiling (Ribo-seq) data from samples collected under varying experimental conditions. A consistent framework for integrating these diverse omics datasets was lacking, underscoring the need for a broadly applicable computational model.

Despite more than six decades of research into the mechanistic details of protein synthesis, the interplay between key regulatory factors remains poorly understood. This thesis aims to address this gap by developing an *in silico* integrative model, implemented as two open-source computational tools freely available to the research community. One tool, designed for didactic purposes, includes visualization features to aid in training newcomers, while the other is optimized for deployment on high-performance computing platforms to support research applications.

At the intersection of multiple disciplines, this PhD research advances the development of a modeling and simulation framework for protein synthesis. By integrating concepts from biophysics, biochemistry, enzymology, chemical kinetics, molecular biology, X-ray crystallography, structural biology, synthetic biology, statistical physics, statistics, bioinformatics, and computational modeling, this work enhances our ability to analyze and interpret the complexities of ribosome function.

Abstract

Translational regulation through synonymous codon usage has recently been shown to play an important role in health and disease. Modified tRNAs are important actors involved in regulating protein expression levels by optimizing the decoding of differentially used codons. Nevertheless their contributions in protein synthesis dynamics remains unclear. Agent-Based Models (ABMs), particularly the Totally Asymmetric Simple Exclusion Process (TASEP), have been employed to quantify ribosomal translation rates. Building on this foundation, our work extends ABM and TASEP frameworks to incorporate the effects of tRNA modifications and additional factors impacting translation. We developed a computational stochastic model to quantify protein synthesis rates, utilizing ribosome residence times at each codon, based on transcript codon sequences and relative transcript abundances. The model integrates codon-specific rates of tRNA accommodation (both modified and unmodified), peptide bond formation, and translocation.

Key features include modulation of elongation rates by charged amino acids within the ribosomal exit tunnel, proline transpeptidation at the peptidyl transferase center, and translational slowing due to mRNA secondary structures. Our model enables the comparison of relative protein expression levels, polysome profiling, and Ribo-Seq data across various scenarios, accounting for ribosome pool availability and initiation rates.

Ultimately, this model aims to elucidate how codon usage and tRNA modifications interact dynamically to influence protein synthesis. By systematically disentangling the contributing factors, we seek to evaluate their sensitivity not only on protein output but also on ribosome density and polysome profiles.

Résumé sommaire

La régulation traductionnelle reposant sur l'usage de codons synonymes joue un rôle important en biologie moléculaire et notamment dans la compréhension des mécanismes de résistance aux therapies ciblées contre les cancers. Les ARN de transfert (ARNt) modifiés sont des acteurs essentiels impliqués dans la régulation des niveaux d'expression protéique en optimisant le décodage des codons utilisés de manière différentielle. Néanmoins, leur contribution à la dynamique de la synthèse protéique reste encore mal comprise.

Les modèles multi-agents (Agent-Based Models, ABMs), en particulier le processus d'exclusion totalement asymétrique (Totally Asymmetric Simple Exclusion Process, TASEP), ont été employés pour quantifier les taux de traduction par les ribosomes. En s'appuyant sur ces fondations, notre travail étend les modèles ABM et TASEP afin d'intégrer les effets des modifications des ARNt ainsi que d'autres facteurs influençant la traduction. Nous avons développé un modèle stochastique computationnel permettant de quantifier les taux de synthèse protéique, en utilisant les temps de résidence des ribosomes sur chaque codon, basés sur les séquences des transcrits et leur abondance relative. Le modèle intègre, par codon et pour chaque ribosome individuel, les taux spécifiques d'accommodation des ARNt (modifiés ou non), la formation de la liaison peptidique, et la translocation.

Les caractéristiques clés du modèle incluent la modulation des vitesses d'élongation par les acides aminés chargés dans le tunnel de sortie du ribosome, la transpeptidation de la proline au centre de la peptidyl transférase, ainsi que le ralentissement de la traduction dû aux structures secondaires de l'ARNm. Notre modèle permet la comparaison des niveaux relatifs d'expression protéique, des profils de polysomes, et des données de type Ribo-Seq dans divers scénarios, en tenant compte de la disponibilité du pool ribosomal et des taux d'initiation.

En définitive, ce modèle vise à élucider comment l'utilisation des codons et les modifications des ARNt interagissent dynamiquement pour influencer la synthèse protéique. En démêlant systématiquement les facteurs contributifs, nous cherchons à évaluer leur sensibilité non seulement sur la production protéique, mais également sur la densité ribosomale et les profils de polysomes.

List of Abbreviations and Glossary

A76 Adenosine 76 at the 3'end of a tRNA molecule. The last 3 nucleotides are CCA at the 3'end of all tRNA molecules. The hydroxyl group at C2' or C3' on the ribose of A76 is esterified to the carboxyl group of the cognate amino acid to form the amino acylated tRNA. (page 287)

ABM Agent Based Model. (page 72)

ADAT2 A tRNA modifying enzyme operating in the nucleus of eukaryotic cell, acronym for Adenosine Deaminase Acting on tRNAs, which deaminates adenosine A34 of tRNA molecules, and converts it to inosine I34, expanding the number of synonymous codons that can pair with the modified tRNA anticodon to three: inosine (I34) can base pair with U, A and C but not G. (page 172)

ASL Anticodon stem loop, usually the nucleotides located at and around positions 34, 35, 36 of a tRNA molecule. By contrast, the acceptor stem loop is located at position 74, 75 and 76 and has always the same sequence `CCA`, at the extremity of which the cognate amino acid is acylated to the ribose of adenosine 76. (page 165), (page 302)

Autophagy and autophagosomes Autophagy ("self-eating") is a cytoprotective mechanism found in all eukaryotes. It enables the cell to get rid of superfluous material such as defective organelles and macromolecular complexes that when accumulating would have a damaging effect. In its most frequent form, called macroautophagy, the material to be removed and recycled gets packed into double-membrane vesicles called autophagosomes which are transported to lysosomes where their cargo is broken up. (page 36), (page 319)

BR A Brownian ratchet (BR) is a biophysical mechanism that drives the unidirectional movement of a molecular-scale object, even though thermal energy induces

random motion in both directions. This directionality arises from a coupled chemical reaction that selectively prevents backward movement. In the Brownian ratchet, the system moves back-and-forth spontaneously, driven by thermal energy, along the 'mechanical coordinate', in the Gibbs free energy landscape. However, a chemical transition, that only occurs when the system reaches the forward state, prevents reversal and converts otherwise random motion into directed movement. The Brownian ratchet is typically contrasted with the Power Stroke. (page 300, 301)

CA A Cellular Automaton is a discrete model of computation. It consists of a regular grid of cells, each one in a finite number of states. The grid can be of any finite number of dimensions. For each cell, a set of cells called its neighborhood is defined relative to the specified cell. From an initial state, a new generation of states for the cells is created according to some rule (generally a mathematical function) that determines the new state of each cell in terms of the current state of the cell and the states of the cells in its neighborhood. (page 73)

CDS Coding DNA Sequence. It refers to the portion of a gene's DNA or RNA that is translated into a protein. The CDS excludes introns and untranslated regions (UTRs), focusing only on the exonic sequences that directly code for amino acids. (page 313), (page 337)

Cognate-tRNA Aminoacyl tRNA (aa-tRNA) that complies with the standard Watson-Crick rule of base pairing for the first two bases in a codon and can form either canonical or non-Watson-Crick pairs at the third base or 'wobble' position. (page 24), (page 165)

CRE Cis-Regulatory Element, a sequence element that only regulates the same mRNA where it is located. (page 13)

Cryo-EM Cryo-electron microscopy. (page 286)

Cycloheximide Cycloheximide and emitine are antibiotics acting as translation elongation inhibitors in eukaryotic cells (chloramphenicol for prokaryotes) that 'freezes' ribosomes anywhere on the mRNA, thereby preventing ribosome run off. Cycloheximide occupies the E-site of the eukaryotic 60S subunit and prevents translocation of deacylated-tRNA from the P-site. Cycloheximide tends to freeze the ribosome more frequently near the start codon. An alternative to prevent cycloheximide positional bias is simply by thermal freezing. (page 34), (page 39)

Dhp Deoxyhypusine. (page 286)

DHPS Deoxyhypusine synthase. (page 286)

DOHH Deoxyhypusine hydroxylase. (page 286)

EF-P Elongation factor P facilitating peptide bond formation of poor substrates at the peptidyl transferase center in prokaryotes. (page 283)

eIF5A Elongation factor eiF5A facilitating peptide bond formation of poor substrates at the peptidyl transferase center in eukaryotes and archea. (page 283)

Emergent behavior A hallmark of a complex system made of a collection of constituents where a systemic dynamics is generated that could not have been reasonably inferred from examination of the properties and functions of the individual constituents alone. (page 75)

EpmA EpmA is also known as `YjeA`, `PoxA` and `GenX`, catalyzes the attachment of (R)-β-lysine to the ϵ-amino group of lysine 34 of the translation elongation factor P (`EF-P`) in prokaryotes. (page 283)

EpmB EpmB is also known as `YjeK`, enzyme lysine 2,3-aminomutase converting (S)-α-lysine to (R)-β-lysine in prokaryotes. (page 283)

EpmC EpmC is also known as `YfcM`) recognizes translation elongation factor `EF-P` only in its modified form and hydroxylates the C5(δ) position of lysine 34 in prokaryotes. (page 283)

Flocking Flocking or murmuration of starling birds. Starlings are small passerine birds in the family of Sturnidae, genus *Sturnus*. Most species associate in flocks of varying size. Murmuration describes the flocking of starlings, including swarm behavior of their large flight formation. (page 71)

Harringtonine and lactimidomycin Harringtonine is a natural alkaloid from the coniferous species *Cephalotaxus harringtonia* acting as an antibiotic and used in ribosome profiling protocols to block translation initiation. Harringtonine only binds to vacant 60S subunit or 80S ribosome that have just assembled from the subunits and started translation. It forms an 80S ribosomal complex with the intiator tRNA but blocks aminoacyl-tRNA binding in the A-site and peptide bond formation. Harringtonine stops elongation immediately after the start codon. Lactimidomycin is an antibiotic produced by the bacteria *Streptomyces amphibiosphorus*. Lactimidomycin binds to 80S ribosomal subunits after its assembly on start codons. Like cycloheximide, lactimidomycin binds to the E-site and prevents translocation of a deacylated-tRNA from the P-site. Lactimidomycin cannot bind to actively translating ribosomes whereas cycloheximide can. Both harringtonine and lactimidomycin specifically block translocation at the start codon by occupying the E-site (exit site) of the ribosome when methionyl-tRNA is at the P-site (AUG start codon) and any aminoacyl-tRNA is at the A-site. At the

same time, previously initiated ribosomes continue translation, which results in only one 80S monosome remaining on the mRNA at the beginning of the coding region. This makes harringtonine and lactimidomycin useful tools in ribosome profiling (Ribo-seq) for mapping start codons on a genome-wide scale. (page 39)

Hpu Hypusine. (page 286)

Inductive bias Assumptions and previous knowledge enforced by a priori construction in a model. Inductive bias is appropriate when a model is to be tailored for solving a particular task more efficiently. It can also be detrimental if it limits the scope of the potential outcome or if it (unintentionally) constrains the exploration space of the model prediction. (page 66)

Isoacceptor Member of a family of tRNAs who differ in the anticodon sequence but share the same amino acid at the acceptor stem loop. Due to the degeneracy of the genetic code, the multiple tRNA isoacceptors share the same amino acid but have different anticodons that will pair with the mRNA synonymous codons of this amino acid. (page 53), (page 164), (page 166)

Isodecoder Member of a family of tRNAs who share the exact same anticodon sequence located at nucleotide 34, 35, 36 but differ elsewhere in their 76 nucleotides body sequence. The isodecoders also necessarily share the same amino acid at the acceptor stem. (page 164)

Kozak sequence Sequence of nucleotides flanking the start codon of an ORF, usually from -6 to +4, given that the first nucleotide of ORF is +1 (GCCRCCAUGG, where R can be A or G); named after Marylin Kozak who discovered the existence of these sequences in Eukarya. Kozak and Shine-Dalgarno sequences are distinct sequences found in eukaryotic and prokaryotic organisms respectively and play different roles in translation initiation. The Shine-Dalgarno sequence, 5'-AGGAGG-3' (or similar) usually -5 or -10 from ORF, base-pairs with a complementary sequence of 16S-rRNA of the small ribosomal subunit helping the ribosome to position correctly at the start codon. In Eukarya, the ribosome begins scanning from the 5' cap of the mRNA, and the Kozak sequence provides a context that enhances the recognition of the AUG start codon. (page 13)

LSU Large Subunit of the ribosome. (page 10), (page 163), (page 286)

mcm5s2-tRNA 2-thio, 5-methoxycarbonyl-methyluridine34 -tRNA: an enzymatically modified tRNA whose elongation cycle kinetics is changed and targetting specific cognate codons. (page 166)

mTOR Stands for mammalian (or mechanistic) Target Of Rapamycin. It is a signaling pathway integrating both intracellular and extracellular signals and serving as a central regulator of cell metabolism, growth, proliferation and survival. The mTOR protein is a 289-kDa serine-threonine kinase that belongs to the phosphoinositide 3-kinase (PI3K)-related family which is conserved throughout evolution. mTOR nucleates at least two distinct multi-protein complexes, mTOR complex 1 (mTORC1) and mTOR complex 2 (mTORC2). mTOR plays a key role in ribosomal protein synthesis and ribosome biogenesis. (page 36), (page 318)

NAD^+ Nicotinamide adenine dinucleotide, oxydized form of cofactor (electron acceptor). (page 286)

Near- and non- cognate tRNA Aminoacyl tRNA (aa-tRNA) for which there is at least one mismatch in the base pairing of the anticodon to the codon, at either of the three positions and leading to an amino acid misincorporation in the protein sequence although the mRNA sequence is correct. (page 24), (page 165)

ODE Ordinary Differential Equation. (page 71)

OOP Object-oriented programming. (page 78)

ORF Open reading frame, a sequence of triplets encoding amino acids. An ORF begins with a start codon and is bounded by a stop codon. Most of ORFs are sequences counting an integer multiple of three nucleotides under the assumption that there is no frameshifting. (page 13), (page 24), (page 38)

PB A petabyte (PB) is a unit of digital information storage corresponding to 10^{15} (B), or 1 million gigabytes (GB). In practical terms, 1 PB is roughly equivalent of $1,000,000$ hours of video or 100 million photos. In RNA-seq, a typical 10-30 million paired-end reads in FASTQ format could result in a file size around 5-10 GB. Similarly a Ribo-Seq raw dataset with 20 million reads might produce a compressed FASTQ file of 5-10 GB. You may expect that one petabyte corresponds to the storage capacity of roughly $50,000$ to $100,000$ Ribo-seq samples from $25,000$ to $50,000$ paired case-control experiments. (page 402)

PDE Partial Differential Equation. (page 71)

PS A Power Stroke (PS) is a biophysical mechanism that drives a mechanical movement, occuring most of the time as a burst, along a 'mechanical axis' that is coupled to a downhill chemical reaction progressing along a 'chemical axis'. In a Power Stroke the system moves diagonally in the Gibbs free energy landscape along both axis simultaneously. The Power Stroke mechanism is typically contrasted with the Brownian ratchet. (page 302)

PTC Peptidyl Transferase Center, the catalytic shell pocket where the peptide bond formation is catalyzed in the large subunit of the ribosome. (page 281), (page 283)

Rapamycin Rapamycin is an antibiotic from the bacteria *Streptomyces hygroscopicus* which, in eukarya cells, interacts with a peptidylprolyl cis-trans isomerase (immunophilin FKBP12), and forms a complex that is a highly specific inhibitor of the protein kinase mTOR. Rapamycin and its derivatives (rapalogues) are inhibitors of mTOR. They are drugs approved for use in cancer therapies. (page 317)

Ribo-seq Ribo-seq, or ribosome profiling or ribosome footprinting, is an experimental technique delivering a snapshot of ribosome positions along all transcripts in a bulk of cells (typically between 5 and 15 million cells) at a given condition. Its first version was developed at the end of the 1960s to study translation initiation. It has been extended in the 1980s to investigate the role of slow codons and ribosome pausing. In 2009, Ingolia *et al.* revamped this technique to exploit the next generation sequencing, making Ribo-seq the state-of-the art technique for studying gene expression at the level of translation. Ribo-seq is based on the principle that a translating ribosome protects a short stretch of mRNA within its structure. Once ribosomes are 'frozen' in the act of translation using translation elongation inhibitors, RNA-digesting enzymes known as RNases can be added to destroy any mRNA that is unprotected by the arrested ribosomes. After RNase digestion, ribosomes are enriched and the ribosome-protected mRNA is then isolated and converted into NGS(Illumina)-compatible cDNA libraries by reverse transcriptase. These ribosome-protected mRNA fragments are commonly called RPFs or ribosome footprints. Mapping these sequenced RPFs to the transcriptome provides a 'snapshot' of translation that reveals the position and densities of ribosomes on individual mRNAs transcriptome-wide. This snapshot can help determine which proteins were being synthesized in the cell at the time of the experiment. Ribo-seq enables the identification of alternative mRNA translation start sites, the confirmation of annotated ORFs, as well as upstream ORFs (uORFs) that may be involved in the regulation of translation, the distribution of ribosomes on an mRNA and the rate at which ribosomes decode codons. (page 38)

Ribophagy Ribophagy is a selective autophagy process specifically targeting ribosome turnover, well characterized in yeast and mammals. Triggered by nutriment deprivation, ribophagy is the process by which mature ribosomes are delivered to the vacuole or lysosome in an autophagy-dependent fashion, followed by rapid degradation by vacuolar enzymes. In yeast, ribophagy turns over 60S ribosomal proteins. (page 320)

Ribosome pool ratio Ribosome pool ratio. The ribosome pool ratio is the ribosome-to-transcript ratio, i.e., the ratio of the number of ribosomes to the number

of transcript copies. In our ABM model of protein translation, it is a tunable parameter that is fixed by the user before a simulation is started. (page 326)

Second genetic code A hypothetical theory recognizing that additional information is encrypted in the mRNA sequences where the synonymous codons are not actually synonymous as they would actually impose different elongation rates for co-translational folding purposes and/or for enabling tRNA modifying enzymes to reprogram the elongation rate to favor the translation of those transcripts for which the codon usage is enriched in the specifically targeted synonymous codons. (page 172)

Secondary structure mRNA secondary structures are two-dimensional, non-covalent interactions within an mRNA molecule (or any RNA molecule that is otherwise expected to be single-stranded), consisting of contiguous base pairs and loops. RNA molecules contain both canonical Watson-Crick base pairs and many non-canonical base pairs. (page 15), (page 18), (page 26), (page 44), (page 293)

SSU Small Subunit of the ribosome. (page 10)

TASEP Totally Asymmetric Simple Exclusion Process refers to a statistical physics inspired diffusion process modeled by a lattice of discrete sites in one dimension. In this process, autonomous particles can move by unit steps from one site to the next neighboring site, in a single direction (asymmetric), if and only if, the next site is free of other particles (simple exclusion). The model implements rules of intiation, elongation and termination that are stochastic. A free particle that is not on the lattice may jump on the first site of the lattice following an exponential probability density distribution parametrized by an initiation rate. A particle already engaged in the lattice may jump from one site to the next neighboring one at an elongation rate defined by another probability density distribution and may leave the last site of the lattice by a third probability density distribution parametrized with a termination rate. In homogeneous TASEP, all the elongation probability density distributions are equal across the inner sites of the lattice. In heterogeneous TASEP, each site may have its own elongation probability density distribution. In extended TASEP, the free particles can engage on array-like (multiple) 1D-lattices. The extended heterogeneous TASEP model is used to model the competitive translation of multiple mRNAs by a pool of ribosomes. (page 79)

TB A terabyte (TB) is a unit of digital information storage corresponding to 10^{12} bytes (B), or $1,000$ gigabytes (GB). In practical terms, 1 TB is roughly equivalent of $1,000$ hours of video or $100,000$ photos. In RNA-seq, a typical 10-30 million paired-end reads in FASTQ format could result in a file size around 5-10 GB. Similarly a Ribo-Seq raw dataset with 20 million reads might produce a

compressed FASTQ file of 5-10 GB. You may expect that one terabyte corresponds to the storage capacity of roughly 50 to 100 Ribo-seq samples from 25 to 50 paired case-control experiments. (page 402)

TE The translation efficiency (TE) in the protein synthesis of a particular transcript was initially defined by Ingolia *et al.*, 2009, as the ratio of the number of ribosome footprints in a given transcript to the number of transcript copies of that transcript, or equivalently, by the Ribo-seq fragment reads divided by the RNA-seq fragment reads mapped on a given transcript. The scoring can be normalized (or not) by the length of the transcript. TE scores are often the log (in basis 2)-ratio between polysome-associated mRNA data and cytosolic mRNA data. This metrics suffers spurious correlation. (page 44)

Torin1 Torin1 is an effective inducer of autophagy, as inhibition of mTOR, mimics cellular starvation by blocking signals required for cell growth and proliferation. (page 36)

UML UML stands for Unified Modeling Language. It takes the form of standardized visual diagrams to show the behavior and structure of a system. It helps describe how various parts work together. In Object Oriented Object programmation, UML represents classes, with their attributes, methods and functional links between interacting objects. (page 82)

Universal scaling Universal scaling in the context of protein synthesis is the observation that protein synthesis times scale linearly with coding-sequence length [Sharma et al. 2018]. Across transcripts in an organism, the average codon translation rate exhibits minimal variation, implying a nearly constant elongation rate per transcript. This allows the synthesis time of a protein to be estimated simply by multiplying the transcriptome-wide average elongation rate by the coding sequence length. This is also related to the notion that the dwell time of a ribosome on any given codon is the same within a transcript, on average, across codons of the same type. This is also true for the average dwell time of the ribosome on any given codon across different transcripts throughout the transcriptome. The causes of this universal scaling are multiple:
(i) Despite local fluctuations, codon translation rates are near-randomly distributed across transcripts, meaning slow-translating codons are offset by fast-translating ones. This statistical balance leads to a consistent average translation rate per transcript;
(ii) The law of large numbers–e.g., large number of codons in a transcript–explains the mathematical origin of scaling. The average codon translation rate per transcript follows the law of large numbers, converging toward the transcriptome-wide mean as gene length increases. This explains why longer transcripts exhibit translation times that are highly predictable based on their length;

(iii) Key molecular determinants of translation speed, such as codon usage, tRNA availability, proline content, and charged residues, scale proportionally with transcript length. This suggests that randomness in these factors drives the overall scaling behavior. This universal scaling confers robustness to mRNA translation and protein elongation, ensuring stability despite biological complexity. (page 65), (page 387)

Wobble base A single tRNA molecule can recognize more than one codon. The first two bases in the codon-anticondon comply with the standard Watson-Crick base pairing rule: one to one mapping of G to C and A to U. The recognition of the third base is less stringent than of the first two. The pairing of the third base of the codon with the first base of the anticodon at position 34 shows some steric freedom ("wobble"). This wobble (imprecision) capacity allows a single tRNA molecule to match different synonymous codons and to explain why the number of different tRNA isoacceptors in most species across the three domains of life can be (much) smaller than the 61 sense codons. (page 24), (page 52), (page 164)

Contents

Abstract iii

Résumé sommaire v

List of Abbreviations and Glossary xv

List of Symbols xvii

Contents xvii

List of Figures xxvii

List of Tables xxxi

Acknowledgements 1

1 Introduction to translation, protein elongation cycle, ribosome profiling analysis and translational control 5

1.1 The flux of information in modern molecular biology practice 5

1.2 Translation by ribosomes 6

1.2.1 Ribosomes 7

1.2.2 Ribosomes subunits and their compositions 10

1.2.3 Templates used in translation: transcripts 11

1.2.4 Initiation 13

1.2.5 Initiation in procaryotes 15

1.2.6 Initiation in eucaryotes 18

1.2.7 Energy budget of initiation 24

1.3 Ribosome elongation cycle 24

1.3.1 Mechanism of the peptide bond formation at the LSU catalytic site 26

1.3.2 Factors affecting the elongation cycle 26

1.4 Termination 28
1.5 Energy budget of an elongation cycle 28
1.5.1 Aminoacyl-tRNA synthetases first activate amino acids and bind them to specific tRNAs 29
1.5.2 Two elongation factors with GTP-ase activity drive the elongation cycle in the ribosome 30
1.6 Polysome versus ribosome profiling 32
1.6.1 Differential gene expression and translational expression analysis 32
1.6.2 RNA-Seq 33
1.6.3 Polysome fractionation profiling 34
1.6.4 Ribosome profiling (Ribo-Seq) and ribosome footprint density maps 38
1.6.5 Dry and wet lab ribosome profiling data analysis 40
1.6.6 Translation efficiency 44
1.6.7 Spurious correlation in translation efficiency scores 45
1.7 Scoring metrics of codon usage of synonymous codons and tRNA isoacceptors relative abundance 49
1.7.1 Codon usage bias and codon adaptation index 49
1.7.2 tRNA gene copy number 51
1.7.3 tRNA adaptation index 51
1.8 Summary of the challenges in ribosome profiling and the need for improved models 56

2 Thesis objectives and methodology 57
2.1 Background and problem statement 58
2.2 Thesis objectives and methodology 60
2.2.1 Research questions and methodology 60
2.2.2 Limiting assumptions and inductive bias 65
2.2.3 Thesis legacy and open source delivered tools 66
2.3 Structure and organization of the thesis 67

3 Building the ribosomer framework using a TASEP approach 69
3.1 What is a Model in Science? 70
3.2 Modeling in computational biology 70
3.3 Agent-Based Models 72
3.4 Agent-Based Modeling Platforms and Tools 77
3.5 Object Oriented Programming for Agent-Based Model Implementation 78
3.6 TASEP: Totally Asymmetric Simple Exclusion Process: a model for mRNA translation 79
3.7 A digital twin of the protein elongation cycle with its modulating factors 80
3.8 Agents in the model of mRNA translation by ribosomes and ABM structure 82

3.9 Computer cluster resources and methods for computer intensive agent-based model simulations . . . 93
3.10 Summary of main findings and insights . . . 94

4 Mechanobiochemistry in the elongation cycle and queueing time statistical theory of the ribosome on a codon 95
4.1 Core overview and connection to the thesis backbone . . . 97
4.1.1 Key contributions . . . 98
4.1.2 Key outcomes included in the ribosomer model framework . . 102
4.2 A simple geometrical model of the electrostatic environment around the catalytic center of the ribosome and its significance for the elongation cycle kinetics (CSBJ01) . . . 105
4.3 Introduction . . . 106
4.4 Material and Methods . . . 108
4.5 Theory and calculations . . . 111
4.6 Results . . . 125
4.7 Discussion . . . 151
4.8 Concluding remarks and future perspectives . . . 155

5 tRNAs pool and tRNA modifications 157
5.1 Transfer RNAs, tRNAs, key adaptor intermediates in protein synthesis 159
5.2 Introduction to tRNA modifications . . . 165
5.3 ELP3-TRM9 and URM1: tRNA modifying enzymes of U34 . . . 167
5.4 Adenosine Deaminase Acting on tRNAs (ADAT): a tRNA modification enzyme expanding the decoding capacity of multiple tRNAs and the translation efficiency . . . 172
5.5 Implementation and calibration of the tRNA modifications factor in the Agent-Based model . . . 174
5.6 Summary of main findings and insights . . . 183

6 Ribosome exit tunnel electrostatic interaction 185
6.1 Core overview and connection to the thesis backbone . . . 187
6.1.1 Key contributions . . . 187
6.1.2 Key outcomes included in the ribosomer model framework . . 189
6.2 Ribosome exit tunnel electrostatics (PRE) . . . 192
6.3 PRE introduction . . . 192
6.4 Geometrically idealized electrostatic models of the ribosome exit tunnel 197
6.4.1 General modeling approach and assumptions . . . 197
6.4.2 Hollow straight cylinder . . . 201
6.4.3 Normally truncated straight cone model . . . 201
6.4.4 Normally truncated cone concatenated to a cylinder . . . 203
6.4.5 Empirically improved electrostatic potential profile of cylindrical shape best fitted to experimental point measurements . . 206

6.4.6 Empirically improved electrostatic potential profile of truncated cone combined with cylindrical shape best fitted to experimental point measurements . . . 208
6.5 PRE models assumptions . . . 212
6.5.1 Mapping the negatively charged phosphate groups close to the tunnel inner wall and counting the number of charged groups contributing to the formal surface charge density of the ribosome tunnel inner wall from a X-ray solved structure of the large ribosomal subunit . . . 212
6.5.2 Dielectric response of the medium prevailing in the tunnel and mobile ions . . . 218
6.5.3 Polarization surface charge densities on the tunnel wall due to discontinuity between dielectric media: image charge effect with dielectrics . . . 221
6.5.4 Water screening inside the tunnel wall medium . . . 223
6.6 Application of the ribosome exit tunnel model . . . 229
6.6.1 Computing the electrostatic interaction variations of the ribosome exit tunnel for different amino acid sequences in nascent polypeptides . . . 229
6.6.2 Comparing the electrostatic interaction profiles when passing through the ribosome exit tunnel for different amino acid sequences . . . 234
6.7 PRE discussions and future perspectives . . . 242
6.7.1 Summary of results . . . 242
6.7.2 Reliability of model assumptions and model limitations . . . 245
6.7.3 Comparison to state of the art and literature in the field . . . 250
6.7.4 Future perspectives . . . 252
6.8 Reversing the relative time courses of the peptide bond reaction with oligopeptides of different lengths and charged amino acid distributions in the ribosome exit tunnel (CSBJ02) . . . 253
6.9 CSBJ02 introduction . . . 253
6.10 CSBJ02 materials and methods . . . 258
6.10.1 Effect of mechanical forces on chemical reaction kinetics . . . 258
6.10.2 Modified Michaelis-Menten Kinetics . . . 259
6.11 CSBJ02 results . . . 263
6.11.1 Axial force profiles and forces transmitted to the carboxy-terminal end of the tRNA at the P-site . . . 263
6.11.2 Peptide bond formation reaction rates with 10, 22 and 40-mer oligopeptide and puromycin. . . . 264
6.11.3 Estimated theoretical uncertainty for the median time course of the peptide bond formation reaction . . . 267
6.12 CSBJ02 discussion . . . 271

6.12.1 Connections to previous studies . . . 272
6.12.2 Assumptions, limitations and generalization . . . 273
6.12.3 Theoretical uncertainty calculated on model output based on input and experimental uncertainties . . . 274
6.12.4 Synthetic oligopeptides to be used as electrostatic probes and future perspectives . . . 275

7 Slow peptide bond formation by proline residues 277
7.1 Proline is an imino acid . . . 279
7.2 Structural data show that diprolyl-peptide is badly positioned at the P-site for peptide bond formation . . . 281
7.3 Specific elongation factors are necessary to enhance proline peptide bond formation . . . 283
7.3.1 Post-translational β-lysinylation of bacterial EF-P elongation factor . . . 283
7.3.2 Post-translational hypusinylation of eukaryotic eIF5A elongation factor . . . 284
7.4 The modified elongation factors rescue the peptide bond formation involving proline residues . . . 286
7.5 Elongation rates for proline tRNAs and their calibration . . . 288
7.6 Summary of main findings and insights . . . 289

8 mRNA secondary structures 291
8.1 An introduction to mRNA secondary structures and their effects on the elongation cycle of the ribosome . . . 293
8.2 Predicting secondary structures in single-stranded RNA from their primary sequence . . . 294
8.2.1 Bioinformatics of RNA secondary structures . . . 294
8.2.2 Nearest neighbour energy model . . . 295
8.2.3 Free energy minimization . . . 295
8.2.4 Thermodynamic ensemble of structures . . . 296
8.2.5 Reliability and prediction performance . . . 296
8.2.6 The Vienna RNA package or the Vienna Websuite . . . 299
8.3 More details on the third step of elongation: translocation and the Brownian ratchet mechanism . . . 299
8.4 Unwinding or melting mRNA secondary structures requires forces . . . 303
8.5 Ribosome gear box bifurcation and unwinding forces . . . 304
8.6 Implementation of the effect of secondary structures and its calibration in the agent-based model . . . 308
8.6.1 Modulation of the translocation rate (k_3) . . . 308
8.6.2 Pseudocode for mRNA secondary structure effect on translocation rate . . . 309
8.6.3 Calibration . . . 312

8.7 Summary of main findings and insights 312

9 Limited or abundant ribosomal resources and non-uniform initiation rates 313
9.1 Ribosome biogenesis and ribophagy: a brief introduction 316
9.2 mTOR and MAPK: central nodes in the signaling cascade of ribosome biogenesis and protein synthesis 317
9.2.1 Cellular effects of mTOR 318
9.2.2 Ribosome biogenesis 318
9.2.3 mTORC1 signaling pathway is connected to translation 319
9.2.4 Autophagy and ribophagy 319
9.2.5 MAPK signaling and melanoma cancer 321
9.3 The pool of ribosomes 322
9.3.1 Ribosome lifetime 322
9.4 Non-uniform initiation rates 322
9.5 Parameter calibration of the ribosome pool and the initiation rates . . 325
9.5.1 Calibration of the ratio of the ribosome pool over the number of transcripts 326
9.5.2 Calibration of the initiation rates 327
9.6 Summary of main findings and insights 331

10 Model output and flexibility offered by the input data and parameters 333
10.1 Summary of the model and its input data and parameters 334
10.1.1 Ribosomer at a glance 334
10.1.2 List of model input data and parameters 336
10.1.3 Steady state achievement and sampling rate used to build aggregated metrics 339
10.2 Predicted output variables of the model 340
10.2.1 Distribution of free and translating ribosomes 340
10.2.2 Protein relative abundance 343
10.2.3 Translation efficiency 344
10.2.4 Polysome fragmentation profile (global) 344
10.2.5 Polysome fragmentation profile (per transcript) 345
10.2.6 Ribosome density maps per transcript 345
10.3 Verification and sanity checks 349
10.3.1 Verification of consistent output across the two language implementations 349
10.3.2 Time-step adequacy and real-time compatibility in simulations 350
10.4 Summary of main findings and insights 351

11 Sensitivity and uncertainty analysis 353
11.1 Threshold for effect size in the comparison of paired simulations . . . 354
11.1.1 Design of computational experiments 355

11.2 Sensitivity and uncertainty analysis . . . 356
11.2.1 The curse of multiple testing . . . 358
11.2.2 List of model parameters . . . 359
11.3 Results overview of the sensitive parameters on the model predicted outcomes . . . 360
11.3.1 Transcript copy number . . . 361
11.3.2 Ribosome pool size and general initiation rate . . . 363
11.3.3 Sensitivity of ribosome pool size and general initiation rate on the ribosome fragmentation profile . . . 372
11.3.4 Codon usage bias and U34 tRNA modifications . . . 383
11.3.5 Ribosome recruitment on privileged transcripts: fold change in individual initiation rates . . . 388
11.4 Summary . . . 391

12 General discussion **393**
12.1 Summary of the work and main contributions . . . 395
12.1.1 Chapters summary . . . 395
12.1.2 Main contributions, key insights and thesis legacy . . . 397
12.2 General reflections on computational biology . . . 400
12.2.1 The importance of uncertainty quantification . . . 400
12.2.2 The importance of sharing and accessing data . . . 402
12.2.3 The importance and role of *in-silico* modeling as an integrative framework for theory building and mechanisms elucidation . . . 404
12.3 Perspectives and future work . . . 405
12.3.1 Perspectives and biotechnology applications . . . 405
12.3.2 What comes next for the RIBOSOMER integrative platform? . . . 405

A Queueing time statistical theory **407**
A.1 Sum of exponentially distributed random variables with arbitrary parameters, hypo-exponential, gamma and exponentially modified Gaussian density distributions . . . 407
A.1.1 Probability density function for the sum of random variables as a convolution product of the probability density functions of the terms in the sum . . . 408
A.1.2 Poisson process, exponential distribution and the memoryless property . . . 409
A.1.3 Erlang and Gamma distributions and the loss of the memoryless property . . . 411
A.1.4 Hypo-exponential density as a sum of independent exponentials having arbitrary pairwise distinct parameters . . . 413
A.1.5 Exponentially modified Gaussian density . . . 414

A.1.6 A note on the number of the required parameters to determine the density of a distribution, their relation to the mean, variance, skewness and definition domains 415
A.1.7 Brute force mutual comparison of distributions and quality of the fit of a (shifted) hypo-exponential or a (shifted) Gamma to the exponentially modified Gaussian density 416

B Supplementary material of the CSBJ01 paper 421
B.1 Analytical solution for the area of the truncated prolate spheroid as a surface of revolution of a truncated ellipse 421

C Supplementary material of the PRE paper 425
C.1 Hollow straight cylinder model 425
C.2 Normally truncated straight cone model 430
C.3 Normally truncated straight cone model complete derivation 436
C.3.1 Scalar potential 436
C.3.2 Electric field 441
C.4 Material and methods for mapping the charged chemical groups on the tunnel inner wall from X-ray solved structures of the ribosomal large subunit 444
C.5 Comparison of coarse grained electrostatic potential calculated from the observed structural data with the geometrically idealized model . . 445
C.6 Specific effect of the tunnel electrostatic interaction on the elongation rate 450
C.7 Energy sources available for the ribosome 452

D Supplementary material of the CSBJ02 paper 455
D.1 Uncertainty quantification of the theoretical peptide bond formation rate 455

E BASH Script for HPC Job Scheduling and Resource Management 461

F Fully crossed factorial design (fixed effects model) 465
F.1 Proportion of free ribosomes in the pool 465
F.1.1 Analysis of variance for the two-factor fixed effects model . . 465
F.1.2 Diagnostics case statistics and model adequacy checking . . . 467
F.1.3 Conclude on hypothesis tests 469
F.1.4 Model parameters estimation for proportion of free ribosomes response 472
F.2 Proportion of free transcripts in the transcriptome 474
F.2.1 Analysis of variance for the two-factor fixed effects model . . 474
F.2.2 Diagnostics case statistics and model adequacy checking . . . 474
F.2.3 Conclude on hypothesis tests 475
F.2.4 Model parameters estimation for proportion of free transcripts response 477

F.2.5 Fitting Response surfaces 477

Bibliography **481**

Publications and curriculum vitae **499**

F.3 Publications 499

F.3.1 Original contributions 499

F.3.2 Additional publications 500

F.4 Curriculum vitae 501

List of Figures

1.1 Translatomics filling the gap between transcriptomics and proteomics 6
1.2 Electron micrograph of rough and smooth endoplasmic reticulum . . 8
1.3 Ribosomes in translation 9
1.4 Ribosome subunits and composition 10
1.5 Structure of caps at the 5' end of eucaryotic mRNAs 12
1.6 General structure of UTR leader sequence of eucaryotic mRNA 14
1.7 Model of initiation in procaryotes 16
1.8 Canonical cap-dependent scanning model of eucaryotic initiation . . . 20
1.9 mTORC1 sensitive TOP motif mRNA initiation 23
1.10 Ribosome elongation cycle sub-steps 25
1.11 Mechanism of proton transfer in the peptide bond formation 27
1.12 Aminoacyl-adenylate intermediate 29
1.13 Simplified workflow of the ribosome profiling protocol 35
1.14 Polysomal profiles in MCF7 cells to study global translation 37
1.15 Relation between ribosome dwell times on codons and ribosome occupancy map 39
1.16 Ribosome footprinting theoretical scenarios leading to observed ribosome occupancy pattern 46

2.1 Summary of factors affecting the ribosome elongation cycle 61
2.2 Speed of translation along a transcript 64
2.3 Wireframe flowchart of the thesis 68

3.1 Statistical physics inspired diffusion model in one dimension: TASEP 79
3.2 Ribosomer settings user interface 81
3.3 UML diagram of Ribosomer ABL 84
3.4 Illustration of TASEP model contextual factors 86
3.5 Ribosomer snapshot of a simulation run 90
3.6 Pool of free ribosomes 92

4.1 Focus on the sequential steps of the ribosome elongation cycle 96
4.2 Gibbs free activation energy in Ehring's theory of transition state . . . 99
4.3 Queueing time statistical distribution of a ribosome on a codon 103
4.4 Fragment of the secondary structure of 23SrRNA 109
4.5 Gibbs free energy activation barrier 119
4.6 Ribosome exit tunnel and cavity around the PTC 126
4.7 Atlas of the tri-dimensional PTC shell cavity for 5 species 128
4.8 Electrostatic potential profiles and electric field 131
4.9 X-ray solved tRNA structures 133
4.10 Electrostatic potential contributed by *Thermus thermophilus* PDB code 4Y4P 135
4.11 Electrostatic potential around PTC cavity 139
4.12 Hypo-exponential distribution 142
4.13 Elongation minimal case 144
4.14 Ribosome residence time distribution on specific codons 149
4.15 RRT in *Saccharomyces cerivisiae* and *Esherichia coli* 150

5.1 Graphical abstract on the impact of tRNA modifications on the ribosome elongation cycle 158
5.2 tRNA symbolic structure and mode of attachment of amino acid . . . 160
5.3 Base sequence and clover leaf secondary structure of yeast alanine tRNA 161
5.4 tRNA biogenesis and life cycle 166
5.5 U34 post-transcriptionally modified structures 167
5.6 Simplified pathways and enzymes modifying wobble U34 tRNAs . . . 169
5.7 Adenine editing to inosine and decoding expansion of targeted tRNAs 173
5.8 Inosine pairing with C, A and U 173
5.9 ADAT silencing effects on alanine synonymous codons queueing times 179
5.10 ADAT silencing effects on GCC GCU alanine codons queueing times 181
5.11 ADAT silencing effects on GCG GCA alanine codons queueing times 182

6.1 Graphical abstract of the impact of the ribosome exit tunnel electrostatic interaction on the elongation cycle 186
6.2 Bioinformatics algorithm for the computation of electrostatic axial forces acting on the nascent chain in the ribosome exit tunnel 191
6.3 Ribosome exit tunnel structure 196
6.4 Idealized and realistic ribosome exit tunnel models 199
6.5 Electrostatic scalar potential 211
6.6 2D Mapping of phosphorus atoms and charged amino acid residues . . 216
6.7 Polarization surface charge densities on tunnel wall 221
6.8 Phosphate moieties and water screening 224
6.9 Protrusion of positively charged amino acids close to the constriction site 229
6.10 Algorithm for computing axial forces on the nascent peptide 232
6.11 Synthetic axial force profiles 236

6.12 HTT . . . 239
6.13 Real axial profiles for huntingtin . . . 239
6.14 Electrostatic potential and axial force profiles around the ribosome PTC and the exit tunnel . . . 261
6.15 Elongation extends case: 10-mer . . . 266
6.16 Elongation extends case: 22-mer . . . 268
6.17 Elongation extends case: 40-mer . . . 270

7.1 Graphical abstract on the impact of proline incorporation on the ribosome elongation cycle . . . 278
7.2 Imino acid proline . . . 279
7.3 Proline peptide bond formation . . . 280
7.4 Proline side chain atypical position in the ribosomal A-site . . . 280
7.5 β−lysinylation of EF-P elongation factor . . . 282
7.6 Polyamines pathway and hypusinylation of eIF5A elongation factor . . . 285
7.7 EF-P and eIF5A elongation factors structures and functions . . . 287

8.1 Graphical abstract of the impact on the elongation cycle of secondary structures downstream the mRNA . . . 292
8.2 Forces to unwind and to melt secondary structures . . . 293
8.3 Minimum free energy mRNA secondary structure . . . 298
8.4 Gibbs free energy landscape for translocation . . . 301
8.5 Gear box bifurcation pathways . . . 305
8.6 Translocation gear shift pathways . . . 307

9.1 Graphical abstract of chapter on the ribosome pool and initiation rates 314
9.2 Electron micrograph of ribophagy, showing autophagosomes containing ribosomes . . . 320
9.3 Electron micrograph of macroautophagy, showing autophagosomes containing mitochondria . . . 321
9.4 Initiation rate and ribosome elongation rate on transcripts . . . 324

10.1 Multiple lattices extended TASEP . . . 334
10.2 Summary of Ribosomer model and contextual factors . . . 335
10.3 Free and translating ribosomes . . . 342
10.4 Protein relative abundance . . . 343
10.5 Protein translation efficiency . . . 344
10.6 Polysome profiles predicted output . . . 346
10.7 Per-transcript polysome profiles predicted output . . . 347
10.8 Ribosome density map of RPL4 . . . 348

11.1 Transcripts copy number sensitivity . . . 362
11.2 Heatmaps of proportions of free ribosomes and free transcripts . . . 364

11.3 Main effects on proportion of free ribosomes 367
11.4 Main effects on proportion of free transcripts 369
11.5 Contour plot and response surface of free ribosomes and free transcripts 370
11.6 Polysome fragmentation profile in restricted ribosome pool 376
11.7 Polysome fragmentation profile in well-supplied ribosome pool 377
11.8 Codon usage and U34 tRNA modification analysis 386
11.9 Ribosome recruitment on privileged transcripts analysis 390

12.1 Ribosome cycle recapitulative schematic overview 394

A.1 Best least squared fit of shifted hypo-exponential 418

C.1 Electrostatic potential for a cylinder 429
C.2 Electrostatic potential for truncated cone 435
C.3 Electrostatic potential calculated from exact positions of 94 charged atoms 448
C.4 Maxwell Boltzmann factors 451

F.1 ANOVA Fixed effects model adequacy checking 468
F.2 Main effects on proportion of free ribosomes 470
F.3 Interaction effects on the proportion of free ribosomes 471
F.4 ANOVA Fixed effects model adequacy checking 474
F.5 Main effects on proportion of free transcripts 475
F.6 Interaction effects on the proportion of free transcripts 476

List of Tables

1.7.1 Crick's wobble rules for calculating the absolute adaptiveness of a codon to the tRNA pool 53
1.7.2 Weight factors for base pairings in codon-anticodon interaction for yeast 54
1.7.3 Weight factors for base pairings in codon-anticodon interaction for the 3 domains of life 55

4.4.1 Universally most conserved 23S/28S rRNA sequences 110
4.6.1 Experimentally measured dipeptityl transfer rate constants 145
4.6.2 Maxwell-Boltzmann factors modulating the dipeptidyl transfer rate constants 146
4.6.3 RRT distribution deconvolution 148

5.1.1 Wobble base rule 163
5.3.1 Codon-anticodon relation for U34 tRNA modification affecting amino acid K, Q, E 171
5.5.1 Dictionary of keys:values upon ADAT-silencing 175
5.5.2 Codon-anticodon relation for ADAT-related amino acid T, A P 176
5.5.3 Codon-anticodon relation for ADAT-related amino acid I and V 177
5.5.4 Codon-anticodon relation for ADAT-related amino acid S, L, R 178

6.5.1 Screening lengths 226
6.11.1 Maxwell-Boltzmann factors modulating the x-mer transfer rate 265

7.5.1 Peptide bond formation kinetics parameters for prolyl-tRNA 288

9.5.1 Parameter sample space for the ribosome pool ratio 326
9.5.2 Parameter sample space for the mean initiation rates 330

11.3.1 Analysis of variance table for the proportion of free ribosome 366
11.3.2 Analysis of variance table for the proportion of free transcripts 368
11.3.3 Statistics of number of ribosomes in polysome fragments 374

C.5.1 Phenomenological parameter values in the tunnel 447

D.1.1 Numerical values and uncertainties of the input variables of the model 457
D.1.2 Numerical values and calculated uncertainties 460

Acknowledgements

Before starting my scientific career, I was involved in corporate activities, working for a variety of small and medium enterprises. My curiosity and appetite for learning drove me to enroll in multiple STEM academic programs in parallel, all while working full-time as an employee. Students usually engage in *extra-curricular* activities—I, on the other hand, was a corporate employee engaging in *extra-corporate-lar* activities, expanding both my skill set and my overtime hours. One might say I had a *double major* in business and curiosity.

During this long corporate journey, I had the chance to meet highly skilled professionals, committed employees, and even business angels. I was never bored and always enjoyed looking up at the sky—while keeping my feet firmly on the ground or, quite literally, in the mud, as I worked with wastewater treatment sludge and landfill leachates.

Many former corporate colleagues and bosses have described me as someone who blends a highly theoretical and conceptual education with practical, hands-on expertise. At my core, I find true fulfillment in learning and satisfying my insatiable curiosity for new knowledge.

After corporate workers and business angels, I met my *research angel*: Professor Liesbet Geris. Lies, I wish to express my deepest gratitude for being the first person to take my unconventional academic path seriously and for having the courage to invest in me—without bias, despite my non-traditional trajectory as a senior candidate, with no scientific seniority, in his fifties. I sincerely appreciate your trust and, most importantly, your generous research funding, which has supported me for seven very happy years. I have told you this before, but I say it again: you are a modern Renaissance patron, championing both science and the arts. Your willingness to take a risk on an interdisciplinary PhD project outside your primary research track demonstrates your vision as a principal investigator. At least we shared a common foundation in modeling in computational biology. I am also deeply appreciative of the guidance, rigor, and scientific integrity you have instilled in me through our discussions—whether in your office or virtually, often extending into evenings, weekends, and even late nights on

Slack. Your expertise in modeling and digital twin development has been invaluable. You were among the pioneers in the field, shaping how researchers conceptualize *in silico* representations of biological systems with increasing complexity. I also recognize and value the effort you have dedicated to teaching me how to write more concise papers. I promise to keep improving in that area! Your strategic insights on selecting editorial boards and navigating peer review have been incredibly instructive. I am proud and grateful to have worked with you and to have witnessed the numerous prestigious grants, awards and *honoris causa* distinctions you have received across Europe.

Selecting the right PhD research topic is of utmost importance, as it will define several years of work. Professor Michel Georges, former director of GIGA, once advised PhD candidates in an introductory lecture at the doctoral school that the most effective way to begin a doctoral project is to identify the boundary between what is known and what remains unknown. Once this frontier is found, the task of the researcher is to move to the edge and push it further.

How did I find the topic and how did I move to the edge ?

In November 2018, during a discussion at the GIGA doctoral school, Professor Pierre Close introduced me to one of his research interests: enzymatic modifications of nucleotide 34 in tRNAs and their effects on protein elongation rates. He explained how transcripts with different codon usage are translated at varying efficiencies, leading to shifts in the proteome landscape. In melanoma and lung cancer contexts, these modifications to U34 or A34-tRNAs alter the oncoproteome. A seemingly small chemical modification in the anticodon loop of a tRNA molecule could influence the translation of the entire proteome.

At first glance, the underlying molecular mechanisms appeared bafflingly complex—at the very least, fascinating. This topic immediately captivated me, as it challenged my background in chemical and bioengineering. Pierre, thank you for bringing me on this journey. After several weeks of literature review, I became convinced that the key to advancing knowledge in this field lay in addressing its complexity through computational modeling—systematically integrating the multiple contributing factors described in the literature. Given that at least five interconnected variables were involved, tackling this question required a truly interdisciplinary and integrative approach. I also want to express my gratitude to Pierre Close coworkers: Dr. Francesca Rapino, Dr. Arnaud Blomme and Marine Leclercq for sharing their datasets.

I would like to express my retrospective gratitude to my high school physics teacher, Madame Marie-Thérèse Henrotin, whose curiosity and passion for teaching left a lasting impression on me. A special thank you goes to my chemistry teacher, Monsieur Alain Capouillez, an autodidact who introduced me to Henry Eyring's transition state theory of catalysis when I was only 16. This was not part of the standard curriculum, and I am immensely grateful that he chose to go beyond it. He would have certainly appreciated

the mechanochemistry embedded in the transition-state theory applied in this thesis to study the ribozyme-catalyzed kinetics of peptide bond formation.

I would also like to extend my sincere thanks to the members of my thesis jury for their time and interest in my research. In particular, I am grateful to Professor Eveline Lescrinier for her co-endorsement—alongside Professor Liesbet Geris—of my future postdoctoral project. I also extend my appreciation to Professor André Matagne, Dr. Frédéric Kerff and Dr. Arnaud Vanden Broeck. A sincere thank you to professor André Matagne for welcoming me as a member of the Belgian Biophysical Society. I truly appreciate the opportunity to be part of this community and to engage with fellow researchers in the field of biophysics. I look forward to contributing to the society's activities and benefiting from the exchange of ideas and collaborations. I warmly thank Professor Gerben Menschaert for the technical advice, valuable discussions, and insightful exchanges we had at Ghent University regarding Ribo-Seq technology and its data analysis. My gratitude also goes to Professor Dick de Ridder for organizing the Bioinformatics Summer School sessions at Wageningen University with his team, as well as for his pioneering work over a decade ago on modeling protein synthesis using TASEP.

Additionally, I would like to acknowledge Professor Tamir Tuller of Tel Aviv University. In these challenging times, researchers around the world face difficulties due to conflicts, the rise of authoritarian regimes or repressive governance beyond their control. I deeply regret the isolation that many scientists endure, ranging from professional exclusion to, in the worst cases, acts of censorship, restrictions on freedoms or even war-related casualties. It is my hope that scientific collaboration and open discourse will continue to transcend political and ideological barriers.

I would like to express my gratitude to all my fellow colleagues, PhD candidates and postdocs in Lies' Biomechanics group with whom I spent joyful moments: Bernard Staumont, Sophie Bekisz, Loic Comeliau, Luiz Ladeira, Alessio Gamba, Margaryta Ivanets, Sophie Nguyen, Ehsan Sadeghian, Fernando Perez Boerema, Bingbing Liang, Satanik Mukherjee, Mohammad Mehrian, Morgane Germain, Niki Loverdou, Raphaelle Lesage, Majid Nazemi, Laura Lafuente Gracia, Mojtaba Barzegari Shankil, Tim Herpelinck, Lisanne Groeneveldt, Sourav Mandal, Oriana de Becker, Ayse Kose, Tom Verbraeken, Mervenaz Sahin, Edoardo Borgiani, Gabriele Nasello, Claire Vilette, Ahmad Alminnawi, Pieter Ansoms, Liesbeth Ory.

I want to thank Isabelle Rausin for being the best mother of our three daughters and for her valuable networking sharing which gave me the opportunity to meet Professor Liesbet Geris.

Finally, and most importantly, I want to express my heartfelt gratitude to my family—my father, and my late mother, whose unwavering enthusiasm and support still inspire me. To Laurence, for her encouragement and patience, and to my three daughters, Emmy,

Sarah and Lara, whose achievements make me so proud and with whom I share a deep and joyful complicity, making them my most cherished and passionate companions in life.

This work was supported by the FWO EOS grant no.30480119 (Joint-t-against-Osteoarthritis) in Belgium, the European Research Council under the European Union's Horizon 2020 Framework Program (H2020/2014-2020) ERC grant no.772418 (INSITE) and the ERC grant no.101088919 INStant CARMA In Silico Trials for Cartilage Regenerative Medicine Applications.

Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under grant no.2.5020.11 and by the Walloon Region.

Chapter 1

Introduction to translation, protein elongation cycle, ribosome profiling analysis and translational control

This chapter is a narrative review of protein synthesis by ribosomes introducing the minimal molecular biology pre-requisites that are necessary to comprehend the need, the scope and purpose of the computational modeling of mRNA translation.

1.1 The flux of information in modern molecular biology practice

For the last seven decades or so, modern biological research has aimed at deciphering the flow of information that occurs in living cells to manage a variety of biological processes and functions. This encompasses a wide range of fields, from fundamental molecular biology to specialized biomedical applications such as cancer therapy, vaccine development, regenerative medicine, and more. No one working in these areas of research can escape the overarching framework governing information management in biology, as illustrated in Figure 1.1. Recent high-throughput techniques have been used to simultaneously study gene expression (transcriptome, RNA-seq) and

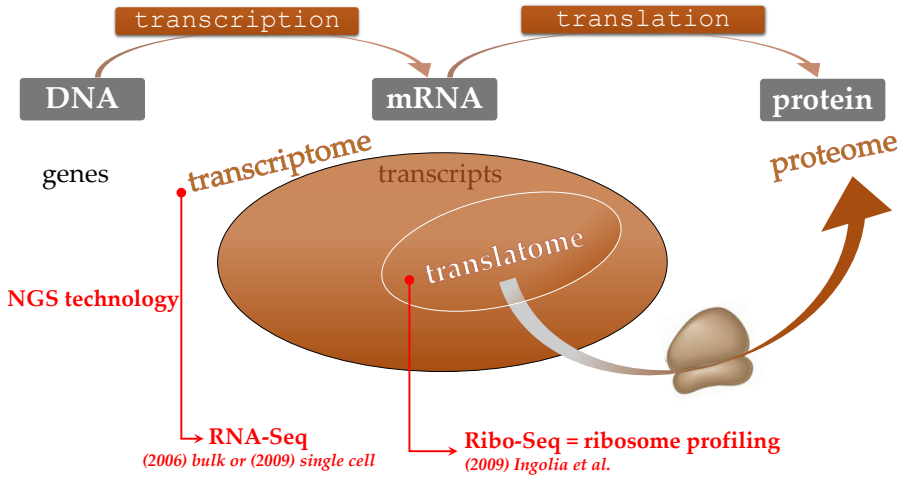


Figure 1.1: Transcriptome, translatome and proteome and their next generation sequencing (NGS) high-throughput technologies.

protein expression (proteome, mass spectrometry) on a genome wide scale. The global association between mRNAs and ribosomes (translatome, Ribo-seq) has been a subject of substantial interest since 2009 as it is a high throughput technology, like RNA-seq, filling the gap between the transcriptome and the proteome. This thesis focuses on the protein synthesis part of this big picture and addresses the study of mRNA translation by ribosomes, or translation for short.

1.2 Translation by ribosomes

Proteins are natural polymers of amino acids taken from a set of twenty amino acids commonly found in all living organisms. When the number of amino acids in the polymer is reduced, the polymer is called a polypeptide. Synthesizing a polypeptide from amino acids using only chemicals in the lab is a very difficult and cumbersome task. Do not even think about synthesizing a protein. Yet, all living organisms regularly synthesize thousands of proteins on their own within a matter of minutes. Protein synthesis is one of the most fundamental biochemical reaction carried out by all living organisms on this planet. Living organisms use a **template** to synthesize proteins. This template is another polymer: a single-stranded polymer of ribonucleotides, that we call the messenger ribonucleic acid, messenger RNA (mRNA). A **mRNA molecule** is also

called a ***transcript***, in the jargon of the big organization shown in Fig. 1.1, as a mRNA is naturally produced by the *transcription* of still another polymer: a double stranded polymer called DNA, made of slightly different nucleotides (deoxyribonucleotides).

1.2.1 Ribosomes

Ribosomes are nano-molecular machines specialized for protein synthesis. These machines are found in all cellular living organisms. The process by which a ribosome synthesizes a protein, using a mRNA molecule as a template is called **translation**. In this thesis, translation mechanisms by ribosomes are studied extensively.

To this date, the origin and biogenesis of ribosomes is still a basic science open question. The adventure of protein synthesis on our planet started around 3.7-4.3 billion years ago, in an era that is called the prebiotic era, long before unicellular life was dated in fossils records [Javaux 2019]. One of the first significant biochemical event in the pre-biotic era preceding the early history of life was when a ribonucleic acid (RNA) molecule folded in such a way as to function as a **catalyst** (a substance that speeds up a chemical reaction up to billion or trillion fold without being used or changed itself) [Morris and Mattick 2012]. This RNA catalyzer facilitated the first peptide bond formation reaction between two amino acids. This emergent biochemical innovation lead to further progress in the chemistry of the pre-biotic era when complex self-assemblies of RNA molecules and proteins allowed the building of these nano-molecular manufacturing machines that we, today, call the ribosomes.

Ribosomes ancestors evolved in parallel with the maintenance of a chemical data server: the genetic information encrypted in messenger RNA molecules. The ribosomes' structure and function co-evolved with other RNA adaptor molecules allowing to translate the language of nucleic acids to the language of proteins. The first encoder-decoder algorithm was materialized by transfer RNA (tRNA) molecules setting up this universal large language model (LLM) at the heart of the natural intelligence of life (pun intended as opposed to artificial intelligence), namely the **genetic code** (Fig. 1.3). One of the most baffling discoveries that humans did, less than 60 years ago, is that this genetic code is universal, at least on our planet Earth. Good and efficient information was useful and survived by natural selection. Bad or wrongly adapted information would go extinct. Today there are up to 10-15 million ribosomes in a single living cell like a hepatocyte in the liver of mammals or humans. The salient feature of the catalytic site of the ribosomes across all kingdoms of life and all extant species we see today, is that it is extremely well conserved. Yet, we do not exactly know or understand all its physicochemical properties. All living species on our planet Earth belong to three *clades* that we call the three domains of life: bacteria, archaea and eukarya[1].

[1] Viruses are not included in this classification.

The translation mechanisms share common features between prokaryotes/archea and eukaryotes although they differ in the details of significant substeps.

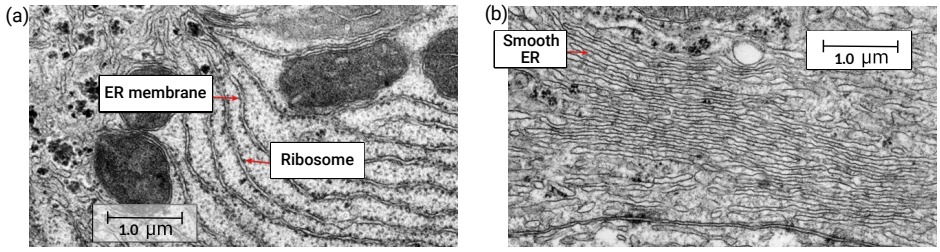


Figure 1.2: Electron micrograph of (a) rough endoplasmic reticulum in a eukaryote cell. Numerous ribosomes are visible. These ribosomes are temporarily attached to the cisternae as the cell translates mRNA into protein that is destined to be membrane-bound or secreted. (b) smooth endoplasmic reticulum revealing the lack of ribosomes on smooth ER. Credits: Yale medcell.org/histology Credits: Yale medcell.org/histology

Ribosomes can lie free in the cytoplasm or can be associated with membranes in the rough endoplasmic reticulum (RER) of eucaryotic cells to produce secreted proteins. All ribosomes in archaea and bacteria are free in the cytoplasm because they lack the membrane-bound structures (like the rough ER) found in eukaryotic cells. Free ribosomes are responsible for the synthesis of cytosol proteins and organelle subunits. They can exist as single free ribosomes or in clusters known as polyribosomes or polysomes. By light microscopy they produce a general low level cytoplasmic basophilia, with more localized and intense basophilia associated with RER (Fig 1.2). With the transmission electron micrography, ribosomes appear as small (25-35 nanometer) electron dense particles in the cytoplasm. The endoplasmic reticulum (ER) is a system of interconnected membranous sacs, channels, or cisternae in the cytoplasm. It has two subtypes: rough endoplasmic reticulum (Fig 1.2 (a)) and smooth endoplasmic reticulum (SER, Fig 1.2 (b)). The RER is a ribbon-like structure surrounding the nucleus near the base of the cell. Its surface appears rough due to the ribosomes attached to its membrane and it is the first organelle into which membrane-bound or extracellular proteins are inserted. SER lacks ribosomes and participates in lipid synthesis and detoxification.

Figure 1.3 shows a protein-synthesizing cell with ribosomes and an illustration of their functional structures.

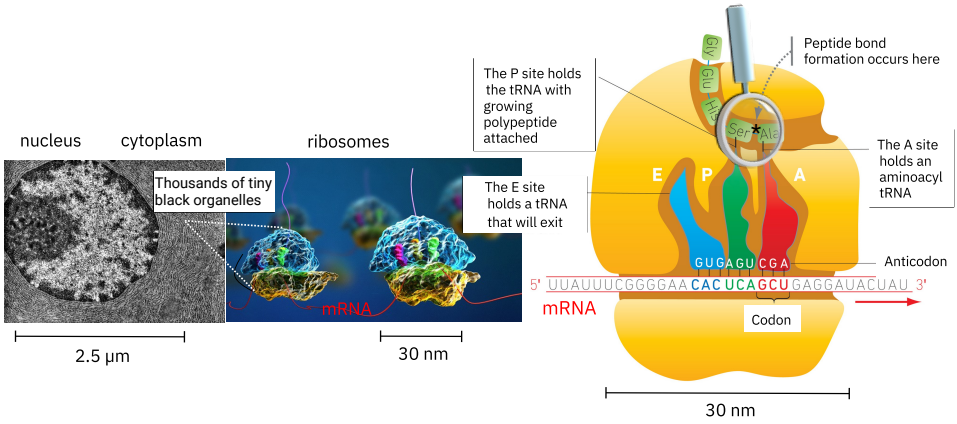


Figure 1.3: Left panel: transmission electron microscope (TEM) micrograph showing the nucleus and cytoplasm of a protein-synthesizing eukaryotic cell. The nuclear envelope, chromatin and nucleolus can be seen. The cytoplasm is full of rough endoplasmic reticulum (RER) studded with ribosomes (tiny black organelles). Central panel: 3D artist view showing ultrastructure of two complete ribosomes during elongation. Right panel: functional anatomy of a ribosome showing where the elongation sub-steps occur. The asterisk is where the peptide bond formation is catalyzed. The mRNA's encoded message is translated into a nascent protein with tRNAs acting as interpreters of the genetic code. [Credits: Jose Luis Calvo. Photograph. and Meletios Verras, Illustrator. Shutterstock. Web. 27 September 2024 and Yale medcell.org/histology].

1.2.2 Ribosomes subunits and their compositions

Figure 1.4 shows the details of the ribosomal subunits and their molecular compositions. Ribosomes are heavy. The weight composition of a ribosome is roughly 80% w/w ribonucleic acids - 20% w/w proteins. Despite the low number of ribonucleic acid molecules ($\sim 3-5$) compared to the high number of proteins in a ribosome ($\sim 55-83$), the large proportion of nucleic acids in the weight is due to the heavy atoms of phosphorus and to the fact that each ribonucleic acid monomer is bridged to the next one by a phosphate moiety.

The eukaryotic ribosome (80S) is composed of a small (40S) and a large (60S) subunit[2]. The small subunit (SSU) is made up of an 18S rRNA and 33 different ribosomal proteins. The large subunit (LSU) contains 28S, 5.8S and 5S rRNA and 50 ribosomal proteins.

[2]The symbol S refers to Svedberg units. It is a non-SI metric unit used for the measure of the sedimentation coefficient in an ultracentrifuge.

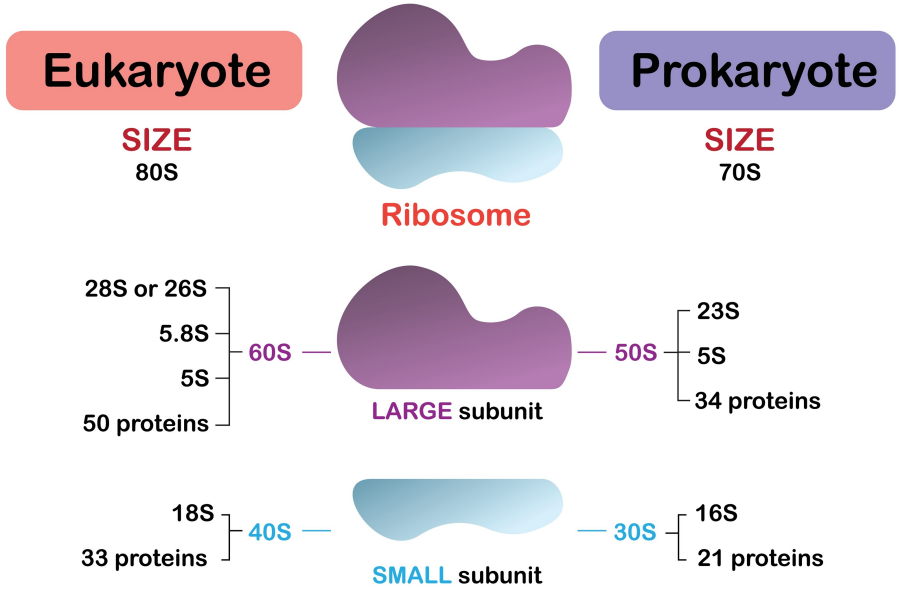


Figure 1.4: Ribosomes can be dissociated into about eighty-three proteins and four RNA molecules in eucaryotes (left panel); fifty-five proteins and three RNA molecules in procaryotes (right panel). They have two subunits: the large subunit and the small subunit [Shutterstock. Web. 27.09.2024].

Procaryotic ribosomes have a lower sedimentation coefficient (70S), are less dense than the eucaryotic ribosomes. Procaryotic ribosomes are larger in size and are a bit less compact than the eucaryotic ribosomes.

Translation of mRNA initiates with the association of a small ribosome subunit (SSU) and a large ribosome subunit (LSU) to form a complete ribosome (monosome). The ribosome moves along from the $5'-$untranslated region ($5' -$ UTR) to the start codon marking the end of initiation. As the ribosome still moves further along the mRNA molecule during translation elongation, additional ribosomes can initiate translation on the same mRNA molecule, forming polysomes. The essentials of mRNA translation are summarized below both for bacteria and eukarya.

1.2.3 Templates used in translation: transcripts

Differences between eucaryotic and procaryotic mRNAs
Important differences between eukaryotic and prokaryotic mRNAs (transcripts), which are key to understanding mRNA translation in different domains of life, are enumerated hereafter [Berg et al. 2019].

1. Primary transcripts in eucaryotes are not used directly as mRNA. Rather, they are extensively processed before being transported from the nucleus to the cytosol. Translation and transcription are spatially and temporally separated in eucaryotes, whereas they are closely coupled in procaryotes.

2. Primary transcripts in eucaryotes range in size from about 2 to 20 kb, and so they have been called heterogeneous nuclear RNA (hnRNA). These primary transcripts are usually several times as long as the mRNAs derived from them. Extensive splicing and cleavage take place in the generation of eucaryotic mRNAs. Very precise splicing enzymes remove introns from primary transcripts of split genes and only the protein coding parts, called exons, are merged. The DNA sequence corresponding to these exons coding parts is called the CDS (coding DNA sequence).

3. Eucaryotic mRNAs are monocistronic, that is, they are templates for the synthesis of a single polypeptide chain. In contrast, many procaryotic mRNAs are polycistronic, e.g., the lactose operon lac mRNA is the template for three polypeptide chains.

4. Eucaryotic mRNAs contain modified nucleotide caps at their 5' ends (Figure 1.5). The 5' end of all known eucaryotic mRNAs, but not of any tRNAs or rRNAs, is modified in a special way. 7-Methylguanosine is joined to mRNA by an unusual 5'-5 pyrophosphate linkage. This highly distinctive structure, called a cap, is acquired by the primary transcript. The 5' triphosphate end of the nascent chain is hydrolyzed to a diphosphate, to which a guanylate unit is transferred from GTP. The N-7 nitrogen of this terminal guanine is then methylated by S-adenosylmethionine to form cap 0. The adjacent riboses may be methylated to form cap 1 and cap 2. These caps contribute to the stability of mRNAs by protecting their 5'ends from phosphatases and nucleases. In addition, caps enhance the translation of mRNA by eucaryotic protein-synthesizing systems. Most eucaryotic mRNAs also have poly-A tails at their 3'ends. A poly-A polymerase adds some 150 to 200 nucleotides to primary transcripts that have a 3'terminal GC and an AAUAA sequence about 20 residues away. Studies have shown that poly-A tail enhances stability of an mRNA but is not required for its translation, and is not required for the transport of mRNA from the nucleus to the cytosol.

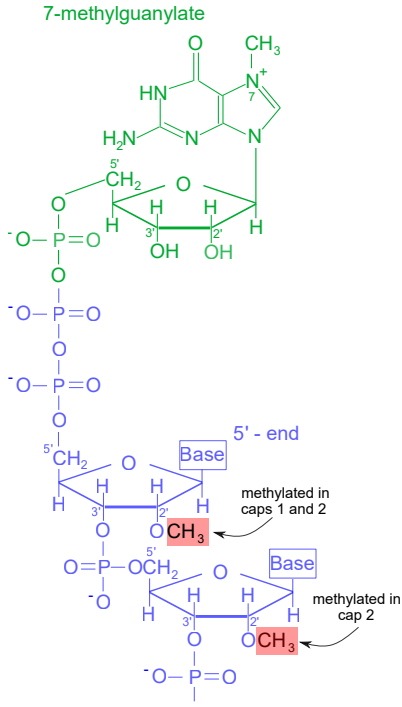


Figure 1.5: Structure of caps at the 5' end of eucaryotic mRNAs. All caps contain 7-methylguanylate shown in green, attached by a pyrophosphate linkage to the 5'end. None of the riboses are methylated in cap 0, one in cap 1 and two in cap 2 [Berg et al. 2019].

1.2.4 Initiation

Translation initiation is the rate-limiting step in mRNA translation and is central to translational control. For most mRNAs, the individual initiation rate constants are unknown. However, if these rates were known, translational efficiency could be predicted on a translatome-wide scale [Rodnina 2018]. A long term goal in translational research is to establish a link between the structures and compositions of the $5'$-untranslated regions upstream of coding sequences and their initiation rate kinetics. This quantitative understanding would explain how different transcripts compete for ribosome recruitment and clarify variations in translation efficiency across the transcriptome.

Upstream open reading frames (uORFs) are regulatory elements that are prevalent in eukaryotic mRNAs. uORFs modulate the translation initiation rate of downstream coding sequences (CDS) by sequestering ribosomes.

The 5' UTR sequence (leader sequence) of prokaryotes consists of the Shine-Dalgarno sequence $5' - AGGAGGU - 3'$. This sequence is found 3-10 nucleotides upstream the start codon of the open reading frame(ORF).

The 5' UTR sequence (leader sequence) of eukaryotes is more complex than prokaryotes. It contains a Kozak consensus sequence (Kozak sequence) flanking the start codon. Figure 1.6 shows a general structure of the 5' UTR leader sequence of eukaryotic mRNA. The figure shows multiple functional domains within the nucleotide sequence. Some of these domains are cis-regulatory elements (CRE) that can interact in multiple ways with ribonucleoproteins or initiation factors to repress or promote initiation [Leppek et al. 2017].

Understanding the detailed mechanisms of translation initiation is still an active domain of research.

Prior to translation elongation, the ribosome must be primed by first binding the unique initiator methionyl-tRNA (Met$-$tRNA$_i^{\text{Met}}$) and locating the proper `AUG` start codon on the mRNA.

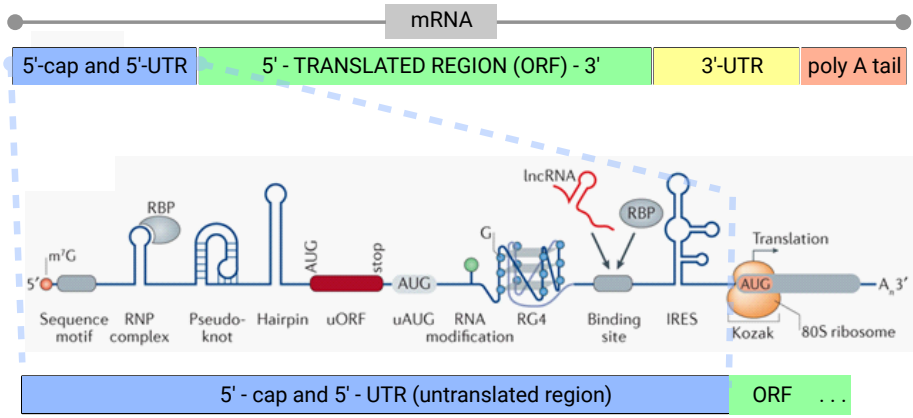


Figure 1.6: The methylguanosine (m^7G)5′ cap structure at the 5′ end of the mRNA and the poly-A tail at the 3′ end stabilize the mRNA and stimulate translation. The 5′ untranslated region UTR, also called leader sequence, contains secondary and tertiary structures and other sequence elements regulating translation initiation. Pseudoknots, hairpins, RNA-G quadruplexes, uUTRs and upstream start codon (uAUGs) inhibit translation. Internal ribosome entry sites (IRES) mediate translation initiation independently of the cap. RNA binding proteins (RBP), long non coding RNAs (lncRNAs) interact with RNA binding sites. The Kozak sequence also regulates translation initiation. The poly-A tail can interact with binding proteins and multiple initiation factors, leading to the circularization of the translating mRNA. Adapted from reference [Leppek et al. 2017].

Translation initiation can be divided into three steps, reviewed in [Dever 2002]:

1. binding Met–$tRNA_i^{Met}$ to the small ribosomal subunit (SSU)
2. binding the small ribosomal subunit (SSU) to the mRNA and `AUG` codon recognition
3. joining of the large ribosomal subunit (LSU) to generate a translationally competent ribosome.

Each of the steps of translation initiation is facilitated by proteins referred to initiation factors.

1.2.5 Initiation in procaryotes

A comprehensive review of translation initiation in procaryotes can be found in reference [Rodnina 2018]. Different types of mRNAs in prokaryotes include those with **Shine-Dalgarno (SD) sequences**, leaderless mRNAs, and mRNAs with internal ORFs. We restrict ourselves here only to the mRNAs containing the Shine–Dalgarno (SD) sequence as they are particularly well studied. They usually have an extended 5' untranslated region (5'-UTR) and an SD sequence located 8–10 nt upstream of the start codon (usually AUG). During SD-led initiation, the small subunit ([SSU], 30S in bacteria) is recruited to the **ribosome binding site (RBS)** through interactions between the SD sequence and the **complementary anti-SD (aSD) sequence in 16S ribosomal RNA (16SrRNA)**. Initiation on SD-led mRNAs is promoted by initiation factors IF1, IF2, and IF3 (inset in Fig. 1.7 (a)). In procaryotes, **only three polypeptides IF1, IF2, IF3** are necessary for translation initiation, as opposed to more than a dozen in eucaryotes.

As depicted in Fig. 1.7, the assembly pathway of the 30S pre-initiation complex (PIC) and 30S initiation complex (30S IC) involves the recruitment of initiation factors in a kinetically preferred sequence, leading to the formation of the mature 70S initiation complex (70S IC). Features affecting translational efficiency in bacteria include the nature of the start codon, the SD sequence, mRNA secondary structure, and specific elements in the RBS. The initiation pathway is conceptualized as a sequence of kinetic checkpoints, where the efficiency is determined by the forward steps towards the mature 70S IC and the rejections steps. The kinetic model explains variations in translational efficiency, with the structure and stability of the RBS affecting mRNA association, start codon stability, and LSU joining rate.

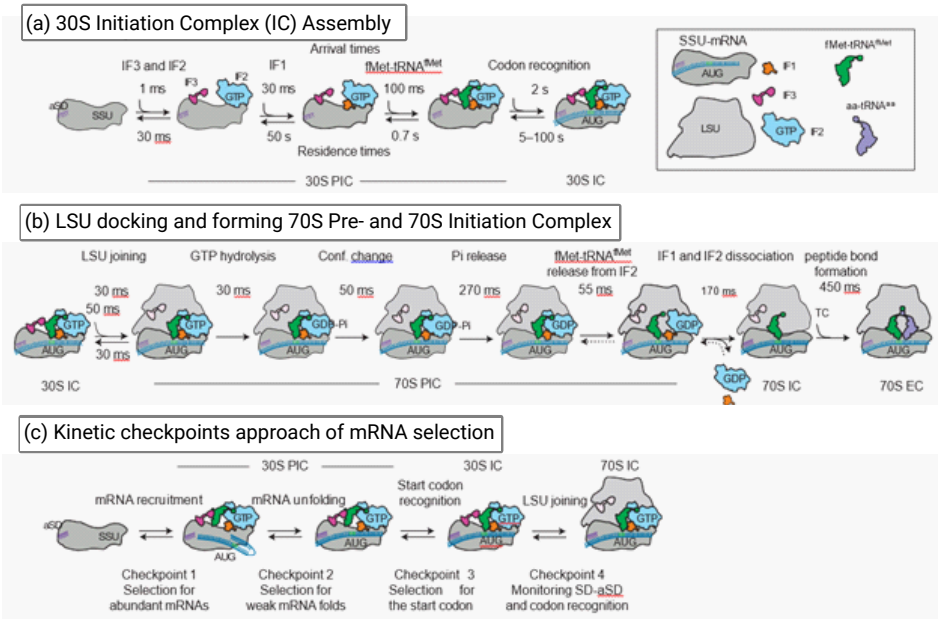


Figure 1.7: Kinetic model of translation initiation in procaryotes. (a) Assembly of the 30S preinitiation complex. The recognition of the start codon ends the transition to the 30S initiation complex. (b) Formation and maturation of the 70S initiation complex after GTP hydrolysis. (c) mRNA-centric and kinetic approach: structured mRNAs can be recruited to the platform of the small subunit (SSU), unfold, and then accommodate in the mRNA-binding channel of the SSU. Credits [Rodnina 2018].

Overview of initiation in procaryotes at a glance [Rodnina 2018].

(1) *Assembly of the pre-initiation (30S PIC) and initiation complex (30S IC).*
The binding of Met−tRNA$_i^{Met}$ to the 30S subunit is primarily the role of IF1, IF2 and IF3, which forms a stable ternary complex with GTP and Met−tRNA$_i^{Met}$. The SSU, IF1, IF2, IF3, and fMet- tRNAfMet form a labile 30S pre-initiation complex (30S PIC). As soon as mRNA is recruited, start codon recognition converts the 30S PIC into the stable 30S initiation complex (30S IC). The assembly pathway of the 30S PIC does not follow a strict order of factor addition. The factors can bind to the SSU independently of each other. However, there is a kinetically preferred sequence of factor association in the order IF3 and IF2, then IF1, followed by the recruitment of fMet-tRNAfMet through IF2 (Fig. 1.7 a). Occasionally, fMet-tRNAfMet can form an IF2•GTP/fMet- tRNAfMet complex, but this complex does not constitute an obligatory delivery pathway for fMet-tRNAfMet. The mRNA can bind to the SSU at any time, independent of the presence of the initiation factors. The association rate depends on the properties of the mRNA, such as the presence of secondary structures in the RBS, as well as the mRNA concentration. Codon recognition changes the conformation of the complex, stabilizes tRNA binding and destabilizes IF3 binding. IF3 changes its position on the ribosome in response to codon recognition.

(2) *Docking of the large subunit (LSU) to form the pre-initiation complex (70S PIC).*
The LSU docks onto the 30S IC. The rate depends on the presence of IF1, IF3, IF2•GTP, and fMet- tRNAfMet. In addition, the rate of subunit joining is attenuated by the mRNA depending on the sequence of the RBS, for example on the strength of the SD–aSD interactions and the length of the spacer between the SD and the start codon.

(3) *GTP hydrolysis and accommodation of Methionyl-tRNA in the P-site.*
After GTP hydrolysis by IF2, fMet-tRNAfMet accommodates in the P site. Displacement of IF3 from its 30S binding site and dissociation of IF1 and IF2 from the complex allows the ribosome to make inter-subunit bridges and leads to formation of the mature 70S IC. The irreversible steps of start-codon recognition and GTP hydrolysis promote conformational changes of the 30S subunit and induce rotation of the two subunits relative to each other. Initiation ends when the first methionyl-tRNA is accommodated at the P-site and ready for the first peptide bond formation with the next aminoacyl-tRNA to be accommodated in the A-site (see elongation below).

According to the **kinetic checkpoints approach** (Fig. 1.7 c), the initiation efficiency is determined by kinetic partitioning between the forward steps on the pathway toward the mature 70S IC, and the backward or rejection steps. The structure and thermodynamic stability of the RBS affect the association (step 1) and unfolding (step 2) of the mRNA. The identity of the start codon determines the stability of the codon–anticodon complex (step 3). Finally, the overall conformation of the 30S IC, which is modulated by the sequence context of the RBS, defines the rate of LSU joining (step 4). The kinetic model can explain any variations in the translational efficiency of different mRNAs. If the rate constants of the elemental steps are known, the translational efficiency can be predicted. In the few cases where such measurements were possible, the calculated value matched well with the directly measured translational efficiency [Milón et al. 2008;Milón and Rodnina 2012].

A key question is which features of the mRNA determine its translational efficiency. In bacteria, the RBS spans nucleotides –20 to +15 around the translation start codon. Translational efficiency is modulated by the nature of the codon used for initiation (AUG, GUG, or UUG), the SD sequence and the spacer between the SD sequence and the start codon, the mRNA secondary structure near the start site, and A/U rich elements in the mRNA that are recognized by the SSU protein bS1. bS1, which is the largest and most acidic ribosomal protein, is required for the binding and unfolding of structured mRNAs.

1.2.6 Initiation in eucaryotes

Initiation in eucaryotes is not fully elucidated yet. We describe two different control mechanisms by which ribosomes are recruited on a transcript to initiate translation in eucaryotes. These two initiation control mechanisms are called the **5m7G Cap-Dependent 5' UTR canonical scanning mechanism** and the **TOP (Terminal Oligopyrimidine Tract) mRNA Translation Initiation mechanism**, respectively. Both mechanisms are cap-dependent and share some common factors but are distinct in the way they recruit ribosomes and their response to nutrient and growth signals through the mTORC1 pathway. The former is a more general mechanism and is less specific in terms of which mRNAs are translated. On the contrary, the latter is more specific and is typically related to the components of the translation machinery, and ribosome biogenesis (ribosomal proteins and elongation factors).

1.2.6.1 Canonical cap-dependent scanning mechanism of initiation

The vast majority of eukaryotic messenger RNAs (mRNAs) initiate translation through a **canonical, cap-dependent scanning mechanism** requiring a free 5' end and 5' cap and several initiation factors to form a translationally active ribosome. In eucaryotes, at least eleven initiation factors (eIFs, where prefix *e* stands for eukaryotic) are necessary, several of which consist of multiple polypeptides.

It is worthwhile to first recall the roles and functions attributed to the translation initiation factors [Dever 2002]. To start translation, the small subunit (40S) recruits an **initiation complex** consisting of at least 11 initiation factors (IF). This assembly occurs in a stepwise manner. In the beginning, the initiation factors eIF1A and eIF3 induce the dissociation of the ribosomal subunits so that mRNA can bind to the ribosomal translation apparatus. The initiation complex recognizes mRNA by a 5'-terminal 7 methylguanosine nucleotide residue called **m^7GpppX cap**, which binds the heterotrimeric initiation factor eIF4F. This initiation factor (eIF4F) is composed of:

- the cap-binding protein eIF4E
- the RNA helicase eIF4A (a DEAD-box ATPase breaking up secondary structures in the mRNA molecule that would otherwise hinder translation. **DEAD-box proteins** are highly conserved ATPases found in pro- and eukaryotes. Their name comes from a characteristic D-E-A-D (Asp-Glu-Ala-Asp) amino acid sequence in their catalytic domain. Most of them are helicases.
- the scaffold protein eIF4G

In cooperation with eIF3, an ATPase made of 12 subunits, the initiation factor complex eIF4F recruits the 40S ribosomal subunit, which has bound additional initiation factors such as eIF1, eIF1A, and eIF5 as well as a complex of initiator tRNA (methionyl-tRNA in eukaryotes) and the GTP-loaded GTPase eIF2, a typical G-protein.

Translation initiation on most eukaryotic mRNAs involves the **m^7GpppX cap** (where X is the primary transcript + 1 nucleotide). Cap recognition by initiation factor eIF4F leads to recruitment of the small ribosomal subunit, in its 43S pre-initiation complex (PIC). eIF4F is composed of three subunits. eIF4E binds the cap structure. eIF4G, a large, multidomain protein, binds eIF4E, mRNA, and eIF4A, a DEAD-box RNA helicase. eIF4G also directly contacts the 43S PIC. The resulting 48S PIC is thought to scan linearly through the mRNA 5′ leader to the start codon. Cap recognition, PIC recruitment and start-codon selection are each major targets for translational control mechanisms. Figure 1.8 shows a schematic representation of initiation in eucaryotes [Zhang et al. 2019].

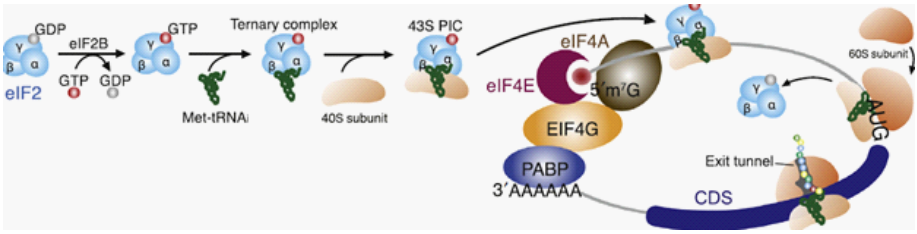


Figure 1.8: Summary of the processes occurring during translation initiation described in reference [Wang et al. 2022] for Eucaryotes. During initiation, the eukaryotic 40S ribosomal subunit forms a 43S pre-initiation complex (PIC) with eukaryotic initiation factors (eIFs) 1, 1A, 3, 5, and the eIF2•GTP•Met-tRNAi ternary complex (TC). The 43S PIC is recruited to the mRNA and must locate the correct start codon, forming a 48S PIC with the start codon base paired to the anticodon of Met-tRNAi in the ribosomal peptidyl site (P site). This is followed by the joining of the catalytic 60S subunit to assemble an 80S initiation complex (80S IC) competent for polypeptide elongation. Extensive biochemical, structural, and genetic analyses suggest that the 43S PIC binds near the 5' end of a capped mRNA facilitated by a multi-protein eIF4F complex (eIF4A, eIF4E, and eIF4G) and moves directionally through the 5' untranslated region (5′ UTR) in search of the first encountered start codon, in a process called scanning. Initiation in cells occurs in the range of ~ 30 s, and scanning ribosomes must navigate a range of 5' UTR lengths with secondary structures that generally impair translation. Translation is stimulated by RNA helicases (e.g., eIF4A and Ded1p in yeast) that couple putative RNA unwinding activities to ATP hydrolysis, but their precise action during scanning and initiation remains enigmatic. Note that a poly-A binding protein (PABD) is bound to the poly-A tail at the 3'end of the mRNA, stabilizing it. Furthemore, upon interaction between PABD with EIF4G, a circularized transcript results which helps ribosome recycling. Reproduced from reference [Zhang et al. 2019].

Canonical cap-dependent scanning mechanism of initiation in eucaryotes at a glance.

Summary of comprehensive reviews of translation initiation in eucaryotes [Dever 2002; Pelletier and Sonenberg 2019]: **(1)** *Forming the ternary complex*. The binding of Met−tRNA$_i^{\text{Met}}$ to the 40S subunit is primarily the role of eIF2, which forms a stable ternary complex with GTP and Met−tRNA$_i^{\text{Met}}$. Interestingly, the Met−tRNA$_i^{\text{Met}}$ is delivered directly to the P-site of the ribosome in contrast to the delivery of aminoacyl-tRNAs to the A-site in elongating ribosomes. Factor eIF1A, an ortholog of the procaryotic factor IF1 which binds to the A-site of the small subunit of the ribosome (SSU), promotes binding of Met−tRNA$_i^{\text{Met}}$ to the 40S subunit by helping to generate a pool of free 40S subunit and by stabilizing ternary complex binding to the 40S subunit. The eIF3 also promotes binding of the eIF2 ternary complex to the 40S subunit.

(2) *Recruiting small ribosome subunit and ternary complex near 5' end mRNA*. The binding of this 40S • Met−tRNA$_i^{\text{Met}}$ • initiation factor complex to the mRNA is promoted by the **eIF4 factors** and by eIF3. The factor eIF4E directly binds the **m^{7}GpppX cap** of the eucaryotic mRNA and, through an interaction with the N-terminus of eIF4G, recruits the cap-binding complex **eIF4F**, a heterotrimeric complex composed of **eIF4E**, **eIF4G** and the DEAD box RNA helicase **eIF4A**. **The eIF4F complex is composed of eIF4A, eIF4E, and eIF4G**. The eIF4A, in conjunction with the RNA binding proteins eIF4B or eIF4H, unwinds RNA secondary structures near the 5′ end in the 5′-UTR of the mRNA, while eIF4G binds to the factor eIF3, which in turn associates with the 40S small ribosomal subunit (SSU). To sum up: the eIF4 factors working together with eIF3 enable the 40S SSU to bind near the 5′ end of the mRNA. The eIF2-dependent binding of Met−tRNA$_i^{\text{Met}}$ to the 40S SSU appears to be required for the subsequent binding of the mRNA.

(3) *Scanning*. Once bound near the 5′ end of an mRNA, the 40S complex scans down the mRNA in a 5′ to 3′ direction searching for the `AUG` start codon. The scanning rate is $\sim$ 100 nt/s. This scanning process is dependent on ATP hydrolysis. It is not known whether eIF4A (or another ATPase) facilitates the scanning process by melting mRNA secondary structures or by actively propelling the ribosome. On most mRNAs, translation initiates at the AUG codon closest to the 5′ end and proper recognition of this first start codon is dependent on base pairing contacts with the anticodon of the tRNA$_i^{\text{Met}}$ as well as factors eIF1, eIF1A, eIF2 and eIF5.

(4) *Ternary complex dissociation*. Upon `AUG` codon recognition, the GTP associated with eIF2 is hydrolyzed to GDP in a reaction that requires eIF5. Following the GTP hydrolysis, many or all the factors will dissociate from the 40S SSU, leaving the Met−tRNA$_i^{\text{Met}}$ in the P-site base paired to the AUG start codon on the mRNA.

(5) *Large subunit joining*. The 60S large ribosome subunit (LSU) joining is catalyzed by eIF5B, an ortholog of procaryotic IF2. Following 60S subunit joining, eIF5B hydrolyzes its GTP and is released from the 80S ribosome, which is now ready to elongate. This is the end of initiation.

1.2.6.2 TOP mRNA mechanism of initiation (mRNAs specific and mTOR sensitive)

TOP (Terminal Oligopyrimidine Tract) mRNAs are a specific class of mRNAs characterized by a cytosine-rich sequence at their 5' untranslated region. The TOP motif begins with a m^7G cap C nucleotide followed by a run of approximately 4-15 pyrimidines often followed by a G rich region. They typically encode components of the translational machinery, such as ribosomal proteins and elongation factors. **All transcripts coding for the human ribosomal proteins have a leader sequence 5' UTR with the TOP motif**. This is also the case for multiple subunits of initiation factors like eIF3, eIF4A, eIF2 and the poly-A binding protein PABP. TOP motif mRNAs are regulated in a cap-dependent but distinct manner compared to the canonical scanning mechanism. They are particularly sensitive to the activity of the mechanistic target of rapamycin complex 1 (mTORC1) signaling pathway [Cockman et al. 2020; Meyuhas and Kahan 2015]. Under favorable growth conditions, mTORC1 is active and phosphorylates downstream targets, including 4E-BP1 (eIF4E-binding protein 1). When 4E-BP1 is phosphorylated, it releases eIF4E, allowing eIF4E to bind to the 5m7G cap of TOP mRNAs and initiate translation. Additionally, specific RNA-binding proteins, like La-related protein 1 (LARP1), can bind to the 5' TOP motif and either promote or inhibit the translation of TOP mRNAs in response to mTOR signaling. The recruitment of ribosomes to TOP mRNAs is also modulated by factors such as eIF4E availability and eIF4G association, but it is particularly responsive to nutrient and growth signals through the mTORC1 pathway [Cockman et al. 2020]. Figure 1.9 schematically illustrates the activation and inactivation mechanisms of TOP mRNA initiation. Note the circularized structure of the mRNA during the initiation process in the figure.

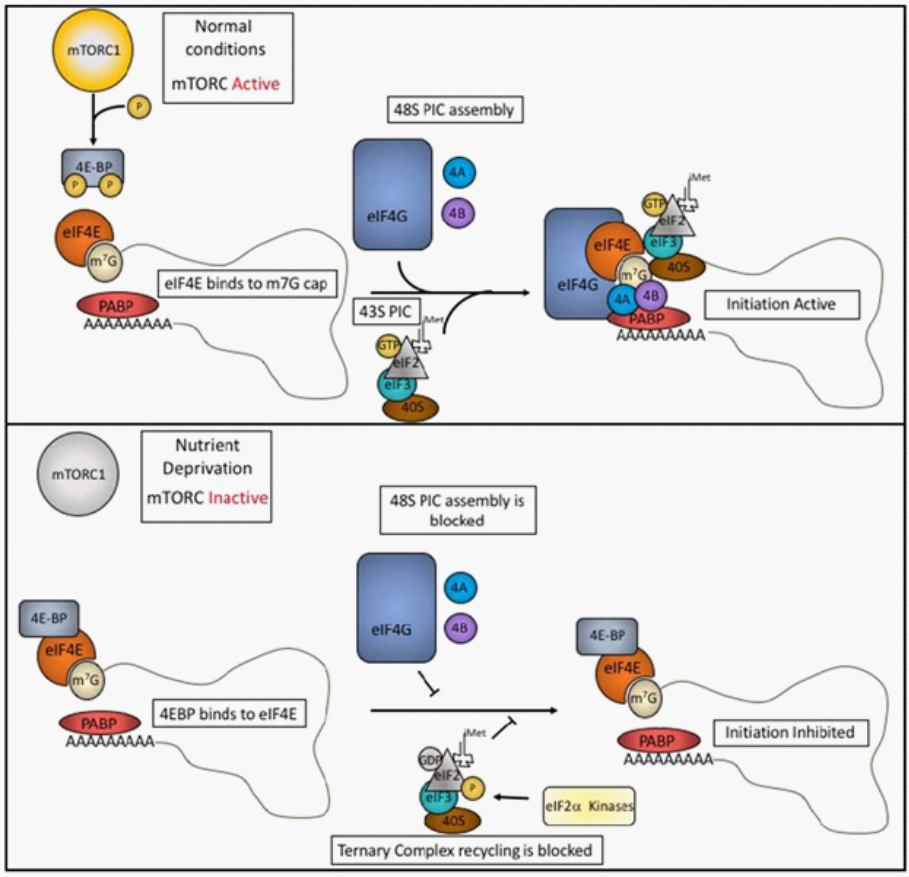


Figure 1.9: Upper panel: Under normal conditions mTORC1 phosphorylates and inactivates 4E-BPs. eIF4E is then able to bind to the m7G-cap of the transcript and to recruit eIF4G, eIF4A (4A), eIF4B (4B), forming the eIF4 cap binding complex. eIF4F recruits the 43S pre-initiation complex (PIC) consisting of eIF3, the ternary complex (eIF2, GTP and the initiator Methionyl-tRNA) and the 40S small ribosomal subunit to form the 48S PIC. Lower panel: During nutriment deprivation, mTORC1 is inactivated resulting in hypo-phosphorylated 4E-BP. 4E-BP can the bind to eIF4E and block assembly of eIF4F, halting translation initiation. The eIF2α kinases are activated during stress responses and phosphorylate eIF2α. This phosphorylation interferes with GDP exchange and renders the ternary complex inactive. Reproduced from reference [Cockman et al. 2020].

1.2.7 Energy budget of initiation

The translation initiation in eucaryotes consumes two GTP molecules per transcript initiation, one each by eIF2 and eIF5B, whereas, in procaryotes, initiation consumes one GTP per transcript [Dever 2002].

1.3 Ribosome elongation cycle

Elongation is discussed in details in reference [Rodnina 2018]. Elongation entails repetitive cycles of **decoding**, **peptide bond formation**, and **translocation**. Elongation begins as soon as the second codon of the ORF becomes accessible for reading by elongator aa-tRNAs and ends when the ribosome arrives at the stop codon. The elongation mechanisms are very similar in procaryotes and eucaryotes. There are three functional sites in the large subunit of the ribosome (LSU) accommodating three tRNA molecules: The A-site (amonoacyl site), the P-site (peptidyl site) and the E-site (exit site), where the letters refer to aminoacylated-tRNA accommodation (A-site), peptidyl-tRNA (P-site) and tRNA exit site (E-site) respectively: Fig. 1.3 and Fig. 1.10.

The three most important homologous elongation factors taking part in the elongation cycle by ribosomes are `EF-Tu/eEF1A`, `EF-G/eEF2`, `EF-P/eiF5A`, for procaryotes/eucaryotes respectively.

The three main sub-steps in elongation sequentially cycle as shown in Fig. 1.10.

(i) **STEP 1: charged tRNA accommodation, decoding and proofreading**: Accommodation involves a ternary complex. In procaryotes, the ternary complex is composed of EF – Tu•GTP•aminoacyl – tRNA (transfer tRNA). In eucaryotes, the ternary complex is an eIF1A • GTP • aminoacyl – tRNA (transfer tRNA). Accommodation at the A-site of the ribosome starts with the sampling-rejection and eventually will result in the binding of the ternary complex. Matching is checked for Watson-Crick (plus wobble base) base pairing (proofreading) between the anticodon loop of the cognate-tRNA or near- and non- cognate tRNA and its cognate or near-cognate codon on the decrypted mRNA. Following GTP hydrolysis and release of an EF – Tu • GDP (procaryotes) or an eIF1A • GDP (eucaryotes) binary complex, the aminoacyl-tRNA is accommodated into the A site, and the binary complex is replenished to a GTP binary complex pool with the help of the exchange factor IF – Ts (procaryotes) or eEF1B (eucaryotes).

(ii) **STEP 2: peptide bond formation** between the peptidyl-tRNA at the P-site and the aminoacylated-tRNA at the A-site. During catalysis of the peptide bond formation, the A- and P-site tRNAs shift into hybrid states with the acceptor ends of the tRNAs moving to the P and E sites, respectively. The reaction on

the ribosome LSU catalytic site is accelerated about 10^7 fold as compared to bare substrates in solution in the absence of ribosomes. Substrate positioning for peptide bond formation is aided by binding of the factor EF-P (procaryotes) or eIF5A and its hypusine addition (eucaryotes) in the E site. The reactivities of natural aminoacids in the peptidyl transferase reaction differ substantially. Nevertheless, the ribosome can make peptides with most amino acid combinations without the help of any additional auxiliary factors. One notable exception is the synthesis of poly-proline stretches.

(iii) **STEP 3**: eviction of the deacylated-tRNA at the P site and ribosome **translocation** to the next codon. In procaryotes, translocation is promoted by EF-G at the cost of GTP hydrolysis. The movement of tRNAs and the mRNA during translocation is a multistep complex process. In eucaryotes, following peptide bond formation, the factor eIF2 • GTP with its diphtamide modification (eucaryotes) binds in the A-site and promotes translocation of the 2 tRNAs (in A and P-sites) into the canonical P and E-sites respectively. Following release of the deacylated tRNA from the E site, the next cycle of elongation can start again with binding of the appropriate ternary complex to the A site.

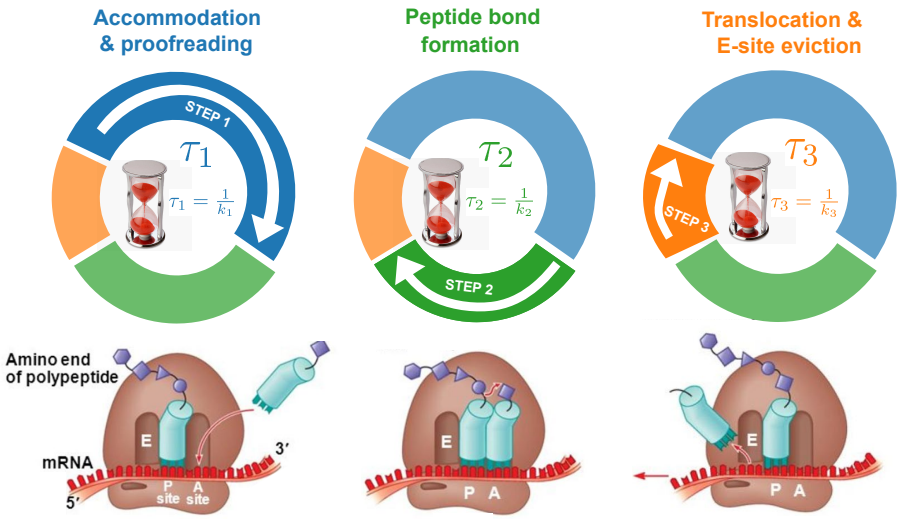


Figure 1.10: The ribosome elongation cycle. Step 1: accommodation & proofreading. Step 2: peptide bond formation. Step 3: translocation and tRNA evection. Each sub-step is approximated by a first order kinetic reaction with rate constants k_1, k_2 and k_3, respectively. The sequential time constants for each sub-step are τ_1, τ_2, τ_3. The three lower panel insets were reproduced from [Campbell et al. 2020].

During decoding, the ribosome engaged in sampling and rejection has to select an aminoacyl-tRNA cognate to the given codon from the pool of different aminoacyl-tRNAs (aa-tRNAs). The fidelity of aa-tRNA selection is high, with error frequencies of the order of magnitude $\sim 10^{-3}$ or even less. It is still an unresolved question how the ribosome responds to codon-anticodon mismatches. The non-cognate tRNAs are rejected, but how is not fully understood.

1.3.1 Mechanism of the peptide bond formation at the LSU catalytic site

The ribosome's active site is composed of rRNA and the ribosome's LSU is the largest known RNA-catalyst and the only known natural ribozyme that has polymerase activity [Rodnina 2018]. The catalysis is mainly entropic [Sievers et al. 2004]. The ribosome facilitates the reaction by ordering water molecules, positioning of rRNA residues and tRNA substrates and precise electrostatic interaction [Sharma et al. 2005; Wallin and Aqvist 2010]. Figure 1.11 shows the two alternative models for the proton transfer during peptide bond formation. The exact mechanism is still a debated question and the possible role of the electrostatic environment around the catalytic site is still unclear. It is worth noting that in the proton wire model [Polikanov et al. 2014], Fig. 1.11 (b), the concerted movement of electrons (or protons) almost align in the same direction. This direction is hypothetically consistent with a local electric field dominantly pointing in the opposite direction of the movement of the electrons.

1.3.2 Factors affecting the elongation cycle

Elongation rates are not uniform across different transcripts nor even within a single transcript. The elongation rates vary as the ribosome traverses an mRNA and ribosomes pause or stall at certain locations. The factors causing the variation in this elongation rate, most cited in the literature, are the mRNA secondary structure, the interaction of the nascent peptide with the ribosome peptide exit tunnel, certain combination of codons that are either poorly adapted to the tRNA pool or such as the ones coding for proline. During the synthesis of poly-proline stretches with three or more consecutive prolines or of distinct XPPX sequences with two prolines flanked by specific amino acids, the ribosome stalls because of very low rate of peptide bond formation. For example, for the PPP motif, the ribosome is stalled after incorporation of the second proline. The stalling is alleviated by EF-P (procaryotes), a specialized translation factor that enters the E-site of the ribosome and acts by entropic steering of the P- and A-sites substrates toward a catalytic productive orientation in the peptidyl transferase center. The eucaryotic homolog of EF-P is eIF5A and also accelerates the formation of poly-proline motifs. Both EF-P and eIF5A are posttranslationnaly modified: lysil-lysine

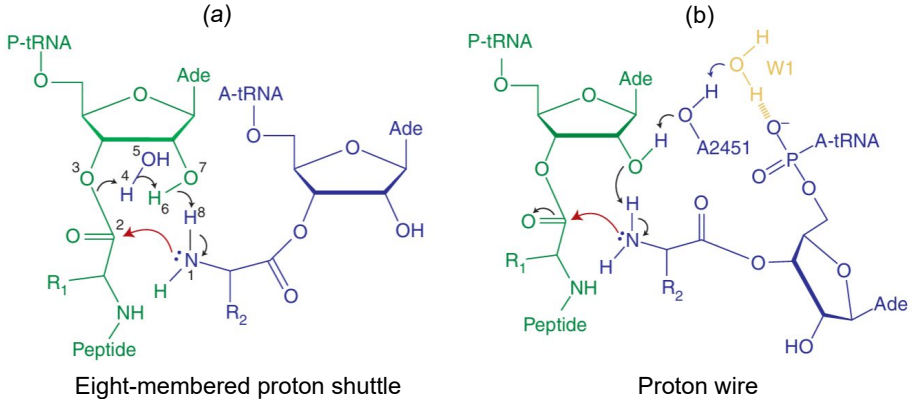


Figure 1.11: Mechanism and reaction scheme for the peptide bond formation at the catalytic pocket of the ribosome. Models for proton transfer during peptide bond formation. The nucleophilic attack is shown by the red arrows. Recall that all the arrows in the figure show the movement of electrons as is the convention in organic chemistry. (a) Eight-membered proton shuttle. In this model, the attack of the $\alpha-$amino group on the ester carbonyl carbon results in an eight-membered rate limiting transition state in which the proton from the $\alpha-$amino is received by the 2′OH group of A76 tRNA at the P-site, which at the same time donates its proton to the carbonyl oxygen of the P-site substrate, via a single water molecule close to the ester bond. (b) Proton wire [Polikanov et al. 2014]. The proton from the $\alpha-$amino group is also received by the 2′OH group of A76, which in turn donates a proton to the 2′OH of A2451 23S/28S rRNA, and then to a water molecule whose polarity is increased by the negative charge of one of the non-bridging oxygen from the phosphate moiety in the last phosphodiester bond of aa-tRNA at the A-site. Adapted from references [Polikanov et al. 2014; Rodnina 2018].

for EF-P and hypusine for eIF5A. Silencing or knocking down the expression of EF-P or eIF5A significantly affects elongation. More on this in chapter 7.

Furthermore, local accumulation of multiple ribosomes due to ribosome stalling or pausing is intertwined with temporary convergence of initiation rates and elongation rates and with high levels in ribosome pools. Each one of these factors is the topic of a dedicated chapter in this thesis.

1.4 Termination

Termination occurs when the ribosome encounters a stop codon in the mRNA. In bacteria, stop codons are recognized by the termination (or relase) factors RF1 and RF2, which read the codons UAG/UAA and UGA/UAA respectively. Another termination factor, RF3, facilitates turnover of RF1 and RF2 but is not required for peptidyl-tRNA hydrolysis. The mechanism of termination entails three steps:

(i) recognition of the stop codon

(ii) hydrolysis of the ester bond of the peptidyl-tRNA

(iii) dissociation of RF1/RF2 with the help of RF3.

The first two steps are accomplished by RF1 or RF2. RF1 and RF2 select the respective stop codons by conserved recognition motifs, PVT in RF1 or SPF in RF2 and the fact that the uracil in the first position of all three stop codons is recognized by the amino terminus of helix $\alpha 5$ of RF1 or RF2. The peptidyl-tRNA hydrolysis is catalyzed by the peptidyl transferase center of the ribosome with the help of the GGQ motif that is conserved in RD1 and RF2. RF3 is required to release RF1/RF2 from the ribosome.

The mechanism of translation termination appears different in prokaryotes and eukaryotes, where only two factors, eRF1 and eRF3 (peptide release requires GTP hydrolysis by eRF3 in eukaryotes) are responsible for termination on all three codons.

1.5 Energy budget of an elongation cycle

The ribosome is a complex macromolecular machine that requires energy to carry out its multiple tasks. During elongation, a ribosome has to translocate the mRNA each time a codon has been paired to its cognate or semi-cognate tRNA and has to push the nascent protein through the exit tunnel. The detailed energy balance (energy sources and uptakes) required for elongation has not been fully resolved [Joiret et al. 2022b; Kaiser and Tinoco 2014]. A single round of the elongation cycle requires two energy rich GTP molecules that are used within the ribosome large subunit. In the cytoplasm, the preliminary aminoacylation of a single tRNA requires two ATP energy rich molecules. The net energy budget for a nascent protein chain to be elongated by one amino acid is four energy unit equivalents of ATP molecules. Two phases are needed to deliver the required energy to the ribosome. The first phase occurs in the cytoplasm, the second in the ribosome itself.

1.5.1 Aminoacyl-tRNA synthetases first activate amino acids and bind them to specific tRNAs

This first phase occurs in the cytoplasm. The formation of a peptide bond between the amino group of one amino acid and the carboxyl group of another is thermodynamically unfavorable [Berg et al. 2019]. This thermodynamic barrier is overcome by activating the carboxyl group of the precursor amino acids. The activated intermediates in protein synthesis are amino acid esters, in which the carboxyl group of an amino acid is linked to either the 2'- or the 3'-hydroxyl group of the ribose unit at the 3'-end of the tRNA. The aminoacyl group can migrate rapidly between the 2'- and the 3'-hydroxyl groups. The activated intermediate is called aminoacyl-tRNA (Figure 5.2 (b) in chapter 5).

The attachment of an amino acid to a tRNA is important not only because this activates its carboxyl group so that it can form a peptide, but also because amino acids by themselves cannot recognize the codons on mRNA. Rather, amino acids are carried to the ribosomes by specific tRNAs, which do recognize codons on mRNA and thereby act as adaptor molecules. More on this in chapter 5.

The activation of amino acids and their subsequent linkage to tRNAs are catalyzed by specific *aminoacyl-tRNA synthetases*, also called *activating enzymes*. For most synthetases, the first step is the formation of an *aminoacyl-adenylate* from an amino acid and ATP. This activated species is a mixed anhydride in which the carboxyl group of the amino acid is linked to the phosphoryl group of AMP (Figure 1.12). For other synthetases, the reaction of ATP, amino acid, and tRNA occurs without a detectable aminoacyl-adenylate intermediate.

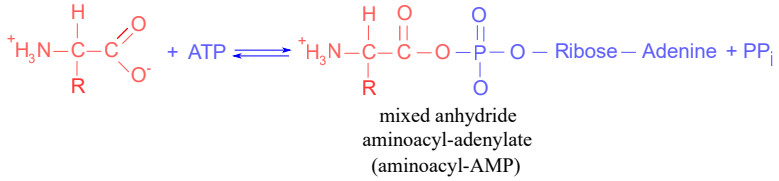


Figure 1.12: First step in the activation of an amino acid before its linkage to a tRNA. A mixed anhydride is formed between the carboxyl group of the amino acid and the phosphoryl group of an AMP molecule which requires the cleavage of an ATP molecule and the release of a pyrophosphate, eventually hydrolyzed into two inorganic phosphates driving the global reaction forward.

The next step is the transfer of the aminoacyl group of aminoacyl-AMP to a tRNA molecule to form aminoacyl-tRNA, the activated intermediate in protein synthesis. Whether the aminoacyl group is transferred to the 2'-hydroxyl or the 3'-hydroxyl group of the ribose unit at the 3'-end of tRNA depends on the particular species. In any cases,

the activated amino acid can migrate very rapidly between the 2'- and 3'-hydroxyl groups.

$$\text{aminoacyl-AMP} + \text{tRNA} \rightleftharpoons \text{aminoacyl-tRNA} + \text{AMP} \tag{1.1}$$

The sum of these activation and transfer steps is

$$\text{amino acid} + \text{ATP} + \text{tRNA} \rightleftharpoons \text{aminoacyl-tRNA} + \text{AMP} + \text{PP}_i \tag{1.2}$$

The ΔG^0 of this reaction is close to 0. This comes from the fact that the free energy of hydrolysis of the ester bond of aminoacyl-tRNA is similar to that of the terminal phosphoryl group of ATP. It is the hydrolysis of pyrophosphate that drives the synthesis of the aminoacyl-tRNA. The sum of these three reactions is highly exergonic:

$$\text{amino acid} + \text{ATP} + \text{tRNA} + H_2O \rightleftharpoons \text{aminoacyl-tRNA} + \text{AMP} + 2\,\text{P}_i \tag{1.3}$$

Thus, two high-energy phosphoryl groups are consumed in the synthesis of an aminoacyl-tRNA. One of them is consumed in forming the ester linkage of aminoacyl-tRNA, whereas the other is consumed in driving the reaction forward.

The activation and transfer steps for a particular amino acid are catalyzed by the same aminoacyl-tRNA synthetase. In fact, the aminoacyl-AMP intermediate does not dissociate from the synthetase.

The correct translation of the genetic message depends on the high degree of specificity of aminoacyl-tRNA synthesases. These enzymes are highly selective in their recognition of the amino acid to be activated and of the prospective tRNA acceptor having the correct anticodon.

1.5.2 Two elongation factors with GTP-ase activity drive the elongation cycle in the ribosome

The second phase occurs in the ribosome. The energy is found in the biochemical reactions taking place in the ribosome with the help of the associated catalytic sites of enzymes like the elongation factors (eEF in eukaryotes) or ribozymes. The elongation factors (EF and EF-G) are GTPases whose activity is controlled by the ribosome. When an aminoacyl group is hydrolyzed from the loaded tRNA, an ester group is broken and energy is released. For each amino acid incorporation cycle, two GTPs molecules are hydrolysed (one with the help of EF in the ternary complex accommodated at the A-site and one with the help of EF-G required for the mechanical translocation). The peptide bond formation itself requires free energy at each chain elongation by one residue. A very rough estimate of the net change in standard Gibbs free energy for the net balance between peptide bond formation and ester hydrolysis at pH = 7, 25°C yields $\Delta G^\circ = -3.7 \pm 1.2\,\text{kcal/mol} = -15.5 \pm 5.0\,\text{kJ/mol}$ [Kaiser and Tinoco 2014; Liu et al. 2014a]. This is known as the transpeptidation Gibbs free energy.

Peptide bond formation the formation of the simplest dipeptide glycylglycine is endergonic and requires 15 kJ/mol (3.6 kcal/mol) per mole of formed peptidic bond:
$\Delta G^\circ = +3.6$ kcal/mol for one residue incorporation (per ribosome cycle).

Hydrolysis of ester bond in aminoacyl-tRNA the hydrolysis of the ester bond in aa-tRNA is exergonic and releases 30.5kJ/mol(7.3, kcal/mol) per amino acid released from the tRNA:
$\Delta G^\circ = -7.3$ kcal/mol (per ribosome cycle).

Hydrolysis of 2 GTPs the hydrolysis of 2 GTPs is exergonic and releases 30.5 kJ/mol (7.3 kcal/mol) per mole of GTP. Hence, per residue incorporation cycle (2 GTPs):
$\Delta G^\circ = -14.6$ kcal/mole (per ribosome cycle).

Net Gibbs free energy available to the ribosome per aa residue incorporation:

$\Delta G^\circ = -18.3$ kcal/mol (per residue incorporation).

The net result is that one ester bond to the 3'-hydroxyl of a ribose has been broken (locally in the ribosome) and one peptide bond in the nascent protein has been formed, two GTPs have been hydrolyzed, the ribosome has shifted forward the **mRNA** by one codon (translocation distance on **mRNA**, $\Delta x \sim 1.4$ nm $(0.9 - 1.8)$, parenthesis indicate 95% confidence limits [Liu et al. 2014a]) and the nascent peptide has advanced in the ribosome exit tunnel by one residue (nascent peptide chain distance displacement in the tunnel at each translocation, $\Delta z \sim 0.25$ nm, which is the estimated distance between two consecutive amino acid $\alpha-$carbons). It is not fully elucidated whether (or how) free energy could be stored in the ribosome and used later to catalyze translocation and possibly assist the progression of the nascent protein through the ribosome exit tunnel when needed. Each step in translation involves intra-subunit or inter-subunit conformational changes [Desai et al. 2019; Kaiser and Tinoco 2014; Liu et al. 2014a]. Such conformational changes could store energy that could be released at a subsequent step, with a thermodynamical yield, providing a conceivable mechanism of harnessing the biochemical energy to use it for mechanical translocation and for moving the nascent peptide through the ribosome exit tunnel when required. As folding begins outside the ribosome exit tunnel, the formation of native structure can generate a mechanical pulling force [Simpson et al. 2020], potentially aiding the emergence of the polypeptide. Optical tweezers assays have opened the way to characterizing the ribosome's full mechanochemical cycle [Desai et al. 2019; Liu et al. 2014a]. Recently, such *in vitro* assays [Desai et al. 2019; Liu et al. 2014a] provided an estimate for the maximal mechanical energy required per translocation step (near stalling on the **mRNA**), 21.2 pN · nm = $5.2 k_B T$, at 296 K, or ~ 3.1 kcal/mol. As estimated above, the Gibbs free energy available from the transpeptidation step (ester hydrolysis and peptide formation without the help of GTP hydrolysis) is $\Delta G^\circ = -3.7 \pm 1.2$ kcal/mol.

The mechanical work for translocation would be around 80% of the Gibbs free energy available from the transpeptidation. Such a high thermodynamic efficiency for conversion of chemical energy to mechanical motion is higher than occurs in most molecular motors [Bustamante et al. 2004]. Instead, efficient translocation would require the hydrolysis of at least one GTP with the help of elongation factor EF-G [Liu et al. 2014a]. EF-G dependent GTP hydrolysis was shown to precede and greatly accelerate translocation [Rodnina et al. 1997]. The mechanical translocation of the ribosome on the mRNA by one codon would take $\frac{3.1\,\mathrm{kcal/mol}}{7.3\,\mathrm{kcal/mol}}$ or 43 % of the Gibbs free energy released by the hydrolysis of one GTP, assisted by elongation factor EF-G. The mechanical energy required to push the nascent peptide chain through the large subunit exit tunnel could be provided by the transpeptidation Gibbs free energy or the hydrolysis of one GTP assisted by elongation factor EF in the ternary complex accommodated in the A site or a combination of both.

1.6 Polysome versus ribosome profiling

The higher the initiation rate as compared to the elongation rate for a given transcript, the higher the density of ribosomes on this transcript. Polysome analysis first uses a high concentration of cycloheximide (elongation inhibitor) to freeze ribosomes on mRNA, followed by size-fractionation by sucrose density gradient centrifugation to separate populations of mRNAs with various numbers of ribosomes loaded. When combined with techniques to detect specific genes, such as real-time PCR and Northern blotting, this polysome profiling can examine the translation level of a specific gene. When combined with transcriptome wide analysis such as RNA-seq, the translation level of each gene in the whole genome can be addressed.

1.6.1 Differential gene expression and translational expression analysis

Molecular biologists studying gene expression analysis intuitively reason as follows. According to the scheme previously shown in Fig.1.1, the protein levels could be regulated in at least the following ways:

1. *transcription* > *translation*. In this mode, the transcription increases the relative number of a subset of specific mRNAs. The ribosome pool and the initiation rate are kept the same before and after the increase of the number of transcripts or across the experimental settings that only affected transcription. The subsequent increase in protein levels is strictly and properly caused by

transcription upregulation. The deciphering of the gene expression analysis could be done by relying only on transcriptomic data, i.e., RNA-Seq.

2. *TRANSLATION* > *TRANSCRIPTION*. In this mode, the absolute size of all subsets of transcripts are kept constant (or are at steady-state). Two possibilities arise:

   i. the relative initiation rate of a subset of transcripts increases as compared to other subsets, while the ribosome pool size is still the same. In this case, the increase in the relative protein levels is caused by upregulation of initiation rates of the subset of the transcripts that are sensitive to the privileged mode of initiation. As an example, this can occur through a major regulating cascade triggered by mTORC1, acting on 4E-BP by phosphorylation/dephosphorylation of the initiation factor like eIF4E in eukaryotes for instance, as described previously, where transcripts having the right upstream 5'-UTR leader sequences are selectively preferentially initiated (TOP-mRNAs).

   ii. the relative initiation rates are unchanged but the ribosome pool size is changed. If the ribosome pool is decreased, the transcripts subsets compete for ribosomes and the relative protein levels are expected to be highly sensitive to the specific initiation rates, to the length of the transcripts, to the differences between initiation and elongation rates, and to the ability of a transcript to circularize the ribosomes after termination, back to initiation of the same transcript. If the ribosome pool is large or infinite, the higher protein levels are expected for those proteins for which the initiation rates are close to the elongation rates.

   The alternatives (i) and (ii) are not necessarily mutually exclusive. If the transcript levels are unchanged across the experimental settings, polysome profiling data alone could be used to draw conclusion on translational control effects.

1.6.2 RNA-Seq

RNA sequencing (RNA-Seq) is the state-of-the-art high-throughput method for transcriptome profiling, enabling the quantification and characterization of mRNA molecules in a given sample. This technique involves isolating RNA, converting it into complementary DNA (cDNA), fragmenting, and sequencing it using next-generation sequencing (NGS) technologies. It allows genome-wide assessment of gene expression with single-nucleotide resolution. The method is widely used to determine **differential gene expression**. A differential gene expression analysis from RNA-Seq data involves a bioinformatics workflow. First, raw sequencing reads undergo quality control using tools like `FastQC` and `Trim Galore` to assess read quality and remove adapters. Next, reads are aligned to a reference genome using aligners such as `STAR` or `HISAT2`, or

assembled de novo if no reference is available. Aligned reads are then quantified using `featureCounts` or `HTSeq` to generate gene expression matrices. Normalization and statistical testing are performed using `DESeq2`, `edgeR`, or `limma-voom` to identify differentially expressed genes based on statistical significance thresholds (e.g., adjusted p-value < 0.05). The results undergo functional enrichment analysis via `Gene Ontology` (GO) or `KEGG` pathway analysis to interpret biological significance. Visualization tools like `ggplot2`, PCA, and heatmaps help in data interpretation. This workflow ensures robust and reproducible identification of transcriptomic changes across conditions or time series.

1.6.3 Polysome fractionation profiling

Ribosome loading on mRNA is determined by the relative rates of translational initiation and elongation, polysome analysis can be used to assess the global translational process [He and Green 2013]. Polysome fractionation profiling by sucrose density gradient centrifugation in eukaryotic cells is a technique used to separate polysomes from monosomes, ribosomal subunits and messenger ribonucleoprotein particles (mRNPs), see first three steps in Fig. 1.13. This enables discrimination between efficiently translated (associated with heavy polysomes) from poorly translated (associated with light polysomes) mRNAs. In this assay, ribosomes are immobilized on the mRNA using translation elongation inhibitors such as cycloheximide and cytosolic extracts are separated on 5-50% linear sucrose density gradients by ultracentrifugation. Subsequent fractionation of sucrose gradients allows isolation of mRNAs according to the number of ribosomes they bind to. RNA extracted from each fraction can then be used to determine changes in the distributions of mRNAs across the gradient between different conditions, whereby translational efficiency increases from the top to the bottom of the gradient. Northern Blotting or quantitative reverse transcription polymerase chain reaction (qRT-PCR) are used to determine levels of mRNA in each fraction. Numerous studies have showed that mRNA translation is regulated mainly at the rate-limiting initiation step [Sonenberg and Hinnebusch 2009]. The proportion of ribosomes engaged in polysomes positively correlates with translation initiation rates.

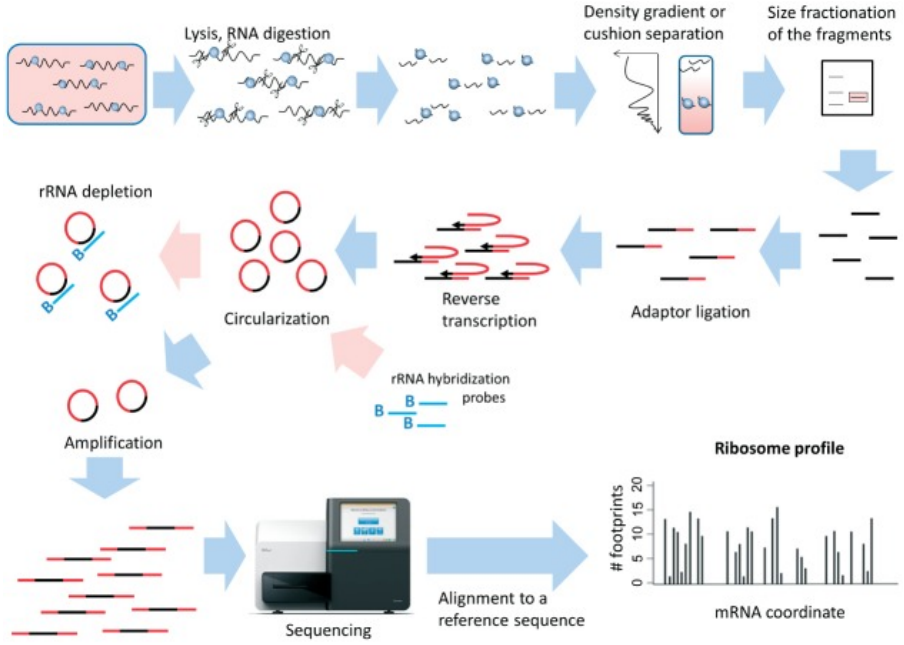


Figure 1.13: Major steps of the ribosome profiling protocol as described in [Ingolia et al. 2012]. Reproduced from [Michel and Baranov 2013].

Figure 1.14 shows an example of representative results of applying the polysome fractionation method to investigate the role of mTOR (see below mammalian target of rapamycin) signaling in mediating the effects of insulin on mRNA translation [Gandin et al. 2014]. In this example, MCF7 human breast cancer cells were maintained in low serum and then stimulated with insulin alone or in combination with the active-site mTOR inhibitor Torin1. Torin1 is an effective inducer of Autophagy and autophagosomes, as inhibitor of mTOR, mimics cellular starvation by blocking signals required for cell growth and proliferation. Non-stimulated cells that were continuously kept in low serum, served as control. mRNPs, monosome (80S) and polysome fractions were separated using the polysome fractionation method. Relative to the control cells, insulin induced an increase in absorbance in gradient fractions corresponding to polysomes, accompanied by a concomitant decrease in absorbance in the monosome fraction, Figure 1.14. These findings show that the proportion of ribosomes engaged in polysomes is increased in insulin treated cells as compared to control cells, indicating that, as expected, insulin stimulates global translation initiation rates. Torin1 reversed the effects of insulin on absorbance profiles, corroborating the findings that mTOR signaling plays a major role in mediating the effects of insulin on the translation machinery.
There is an important distinction between polysome fractionation profiling and ribosome profiling (next subsection) [ibid.]. The example presented above shows how polysome fractionation profiling can be interpreted in terms of variation in translation initiation rates. The example associates higher initiation rates to heavier polysomes. A larger number of ribosomes in transcripts reflects larger initiation rates. This interpretation is made globally for all transcripts in the case of polysome fractionation profiling.

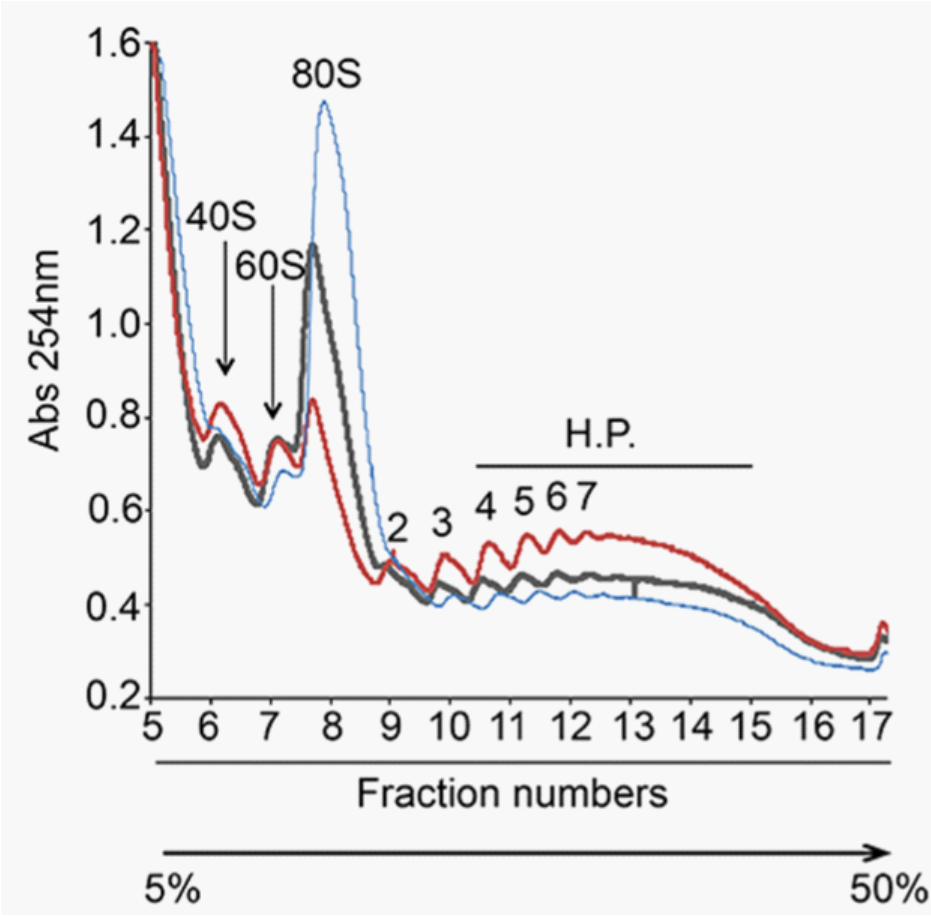


Figure 1.14: Polysome fractionation profiles on sucrose density gradient showing the effects of serum starvation, insulin and mTOR signaling on global translation in MCF7 cells. MCF7 cells were deprived of nutriments for 16 hours and treated with 5nM insulin alone (red line) or in combination with 250 nM Torin1: insulin + Torin1 for 4 hours (blue line). Untreated cells continuously deprived of nutriments were used as a control (black line). Free ribosomal subunits (40S and 60S), monosomes (80S), polysomes and heavy polysomes (HP) where the number of ribosomes in the polysome fractions are indicated. Relative to the control cells, insulin induced an increase in absorbance in sucrose gradient fractions corresponding to polysomes, accompanied by a concomitant decrease in absorbance in the monosome fraction. The proportion of ribosomes engaged in polysomes is increased in insulin treated cells as compared to control cells: insulin stimulates global translation initiation rates. Torin1 reversed the effects of insulin on absorbance profiles. Reproduced from [Gandin et al. 2014].

1.6.4 Ribosome profiling (Ribo-Seq) and ribosome footprint density maps

The salient improvement of **ribosome profiling (Ribo-Seq)** in comparison with polysome profiling is in the ability of the former **to obtain position-specific information regarding ribosome locations on mRNAs** [Kiniry et al. 2019; Michel and Baranov 2013]. The association of an mRNA transcript with ribosomes does not necessarily mean that the main open reading frame (ORF) of this mRNA is translated. Indeed, ribosomes could stall on an mRNA transcript without producing a protein. Besides, translation could occur at ORFs other than the main protein coding open reading frame (pORF). Ribo-seq, or ribosome profiling or ribosome footprinting, is an experimental technique delivering a snapshot of ribosome positions along all transcripts in a bulk of cells (typically between 5 and 15 million cells) at a given condition. Its first version was developed at the end of the 1960s to study translation initiation. It has been extended in the 1980s to investigate the role of slow codons and ribosome pausing. In 2009, Ingolia *et al*. revamped this technique to exploit the next generation sequencing, making Ribo-seq the state-of-the art technique for studying gene expression at the level of translation [Ingolia et al. 2009]. For details of the ribosome profiling experimental protocols and the primary research articles that first used them, see references [Dana and Tuller 2012; Ingolia 2010; Ingolia et al. 2012, 2011; Ingolia et al. 2009; Larsson et al. 2010; Tuller et al. 2010]. Figure 1.13 outlines the major steps of the ribosome profiling protocol as described in reference [Ingolia et al. 2012]. A concise and comprehensive guide for newcomers to the ribosome profiling wet lab and dry lab techniques is given in reference [Power 2022]. Ribo-seq is based on the principle that a translating ribosome protects a short stretch of mRNA within its structure. Once ribosomes are 'frozen' in the act of translation using translation elongation inhibitors, RNA-digesting enzymes known as RNases can be added to destroy any mRNA that is unprotected by the arrested ribosomes. After RNase digestion, ribosomes are enriched and the ribosome-protected mRNA is then isolated and converted into NGS(Illumina)-compatible cDNA libraries by reverse transcriptase. These ribosome-protected mRNA fragments are commonly called RPFs or ribosome footprints. Mapping these sequenced RPFs to the transcriptome provides a 'snapshot' of translation that reveals the position and densities of ribosomes on individual mRNAs transcriptome-wide. This snapshot can help determine which proteins were being synthesized in the cell at the time of the experiment. Ribo-seq enables the identification of alternative mRNA translation start sites, the confirmation of annotated open reading frames (ORFs), as well as upstream (uORFs) that may be involved in the regulation of translation, the distribution of ribosomes on an mRNA and the rate at which ribosomes decode codons [ibid.].

In short, **Ribo-seq** consists in isolating mRNA fragments (called "reads" or **ribosome protected fragments, RPF**) covered by a ribosome engaged in translation (the length of a ribosome footprint is $\sim$ 28 – 31 nt or about 10 codons) in the sample of bulk cells after treatment (or not) with cycloheximide (an inhibitor of elongation in eukaryotes) [Szavits-Nossan and Ciandrini 2020]. The reads or ribosome protected fragments (RPFs) are generated by RNase I treatment and analysed by deep-sequencing platforms. The sequences of the reads are bioinformatically aligned on the (known) transcriptome of the corresponding species in order to build histograms of ribosome occupancy at codon resolution on each transcript. This allows determination of the ribosome position at a single nucleotide resolution for all mRNAs (then called the translatome). Ribo-seq provided an unprecedented view on translation that led to several discoveries. Depending on the study focus, other drugs than cycloheximide can be used such as harringtonine and lactimidomycin or lactimidomycin that block the translation initiation [Dmitriev et al. 2020].

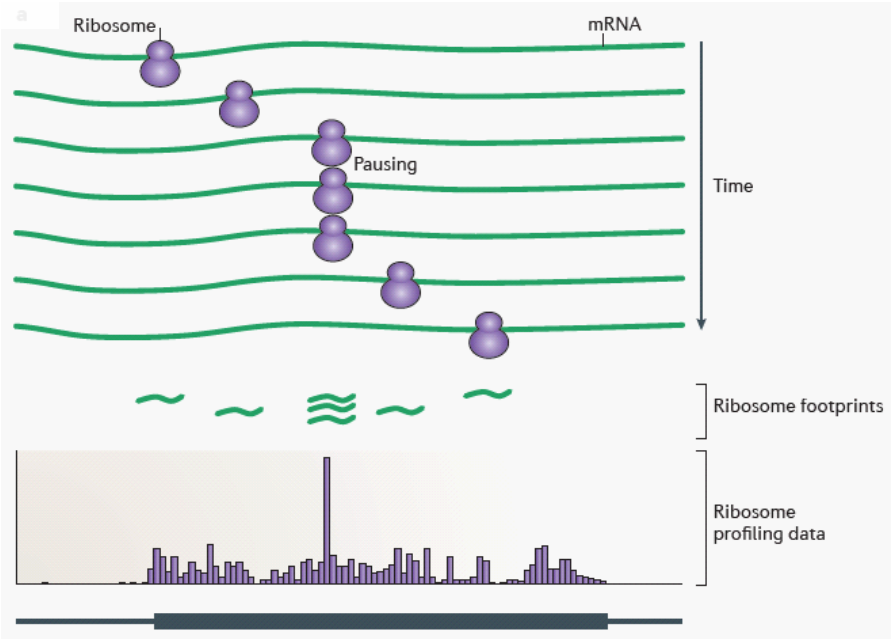


Figure 1.15: Relation between the local elongation rate and the ribosome occupancy (density) map. Reproduced with permission from [Ingolia et al. 2009].

One of the goals of ribosome profiling is to understand how the elongation rate along the transcript depends on the choice of codons. Codon elongation rates are usually estimated assuming that the ribosome density at codon i is proportional to the ribosome

dwell time on that particular codon (Fig. 1.15). This assumption follows from the conservation of the ribosome current assuming no ribosome drop-off. The drop-off rate is so small as compared to the initiation or elongation rates that this assumption is well justified.

1.6.5 Dry and wet lab ribosome profiling data analysis

Bioinformatics and computational approaches are integral to the analysis of ribosome profiling data at multiple stages, forming a core component of the Ribo-seq workflow. However, these methods pose significant challenges, as numerous alternative tools are available, and their outputs often do not converge. An extensive review of computational strategies and tools for ribosome profiling data analysis is provided in reference [Kiniry et al. 2019]. In the box below, we summarize the key steps involved in ribosome profiling analysis and highlight the important dry lab milestones in the workflow [Power 2022]. A thorough understanding of the corresponding wet lab procedures, outlined below, is essential for correctly interpreting the role and implications of these dry lab steps.

Computational methods and tools for ribosome profiling data analysis at a glance.

(1) *Ribonuclease digestion quality check and triplet periodicity check*. Ribonucleases are enzymes that break down and destroy RNA. Here, they are employed to cleave regions of mRNA that reside outside of the ribosome, ideally leaving only the fragment of mRNA that are stored within and protected by the ribosome. The choice of nuclease is important as some ribonucleases are not compatible with certain species. RNase I is a robust ribonuclease that is capable of providing good digestion in human cell types and yeast cell types but cannot be used for digestion in bacterial cell types as it is capable of damaging the ribosome. For bacteria, a different ribonuclease must be used (the much milder micrococcal nuclease must be used in this case). The quality of the ribonucleic digestion can be visualized using plots such as a triplet multiplicity plot or a metagene profile. As codons are encoded in groups of three nucleotides, a strong digestion on a metagene profile would give a very clear repeating pattern where the number of RPFs in one frame would be much higher than the other two frames.

(2) *Determining the position of the decoding center and inferring the position of A-or P-site* A metagene profile of a high-quality ribosome profiling dataset is expected to have a sharp difference in footprint density at the start and stop codons, so that the density is higher downstream of starts and upstream of

stops. Since initiation is slower than elongation, a peak of footprint density is expected at the start codon. The location of the 5′−end peak density indicates the distance between footprints 5′−ends and ribosome P-site codon where the first tRNA-Met$_i$ is being incorporated. This distance is called the offset. The offset is an integer number of nucleotides which is added to the coordinate of the 5′−end of a mapped read to infer the position of the P- or A-site of the ribosome that produced it. Applying a static offset regardless of read length is often sufficient to determine positions of A- or P-sites with an accuracy that is satisfactory for most Ribo-Seq applications. More sophisticated methods of offset determination have been developed that maximize the difference of the estimated dwell time between codons. This assumes that the A-site has a predominant role in influencing the decoding rate.

(3) *Size selection* After nuclease digestion, the samples undergo T4 PNK end repair. This tailors the ends of RNA by removing the 3'-phosphoryl groups generated from RNase I cleavage and adds a hydroxyl group to prepare them for subsequent linker ligation. This step is followed by fractionation on a 15% PAGE-urea gel (urea is added to denature the RNA, thereby preventing any secondary structures from forming). The use of a size selection marker allows for the determination of bands that are within a specific size. RPFs or footprints should be approximately 28-31 nucleotides in length. The size selection marker should contain bands that are just above and below the desired footprint size, e.g., 24 and 32 nucleotides in length. This use of bracketing the samples with the size selection markers allows for easier identification of the region of the gel to excised out. The RNA in these scalpel excised regions is then extracted and purified. All the RNAs in these regions are supposed to have the right size corresponding to a RPF.

(4) *RNA depletion (rRNA and tRNA depletion)* After recovering the RNA from gel slices, the next step is the removal of ribosomal RNA (rRNA). rRNA is indeed the most predominant RNA making up for about 80% of cellular RNA in the cell. This is an issue, as an abundance of rRNA leads to fewer RPFs being sequenced, resulting in less mapping reads, essentially reducing the useful size of the library. Most contaminating rRNAs are generated from ribonuclease digestion of the ribosome, nicking off RPF-sized fragments. One way of depleting rRNA is using biotinylated oligos designed to hybridize to the predominant rRNAs. After mixing the samples of interest with a depletion mix containing the designed oligos, the biotinylated oligo-bound rRNA can be removed using streptavidin conjugated to magnetic beads. After this mode of clean-up, the sample should be depleted of the majority of its rRNA contents, thereby allowing for more RPFs to be sequenced later on.

(5) *Linker ligation and reverse transcription* In order to convert the ribosome footprints to DNA, they must be reverse transcribed. However, reverse transcriptase requires a primer to initiate polymerization. To achieve this, either the RPFs can be tailed with a poly A polymerase or else a single-stranded RNA linker of a known sequence can be ligated to the $3'-$end of the RPF. There are several benefits of using $3'-$ linkers instead of polyadenylation tailing. One is the incorporation of random (but known) nucleotides at the $5'-$end of the linker which can act as unique molecular identifiers (UMIs) to aid removal of PCR duplicates during subsequent analysis. Random nucleotides at the $5'-$end also have the benefit of reducing potential ligation biases. Another benefit of the addition of $3'-$end linkers is that they can be designed to contain unique barcodes for each linker, allowing for multiplexing–combining different samples together into a single pool of samples for deep sequencing, prior to cDNA synthesis. Linker ligation is usually done by using an enzyme known as T4 RNA ligase truncated K227Q in conjunction with adenylated linkers to join the ends of samples to the linkers strands. The ligated product can then be isolated by either running the samples on a 15% PAGE-urea gel or via enzymatic digestion to cleave and remove any non ligated linkers. Purified linket-ligated RPFs can then be converted into DNA by standard reverse transcriptase reaction.

(6) *Polymerase chain reaction and NGS sequencing* Following cDNA synthesis, libraries are amplified by polymerase chain reaction (PCR). PCR is a reaction that amplifies DNA exponentially, causing it to double in size in every cycle. PCR is done in three stages (*denaturation*, *annealing* and *extension*) in which the DNA is subjected to rapid cycles of heating and cooling. The three stages are cycled through until a desired concentration of library is made. This library can then be sequenced to generate a bioinformatic library containing RPF sequenced reads.

(7) *Bioinformatics data analysis* A general flow of ribosome profiling mapping pipeline typically include the use of the following softwares or algorithms:

- ***FastQC*** to determine basic quality metrics like read lengths and sequencing quality (Phred score, base calling, sequencing depth).
- ***Cutadapt*** to demultiplex and cut away adapter sequences added during linker ligation and PCR.
- ***Bowtie and STAR*** [Dobin et al. 2012; Langmead et al. 2009] to bioinformatically first remove remaining rRNA contaminants by using a

short read aligner and then, to align and map the reads to the corresponding organism's annotated transcriptome.

- ***Samtools*** to convert the aligned reads (SAM: Sequence Alignment Map file) into a sorted BAM file (compressed binary version of the sequence alignment file).
- ***HT-seq*** generates a gene count file that counts the number of reads aligned to each gene.

There are a number of online browser-based platforms for visualizing ribosome profiles showing the mRNA positions mapped RPFs allowing for further metadata analysis, often requiring specific files such as Fasta files.

Ribosome profiling (Ribo-Seq) is often used for the three following applications

- **Identification of translated open reading frames:** Detecting translation using Ribo-Seq data is not straightforward as the presence of a footprint in a given genomic region does not necessarily mean that that region is being translated. Not all sequences found in a ribosome profiling cDNA library derive from genuine ribosome protected fragments within the ribosome mRNA channel. Contamination from rRNAs and tRNAs is very abundant. At least seventeen computational tools for identifying translated ORFs exist and briefly reviewed in reference [Kiniry et al. 2019]. The authors recognize that benchmarking studies are difficult to carry out due to a lack of gold standard sets of translated ORF and adequate methodology orthogonal to ribosome profiling.
- **Translation differential expression:** Ribosome profiling analysis is most frequently used for characterization of differential gene expression as part of a time series or control/treatment group. When RNA levels do not change, but the ribosome profiling signal changes, it is reasonable to attribute these changes to changes in translation efficiencies. Often attempts are made to measure differential translation even when RNA levels do change simply by dividing the number of ribosome footprints aligning to an ORF by the number of RNA-Seq reads. Such a procedure has several problems. One of them is that ratios do not carry information on the statistical significance. Indeed, a ratio of $2/4$ is equal to $200/400$. A second one results from the fact that this ratio has the Cauchy distribution as the best fitted statistical distribution. This is hard to model as Cauchy distributions have no defined mean or variance. The third one comes from spurious correlation between such ratios and their components (RNA levels) which lead to both false positive and negative results [Larsson et al. 2010].
- **Footprint density and ribosomal dwell times at codon resolution:** Elongation rates vary as the ribosome traverses an mRNA and ribosomes could pause or stall

at certain locations The causes of reduced or increased rates depend on factors like mRNA secondary structure, interaction of the nascent peptide with the ribosome exit tunnel, certain combination of codons and are due to the inherent differences in ribosome decoding rates at codon resolution or in the rates of peptide bond formation between different pairs of amino acid. Global assessment of footprint density allows for the magnitude of the effects of these factors to be estimated. The assumption is here that a longer dwell time of a ribosome on a codon will give rise to an increased footprint density of ribosomes on this codon, especially when all RPF are referenced on a common decoding center (P-site or A-site) and aggregated per codon type.

1.6.6 Translation efficiency

There are several methodological limitations and difficulties in the interpretation of ribosome profiling that restrict the ability of ribosome profiling to accurately estimate mRNA translation efficiency (TE), Fig. 1.16. Most importantly, whereas ribosome profiling allows direct identification of ribosome position on a given mRNA molecule, the number of ribosomes that is associated with a given mRNA is indirectly estimated by normalizing frequencies of reads in RPFs (ribosome-associated mRNA) over those observed in randomly fragmented mRNAs (total mRNA). For instance, in a sample where four copies of a "B" mRNA molecule (Ba, Bb, Bc, Bd) are occupied by four ribosomes at the position 1, 2, 3, 4, an inherent shortcoming of the ribosome profiling technique will not permit a distinction between a scenario where all 4 ribosomes associate only with a single copy of this "B" mRNA molecule, say only with Ba, at position 1, 2, 3, 4 (a polysome of 4 ribosomes on a single mRNA molecule at 4 distinct positions) and a scenario where Ba, Bb, Bc and Bd mRNAs are each occupied by a single ribosome at the position 1 for Ba, 2 for Bb, 3 for Bc and 4 for Bd (4 single monosomes on 4 separate copies of the same mRNA molecule, with all monosomes at 4 distinct positions) [Gandin et al. 2014]. In contrast, during polysome fractionation profiling, polysome integrity is preserved. This allows isolation of pools of mRNas associated with a defined number of ribosomes Figure 1.14. This important distinction between polysome fractionation profiling and ribosome profiling (Ribo-seq) suggests that whereas the former method can be used to directly compare mRNAs in messenger ribonucleoprotein particles (mRNP), light and heavy polysome (HP) fractions, the latter method will likely fail to capture changes in the translatome that are caused by mRNAs that transition from light to heavy polysomes [ibid.]. The biological significance of the results of the two methods must be interpreted with caution. Polysome fractionation profiling and ribosome profiling (Ribo-seq) are complementary methods that primarily provide information regarding the number of ribosomes associated with mRNA and position of the ribosome on mRNA, respectively. Polysome fractionation profiling can help identify the transcripts for which initiation rates is affected by a given treatment

(versus control). Ribo-seq can help provide information on the time spent by ribosomes on individual codons (individual elongation rates) and how these dwelling times are affected by a given treatment (versus control). It remains essential that the genome-wide data obtained by both procedures are adequately analysed and validated functionally and biochemically.

It is of utmost importance to emphasize that the levels of polysome-associated mRNAs are, in addition to translation, affected by transcriptional and post-transcriptional mechanisms that influence cytosolic mRNA levels. Therefore, to determine differences in translation using geneome-wide data from polysome-associated mRNA, it is necessary to correct for the effects from steps in the gene expression pathway that are upstream of translation. To allow such correction, cytosolic RNA is prepared in parallel with polysome associated RNA from each sample and the genowe-wide steady-state mRNA levels are determined. Currently, so-called translation-efficiency (TE) scores (log2-ratio between polysome-associated mRNA data and cytosolic mRNA data) are often employed to correct for the effects of changes in cytosolic mRNA levels on translation efficiency of a given mRNA.

1.6.7 Spurious correlation in translation efficiency scores

The experimental data **interpretation of differential translation** should be conducted with caution. The translational differential analysis has often been relying on simple intuitive approaches such as the one exposed in the previous section. Correction for cytosolic mRNA levels has generally been achieved by **dividing actively translated mRNA levels** (as quantified in **Ribo-Seq** by the number of ribosome protected fragments on a transcript or by the number of ribosomes bound to an individual transcript in **polysome profiling**) **by cytosolic mRNA levels** (as quantified by **RNA-Seq read counts** for that transcript), obtained in parallel, and taking the logarithm (in basis 2) of that ratio. A rigorous and well-supported argument against the use of log-ratio transformations in genome-wide translational control studies was presented by [Larsson et al. 2010]. This work systematically demonstrates how the statistical properties of difference scores (log-transformed ratios) lead to **spurious correlations**, potentially resulting in **false biological conclusions**. The authors reference Pearson's 1896 work on spurious correlation, emphasizing that the problem has been well-documented for a long time in statistics [Pearson 1896].

We recall here how the problem results from the non-trivial correlation between log ratios (translatomics relative to transcriptomics) and cytosolic mRNA levels (transcriptomics alone). Let us call Y the vector of translational activity data for a specific mRNA (as measured by the number of ribosomes on that mRNA or the number of ribosome protected fragments reads of that mRNA in Ribo-Seq data, i.e., the translatomics data). Z is the vector of the paired cytosolic mRNA data for the same mRNA (as measured

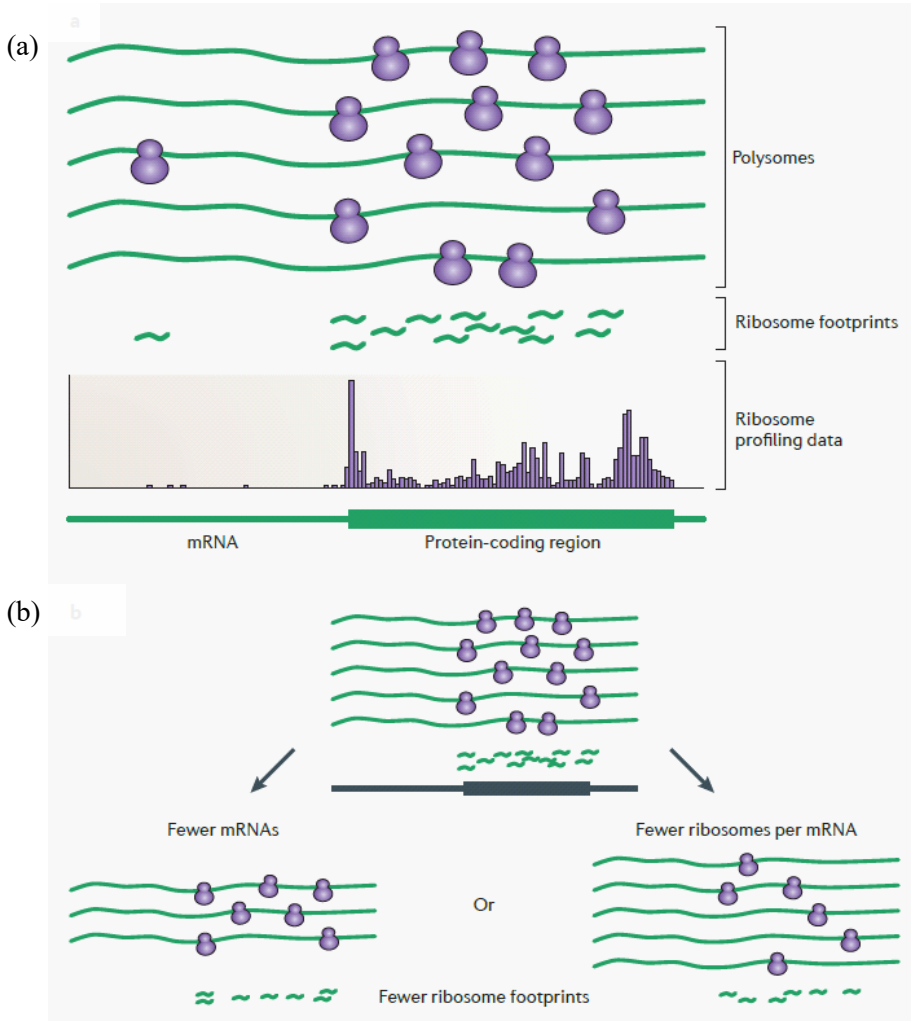


Figure 1.16: Ribosome footprinting theoretical scenarios at the crossroad of transcripts relative abundance, initiation rates, elongation rates, polysome random fragmentation generating an observed pattern of ribosome occupancy (density). Reproduced with permission from [Ingolia et al. 2009].

by RNA-Seq data, i.e., the transcriptomics data). The Pearson correlation coefficient is r and the sample standard deviation are s_Y and s_Z respectively. The correlation

coefficient between $Y - Z$ and Z is mathematically expressed by:

$$r_{(Y-Z)Z} = \frac{r_{YZ}s_Y s_Z - s_Z^2}{s_Z\sqrt{s_Y^2 + s_Z^2 - 2r_{YZ}s_Y s_Z}} \tag{1.4}$$

Equation (1.4) formally expresses the correlation between the difference or log-difference score of translatomics relative to transcriptomics and the transcriptomics score, demonstrating that this correlation is influenced by the standard deviations of both variables and their Pearson correlation coefficient. When translational and cytosolic levels are uncorrelated, Eq. (1.4) simplifies to:

$$r_{(Y-Z)Z} = \frac{-s_Z}{\sqrt{s_Y^2 + s_Z^2}} \tag{1.5}$$

Equation (1.5) shows that when translational activity and cytosolic mRNA levels are uncorrelated, the correlation of the log-difference score with cytosolic mRNA levels depends purely on their relative standard deviations. This makes the origin of these false negatives and false positives clear. When the standard deviations are also equal, Eq. (1.5) yields a correlation of $-1/\sqrt{2} = -0.71$. That is, half (-0.71^2) of the variance associated with the log ratios arises in the more typical situation in which the correlation between translational activity data and cytosolic mRNA is nonzero. When standard deviations are also equal, Eq. (1.4), then simplifies to:

$$r_{(Y-Z)Z} = \frac{1}{\sqrt{2}}\frac{(r_{YZ} - 1)}{\sqrt{1 - r_{YZ}}} \tag{1.6}$$

Equation (1.6) generalizes to the case where translational activity and cytosolic mRNA are correlated, confirming that log-ratio transformations do not effectively remove the confounding effect of cytosolic mRNA levels. As an example, when the correlation between translational activity and paired cytosolic mRNA is 0.60, Eq. (1.6) yields a correlation of -0.45. Thus, under various realistic scenarios, the correlation between log ratios and cytosolic mRNA is non trivial and gives rise to biological false positives and negatives.

Larsson and coworkers extended the study beyond theory by analyzing correlations in real datasets [Larsson et al. 2010]. They found a median spurious correlation of -0.61 across studies, confirming that log-ratio transformations frequently introduce systematic biases. The empirical range between +1 and -1 illustrates that the magnitude and direction of bias vary across genes and conditions, undermining the reliability of log-ratio approaches.

The authors conclude on the biological implications as follows:

- False positives: Certain mRNAs may appear to be under translational control when they are not, due to noise in the cytosolic mRNA data or arbitrary thresholds for inclusion.
- False negatives: Genuine translational regulation may be overlooked when log-ratios obscure independent effects of translational activity.
- Inadequacy of heuristic fixes: The authors dismiss heuristic threshold-based approaches as statistically arbitrary and unreliable.

In conclusions, there appears to be serious concerns about the use of log-ratio transformations in translational control studies. Combining theoretical justification and empirical validation, it was demonstrated that log ratios systematically introduce bias and fail to correct for confounding effects. Alternative methods are necessary for accurate assessment of translational regulation.

1.7 Scoring metrics of codon usage of synonymous codons and tRNA isoacceptors relative abundance

The deciphering of the genetic code has shown that there are 64 different codons (triplets of nucleotides), 61 encoding for amino acids and 3 stop codons that dictates termination; but there are only 20 different translated amino acids. The overabundance in the number of codons (**synonymous codons**), allows many amino acids to be encoded by more than one codon.

The whole genome (or transcriptome) sequencing of a large number of species, revealed that, in a large proportion of cases, the alternative synonymous codons for any one amino acid are not used randomly. The deviation from uniform codon usage is termed **codon usage bias (CUB)**. Further, part of this nonrandom usage is species or taxon specific. Besides, within species, there is considerable heterogeneity between genes, and in a number of model species like *E. coli* and *S. cerevisiae*, there is a clear positive correlation between the degree of codon usage bias and level of gene expression. The differences largely appear to be in the degree rather than the direction of codon usage bias.

A significant correlation has been observed between the codon usage frequency and the **tRNA gene copy number (tGCN)** for any given species. This indicates a general balance between supply (the abundance of cognate or near-cognate tRNAs) and demand (the codon usage frequency) in the protein elongation processes. The strength of this correlation supported the introduction of useful bioinformatics scoring metrics such as the **codon adaptation index CAI**, and the **tRNA adaptation index, tAI**. These scoring metrics can be used to infer the elongation efficiency of a given transcript in any given expression vector.

The details and rationale about these scoring metrics are presented hereafter and will be referred to in later chapters and especially in chapter 5.

1.7.1 Codon usage bias and codon adaptation index

The **Codon Adaptation Index (CAI)** relies on a reference table of **relative synonymous codon usage (RSCU)** of a given species [Sharp and Li 1987]. In some species like *E. coli* and *S. cerevisiae*, very highly expressed genes appear to have the greatest degree of synonymous codon bias. The pattern of codon usage in highly expressed genes can reveal which of the synonymous codons for an amino acid is the most efficient in translation (or more precisely in the protein elongation cycle), and the relative extent to which other codons are disadvantageous.

First, construct a reference table of relative synonymous codon usage values from a thoroughly chosen subset of the very highly expressed genes of the organism in question. Choose housekeeping genes or genes encoding the ribosomal proteins for instance. The RSCU value for a codon is the observed frequency of that codon divided by the frequency expected under the assumption of a uniform distribution in the usage of the synonymous codons for the same amino acid:

$$\mathrm{RSCU}_{ij} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}} \tag{1.7}$$

where X_{ij} is the number of occurrences of the j^{th} codon for the i^{th} amino acid, and n_i is the number of synonymous codons for the i^{th} amino acid, i.e., the level of degeneracy of the amino acid (from one to six). The **relative adaptiveness of a codon**, w_{ij} is then the frequency of the use of that codon compared to the frequency of the optimal codon for that amino acid:

$$w_{ij} = \frac{\mathrm{RSCU}}{\mathrm{RSCU}_{\max}} = \frac{X_{ij}}{X_{\max}} \tag{1.8}$$

where $\mathrm{RSCU}_{\max}$ and $X_{\max}$ are the RSCU and X values for the most frequently used codon for the i^{th} amino acid.

The **Codon Adaptation Index** [Sharp and Li 1987] for a gene or a transcript is then calculated as the geometric mean of the RSCU values (from the species reference table of RSCU) corresponding to each of the codons used in that gene or transcript, divided by the maximum possible CAI for a gene of the same amino acid composition:

$$\mathrm{CAI} = \frac{\mathrm{CAI}_{\mathrm{obs}}}{\mathrm{CAI}_{\max}} \tag{1.9}$$

where

$$\mathrm{CAI}_{\mathrm{obs}} = \Big(\prod_{k=1}^{L} \mathrm{RSCU}_k\Big)^{1/L} \tag{1.10}$$

$$\mathrm{CAI}_{\max} = \Big(\prod_{k=1}^{L} \mathrm{RSCU}_{\max k}\Big)^{1/L} \tag{1.11}$$

These last two equations save computation time and avoid underflow issues in computer calculations. If a certain codon is never used in the reference set, then the CAI for any other gene in which that codon appears becomes zero. To prevent this from happening, a value of 0.5 is assigned to any X_{ij} that would otherwise be zero. Besides, the number of `AUG` and `UGG` codons which are not degenerated (methionine and tryptophan only have one codon each), are subtracted from L, since the RSCU values for these two codons are both fixed at 1.0 and do not contribute to the CAI.

Equation (1.9) is equivalent to:

$$\mathrm{CAI} = \left(\prod_{k=1}^{L} w_k \right)^{1/L} \tag{1.12}$$

$$\mathrm{CAI} = \exp\left(\frac{1}{L} \sum_{k=1}^{L} \ln w_k \right) \tag{1.13}$$

$$\mathrm{CAI} = \exp\left(\frac{1}{L} \sum_{i=1}^{18} \sum_{j=1}^{n_i} X_{ij} \ln w_{ij} \right) \tag{1.14}$$

CAI values reflect the level of gene expression. Ribosomal protein genes are highly expressed, and generally have high CAI values in unicellular organism like *E. coli* and *S. cerevisiae* but mammalian ribosomal protein genes do not seem to show particularly high synonymous codon bias. In multicellular organisms, tRNA abundancies are believed to have important selective constraints on codon usage, and in such organisms, it was also observed that tRNA populations vary among tissues. For these reasons, the two following metrics were introduced: tRNA gene copy number and tRNA adaptation index.

1.7.2 tRNA gene copy number

The tRNA relative abundance in a cell in a given physiological condition is very difficult to assess as tRNA separation and sequencing methods are lagging behind current next generation sequencing technologies. A number of studies have however shown that the tRNA abundance within a cell and with codons preferences in the cell genome is positively correlated with the **tRNA gene copy number (tGCN)**. Today, there are repositories of the tRNA copy number for a number of species. The tRNA copy number-related data reference we used in the thesis is from GtRNAdb database [http://gtrnadb.ucsc.edu/GtRNAdb2/]. The codon usage frequencies database for the collected species is from [https://kasuza.or.jp/codon/].

1.7.3 tRNA adaptation index

The **tRNA adaptation index** (tAI_g) of a gene g is a scoring metric introduced by Reis *et al.* to estimate the amount of adaptation of a gene g to its genomic tRNA pool [Reis et al. 2004]. It is a measure of the tRNA usage by coding sequence inspired by the codon adaptation index (CAI) of Sharp and Li [Sharp and Li 1987]. However the tAI aims to relate only the adaptation of the gene and their codons contents to the elongation

steps which occur in the ribosome via the adaptation to the tRNA pool, the wobble base interaction and properties of the ribosome (more on these wobble rules in chapter 5). The tAI index includes, amongst others, weights that represent wobble interactions between codons and tRNA molecules. These weights were initially based only on the gene expression in *Saccharomyces cerevisiae*. However, the efficiencies of the different codon-tRNA interactions are expected to vary among different organisms [Sabi and Tuller 2014]. For this reason, the tAI index was further improved by Tuller *et al.* and relabelled **stAI** where weighting factors were introduced that are **species specific** [Sabi et al. 2016].

In order to calculate the tAI index, the **absolute adaptiveness** value, $\mathbf{W_i}$, of the i^{th} codon is defined in the following equation:

$$W_i = \sum_{j=1}^{n_i} (1 - S_{ij}) \cdot tGCN_{ij} \tag{1.15}$$

where tGCN_{ij} is the gene copy number of the j^{th} tRNA that recognizes the i^{th} codon, where the underlying assumption holds that the gene copy number of this j^{th} tRNA is a proxy of the j^{th} tRNA level. S_{ij} is a selective constraint on the efficiency of the (affinity) interaction between the i^{th} codon and the j^{th} tRNA, which is cored between 0 (perfect interaction) and 1 (no interaction). Specifically, the S_{ij} are weight factors that can be related to aspect of translation elongation (tRNA, wobble base interaction, properties of the ribosome), as these aspects are expected to affect the efficiency of the codon-anticodon interaction. The W_i values are calculated according to the Cricks's wobble rules for codon-anticodon pairing. The S_{ij} values and the rules to calculate the W_i were detailed in reference [Reis et al. 2004].

From the W_i values, the *relative adaptiveness value* w_i of a codon is obtained as:

$$w_i = \begin{cases} \frac{W_i}{W_{\text{max}}}, & \text{if } w_i \neq 0, \\ w_{\text{mean}}, & \text{else}, \end{cases} \tag{1.16}$$

where W_{max} is the maximum W_i value and w_{mean} is the geometric mean of all w_i with $W_i \neq 0$. The tRNA adaptation index tAI_g of a gene g is defined as the geometric mean of the relative adaptiveness values of its codons

$$\text{tAI}_g = \left(\prod_{k=1}^{l_g} w_{i_{kg}} \right)^{1/l_g} \tag{1.17}$$

where i_{kg} is the codon defined by the k^{th} triplet in the open reading frame (ORF) of gene g and l_g is the length of the gene in codons (except the stop codon).

The last improvement of the tRNA adaptation index scoring metrics introduced in references [Sabi and Tuller 2014; Sabi et al. 2016] is related to the selection of a meaningful set of S_{ij} values.

Researchers have proposed a method valid for any species where the set of values is chosen such that the correlation between the tAI and the expression levels of the genes is maximized [Reis et al. 2004]. They calibrated the set of weight values for yeast. Tuller *et al.* adjusted the weights to any target model organism [Sabi and Tuller 2014; Sabi et al. 2016]. The W_i values are calculated according to Crick's wobble rules for codon-anticodon pairing as displayed in Table 1.7.1.

Table 1.7.1: Crick's wobble rules for calculating the absolute adaptiveness W_i of a codon to the tRNA pool [Reis et al. 2004; Sabi and Tuller 2014].

Crick's wobble rules for calculating W_i				
	Codon third position		Anticodon first position	W_i
i	U	j	I	$(1 - S_{U:I})tGCN_{i,j} + (1 - S_{U:G})tGCN_{i,j+1}$
$i + 1$	C	$j + 1$	G	$(1 - S_{C:G})tGCN_{i+1,j+1} + (1 - S_{C:I})tGCN_{i+1,j}$
$i + 2$	A	$j + 2$	U	$(1 - S_{A:U})tGCN_{i+2,j+2} + (1 - S_{A:I})tGCN_{i+2,j}$
$i + 3$	G	$j + 3$	C	$(1 - S_{G:C})tGCN_{i+3,j} + (1 - S_{G:U})tGCN_{i+2,j+2}$

The 64 codons are clustered in the genetic code into 16 groups, each one consist of four codons. The four codons in each group differ only in their third position (the wobble position). The W_i are calculated based on equation 1.15. The formulas for calculating the W_i for the four groups are explicitly given in the last column of the table. i denotes the index of the codon in the quartet which ends with U, $i + 1$, $i + 2$ and $i + 3$ denote the three other codons which end with bases C, A and G respectively. j denotes the index of the tRNA whose anticodon starts with I; all base pairing between the i^{th} codon and the j^{th} anticodon are Watson-Crick. $j + 1$, $j + 2$ and $j + 3$ denote the three other tRNA whose codons start with bases G, U, C respectively. $tGCN_{ij}$ represents the tRNA gene copy number corresponding to the interaction between the i^{th} codon and the j^{th} tRNA. For each codon, W_i sums over all tRNAs that can pair with the codon.

In this section, all tRNA anticodons (nucleotide triplets) are referenced in the 5'-3' direction (as are all mRNA codons). In yeasts, there are 16 tRNAs genes out of the 275 genes coding for all tRNAs. Out of these 16 alanine tRNAs *iso-acceptors* (see glossary isoacceptor), 11 have the anticodon IGC (able to decode 3 of the 4 synonymous codons of alanine, i.e., GCU, GCC and GCA) and 5 have the anticodon UGC (able to decode 2 of the 4 synonymous codons of alanine, i.e., GCA and GCG). For example, Table 1.7.1 shows that the GCU codon (alanine) which ends with U can either pair with anticodons that start with I (IGC) and generate a Watson-Crick base pairing, or pair with anticodons that start with G (GGC in the alanine example) and generate a wobble base pairing. It should be noted however that there are no tRNA in yeasts or humans that have GGC as an anticodon (that could theoretically match the GCU and GCC codons of alanine). But there are, of course, other tRNAs with their anticodons starting with G that can pair with codons starting with U or C which do not belong to

the synonymous codons of alanine. For alanine, in yeast, there are only two different anticodons, IGC and UGC-tRNAs which are able to decode 4 synonymous codons (as CGC-tRNA does not exist for yeast). This degeneracy allowed by the wobble base explains why the number of different tRNA iso-acceptors, as observed in most species, is less than the total number of the 61 sense codons. More on this in chapter 5.

When applied to yeast, the maximum correlation search method (correlation between tAI and protein expression) suggests to use the S_{ij} weight coefficients that are tabulated in Table 1.7.2 [Reis et al. 2004].

Table 1.7.2: The weight factors S_{ij} values for yeast are given to the pairing between the third position of the i^{th} codon and the first position of the j^{th} anticodon (tRNA). S_{ij} values of standard Watson-Crick base pairing are set to zero and the wobble values are shown in bold [Reis et al. 2004].

Weights of the different base pairings for yeast				
j	i			
	G	U	C	A
I		0	**0.28**	**0.9999**
G		**0.41**	0	
U	**0.68**			0
C	0			
A		0		

Table 1.7.3 is an update of the adjusted weights averaged across species in the three domains of life [Sabi and Tuller 2014; Sabi et al. 2016].

As mentioned above, S_{ij} values are between 0 and 1. A value closer to zero means a good base pairing interaction between codon and anticodon, while a value close to one is a weaker interaction. It can be seen from Table 1.7.3 that cytosine-inosine interaction has the lowest value for all domains of life as compared to the other wobble interactions within the same domain, while the wobble adenosine-inosine have the largest values within any given domain. This suggests a good C:I interaction and an inefficient A:I interaction. This would also suggest that the accommodation and proofreading step of elongation could show higher rates (fast codon) when one codon ends with C, such as in GCC (coding for alanine) as compared to slower rates (slow codon) for a synonymous codon that ends with A, such as in GCA (also coding for alanine). This is indeed the case as was shown by Ribo-Seq meta analysis conducted on *Neurospora crassa* [Lyu et al. 2020] not only for alanine (GCC faster than CGA) but also for the eight ADAT sensitive codons coding for the TAPSLIVR amino acids, where the relative codon occupancy by ribosomes is higher for the A ending codons (slow codons) than for the C ending codons (fast codons). Explicitly, for each of the following amino

Table 1.7.3: The weight factors S_{ij} values averaged across a number of species for the three domains of life are given to the pairing between the third position of the i^{th} codon and the first position of the j^{th} anticodon (tRNA). S_{ij} values of standard Watson-Crick base pairing are set to zero and the wobble values are shown in bold [Sabi and Tuller 2014; Sabi et al. 2016].

Weights of the different base pairings for the three domains of life					
j		i			
		G	U	C	A
	Domain				
I	Eukarya		0	**0.4659**	**0.9075**
	Bacteria		0	**0.4211**	**0.8733**
	Archea		0	**0.3774**	**0.5015**
G	Eukarya		**0.6295**	0	
	Bacteria		**0.698**	0	
	Archea		**0.4363**	0	
U	Eukarya	**0.7861**			0
	Bacteria	**0.6294**			0
	Archea	**0.3898**			0
C	Eukarya	0			
	Bacteria	0			
	Archea	0			
A	Eukarya		0		
	Bacteria		0		
	Archea		0		

acids in *Neurospora crassa* wild type, there is a significant difference in ribosome occupancy density on the A ending codon as compared to the C ending codon for all the 8 ADAT sensitive codons. The sensitive codons are the ones for which there is a wild type natural editing of the A nucleotide at the start position of the anticodon to the I nucleotide (inosine) by enzymatic deamination via ADAT–Adenosine Deaminase Acting on tRNA; more on this in chapter 5.

- Alanine: GCC (fast) has a lower ribosomal footprint as compared to GCA (slow)
- Threonine: ACC (fast) has a lower ribosomal footprint as compared to ACA (slow)
- Proline: CCC (fast) has a lower ribosomal footprint as compared to CCA (slow)
- Serine: UCC (fast) has a lower ribosomal footprint as compared to UCA (slow)
- Leucine: CUC (fast) has a lower ribosomal footprint as compared to CUA (slow)

- Isoleucine: AUC (fast) has a lower ribosomal footprint as compared to AUA (slow)
- Valine: GUC (fast) has a lower ribosomal footprint as compared to GUA (slow)
- Arginine: CGC (fast) has a lower ribosomal footprint as compared to CGA (slow)

1.8 Summary of the challenges in ribosome profiling and the need for improved models

This first chapter provided an overview of protein synthesis by ribosomes, emphasizing the importance of translation regulation and the challenges associated with current ribosome profiling techniques. While ribosome profiling (Ribo-seq) has advanced our understanding of translation dynamics, it suffers from methodological limitations, including **spurious correlations in translation efficiency (TE) scores**, which can lead to misleading interpretations of ribosome occupancy and mRNA translation rates. Additionally, Ribo-seq's inability to distinguish between polysomes and monosomes limits its usefulness in assessing ribosome distribution across transcripts.

The chapter also highlights the complexity of translation, where elongation rates are influenced by intertwined factors such as **codon usage, tRNA adaptation, mRNA secondary structure, nascent peptide interactions, and the presence of slow donor or slow acceptor amino acids as substrates for peptide bond formation at the LSU catalytic site**. These factors contribute to ribosome queueing and potential stalling. The shortcomings of existing experimental methods underscore the need for computational modeling to disentangle their contributions to translation regulation.

This thesis addresses these gaps by **integrating physical and statistical approaches into an agent-based model (ABM)** to provide a mechanistic understanding of ribosome kinetics and translational control.

Chapter 2

Thesis objectives and methodology

What I cannot create I do not understand.
Note found written on Richard Feynman's blackboard at his Caltech office when he passed away in 1988.

This chapter outlines the objectives and methodology of the thesis, which focuses on disentangling the key factors influencing mRNA translation by ribosomes from a systems biology perspective. Recognizing the limitations of existing experimental methods in isolating individual factors, the thesis employs an agent-based model (ABM), named **`Ribosomer`**, to simulate and study the effects of five main factors: tRNA abundance/modifications, electrostatic interactions in the ribosome exit tunnel, proline incorporation, mRNA secondary structures, and ribosome congestion. The methodology combines mechanistic modeling and extensive computational simulations to replicate experimental observations and address how these factors shape translation efficiency and protein synthesis. Secondary objectives include generating ribosome occupancy maps and exploring parameter spaces inaccessible experimentally. The chapter concludes by emphasizing the open-source release of the **`Ribosomer`** tool for future research.

2.1 Background and problem statement

The gene expression pathway leading to protein production comprises multiple mechanistic layers subject to regulation [Larsson et al. 2012]. These regulatory processes are generally categorized into transcriptional regulation, which involves the conversion of DNA sequences into mRNA sequences—often mediated by transcription factors—and post-transcriptional regulation, which affects translation, the process of converting mRNAs into proteins. Over the past two decades, research has predominantly focused on transcriptional regulation, with genome-wide analyses of steady-state mRNA levels being used to infer gene expression profiles. However, recent studies indicate that steady-state mRNA levels correlate only loosely with the composition of the proteome. This finding highlights the critical role of post-transcriptional mechanisms, including mRNA translation, in gene expression regulation.

Some post-transcriptional mechanisms, such as RNA splicing (in eukaryotes) and mRNA editing, determine the identity and activity of protein products, while others regulate protein levels and function by controlling mRNA transport from the nucleus to the cytoplasm (in eukaryotes), mRNA export, localization, stability, silencing, translation, and post-translational modifications.

Translational control of gene expression can be achieved by modulating translation initiation, elongation, and termination [Larsson et al. 2010]. Translation can be influenced by changes in ribosome biogenesis, ribophagy, tRNA modifications, or interactions with translation factors (such as initiation or elongation factors), which themselves are regulated by complex signaling cascades. Differential translation typically involves changes in the number of ribosomes bound to each mRNA, thereby altering the rate of protein synthesis per mRNA molecule over time. This regulation can occur through specific mechanisms targeting individual mRNAs or subsets of mRNAs, or through global processes affecting most mRNAs equally.

Understanding how post-transcriptional mechanisms contribute to protein levels and activity, and how translational regulation shapes the final outcomes of gene expression, has been a major focus of research in recent decades. Recent high-throughput techniques have been used to simultaneously study gene expression (transcriptome, RNA-seq) and protein expression (proteome, mass spectrometry) on a genome wide scale. The global association between mRNAs and ribosomes (translatome, Ribo-seq) has been a subject of substantial interest since 2009 as it is a high throughput technology, like RNA-seq, filling the gap between the transcriptome and the proteome.

Ribosome profiling quantifies the genome-wide ribosome occupancy of transcripts. With the integration of matched RNA sequencing data, the translation efficiency (TE) of genes can be calculated to reveal translational regulation. This layer of gene-expression regulation is otherwise difficult to assess on a global scale and generally

not well understood. Current statistical methods to calculate differences in TE have low accuracy, cannot accommodate complex experimental designs or confounding factors, and do not categorize genes into house-keeping, intensified, or exclusively translationally regulated genes.

In the translation of mRNAs by ribosomes, depending on the number of free available ribosomes, the set of transcripts in a cell compete for ribosomes. The translation of a given transcript results from a complex interplay between the pool of ribosomes available, the relative initiation rates across distinct transcripts and the interdependence between elongation rates and initiation rates. The elongation rate of a single ribosome itself also depends on intertwined factors described in the literature since long. How these factors, individually and concomitantly contribute during protein synthesis is poorly understood. In general, at the moment, the study of each individual factor is not experimentally achievable; at least not independently or separately from the other factors. Five main factors have been documented in the literature from the start of the thesis work:

- tRNA relative abundance and effects of tRNA modifications at the wobble base of the anti-codon
- electrostatic interaction of the nascent protein elongated by the ribosome as it progresses through the ribosome exit tunnel
- occasional incorporation of proline amino acid during the elongation cycle
- mRNA secondary structures hampering translocation in the mRNA sequence downstream a ribosome
- ribosomes congestion due to stalling events

Genome-wide high throughput experimental tools feed their results to big data repositories at high pace. These transcriptomics and translatomics protocols are extensively used worldwide in laboratories working in a variety of fields from fundamental molecular biology to cancer research and the results are available worldwide in the open source and open science community.

Polysome fractionation profiling and ribosome profiling (Ribo-seq) have led to the development of translation scoring metrics, such as translation efficiency (TE). However, these metrics are often challenging to interpret, and Ribo-seq results can be difficult to reconcile with RNA-Seq and proteomics data across different experimental conditions.

A key challenge in interpreting these metrics arises from the lack of direct experimental access to critical biological parameters, which are either not routinely measured or remain unknown. For instance, the initiation rates of ribosomes on individual transcripts and the relative sizes of the free ribosome pool and transcriptome are rarely quantified.

Yet, this information would be essential for accurately interpreting translation efficiency and related metrics.

2.2 Thesis objectives and methodology

2.2.1 Research questions and methodology

2.2.1.1 Main research question and general objective

How is the translatome processed from a systems biology perspective ? As discussed previously, several intertwined factors have been identified and studied for decades. An **essential objective of the thesis is to try to disentangle the individual contributions of these factors**. Figure 2.1 represents the ribosome elongation cycle at the center of the protein synthesis pathway. The figure provides a summary of the main factors affecting the ribosome elongation cycle, the rate and output of protein synthesis as mentioned in the previous section. Each of these 5 factors acts at a particular substep in the cycle. The mechanisms by which each factor is working are addressed in a separate chapter of this PhD dissertation. These chapters detail and explain the building blocks of the protein synthesis model which is the core of the thesis (**chapter 3**).

The elongation cycle entails three main steps.

Step 1 is the accommodation and proofreading of aa-tRNA at the A-site by the ribosome.

Step 2 is the peptide bond formation between the carboxyl-terminal end of the peptidyl-tRNA at the P-site and the aminoacyl-tRNA at the A-site.

Step 3 is the translocation and eviction of the deacylated-tRNA (free tRNA) from the E-site.

As a cyclic process, the total elongation time is divided into three distinct time steps, each with a different duration. The sequential combination of these steps is analyzed using queueing theory in probability and statistics (**Chapter 4**). While Step 1 generally has the longest duration, no single step can be definitively considered the rate-limiting factor within the cycle. This dissertation explores various factors influencing each step. Step 1 is affected by the availability of tRNAs and modifications at the anticodon wobble base position (**Chapter 5**). Step 2 is influenced by electrostatic interactions of the nascent polypeptide as it progresses through the ribosome exit tunnel (**Chapter 6**) and by the occasional incorporation of proline residues (**Chapter 7**). Step 3 is impacted by the secondary structure of mRNA downstream of an elongating ribosome (**Chapter 8**).

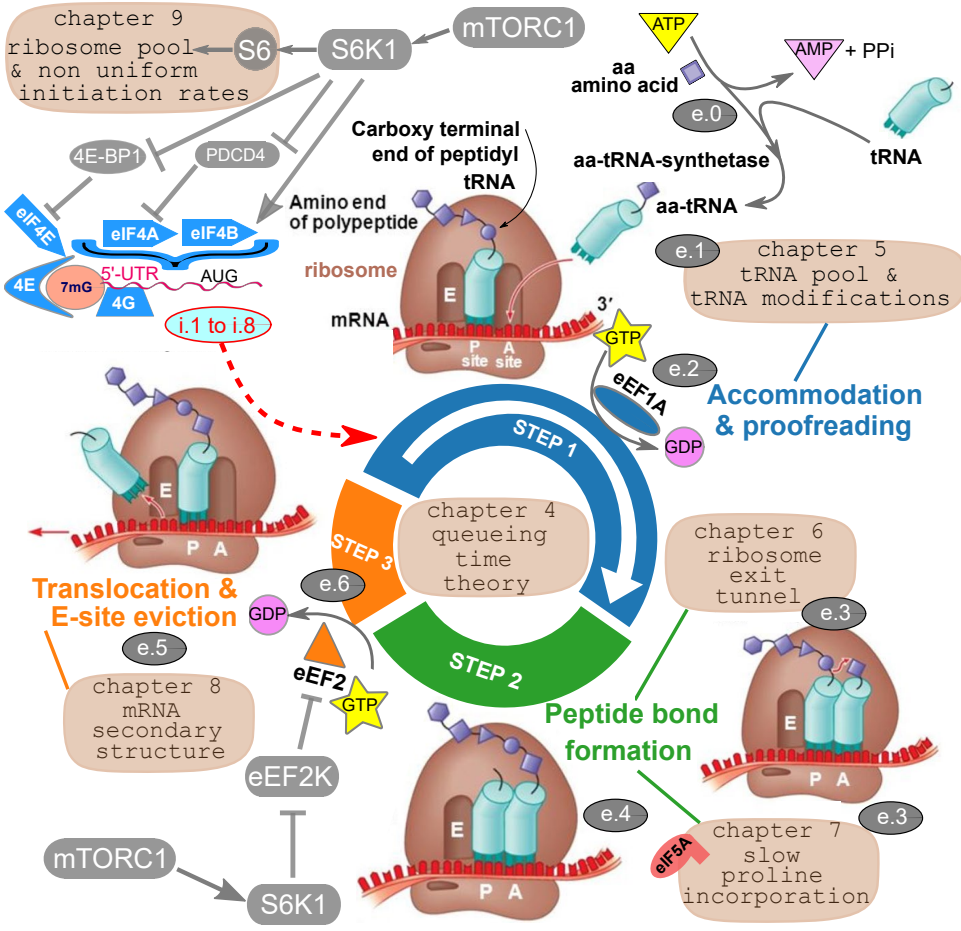


Figure 2.1: Representation of the ribosome elongation cycle in protein synthesis. Schematic view of the main factors that are studied in this PhD thesis and how they affect the elongation cycle. The elongation cycle entails three main steps. Before a ribosome can start an elongation round, initiation must occur. Light blue oval (`i.1` to `i.8`): initiation sub-steps. Non-uniform initiation rates of the ribosome pool strongly affect the protein landscape. Gray ovals show the elongation substeps from tRNA-acylation to peptide incorporation: `e.0`, tRNA aminoacylation; `e.1`, tRNA accommodation and decoding; `e.2`, eEF1A/GDP dissociation after delivery of aminoacyl-tRNA (aa-tRNA) at the A-site; `e.3`, peptide bond formation combined with the polypeptide progression through the ribosome exit tunnel and with proline incorporation;`e.4`, peptidyl transferase reaction combined with the preceding substeps; `e.5`, translocation; `e.6`, eEF2/GDP dissociation after translocation and tRNA E-site eviction. Ribosome termination and recycling are not represented. Modulators of signaling cascades: `mTORC1`, `S6K1`, `S6`, `PDCD4`, `4E-BP1`, `eEF2K`.

The fifth factor–ribosome congestion due to stalling events– also impacts Step 3. However this fifth factor is not the topic of a separate chapter on its own. Instead, this factor is directly integrated as a behavioral rule for all ribosome agents described in the computational model.

Before a ribosome can sequentially cycle through each codon of the open reading frame of a specific mRNA molecule, it must first be assembled. Initiation must occur on this mRNA, a process that is highly regulated and considered the rate-limiting step of translation. The determinants of non-uniform initiation rates and modulation of the ribosome pool are examined in **Chapter 9**.

To achieve the essential objective above, the method of the thesis is to build an **agent based model** (ABM) to incorporate each factor separately and then study, through extensive computational simulations, how these factors individually or concomitantly contribute to the protein landscape.

An important step towards meeting the aforementioned objective is theory building. Our goal is to identify the key factors that explain the output patterns observed in simulations and assess how closely these outcomes align with experimental data. Richard Feynman's statement, "*What I cannot create, I do not understand*," suggests that to fully comprehend translation and translatome processing, we must construct a surrogate computational model of the phenomenon and generate synthetic outputs that closely replicate real experimental observations. This approach embodies the essence of *in-silico* research in biomedical science and illustrates the fundamental purpose of a digital twin of a biological system in computational biology.

Complex computational simulations are not merely tools for integrating and comparing theory with experiment; more importantly, they play a critical role in theory construction. They help identify key components, uncertainties, and inherent stochasticities, generate and evaluate new hypotheses, suggest novel experiments and data collection strategies, and enhance decision-making [An et al. 2009].

It is important to emphasize that agent-based models (ABMs) are not inductive models, as most statistical models are, since they are not derived from patterns in data. Inductive models begin with a dataset and make data-driven inferences about the mechanisms that may have generated the observed data. In contrast, ABMs start with predefined mechanisms or behavioral rules—hypotheses in themselves—and attempt to reconstruct observed data patterns through their computational implementation.

This distinction is methodologically significant: if the objective of a modeling endeavor is to identify patterns within an existing dataset, agent-based modeling is likely not the appropriate approach. However, if the goal is to test the plausibility of a set of identified or presumed mechanisms within a system, ABMs can be highly effective [ibid.]. In this work, our methodology specifically aims to evaluate the credibility—if not the veracity—of a set of presumed mechanisms governing mRNA translation by ribosomes.

This PhD dissertation proposes a set of **mixed mechanistic-stochastic rules governing the agents and their behavior** within the developed model. A key contribution of this modeling effort is the integration of tRNA modification effects amongst the 5 aforementioned factors, enhancing the model's biological relevance and predictive capacity.

The key contribution of this thesis was the creation of a collection of digital twins representing the biological entities involved in mRNA translation and protein synthesis at both the subcellular and molecular scales. The created computational model is called `Ribosomer`. This computational model was used, in the final months of the thesis, to run tens of thousands of simulations, exploring the effects of each biological agent, either individually or in combination. Exploring the parameter space is crucial for understanding the biological significance of interactions (synergy) among the contributing factors. The overarching goal of systems biology in this context is to uncover how emergent properties of this complex biological process are generated.

Chapter 3 is an extensive introduction to the agent based models (ABM) and details the content of the model of protein synthesis by ribosomes. Chapter 3 also introduces the agents and rules on which the model is built. It also describes how the model is computationally implemented.

2.2.1.2 Secondary research questions and specific objectives

The elongation rate of a ribosome on a given transcript is far from uniform. It depends on codon usage, the relative abundance of tRNAs, and the other aforementioned contextual factors, Fig. 2.2. How much time does a ribosome spend on a codon? How is it different from codon to codon? What are the origin and significance of slow and fast codons? What are the factors affecting the ribosome elongation cycle and what are their consequences on the final protein landscape output?

Our secondary research questions are:

- How tRNA modifications in the wobble base at nucleotide 34 of different tRNAs impact on elongation and on protein synthesis is one of the (secondary) research questions that we aimed to address during this PhD thesis. The tRNA modifications mostly impact step 1. tRNA modifications can be seen as a way to perturb the system of translating ribosomes collectively and test the credibility of the model predictions.
- A specific objective of the thesis is to synthetically generate the ribosome occupancy maps through computer simulations and fill the gap between transcriptomic and proteomic results. The ribosome occupancy maps for each transcript in the translatome, will depend on the statistical aggregation of every

instance of ribosome residence time on each codon. We aim at producing polysome profiles, and ribosome profiles, at individual transcript resolution.

- Conducting exhaustive exploration of the model's parameter space to illuminate the biological significance of factors that are currently inaccessible to experimental measurements or are not routinely measured (e.g., initiation rates on individual transcripts; free ribosome pool size). This is a computationally intensive task that this PhD pursues.

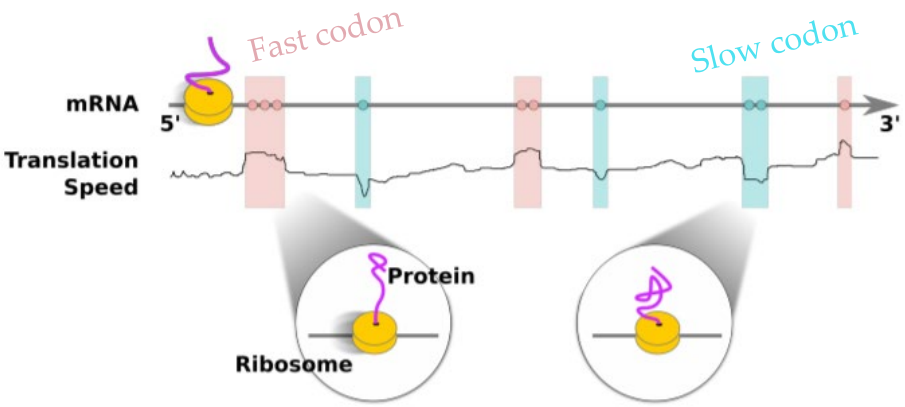


Figure 2.2: Speed of translation along a single transcript. The elongation rate along a transcript is not uniform. There appear to be fast and slow codons. One of the research question addressed in the thesis is to investigate the origin and significance of the variable elongation speed. Adapted from reference [Novoa and Pouplana 2012].

2.2.2 Limiting assumptions and inductive bias

The computational model of protein synthesis and mRNA translation by a pool of ribosomes is an agent-based model which is built on incomplete knowledge of the 'ground truth' exact mechanisms to which the real process obeys. The model is also based on a number of assumptions that are enumerated hereafter. Each of our computing simulation complies with the following explicit assumptions:

- The population of ribosome is homogeneous during a simulation run
- Absence of mRNA decay during a simulation run
- Absence of protein decay during a simulation run
- No amino acid limitation (no fasting)
- No limitation in any amino acid acylation on tRNAs by aa-synthetase (no limiting loading rates of tRNAs in the cytoplasm)
- No limitation in the pool of tRNAs. The tRNAs' lifetimes are much longer than a simulation runtime.
- Absence of frameshifting during elongation
- Each codon has an elongation rate complying to the same rules across all transcripts, meaning that the time a ribosome spends decoding a specific codon is consistent, regardless of the transcript in which it appears. This assumption allows for the estimation of relative decoding rates for each codon type. This strong assumption is connected to the universal scaling behavior observed in ribosomal protein synthesis [Sharma et al. 2018].
- No ribosome drop-off events or ribosome read-through (run-off) events
- No leaky scanning of the upstream untranslated region of the mRNA (5'UTR region) by the small ribosomal subunit. The possibility of alternative start codon initiation is not considered
- Internal Ribosome Entry Sites (IRES) are not taken into consideration
- Parameter stationarity assumption
- The initiation rate is constant over time on any given transcript during a simulation
- The total ribosome pool (free ribosomes + translating ribosomes) is constant over time. The ribosome dynamical biogenesis and decay are ignored. A ribosome has an infinite lifetime during a simulation

- There is no interfering short hairpin RNA silencing translation in the mRNA coding region

An agent-based model inherently incorporates inductive bias, as it is built upon predefined rules and assumptions about the behavior of individual agents and their interactions. This bias can be advantageous when it aligns with well-established biological principles or facts, allowing for efficient simulations and insightful predictions. However, it can also be a limitation if the underlying assumptions oversimplify complex processes or fail to capture emergent behaviors, potentially leading to misleading interpretations of the system's dynamics. Other implicit assumptions possibly remained unnoticed.

2.2.3 Thesis legacy and open source delivered tools

By the end of this thesis, two software tools were developed, each serving as a digital twin–called "`Ribosomer`"– of a collection of molecular-biological agents involved in mRNA translation and protein synthesis at subcellular and molecular scales.

- A Javascript (p5.js) implementation of `Ribosomer` designed for didactic and visualization purposes.
- A Python implementation of `Ribosomer` refactored for computationally intensive tasks, intended for execution in high-performance computing (HPC) environments. This version is designed for research applications, particularly for exploring parameter spaces through extensive simulations.

The didactic JavaScript version is designed to train new researchers, helping them develop a mechanistic understanding of the factors influencing mRNA translation rates. It provides near real-time visualization of how these factors impact a given proteome, allowing users to explore different scenarios, compare simulation outcomes with their expectations, and validate findings against experimental ground truth data.

The computational model is intended for future use, and we encourage researchers to further develop and refine it. Both versions of `Ribosomer` are released as open-source tools to support translatomics studies, facilitating the exploration of the collective behavior of molecular agents involved in protein synthesis.

The tools are available in the following GitHub repositories: https://github.com/MasterCube and specifically:

- https://github.com/MasterCube/Ribosomer-in-Javascript
- https://github.com/MasterCube/Ribosomer-in-python.

2.3 Structure and organization of the thesis

The PhD dissertation is structured into twelve chapters, as illustrated in the wireframe flowchart in Figure 2.3. The brown color code for **chapters 4 to 9** is consistent with the color code retained in Figure 2.1.

Chapter 3 introduces `Ribosomer`, the modeling and simulation tool that forms the core of this thesis. This chapter outlines the framework of the `Ribosomer` agent-based model, which is constructed upon the mechanochemical principles and queueing time statistical theory of the elongation cycle—the focus of **Chapter 4**.

The framework of **Chapter 3** is built upon the conceptual 'bricks' provided by **Chapters 4 to 9**, with each chapter contributing a detailed mechanistic analysis of factors deemed relevant and influential to the elongation cycle. **Chapter 10** presents the model outputs, while **Chapter 11** conducts an uncertainty and sensitivity analysis, exploring the model's parameter space. Finally, **Chapter 12** offers a comprehensive discussion and future perspectives.

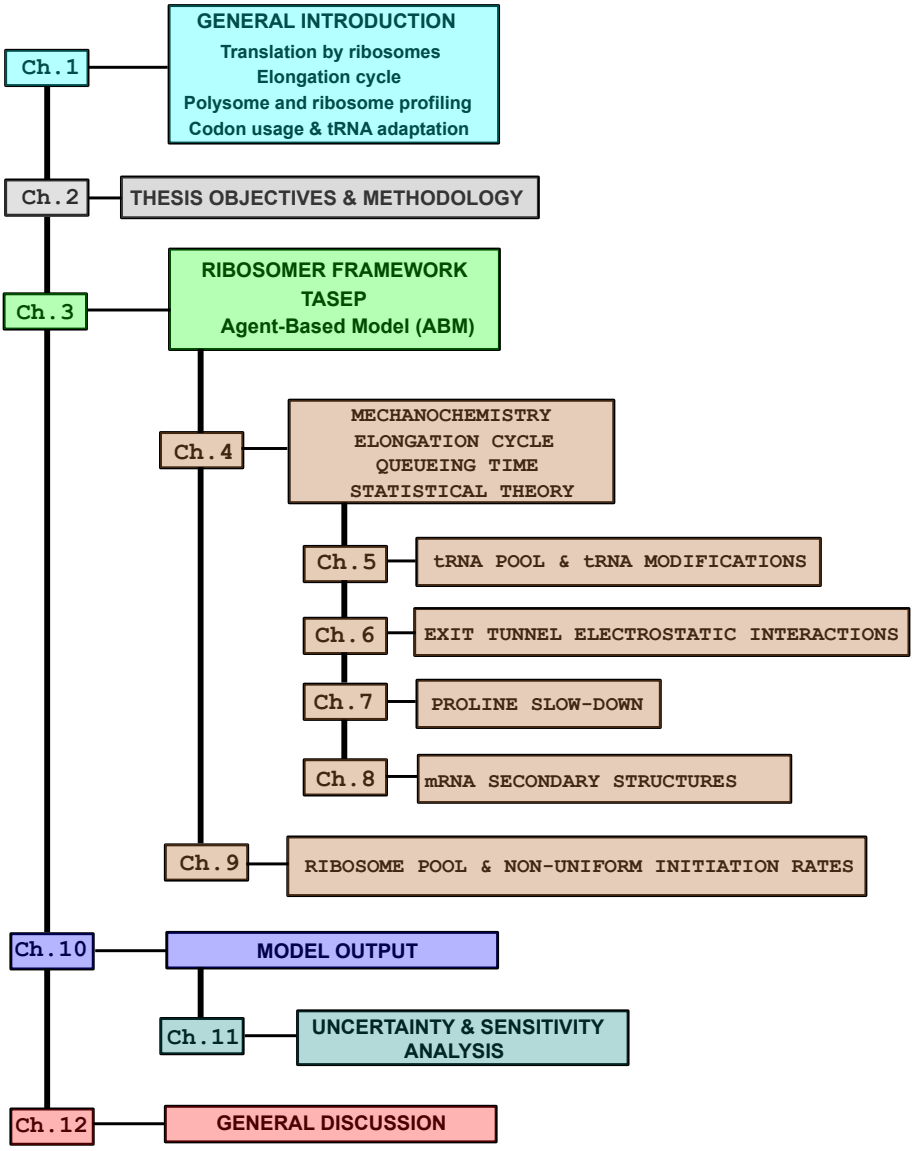


Figure 2.3: Schematic diagram showing the overall organizational structure of the thesis.

Chapter 3

Building the ribosomer framework using a TASEP approach

The sciences do not try to explain, they hardly even try to interpret, they mainly make models. By a model is meant a mathematical construct which, with the addition of certain verbal interpretations, describes observed phenomena. The justification of such a mathematical construct is solely and precisely that it is expected to work.
John von Neumann

A model is an invention, not a discovery.
Massoud *et.al*, cited in [Viceconti and Emili 2024]

This chapter discusses the overall structure of the agent-based framework developed in this thesis. The chapter starts by providing a broadly accessible introduction to agent-based modeling concepts. We explain the rationale behind developing an agent-based model as a research tool to investigate the competitive translation of multiple transcripts by a shared pool of ribosomes. Given the inherently multidisciplinary nature of this thesis, the chapter alternates between sections on computational science concepts and sections introducing key aspects of mRNA translation and the protein elongation cycle in molecular biology. The final sections outline the agents' attributes and rules within the mRNA translation agent-based model, using standard diagrams. These diagrams

visually represent the object-oriented system architecture and employ the Unified Modeling Language (UML) of object-oriented programming. The chapter concludes with a discussion of the resources and methodologies used to execute thousands of agent-based model simulations in parallel on a high-performance computing cluster. The mechanistic principles underlying the rules governing different agents will be explored in detail in the subsequent chapters.

3.1 What is a Model in Science?

In general terms and with a wide scope of applications in Science, a useful definition is borrowed from [Viceconti and Emili 2024]: "*Models are finalised cognitive constructs of finite complexity that idealise an infinitely complex portion of reality through idealizations that contribute to the achievement of a knowledge on that portion of reality that is reliable, verifiable, objective, and shareable*". Models are a way that we use to represent the world. In Science, models idealise a quantum of reality with three purposes: (i) to memorise and logically manipulate quanta of reality (Descriptive models); (ii) to combine our beliefs on different quanta of reality in a coherent and non-contradictory way toward the progressive construction of a shared vision of the world (Integrative models); (iii) to establish causal and quantitative relationships between quanta of reality (Predictive models).

The credibility of a predictive model is defined as its ability to predict causal and quantitative relationships between quantities in the natural phenomenon being modeled, as measured experimentally. The first foundational aspect of a model's credibility is the complex relationships that predictive models have with controlled experiments.

3.2 Modeling in computational biology

Advances in high-throughput sequencing techniques, i.e. next generation sequencing (NGS), opened the possibility of a joint study of the trancriptome (RNA-Seq), the translatome (Ribo-Seq), the proteome (Mass spectrometry) in different contexts, both in vitro and in vivo, in different experimental settings, and for any species. Besides, advances in data generation in synthetic biology and in computer hardware and software have transformed the life sciences from being data-poor to being data-rich [An et al. 2016]. Computation is an essential component of much research in biology, and has become ubiquitous across biomedical research and healthcare. A great deal of recent progress in life sciences relies on computation. This has come to be recognized as a "third pillar of science", together with theory and experimentation. This third pillar

of research in biology, medicine and healthcare added *in silico* to the traditional *in vitro* and *in vivo* research practice. From a methodological perspective, computational models are constructed for at least two distinct but overlapping purposes [ibid.]:

1. models to increase understanding of the system being modeled;
2. models to inform decisions to be made about the system being modeled.

Complex computational simulations are not only a means of integrating and comparing theory and experiment but, more importantly, are used to aid in **theory construction**, to determine crucial components, uncertainties or inherent stochasticities, to generate and examine new hypotheses, to suggest new experiments and data collection efforts and strengthen decision-making [An et al. 2009].

A big challenge of the objects being investigated in biological sciences is the complexity of their interactions among various internal sub-components or with neighbouring objects and their vast range of spatial and temporal scales or metamorphic evolutive trajectories.

In addition to complexity, other challenges for modelers are the relationship opposing spatial resolutions and time scales, and opposing continuous media to discrete and distinct sub-components with sharp discontinuities. As a way out of these challenges, two categories of modeling techniques are employed: the macroscopic and the microscopic, each focusing on one of the two goals: high spatial resolution or extended time scales [Soheilypour and Mofrad 2018].

A well-established macroscopic method for modeling cellular or intracellular processes is to use bulk property models such as ordinary differential equations (ODE) of reaction rates that quantify concentration over time [ibid.]. ODE or coupled ODE representations of molecular reaction networks make the assumption that concentrations are high and the system is well mixed, exhibiting a statistical ensemble averaged behavior. In some systems, the correlation length, or the length at which spatial homogeneity of reactants can be assumed, is small. This is the case if reactions occur faster than the products species can diffuse to satisfy a well mixed assumption. In such cases, spatial details should be considered through the use of partial differential equation (PDE) models and anisotropy in the medium may have to be considered. Both ODE and PDE models are well suited for systems with high concentrations that comply with the continuous hypothesis.

Biological systems often contain a discrete number of particles, or sub-components, or a collection of discrete and even disjoint individuals like in the starling birds flocking problem. Both in the spatial and time scales, the continuum hypothesis is not valid. Deterministic models such as ODE and PDE are not well suited for such systems.

Different modeling alternative approaches were developed to address problems involving a discrete number of particles such as the chemical master equation (CME), the reaction diffusion master equation (RDME), the Gillespie algorithm or the stochastic simulation algorithm (SSA), next reaction method and reaction diffusion SSA.

CME and RDME are sets of deterministic ODEs describing the time-evolution of a molecular system that is well-mixed or locally well-mixed (dividing the domain into sub-volumes and assuming each to be well mixed). These models capture the stochastic nature of such systems but apply the stochasticity to the population, not individuals. They cannot provide detailed spatial information about individual particles and do not allow tracking of individual particle. There are also computationally expensive Brownian dynamics (BD) or molecular dynamics (MD) techniques, but which are limited in the number of molecules they can handle.

The gap between the capabilities of computationally efficient macroscopic models such as ODE, PDE, CME and RDME and more detailed models such as BD and MD creates a **need for mesoscopic or multi-scale modeling techniques**, which can be satisfied using **agent-based models**. Agent-based models also offer a paradigm shift in modeling possibilities and their generic capacities are well suited to describe biological processes [Soheilypour and Mofrad 2018].

3.3 Agent-Based Models

Agent-Based Modeling (ABM) is a complex systems approach for simulating the interactions between multiple independent entities, termed '**agents**', with the objective of assessing their individual effect on the overall system and predicting subsequent emergent phenomena [An et al. 2009; Marchi and Page 2014]. ABM is a bottom-up and deductive approach, that models a complex system from the perspective of its constitutive components, at multiple levels of hierarchy.

Governing **rules** define how each individual agent behaves, moves or interacts, at discrete-event and discrete-time, leading to reproduction of a complex phenomenon.

As discussed in the previous chapter, when the goal is **to test the veracity of a set of identified or presumed mechanisms in a system, then ABM can be extremely useful** [An et al. 2009].

There are several characteristics of agent-based modeling that set it apart from other object-oriented, rule-based modeling systems (such as networks models). Several of these characteristics of ABM make it a perfect candidate for stochastic modeling of molecular biology systems [ibid.]:

1. ABMs easily incorporate ***space***. Agent-based modeling has its origins in two-dimensional cellular automata (CA). As a result, many ABMs are 'grid-based'. This legacy allows spatial representation of the structural relationships with a system as the two-dimensional grid has been expanded into three (or more) dimensions. The spatial nature of ABMs also supports modeling agents with 'bounded knowledge', i.e., input constrained by locality rule that determine its immediate environment. The emphasis on behavior driven by local interactions also matches closely with mechanisms of stimulus and response often observed in biology. However, not all ABM use physical space to define their virtual environments. For instance in the mRNA translation model, instead of using a physical space, a 2D virtual space is used where the first dimension is spanned by all the different transcripts and the second dimension is spanned by the number of codons of a particular transcript. Note that this is almost equivalent to a matrix abstract type but where the column dimensions are not always the same across the rows. In our example the number of codons of a transcript depends on the transcript itself.

2. ABMs easily support ***tracking over time*** or ***memory recording rules***. Most often, in classical dynamic mathematical modeling with ODEs or PDEs, the variables of interest are computationally updated at each time step using only their values at the previous time step, starting from the initial conditions or boundary conditions that must have been provided. Computations are executed on the fly without preserving memory of prior states. Mathematical modeling techniques such as adaptive feedback control with memory windows or memory forgetting factors are exceptions. Agent-based models easily allow to keep track of past states or to keep track of the local and past time span window for any agent of interest. For instance, in the mRNA translation model, the memory is kept of the amino acid sequence of the last stretch of polypeptide that is embedded in the ribosome exit tunnel spanning a window of 50 or more amino acid residues. The purpose is to recall the charged amino acid spatial distributions in the tunnel to calculate the electrostatic interaction currently occurring in the tunnel and due to past elongated amino acids. In summary, ABM easily supports local and temporal context integration.

3. ABMs is compatible with ***multiple instantiations***. In an ABM, each agent class has multiple manifestations (called instantiations in the language of object-oriented programming) as autonomous computational objects forming a population of agents that interact in a usually emulated parallel processing environment. Differing local conditions or differing past recollected specific events lead to different behavioral trajectories of the individual agents, such that the heterogeneous behavior of an individual agent within a population of agents results in aggregated system dynamics. The population dynamics are one of the output of the ABM that represent higher-level system behavior. A known

classical example of this phenomenon is the behavior of the flock of starling birds, in which simulations utilizing a relatively limited set of interaction rules among birds (separation, alignment and cohesion rules) can lead to sophisticated flocking patterns without an overall controller, see the 'Boids' algorithm by Craig Reynolds [Reynolds 1987] and a nice recent javascript implementation by Daniel Shifman [Shifman 2018].

4. ABMs incorporate ***stochasticity***. Many systems, particularly biological ones, include behaviors that appear to be random. 'Appear to' is the important distinction, since what may appear to be random at an observational level may be fully deterministic from a mathematical standpoint. However, despite the fact that a particular system may follow deterministic rules, at the observational level it is difficult or impossible to define the rules or initial conditions within the system with only one observation. ABMs address this issue via the generation of population of agents. Probabilities of a particular outcome are determined for the population as a whole, complying with a probability function that governs the behavior of individual agents. This function can be incorporated into each agent's rules. As the simulation runs and a population of agents is executed, each agent follows a unique behavioral trajectory, with its probabilistic rules gradually collapsing (i.e., being resolved) over time. This process enables the generation of a population of behavioral outputs from a single ABM, producing system behavioral spaces consistent with population-level biological observations. The stochastic behavior of agents can be nested in different levels amplifying the resulting apparent variability in the population outcome. For instance, in the mRNA translation model, the rate of a single protein elongation is codon dependent. A ribosome residence time on a particular codon is sampled from a probability distribution of the queueing time for that codon to be paired with the cognate tRNA anti-codon. There are 61 different probability distributions for the 61 codons. Across simulation runs, the collapsed and aggregated sampled times are different and provide different outcome for the population of protein rates. This stochastic loop is embedded within another sampling rule that ribosomes follow: initiation on a given transcript is sampled from a multinomial distribution. As a result, the outcome space at the protein population level is vast and appears entirely unpredictable. However, by running thousands or even millions of simulation replicates on a computing cluster across a given range of model parameters, collective behavior patterns can be identified and subjected to statistical analysis. Evolutionary trajectories can be tracked and analyzed to distinguish parameter domains that exhibit more deterministic or more stochastic behavior. One key objective of conducting large-scale simulation replicates is to detect emergent transitions or bifurcations and determine the conditions under which they occur—such as the emergence of an 'oncoproteome'—as observed in real biological specimens.

5. ABMs have a ***modular and reusable structure***. This feature is common with all object-oriented programs. The behavior of an ABM is largely defined by the rules of its constituent parts. New information can be added either through the introduction of new agent-types or by modification of existing agent rules without having to re-engineer the entire configuration of the model. In the mRNA translation model, various contributing factors can be incorporated modularly. For example, the elongation rate of a ribosome at codon resolution may be modulated by the electrostatic interactions of the nascent protein within the ribosome exit tunnel. This is achieved by modifying the rule governing the ribosome agent's elongation rate: the queueing time sampling is either multiplied by a Maxwell-Boltzmann factor or left unchanged. This factor accounts for variations in the energy barrier that must be overcome during the catalytic transition state of peptide bond formation. When electrostatic interactions are included, the energy barrier is altered by the mechanical forces experienced by the peptide within the tunnel.

   Similarly, additional factors can be integrated, such as the proline-induced slowdown or the effects of mRNA secondary structure. In the latter case, short hairpin loops formed by self-pairing of complementary nucleotides within the single-stranded mRNA molecule can impede ribosome translocation along the transcript. The usefulness of the modular and reusable feature of an ABM will clearly appear in the next chapters.

   Finally, multiple agent-based models (ABMs) can be combined, provided their interfaces and interactions remain consistent across the incorporated models.

6. ABMs reproduce ***emergent properties***. A central hallmark of agent-based modeling (ABM) is the emergence of systemic dynamics driven by parallelism, competition among agents for limited resources, non-equilibrium conditions, inherent stochasticity, and locally constrained agent rules. These dynamics cannot be reasonably inferred from an examination of individual agent rules alone—a phenomenon known as *emergent behavior*.

   For example, in the case of starling flocking behavior, a superficial observation might suggest the necessity of a central leader orchestrating the coordinated movements of the entire flock. This assumption would imply the existence of predefined, flock-wide command and control mechanisms, whereas in reality, complex group behavior emerges from simple local interaction rules followed by individual birds. This, however, is not nature's way. Birds function on a series of interacting rules at the scale of the individuals, that are locally-constrained. The flocking (also called murmuration) behavior emerges from the collective aggregation of these interactions. The capacity to generate emergent behavior is a remarkable advantage of using ABM for conceptual model verification, as it is often the paradoxical, nonintuitive nature of emergent behavior that breaks a conceptual model.

7. ABMs ***do not require complete knowledge***. ABMs can be constructed in the absence of complete knowledge, keeping the rules as simple and verifiable as possible, even at the expense of some detail. This modeling workflow also forces the question of how much is really known on the system being studied. Hypothesis testing via ABM provides qualitative verification of possible outcomes given incomplete knowledge. Since mechanisms in biology are to some degree always incompletely known, ABM fits well to the means by which experimental biomedical knowledge is currently expressed. Spatially influenced combinatorial freedom among agents means that the emergent behavior has a range of stochasticity, similar to real biology. These outcomes can be subjected to statistical analysis to assess qualitative trends based on the available prior knowledge. In general, a more detailed ABM will lead to a stronger correlation to the real-world and a greater confidence in the ability of the ABM to describe observable phenomena. However, given the inherent incompleteness of biological knowledge, the relationship between the ABM and the reference system's behavior will likely never rise to the level of strictly accurate prediction.

Agent-based simulations of molecular systems enable the exploration of extended time scales while incorporating multi-scale spatial resolution. For example, transcripts of varying lengths and sequences (mRNA molecules) consist of discrete codons, whose decoding dynamics operate on a nanometer spatial scale and a millisecond time scale. These transcripts are translated by a shared pool of ribosomes within the subcellular cytoplasmic compartment, where protein synthesis unfolds over minutes to hours.

ABMs inherently account for spatial constraints and environmental heterogeneity. For instance, transcripts are organized into open reading frames (ORFs), consisting of non-overlapping triplets of nucleotides (codons) that include a start and stop codon. In specific cellular or tissue contexts, the ribosome pool is a limited resource. When ribosomes are scarce, differential protein production arises from transcript competition, where initiation probabilities can follow arbitrary distributions. ABMs also allow for the incorporation of local and non-homogeneous molecular distributions, a key feature of many biological processes.

A major advantage of ABMs is their ability to track individual particles or objects over time, making them particularly suited for *in-singulo* process tracking. For instance, the trajectory of a single transcript or ribosome can be monitored throughout a simulation, with the individuality and integrity of each agent preserved. Furthermore, ABM simulations predict emergent behaviors in complex molecular or subcellular systems by applying rules governing individual components.

In molecular systems biology, a primary objective is to understand the overall functionality of a molecular system and how different parameter values influence its behavior. While some parameter values may be experimentally measurable, others may not. In such cases, ABM simulations offer indirect insights into biologically

relevant phenomena that may not be directly observable but still contribute to system-level understanding. The set of values of these parameters may be experimentally biologically measurable or not. If not, ABM simulations provide indirect insights into pieces of information that could be biologically relevant although not directly measurable.

3.4 Agent-Based Modeling Platforms and Tools

The study of autonomous agents interacting within a virtual environment dates back to the start of computer science and the self-replicating machines proposed by von Neumann in the 1940s. The first use of the term 'agent' with the same meaning as in 'Agent-based models' (ABM) appeared in the 1980-1990s. During this period, increases in computing power made it feasible to simulate systems of a useful size, and rapid growth was seen in the number of tools available to support researchers in this area. Some of the most popular were based around the Logo programming language (e.g. NetLogo). Because of its simplicity, Logo was perfectly suited to allow anyone to define the rules of an agent based simulation and study the emergent behavior that could arise. The use of ABM has continued to expand in the fields of economics, social behavior, statistical physics, ecology, ethology, microbiology, epidemiology, system biology, molecular biology as well as many others.

Many ABMs are created using existing, general-purpose ABM development environments. These software packages are aimed at striking a balance between representational capacity, computational efficiency, and user-friendliness. Among the most populat ABM toolkits are Swarm, Mason, RePast, NetLogo, StarLogo and PhysiCell or CompuCell3D used in biomedical engineering. All these platforms represent some trade-off among the goals mentioned above. A review and comparison of agent-based modeling toolkits can be found in [Railsback et al. 2006]. Some issues encountered with common ABM toolkits include means of implementing highly intensive and parallel computing capacities to meet concurrency and the need of using schedulers to account for the multiple replications of the simulations that is required to explore the parameter space of the ABMs.

3.5 Object Oriented Programming for Agent-Based Model Implementation

In computational and data science, programming comes in two flavors. Typically, there are two generic methods available for programmers [Agarwal and Gaddis 2014]:

- ***procedural*** programming
- ***object-oriented*** programming

The earliest programming languages were procedural, meaning a program was made of one or more procedures. A procedure can simply be thought of as a function that performs specific tasks sequentially such as gathering input from the user or from an input file, performing calculations, writing an output file, displaying output in a graph, and so on. Procedures operate on data items that are separate from the procedures. In a procedural program, the data items are passed from one procedure to another. There is a clear separation of the data and the code that operates on the data. As the program becomes larger and more complex, this separation of data and the code that operates on them can lead to problems.

Whereas procedural programming is centered on creating procedures (functions), object-oriented programming (OOP) is centered on creating objects. An ***object*** is a software entity that contains both data and procedures. The data contained in an object is known as the object's data *attributes*. Object's data attributes are simply variables that reference data. The procedures that an object performs are known as *methods*. Object's methods are functions that perform operations on the object's data attributes. The object is, conceptually, a self-contained abstract unit that consists of data attributes and methods operating on the data attributes. OOP addresses the problem of code and data separation through encapsulation and data hiding. *Encapsulation* refers to the combining of data and code into a single object. *Data hiding* refers to an object's ability to hide its data attributes from code that is outside of the object. Only the object's methods may directly access and make changes to the object's data attributes. In OOP, objects are created as instances of an object's type called a class. A *class* is a piece of generic code that specifies the data attributes and methods for a particular type of objects. A class can be thought of as a blueprint that objects may be created from. A class is a description of an object's characteristics. When a program is running, it can use the class to create, in memory, as many objects of a specific type as needed. Each object that is created from a class is called an *instance* of the class.

Object-oriented programming, as a programming method, is actually conceptually very close to agent-based modeling and this programming method is well suited to implement agent-based models computationally. There is indeed a natural correspondence between

agents, their attributes and their rules on the one hand, with objects, their attributes and methods on the other hand. This makes OOP a very good tool to build an ABM.

In the example of our ABM for mRNA translation, a class 'ribosome' is defined and multiple ribosomes objects are created as multiple instances of the class ribosome. We also have a class to define a 'transcript' and have created many mRNA molecules objects as many instantiations of the class transcript. The objects ribosomes can interact with the objects transcripts. Another class is used to define a 'codon'. Codon objects are instances of the class 'codon'. An object from the class 'transcript' incorporates codon objects.

Both Javascript (and its P5.js processing) and Python are high-level programming language with object-oriented programming functionalities and including emulated multi-threading or multi-processing modules. This makes these languages well suited to build ABMs.

3.6 TASEP: Totally Asymmetric Simple Exclusion Process: a model for mRNA translation

The Totally Asymmetric Simple Exclusion Process (TASEP) refers to a statistical physics inspired diffusion process modeled by a lattice of discrete sites in one dimension, Fig. 3.1.

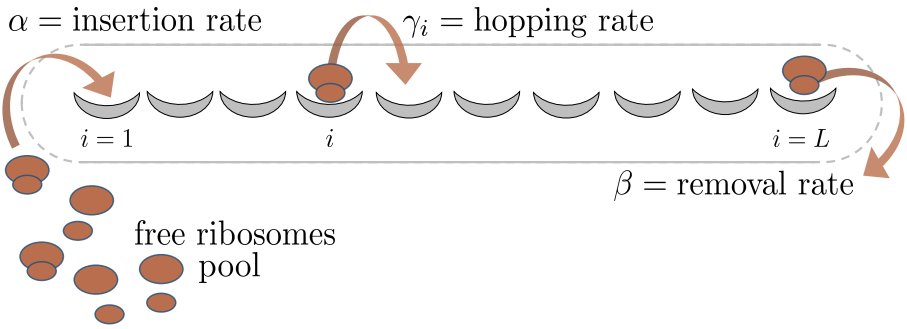


Figure 3.1: Diffusion process of particles in a lattice of discrete sites in one dimension, called Totally Asymmetric Simple Exclusion Process (TASEP). The lattice and its sites correspond to the transcript and the codons. The rates of insertion, hopping to the next site, and termination correspond to initiation, elongation and termination rates, respectively, in the mRNA translation by ribosomes.

Autonomous particles can move by unit steps from one site to the next neighboring site of the lattice in a single direction (asymmetric), if and only if, the next site is free of other particles (simple exclusion). The model implements rules of initiation, elongation and termination that are stochastic. A free particle that is not on the lattice may jump on the first site of the lattice following an exponential probability density distribution parametrized by an initiation rate. A particle already engaged in the lattice may hop from one site to the next neighboring one at an elongation rate defined by another probability density distribution and may leave the last site of the lattice by a third probability density distribution parametrized with a termination rate.

In homogeneous TASEP, all the elongation probability density distributions are equal across the inner sites of the lattice. In heterogeneous TASEP, each site may have its own elongation probability density distribution. In extended TASEP, the free particles can engage on array-like (multiple) 1D-lattices. The extended heterogeneous TASEP model is used to model the competitive translation of multiple mRNAs by a pool of ribosomes. `Ribosomer` is our ABM of protein translation based on the TASEP model and the aforementioned thesis legacy.

3.7 A digital twin of the protein elongation cycle with its modulating factors

The construction of the object classes in `Ribosomer` stems from an abstraction of the elongation cycle of a ribosome on a transcript, as previously introduced in Chapter 1. Multiple instances of ribosomes can either stay free in the ribosome pool, or engage in translation, each on a different instance of a transcript object. Multiple ribosomes instances can engage translation of the same transcript (polysome). Each transcript object is made of a particular sequence of other objects belonging to the codon class. **The ribosomes interact with the transcripts according to rules that are context dependent**. The rules are determined by the codon class. **Codon attributes assign stochastic properties for the behavior (rule) of the ribosome instance**.

When a ribosome instance is in elongation mode, it goes through three substeps: accommodation, peptide bond formation and translocation. The ribosome instance has a **timer** attribute with **queueing time setpoint** attributes that are sampled from predefined probability density functions specific to instances of the codon class. The parameters of these probability density functions are determined by the P-site or A-site codon the ribosome is dwelling on.

Figure 3.2 shows the toggle buttons in the setting page of the `Ribosomer` user interface (Javascript version). The setpoints of the timers also depend on different modulating factors that are context dependent and can be toggled on or off.

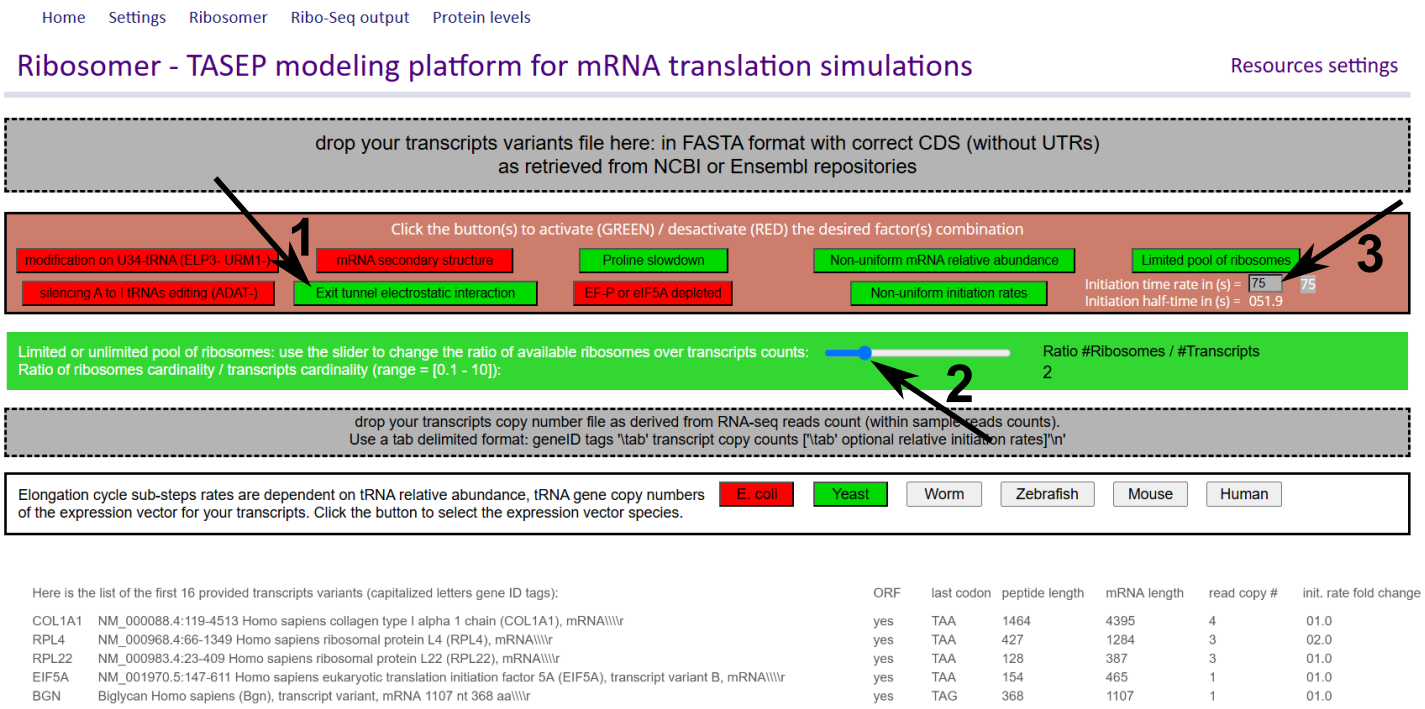


Figure 3.2: Input sheet for the settings of `Ribosomer`. Toggle buttons allow users to individually or simultaneously activate the factors deemed relevant in protein synthesis by ribosomes before running a simulation (arrow 1). A slider allows for the parametrization of the ribosome pool ratio (arrow 2), and an input box enables the user to specify the general initiation rate (arrow 3) associated to the transcript input files (dashed grey boxes).

The toggle buttons allow users to individually or simultaneously incorporate the factors deemed relevant in protein synthesis by ribosomes before running a simulation. This functionality allows the user to compare, through *in silico* simulations, the effect of each of these factors individually or collectively. The contribution to the uncertainty or noise can be further explored by statistical analysis comparing case with control for as many simulation replicates as desired in each case and control. To conduct these *in silico* simulations, the working pipeline is implemented through a scheduler that will use the Python version of `Ribosomer`. This implementation is best fit to *embarrassingly parallel* execution on a high performance computing cluster platform such as `CéCI`[1] (section 3.9 below).

3.8 Agents in the model of mRNA translation by ribosomes and ABM structure

It is common to describe the organization of the classes with formatted diagrams called UML, Unified Modeling Language. In object-oriented programming (OOP), a class serves as a blueprint for creating objects. It encapsulates attributes (data members) and methods (functions) that define the behavior and state of the objects instantiated from it. Among the essential methods in a class are constructors, getters (accessors), and setters (mutators), each serving a distinct role in object management.

i. **constructor attributes** :
   A constructor is a special method invoked automatically upon object instantiation. It initializes an object's attributes, often assigning default values or parameters provided during object creation. In many programming languages, constructors share the class name (e.g.,__init__ in Python).

ii. **getters (or accessors) methods**:
   Getters, also known as accessors, are methods designed to retrieve the value of private or protected attributes. They enforce encapsulation by allowing controlled access to an object's internal state without exposing its implementation details. Typically, a getter follows a naming convention such as get_AttributeName() in Javascript or in Python.

iii. **setters (or mutators) methods**:
   Setters, or mutators, are methods that allow controlled modification of private or protected attributes. They validate and update an attribute's value, ensuring

[1]Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

consistency and preventing unintended modifications. Setters often follow a naming convention like set_AttributeName(value).

The main program in the `Ribosomer` ABM consists of four principal parts: (i) instantiating the objects of the classes; (ii) defining the functions governing interactions between objects; (iii) backtracking past events; and (iv) collecting and aggregating statistics such as protein counts, number of ribosome per transcript (polysome profiling), footprint sequences, ribosome footprint counts to provide ribosome density maps for any chosen transcript. The main attributes and methods of the Lattice, Codon, and Ribosome classes are represented in a simplified UML diagram in Figure 3.3.

1. **class** `Lattice`: the transcript class.
   The term *lattice* is a reminder of the TASEP inspired approach to describe a transcript as a one dimensional lattice that is processed asymmetrically from left to right by a ribosome. The transcript class and each transcript object has 12 attributes. At least five of these attributes need to be initiated at the creation of a particular transcript instance. A transcript has a length in codon scale (`self._length`), a sequence of nucleotides (`self._sequence`), a mRNA length (in nucleotides scale), a number of copies i.e., number of copies of this transcript as known from RNA-seq (`self._copy`), as well as a name i.e., the gene tag name, a numerical ID, a unique transcript ID (`self._uniqueID`), all borrowed from the FASTA format input file.

   Some of the attributes can be updated based on interactions with other objects, such as instances of the ribosome class. Specifically, a given transcript can be partially occupied, with a ribosome protecting a 31-nucleotide fragment, by a single ribosome or multiple ribosomes (polysomes). At least four attributes of each transcript object instance are updated at each computing cycle of the program. The attribute `self._currentFootPrint` is an array recording the P-site location(s), in codon numbering, of all ribosomes that are elongating this instance of the transcript object. If there is no ribosome engaged in elongation on this transcript instance, the array is default to a single value of -1.

   The number of times that this specific copy of the transcript instance has been translated since the start of the simulation is updated in the `.self_readTranslated`. The aggregated number of proteins produced by translation of the transcript instances having the same geneID tag is updated in the `self._countTranslated` attribute. The number of ribosome protected fragments on the transcript instance is updated in the `self._rpfCount` attribute.

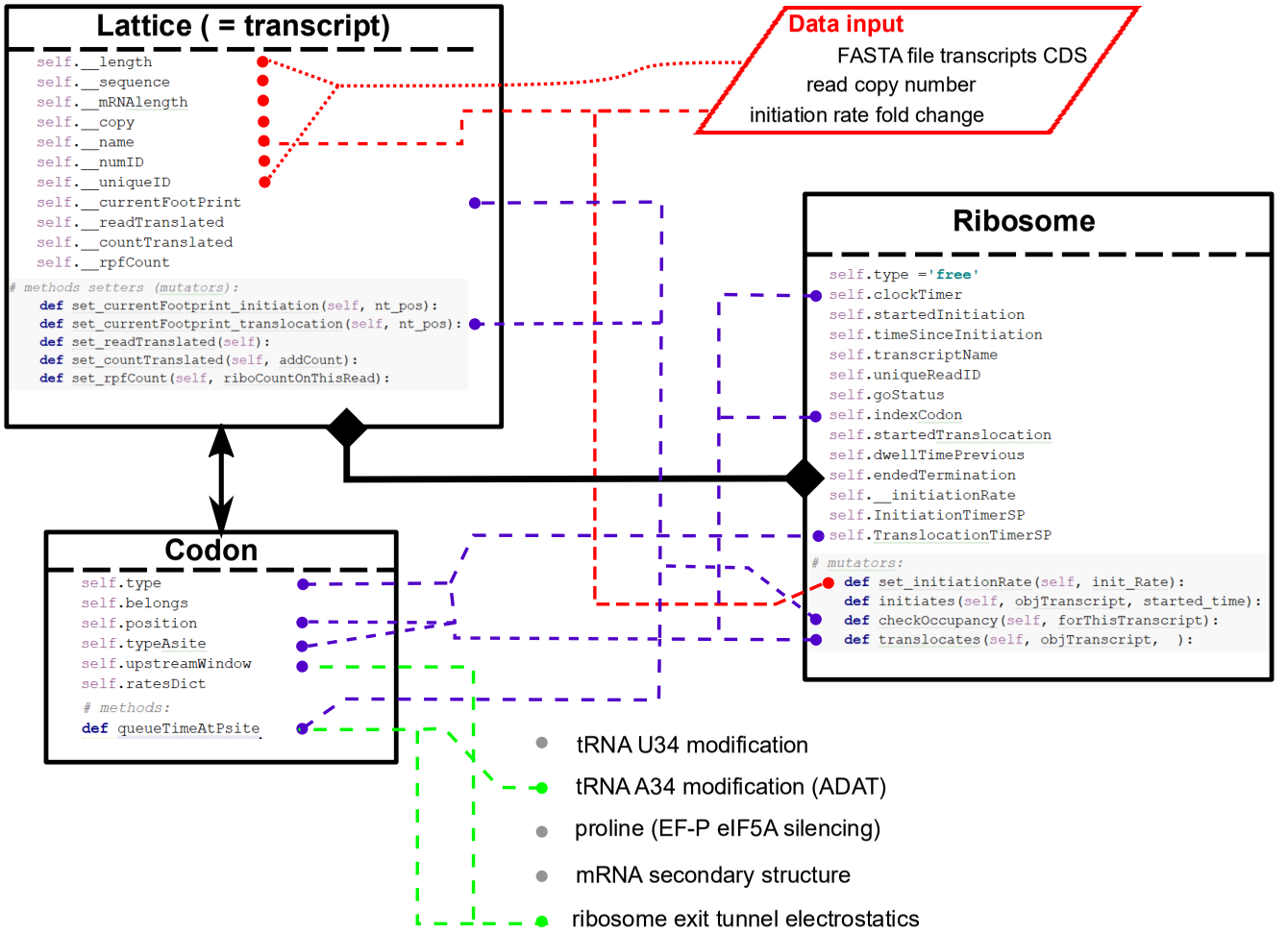


Figure 3.3: Unified Modeling Language (UML) simplified diagram of the Ribosomer agent-based model. The diagram depicts the three classes with their main attributes and methods. Colored dashed lines represent a subset of the functional links. Explanations are provided in the main text.

The reading access to the previous attributes is possible through the associated getters (accessors). To change or to update the attributes of the transcript object instance, only the following setters (mutators) affect the allowed (mutable) attributes of this transcript:

- `transcript.set_currentFootprint_initiation(nt_pos)` where the position of the initiating ribosome will be recorded on the transcript. This is associated to the event of a *free* ribosome engaging in translation and becoming a *translating* ribosome.
- `transcript.set_currentFootprint_translocation(nt_pos)` where the position of the current ribosome elongating this transcript will be updated upon a translocation event. The ribosome keeps his translating status.

2. **class** `Codon`: the codon class.
The codon class and each codon object has six attributes. These attributes are initiated upon creation of an instance of a codon object and can be updated according to interactions between a ribosome object and a transcript object. The codon attributes that are assigned at creation (constructors) are: `self.type`, `self.belongs`, `self.position`, `self.typeAsite`, `self.upstreamWindow`, `self.ratesDict`.

   The `type` attribute retrieves the triplet sequence of the transcript occupied by a ribosome at its current P-site position. The type `belongs` refers to the transcript name tag to which the codon belongs. The `position` refers to the codon ordered number within the transcript it belongs. The `A-site` attribute records the triplet just downstream (A-site) the codon type (P-site). The `upstreamWindow` attribute records the current 150 nucleotides (50 codons) that are upstream the current codon at the P-site. This last piece of information is useful to eventually compute the interaction of the nascent chain with the ribosome exit tunnel (chapter 6). The `ratesDict` attribute records the array of three parameters in the defaulted statistical probability density function from which a queueing timer will be sampled. The codon type at the P-site will assign a specific queueing time probability density function for the time a ribosome will stay on this specific codon of this specific transcript (chapter 4).

   Finally, there is one mutator method in the codon class: `queueingTimeAtPiste` that will compute the exact hypo-exponential probability density function depending on the context (chapter 4). Depending on the toggle buttons status (activated or deactivated), the different factors influencing the elongation rate will be incorporated or not. These factors (Fig. 3.4) are:

   i. the tRNA U34 or A34 (ADAT) enzymatic modifications (chapter 5)
   ii. the electrostatic interaction with the ribosome exit tunnel (chapter 6)
   iii. the proline slowdown effect, with or without `EF-P/eIF5A` silencing effect (chapter 7)

iv. the mRNA secondary structure roadblocks (chapter 8)

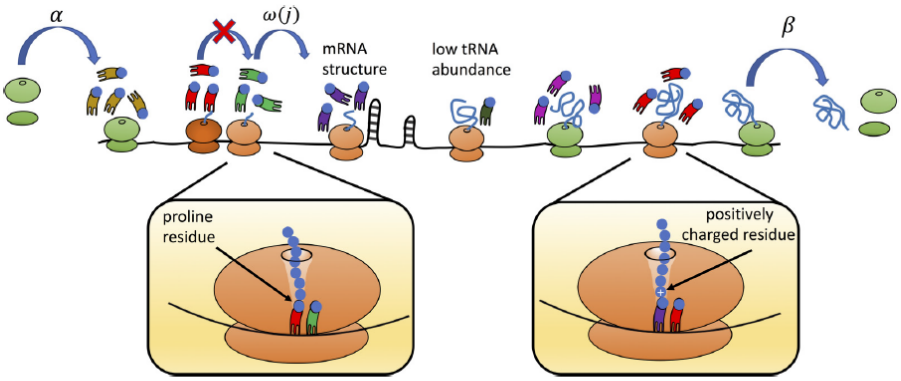


Figure 3.4: Illustration of the contextual factors in the TASEP model. A ribosome initiates translation with rate α when the first six codon positions after the start codon are not blocked by another ribosome. The ribosome translates the j^{th} codon position with elongation rate ω_j when no downstream ribosome occupies the $(j+10)^{\text{th}}$ codon position and terminates the translation process with rate β. Contextual factors influence the elongation rate: tRNA abundance, tRNA enzymatic modifications, electrostatic interaction with the ribosome exit tunnel, proline residues at the A- or P-sites, mRNA secondary structure roadblocks are accounted for in the model. Ribosomes in green and light-brown are translating fast and slow codons, respectively, whereas the ribosome colored in dark-brown is sterically blocked by a downstream ribosome (traffic congestion). Reproduced from [Sharma et al. 2018].

3. **class** `Ribosome`: the ribosome class.
   The Ribosome class and each ribosome object possess more than ten attributes. These attributes are initialized upon the creation of a ribosome object and can be dynamically updated based on interactions between the ribosome, the transcript it is translating, and the codon it is currently occupying–unless it remains *free* in the non-translating ribosome pool. Key attributes include `self.type`, `self.clockTimer`, `self.transcriptName`, `self.uniqueReadID`, `self.indexCodon`, `self.goStatus`. The `type` attribute of the ribosome specifies if the ribosome is in the pool of *free* ribosomes before initiation. Upon creation, all ribosome objects are initially *free*. They initiate translation on a randomly selected transcript copy at a rate governed by first-order kinetics, determined by the general initiation rate (arrow 3) in Figure 3.2). The initiation rate for specific transcripts can be modulated according to a predefined fold-change prescription, specified at the start of the simulation in

the input file (Figure 3.3). This file assigns both transcript copy numbers and the fold-changes relative to the standard initiation rate. By default, the fold-changes are all set to 1, meaning all transcript copies have an equal probability of being initiated by ribosomes, with initiation events uniformly distributed across transcripts. However, if a fold-change other than 1 is specified for particular gene tags, the initiation probabilities are adjusted accordingly. In this case, the likelihood of a given transcript being initiated follows a multinomial distribution. More on this in chapter 9. Once initiated, the ribosome type is not *free* anymore and becomes *translating*. The ribosome instance updates attributes such as the transcript `name`, its `uniqueID`, and the `indexCodon` it is currently occupying at P-site. Additionally, it assigns a setpoint (SP) to its `translocationTimerSP` attribute by sampling from the appropriate queueing time probability density distribution, which depends on both codon identity and contextual factors.

It may take several computing cycles before the ribosome's `clockTimer` reaches the setpoint, allowing translocation to the next codon on the same transcript instance. A key attribute is `goStatus`, which, along with `clockTimer` and `translocationTimerSP`, plays a critical role in the ribosome object's methods. For each cycle in the main program, all ribosomes can take the following actions (methods of the class):

i. `ribosome.checkOccupancy()`
   This method verifies whether the next six codons (18 nucleotides) downstream of the current P-site are unoccupied, allowing for potential initiation at a start codon if the ribosome object type attribute is *free*. If the ribosome is actively *translating*, the `checkOccupancy` method instead examines the next 10 codons. The method returns a boolean variable, `freeToGo`, which serves as a prerequisite for either initiation or translocation. If `freeToGo` is `False`, a **traffic jam** occurs, causing the ribosome **to stall** for the duration of the current computational time step or multiple steps until the **congestion** is resolved.

ii. `ribosome.initiates()`
   Only *free* ribosomes can initiate transcripts. This initiates method is conditional on a free to go status checked by the `checkOccupancy` method.

iii. `ribosome.translocates()`
   The ribosome `translocates` method receives in arguments the transcript read copy object on which it operates. The ribosome monitors its timer and waits until the setpoint for translocation is reached. Once the elapsed time exceeds the setpoint, it verifies whether it is free to proceed using the `checkOccupancy` method. If the condition is met, the ribosome translocates; however, two possible different subsequent events may follow.

   (a) If the next codon is a stop codon, the translocation will be the final one, and the ribosome will detach from the transcript after a time span

determined by the termination rate. The ribosome then returns to the *free* pool, with its attributes reset to those of a *free* ribosome.
Additionally, the attributes of the transcript instance that has just undergone termination are updated: a ribosome footprint count is decremented, the current footprint loses the last ribosome footprint, the protein count for the corresponding transcript tag name is increased by one unit, and the total protein output from this transcript read copy is also incremented.

(b) If the next codon is a sense codon, the translocation method is executed once the queueing time reaches the `translocationTimerSP` setpoint. This method (a mutator method) updates several attributes, not only of the ribosome instance but also of the transcript (Lattice instance) it is translating.
Specifically, the current ribosome footprint on the transcript shifts one codon to the right, the codon index increments by one, and the `upstreamWindow` is updated accordingly. Additionally, the `translocationTimerSP` samples a new setpoint from the probability density function, parameterized based on the newly occupied codon key. This codon key corresponds to a specific set of elongation rates retrieved from the rate dictionary of the 61 sense codons. Depending on the surrounding context, these rates may be modulated by the appropriate toggle factors discussed previously.

The complete codes of the classes and the main programs are provided in the GitHub repository, in Javascript for the didactic tool and in Python for the high performance computing simulations, respectively. The names of the source files are:

I. Javascript (p5.js) implementation: `objectsClasses.js`, `settings_sketch.js`, `protein_sketch.js`, `riboSeq_sketch.js`

II. Python implementation: `myClasses.py`, `myProgram.py`.

The simulation proceeds in discrete time steps until a predefined total simulation time is reached. A realistic simulation run time should span from approximately one hour to several hours, reflecting the physiological time scale of protein synthesis in active cells. The limiting time scales involved in protein synthesis range from 10 to 200 seconds per initiation event (initiation rates). Furthermore, the production of a small protein by a single ribosome takes about 1.5 minutes. In eukaryotes, for instance, translating a sequence of 500 codons at an elongation rate of 5.6 amino acids per second requires approximately 1.5 minutes of total elongation time. However, simulation times of only a few minutes are insufficient for ribosomes to collectively reach a steady state. To ensure that a steady state is achieved, the simulation time must be at least 30 to 50

times longer than the physiological time scale. Consequently, the total simulation run time was arbitrarily set to a minimum of ~ 108 min., which is longer than 50×1.5 min. Preliminary pilot simulations confirmed that these simulation run times are long enough to achieve steady state.

At the end of the simulation, aggregated statistics are generated, including polysome profiles–either across the entire transcriptome (all transcripts specified in the input data file) or for a selected transcript–Ribo-Seq density maps for any chosen transcript, protein relative abundance histograms, and translation efficiency histograms.

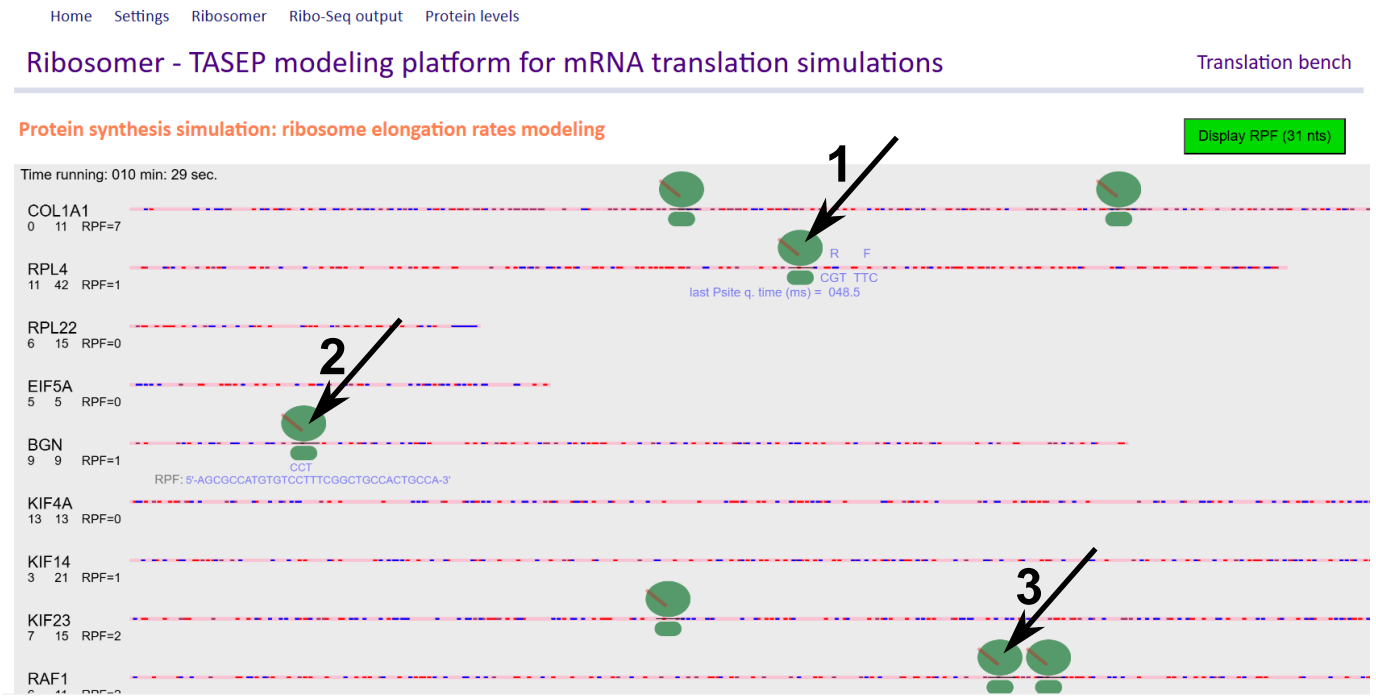


Figure 3.5: Ribosomer snapshot of a simulation run after 10 min. and 29 s. Arrow 1: a monosome on the **RPL4** transcript with P and A site occupied with **CGT** and **TTC** codons respectively. Arrow 2: monosome and its 31 nucleotides protected fragment sequence on the **BGN** transcript. Arrow 3: ribosomes congestion (disome). Explanations in main text.

Figure 3.5 shows a snapshot from a simulation run at time $t = 10, \text{min}, 29, \text{sec}$. At this moment, the transcript copy `RPL4` is footprinted by a single ribosome (RFF count = 1 next to the right of the gene ID tag in the figure). The current codons at the P- and A-sites are `CGT` and `TTC` (arrow 1), respectively, which correspond to arginine (R) and phenylalanine (F) in the codon-to-amino acid translation (using CDS nucleotide nomenclature from the FASTA file). Since the beginning of the simulation, this particular `RPL4` copy has been translated 11 times. Given that multiple copies of this transcript exist, the total number of proteins produced by all `RPL4` copies is 42, as indicated under the `RPL4` tag in the figure. There is one ribosome translating a single copy of transcript `BGN` for which 9 proteins were already produced. On this transcript, the associated ribosome protected fragment has a sequence of 31 nucleotides which is displayed explicitly on the figure (arrow 2). If a Ribo-seq experiment were conducted at this snapshot instant, this sequence would correspond to the precise Ribo-seq read. On the `RAF1` transcript, at the time of the simulation, two ribosomes are almost in a traffic jam situation (stalling), as shown in the figure (arrow 3). In the figure, transcripts are represented as straight lines, with their lengths proportional to the corresponding mRNA lengths. Each nucleotide is represented by one pixel (three pixels per codon) on the canvas. The color code used is as follows: dark pink for a triplet coding for proline (P), blue for a triplet coding for negatively charged amino acids (D, E), and red for a triplet coding for positively charged amino acids (K, R). The memory of the last 50 amino acids is stored in a mobile window, which tracks the nascent chain embedded in the ribosome exit tunnel. More details on this will be provided in the following chapters.

Figure 3.6 shows the pool of free ribosomes (not engaged in translation) at the same snapshot instant of the simulation run as in the previous figure. The ribosome pool ratio was set to 2.0. For 58 transcripts, this amounts to a total number of ribosomes equal to 116. The figure shows, at this moment, 25 (= 21%) *free* ribosomes while 91 (79%) are actively engaged in translation. This complies with a dynamics for which the global initiation rate was set to a mean time value of 75 sec between initiation. This corresponds to a half-life time of the initial pool of free ribosomes of 51.9 sec. On average, it will take around 51.9 sec. before half of the initially *free* 116 ribosomes become engaged in translation of the 58 transcripts copies. Across this transcriptome, the average distance between two ribosomes is around $\sim$ 366 codons ($\sim$ 1100 nts) as shown above the pool of free ribosomes in the figure.

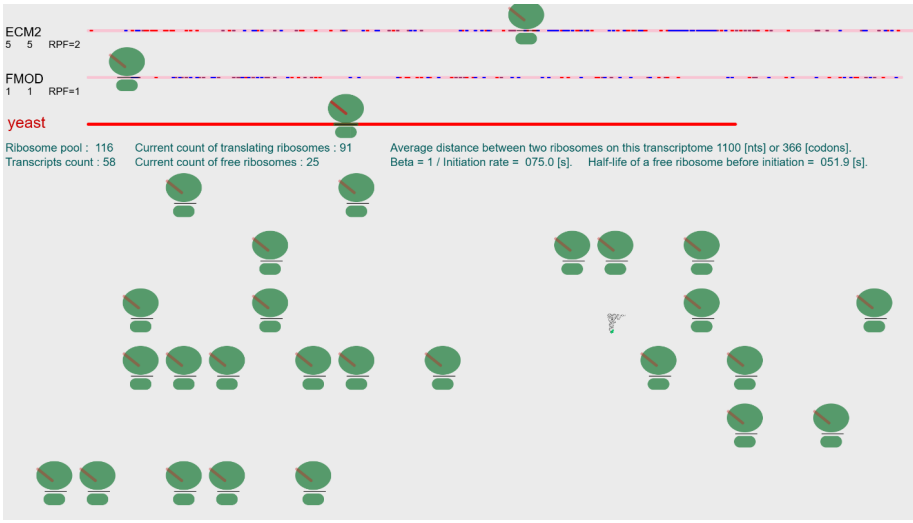


Figure 3.6: Pool of free ribosomes awaiting transcript initiation, governed by a first-order kinetics initiation rate.

3.9 Computer cluster resources and methods for computer intensive agent-based model simulations

A major benefit of using ABMs is their ability to generate, through simulation, non-linear transitions between multiple scales of organization and emerging patterns. This can also be a downsize for uncertainty quantification requiring mathematical continuity and smoothness. Determining parameter settings that lead to different patterns can be very demanding in terms of computing resources. Exhaustive simulation to investigate bifurcations, stability or instability, though cheaper and faster than real-world experimentation, can be expensive in terms of computing cycles and the resources needed to execute them. The high flexibility of ABM comes with the cost of an expansion of the parameter space of the model. The parameter spaces in many ABMs are quite large compared with conventional models. Exploring the simulation output space of an ABM may require exhaustive search of the set of possible values and actions in the parameter space. Broader parameter sets in ABM impose to examine the effects of different parameters using the tools of applied statistics [Marchi and Page 2014] and to resort to the statistical methods of design of (computer) experiments. The principal method of analysis for ABMs remains extensive simulation. A comprehensive and exhaustive search of model-response space calls for a so-called ***embarrassingly parallel simulation***. This requires executing job arrays with a job scheduler for the management of resource sharing in a high-performance cluster dedicated to scientific computing.

The CECI ('Consortium d'équipements de calcul intensif') high-performance computing cluster was used to run roughly $6,000$ simulations, for a total of $\sim 18,000$ core hours. These simulations were part of a multi-parametric sensitivity analysis, designed according to design of experiment principles, using a fully-crossed factorial design (more details in chapter 10 and 11).

Specifically, the NIC5 cluster, hosted in University of Liège, was used. This cluster has 4672 central processing units (CPUs), operating at a 2.9 GHz frequency and organized in 73 nodes of 64 CPU per node. The random access memory (RAM) per node is in the range 256 GB-1 TB. Running software on such a high-performance computer cluster lead us to code the executable main program of the ABM, not with general purpose ABM available toolkits such as NetLogo, but in a high-level language broadly known by the community of computational biologists and bioinformaticians, namely Python. This is one of the reasons `Ribosomer` was re-factored from Javascript (processing p5.js) to Python.

3.10 Summary of main findings and insights

The chapter 3 introduced the `Ribosomer` framework, an agent-based model (ABM) designed to simulate mRNA translation by ribosomes. The chapter provided an accessible introduction to agent-based modeling and its relevance to computational biology. It explained the rationale for using an ABM to investigate competitive translation dynamics, where multiple transcripts are translated by a shared ribosome pool. The model is built on the **Totally Asymmetric Simple Exclusion Process (TASEP)**, a statistical physics-inspired approach that captures ribosome movement along mRNA sequences. The chapter describes the object-oriented structure of the model, detailing the attributes and interactions of agents such as ribosomes, transcripts, and codons. Visual representations, including **Unified Modeling Language (UML) diagrams**, illustrate the architecture of the digitally modeled process. The final sections discussed computational resources and methodologies, emphasizing the execution of large scale simulations on high performance computing clusters. The mechanistic principles underlying the elongation cycle and ribosome queueing dynamics on codons will be explored further in subsequent chapters.

Chapter 4

Mechanobiochemistry in the elongation cycle and queueing time statistical theory of the ribosome on a codon

The inherent probabilistic nature means that we can never precisely predict stochastic outcomes. This might be why we opt to use the more pretentious-sounding word "stochastic" instead of just saying "random". At least in common parlance, "random" has connotations of hopelessness. To try to understand something random feels futile; it's just random. But of course, we can understand many things about stochastic-random processes, despite this inherent unpredictability, by studying the probability distributions.

Professor Karen Abbott
Case Western Reserve University, Department of Biology, Theoretical Ecology and Evolution, Cleveland , Ohio, USA.

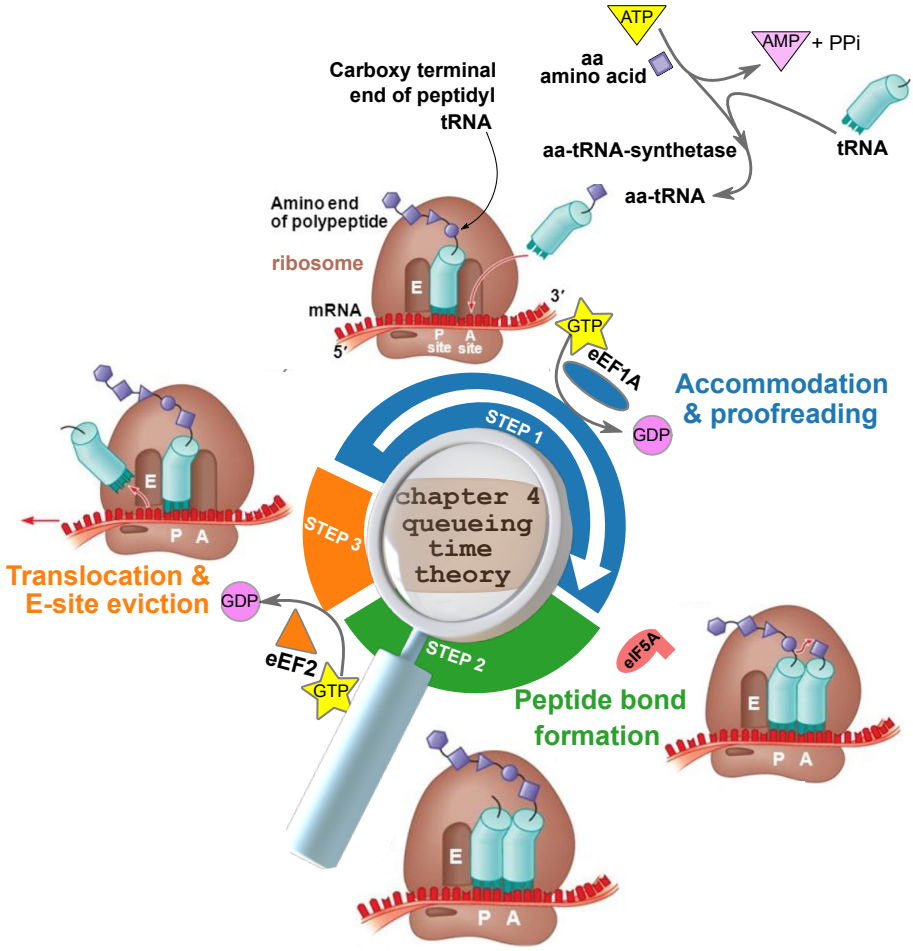


Figure 4.1: Representation of the ribosome elongation cycle in protein synthesis. The queueing time statistical theory is a central contribution to this PhD thesis. The elongation cycle consists of three main steps. Step 1 (dark blue) involves the accommodation and proofreading of aa-tRNA at the A-site; step 2 (green) is peptide bond formation between the carboxyl-terminal end of the peptidyl-tRNA at the P-site and the aminoacyl-tRNA at the A-site; step 3 is the translocation and eviction of the deacylated-tRNA from the E-site. Each step's reaction rate approximately follows first-order kinetics. The total elongation time is the sum of three randomly sampled queueing times, one for each step. It is a known mathematical property that the sum of these three random times follows a probability density function represented by the convolution of their individual distributions. We model the inherent stochasticity of the time spent by a ribosome on a codon using the resulting hypo-exponential distribution, an original contribution to the field.

This chapter was previously published as:
M. Joiret et al. (2023a). "A simple geometrical model of the electrostatic environment around the catalytic center of the ribosome and its significance for the elongation cycle kinetics". In: *Computational and Structural Biotechnology Journal* 21, pp. 3768–3795. DOI: 10.1016/j.csbj.2023.07.016. This paper is referred to as CSBJ01.

The core overview section highlights the key results of the published article that are directly aligned with the central framework of the thesis. This introduction provides sufficient context for understanding the `Ribosomer` model and maintaining continuity with the thesis's core narrative.

4.1 Core overview and connection to the thesis backbone

A solid mathematical description of stochasticity and a biological or biochemical understanding of stochasticity are two different things. In this chapter, our ambition is to describe both and provide a link between these two facets. What is the origin of the inherent stochasticity in the time spent by a ribosome on a codon during protein elongation? We explore in this chapter how certain mechanistic and deterministic processes contribute to stochasticity.

The central idea examines the **incorporation of mechanical effects into Eyring's transition state theory and the influence of energy barrier modulation on reaction rate constants**. It also outlines the partitioning of the elongation cycle into three sequential steps and applies queueing theory to model ribosome kinetics at single-codon resolution, highlighting key concepts and principal findings.

The crucial contribution is the **implementation of the queueing time theory within the agent-based model**. It explains how the ribosome's residence time on each of the 61 sense codons follows a probability density function calibrated through metadata analysis of published studies. Additionally, it describes how mechanochemical principles inform the modeling of elongation kinetics, concepts that are further explored in the subsequent chapters.

Modeling practitioners and experts are often concerned about uncertainties in the parameter values of the models they employ. A **crucial point** to understand in the context of the agent-based model described in this thesis is that **the vast majority of its parameters are not fixed values but are instead sampled from statistical distributions**. As a result, their values are **inherently uncertain**. **These uncertainties are quantified by the variance or standard deviation of the statistical distributions**

from which the model parameters are sampled. The probability density functions, along with their defining parameters (i.e., mean and variance), were **calibrated through fitting to published experimental data** (shared in public repositories). The details of this calibration process are presented in this chapter, which was published as a research paper (CSBJ01).

4.1.1 Key contributions

This thesis produced two significant advances, built upon two conceptual approaches:

(i) A theory of catalysis that describes how mechanical work modulates the Gibbs activation energy barrier. This theory establishes a connection between locally varying, context-dependent (electro-)mechanical forces acting on the transition state and the rate constant of the catalyzed (bio)chemical reaction.

(ii) A queueing time statistical theory that established a connection between aggregated ribosome footprint density counts per codon type and the rate constants of the elongation cycle's sub-steps.

(i) Resorting to mechanobiochemistry in the Eyring's equation of transition state theory of catalysis. The theory of kinetics of catalysis relies on the transition state theory that was introduced by Eyring who linked the rate constant of a chemical reaction to the Gibbs free activation energy of the transition state, TS [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983]. In this representation, a catalyzer accelerates a (bio)chemical reaction through a significant reduction in the transition Gibbs free energy barrier that the reactants have to overcome as shown in Figure 4.2.

In our work, we hypothesized that physical forces play a role in the reduction of the Gibbs free energy barrier of the transition state. The mechanical work of these physical forces affects the activation Gibbs free energy of the transition state. The modulation of the Gibbs free energy activation barrier changes the reaction rate constant through Eyring's relation [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983].

We took the catalyzed reaction scheme for the deacylation and peptide bond formation (transpeptidation) between the two previously accommodated substrates by the ribosome. The overall reaction scheme can be viewed as a nucleophilic substitution (S_N2-like scheme) and there is a single transition state as shown in Figure 4.2. The overall reaction can be written:

$$\underbrace{\text{peptidyl tRNA}}_{P-site} + \underbrace{\text{aminoacyl tRNA}}_{A-site} \rightarrow \underbrace{\text{empty tRNA}}_{P-site} + \underbrace{\text{peptidyl(+1) tRNA}}_{A-site}. \tag{4.1}$$

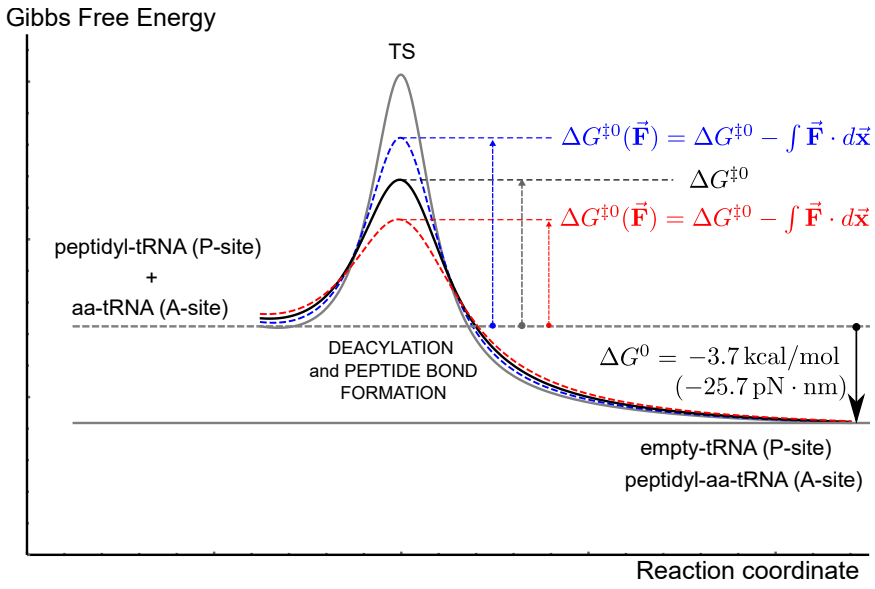


Figure 4.2: Gibbs free energy activation barrier to overcome at the transition state (TS) for the peptide bond formation at the peptidyltransferase center of the ribosome as a function of the reaction coordinate. The activation energy for the catalyzed reaction without mechanical forces (thick black line) is $\Delta G^{\ddagger 0}$ as compared to the uncatalyzed transition state (thin black line). The activation energy for the catalyzed reaction can be higher (dashed blue line) or lower (dashed red line) when the mechanical work exerted by any applied force on the peptidyl-tRNA substrate at the P-site during the reaction is negative or positive respectively.

This reaction is exergonic (thermodynamically favourable) as the change in standard Gibbs free energy is $\Delta G^0 \sim -3.7\,\mathrm{kcal/mol} (= -25.7\,\mathrm{pN \cdot nm})$ [Kaiser and Tinoco 2014]. The rate constant depends on the Gibbs activation energy of the transition state $\Delta G^{\ddagger 0}$. Applying external forces on molecules involved in catalyzed or uncatalyzed chemical reactions affects the kinetics of the reactions. The mechanical work of these applied mechanical forces can quantitatively be incorporated in the calculation of the activation Gibbs free energy of the transition state as already conceptually introduced by Bell [Bell 1978], and others [Bustamante et al. 2004; Ribas-Arino and Marx 2012]:

$$\Delta G^{\ddagger 0}(\vec{\mathbf{F}}) = \Delta G^{\ddagger 0}(\mathbf{0}) - \int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}} \tag{4.2}$$

where $\Delta G^{\ddagger 0}(\vec{\mathbf{F}})$ is the activation energy for the transition state in the presence of an

applied force acting on the system, $\Delta G^{\ddagger 0}(\mathbf{0}) \sim +14\,\text{kcal/mol} = +97.2\ \text{pN}\cdot\text{nm}$ is the activation energy for the transition state without any applied force [Bustamante et al. 2004], and $W = \int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}}$ is the mechanical work exerted by the force upon a test body along its curvilinear path. The mechanical work W is algebraically positive if the force and the displacement are parallel or negative if they are antiparallel. In the former case, $\Delta G^{\ddagger 0}(\vec{\mathbf{F}})$ is smaller than $\Delta G^{\ddagger 0}(\mathbf{0})$, whereas it is larger in the latter case as can be seen in Figure 4.2 (dashed red and dashed blue line respectively). In turn, the modulation of the Gibbs free energy activation barrier changes the reaction rate constant through Eyring's relation [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983]:

$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot\ e^{-\Delta G^{\ddagger 0}(\vec{\mathbf{F}})\,/\,N\,k_B\,T} \tag{4.3}$$

$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot\ e^{-\left(\frac{\Delta G^{\ddagger 0}(\mathbf{0})}{N\,k_B\,T} - \frac{\int \vec{\mathbf{F}}\cdot\mathbf{d}\vec{\mathbf{x}}}{k_B\,T}\right)} = k(\mathbf{0}) \cdot e^{\frac{\int \vec{\mathbf{F}}\cdot\mathbf{d}\vec{\mathbf{x}}}{k_B\,T}} \tag{4.4}$$

where $k(\vec{\mathbf{F}})$ is the reaction rate constant of the rate limiting step in the presence of an applied force upon the substrate at the P-site, $k(\mathbf{0})$ is the reaction rate constant in the absence of applied force. N, k_B, h and κ are Avogadro's number, Boltzmann's constant, Planck's constant and transmission coefficient respectively [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983].

The key result that appears in equation (4.4) is that the rate constant of a (bio)chemical reaction can be modulated by a mechanical work through the **Maxwell-Boltzmann factor** in the rightmost part of the equation. A factor like $e^{\vec{F}\cdot\vec{x}/k_BT}$ is also called an **Arrhenius factor** by other authors, notably Carlos Bustamante, Harry Noller and Ignacio Tinoco [Liu et al. 2014b]. The Maxwell-Boltzmann (or Arrhenius) factor can be larger than 1 and will increase the reaction rate, or smaller than 1 and will decrease the reaction rate depending on the algebraic sign of the mechanical work.

(ii) Resorting to queueing time statistical theory. Our approach assumes from the outset that the time a ribosome spends on a codon is inherently stochastic. In probabilistic terms, the ribosome dwell time on a codon is treated as a random variable. The key objective is to determine its statistical distribution—specifically, the probability density function (PDF) governing the time spent on each codon. Since all 61 sense codons are distinct, a unique probability density function is required for each. The full derivation of the queueing time statistical theory is given in appendix A of the thesis.

The elongation cycle consists of multiple substeps. As described in chapter 1, a complete cycle follows a sequence of three main substeps. Each substep has a rate limiting reaction, which can be approximated by a first-order kinetic process. More

explicitly, in its simplest form[1], a first-order kinetic equation is written:

$$\frac{d[X]}{dt} = -k \cdot [X] \tag{4.5}$$

where $[X]$ represents the concentration of a reactant consumed during the reaction. The solution of this first order ordinary differential equation is:

$$[X(t)] = [X(t_0)] \cdot e^{-k\cdot(t-t_0)} \tag{4.6}$$

The rate constant k contains all information about the rate of the reaction. It has units of s^{-1} and its inverse is proportional to the lifetime of the reactant. The time $\tau_{1/2}$ required for half of the reactant to be transformed by the reaction, known as the half-life time is given by $\tau_{1/2} = \frac{\ln 2}{k}$. The kinetic equation can be normalized such that the area under the curve is equal to 1. In this form, the equation is equivalent to a probability density function, allowing the reaction process to be reinterpreted within the framework of probability theory. If the independent variable t represents time, the reaction half-time corresponds to the waiting time at which there is a 50% probability that the reaction has occurred. This is analogous to the median queueing time. Recall that if a random variable follows an exponential probability density function f_{EXP}:

$$f_{\text{EXP}}(t) = \lambda e^{-\lambda\cdot t} \tag{4.7}$$

the mean of the random variable is $\frac{1}{\lambda}$ and the median is $\frac{\ln 2}{\lambda}$.

The next step in this line of reasoning is to consider that each substep in the elongation cycle can indeed be approximated by a first-order kinetic reaction. Equivalently, each substep can be viewed as a process in which the queueing time follows an exponential distribution. As a result, the total elongation time can be described as the sum of three exponentially distributed queueing times, each with a different rate parameter (k_i or λ_i). A fundamental result in the mathematical theory of probability states that a probability density function f_X of a compound random variable X, defined as the sum of a number, e.g., 3, of independent random variables –here, $X = X_1 + X_2 + X_3$– is given by the convolution of their individual probability density function, f_1, f_2, f_3:

$$f_X = f_1 * f_2 * f_3 \tag{4.8}$$

[1]More generally, a chemical reactions such as

$$X + Y \underset{\nu_1}{\overset{k}{\longrightarrow}} Z$$

obeys the Guldberg and Waage's law of mass action and the forward reaction rate ν_1 would write

$$\nu_1 = \frac{d[X]}{dt}$$

$$\frac{d[X]}{dt} = -k[X][Y]$$

Note that when $[Y]$ is constant, a simple rescaling of k yields a first-order kinetic equation for the reactant X.

where the $*$ sign means the convolution product operation between properly behaved functions or distributions.

When applied to the sum of three exponentially distributed queueing times, one for each of the three substeps of a complete elongation cycle, the resulting total queueing time is hypo-exponentially distributed. The details are in the appendix A (`A.2` of CSBJ01). The dwell time of a ribosome, or the ribosome residence time (RRT), on any single codon during elongation follows a hypo-exponential distribution.

The hypo-exponential distribution has three parameters, i.e., the three rate constants of the three substeps: accommodation, peptide bond formation and translocation. Figure 4.3 shows the profile of a hypo-exponential distribution (red dots).

A reasonable assumption in ribosome profiling analysis is that the number of ribosome-protected fragment (RPF) reads observed on a codon is proportional to the ribosome residence time on that codon. This relationship holds for data obtained from Ribo-Seq experiments.

Under this assumption, the statistical distribution of normalized footprint counts of ribosomes on codons –aggregated per codon type in ribosome profiling meta-analyses– is expected to be strictly equivalent, up to a scaling factor, to the statistical distribution of ribosome residence time.

This assumption aligns with multiple Ribo-Seq meta-analytical findings, as supported both in this chapter and in previously published literature.

4.1.2 Key outcomes included in the ribosomer model framework

A key outcome of this approach was the fitting of 61 hypo-exponential distributions to the normalized footprint counts observed in different model species of interest, namely *Saccharomyces cerevisiae* and *Escherichia coli*. Furthermore, an informed deconvolution of the fitted distributions from the Ribo-Seq normalized footprint counts per codon enabled the inference of the three rate constants corresponding to the substeps of the elongation cycle at codon resolution.

This procedure provided an estimated calibration of the three rate constants, i.e., accommodation, peptide bond formation, and translocation, for each of the 61 sense codons in the species of interest. The calibration was based on the published meta-analysis by Dana and Tuller [Dana 2014; Dana and Tuller 2014].

An important consideration regarding these results is that relative time ratios can be derived by comparing two distributions of ribosome residence time on a codon. However, drawing conclusions about absolute time differences from such distributions

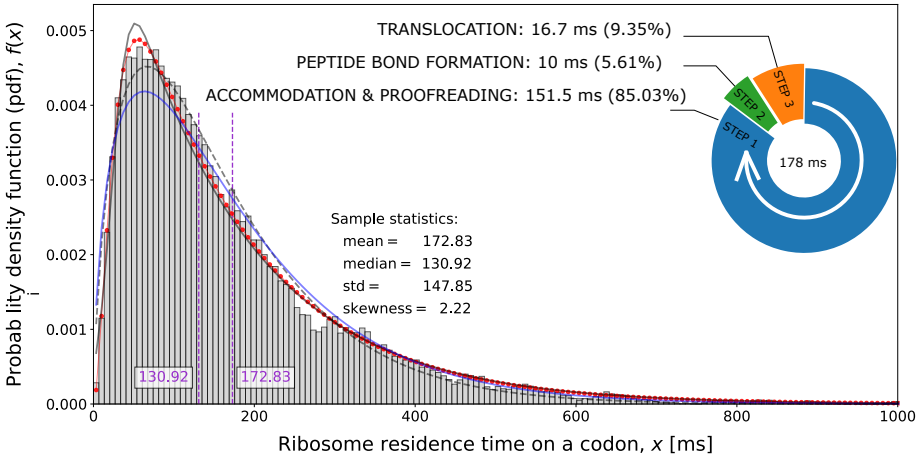


Figure 4.3: Ribosome elongation cycle simulation in three sequential time steps. Hypo-exponential distribution with $1/\lambda_i = 151.5, 10.0, 16.7$ [units in ms (milliseconds)]: red dots. Grey bars: frequency distribution sampled from the hypo-exponential distribution. Exponentially modified Gaussian distribution least sq. fit to hypo-exponential: grey line. Gamma distribution fit to frequency distribution, least sq.: dashed line; max. likelihood estimation: blue line.

may be misleading. Comparisons of RRTs between codons should be expressed in terms of fold changes rather than absolute time differences.

The paper also demonstrates that the following families of probability density functions (PDF) convey highly similar information:

(i) Hypo-exponential distribution (with three rate parameters)

(ii) Exponentially modified Gaussian distribution (with three parameters: one rate parameter for exponential component, plus mean and variance for the Gaussian component)

(iii) Gamma distribution (with two parameters) or shifted Gamma distribution (with three parameters)

(iv) Log-Normal distribution (with two parameters)

The probability density functions, commonly used in the literature, exhibit strong similarities in that differences between the chosen probabilistic model and empirical

data are negligible. This is quantitatively demonstrated using the Kullback-Leibler divergence criterion from information theory.

Among these distributions, the hypo-exponential distribution offers a key advantage: it directly corresponds to the three substeps of the elongation cycle, making it particularly suitable for modeling ribosome residence time.

The key result expressed in equation (4.4) is general and can be applied to estimate changes in the rate constant of any process. In our context, it is used to analyze a biochemical reaction catalyzed by the ribosome, where a transition state is influenced by mechanical force.

The elongation cycle is modeled as a sequence of three substeps (indexed by $i \in 1, 2, 3$), each characterized by its own rate constant k_i. Equivalently, as previously discussed, each substep can be represented by a mean queueing time τ_i, where $\tau_i = 1/k_i$.

The inset of Figure 4.3 illustrates an example of three specific values for the mean queueing time of each individual substep. The main figure presents the statistical distribution of the total queueing time, corresponding to the total time required by the ribosome to complete a full elongation cycle on a single codon.

The mechanochemical modulation of the Gibbs free activation energy barrier for the transition state, as introduced in Equation (4.2), is directly linked to the rate constant of chemical reactions, as derived in Equation (4.4). Each individual rate constant within the elongation cycle can be influenced by the local biochemical context.

A local, context-dependent modulation of the activation energy barrier for a given biochemical process –whether accommodation, peptide bond formation, or translocation– will consequently affect the corresponding rate constant. The quantitative relationship between changes in activation Gibbs free energy and the rate constant can always be modeled using Equation (4.4).

In CSBJ01 [Joiret et al. 2023a], presented in this chapter, we applied Equation (4.4) to the peptide bond formation. We demonstrated how the nature of the charged amino acid at the carboxy-terminal end of the peptidyl-tRNA in the P-site influences peptide bond formation rates through electrostatic forces acting within the ribosome's catalytic center. In Chapter 6, we will extend this approach to account for the effects of electrostatic interactions between the nascent chain and the ribosome exit tunnel on peptide bond formation rates. In Chapter 8, the same framework will again be applied to examine the influence of mRNA secondary structures on the translocation rate.

Our modeling approach integrates causal (mechanistic) and local effects into the mean rate constant of the processes being studied while preserving the inherently stochastic nature of the probability density function governing ribosome dwell time on a codon. Conceptually, the ribosome can be viewed as an agent that continuously samples from the available tRNA pool for cognate-tRNA accommodation or waits to overcome energy

barriers associated with peptide bond formation and translocation as it progresses through each elongation cycle.

4.2 A simple geometrical model of the electrostatic environment around the catalytic center of the ribosome and its significance for the elongation cycle kinetics (CSBJ01)

Subsections `4.5.4`, `4.5.5`, and `4.6.7` below are the key elements required for the context of the `Ribosomer` model (respectively sections `3.4`, `3.5`, and `4.7` of the published paper). The paper's reviewers mentioned Appendix A to be particularly insightful for understanding the stochastic nature of ribosome dwell times on individual codons.

CSBJ01 abstract

The central function of the large subunit of the ribosome is to catalyze peptide bond formation. This biochemical reaction is conducted at the peptidyl transferase center (PTC). Experimental evidence shows that the catalytic activity is affected by the electrostatic environment around the peptidyl transferase center. Here, we set up a minimal geometrical model fitting the available x-ray solved structures of the ribonucleic cavity around the catalytic center of the large subunit of the ribosome. The purpose of this phenomenological model is to estimate quantitatively the electrostatic potential and electric field that are experienced during the peptidyl transfer reaction. At least two reasons motivate the need for developing this quantification. First, we inquire whether the electric field in this particular catalytic environment, made only of nucleic acids, is of the same order of magnitude as the one prevailing in catalytic centers of the proteic enzymes counterparts. Second, the protein synthesis rate is dependent on the nature of the amino acid sequentially incorporated in the nascent chain. The activation energy of the catalytic reaction and its detailed kinetics are shown to be dependent on the mechanical work exerted on the amino acids by the electric field, especially when one of the four charged amino acid residues (R, K, E, D) has previously been incorporated at the carboxy-terminal end of the peptidyl-tRNA. Physical values of the electric field provide quantitative knowledge of mechanical work, activation energy and rate of the peptide bond formation catalyzed by the ribosome. We show that our theoretical calculations are consistent with two independent sets of previously published experimental results. Experimental results for *E.coli* in the minimal case of the dipeptide

bond formation when puromycin is used as the final amino acid acceptor strongly support our theoretically derived reaction time courses. Experimental `Ribo-Seq` results on *E. coli* and *S. cerevisiae* comparing the residence time distribution of ribosomes upon specific codons are also well accounted for by our theoretical calculations. The statistical queueing time theory was used to model the ribosome residence time per codon during nascent protein elongation and applied for the interpretation of the `Ribo-Seq` data. The hypo-exponential distribution fits the residence time observed distribution of the ribosome on a codon. An educated deconvolution of this distribution is used to estimate the rates of each elongation step in a codon specific manner. Our interpretation of all these results sheds light on the functional role of the electrostatic profile around the PTC and its impact on the ribosome elongation cycle.

4.3 Introduction

Ribosomes are the cells' manufacturing tools for building up proteins. They decode the 61 sense codons from a primary message encrypted in a messenger RNA (mRNA) single molecule. They translate it with the help of a set of transfer RNAs (tRNAs) into 20 amino acids to be sequentially polymerized in a nascent polypeptide that will eventually pass through the ribosomal exit tunnel and fold into its final structure.

X-ray solved structural representations of ribosomes have been publicly available for different species at atomic resolution for more than 20 years. The peptide bond is formed between the nascent protein chain and a newly incorporated amino acid at the ribosomal large subunit catalytic center, whose salient feature is that the peptidyl transferase center (PTC) is not an enzyme but a ribozyme, composed of ribosomal RNA (rRNA) [Noller et al. 1992]. There are no ribosomal protein components within a 15 radius of the catalytic center [Rodnina et al. 2006; Simonovic and Steitz 2009]. The electrostatic environment of the PTC is largely determined by the presence of the negatively charged phosphate moieties belonging to the 23S rRNA backbone in archae or bacteria, and to the 28S rRNA backbone in eukarya. The PTC cavity itself is made of rRNA and the substrates that are processed in the cavity also are RNA molecules: the single amino-acylated tRNA at the A site and the peptidyl-tRNA at the P site. It is thus necessary to quantitatively assess the relative contributions of the 2 tRNAs at the P and A site with respect to the contribution of the 23S rRNA (28S rRNA) to the electrostatic potential profile around the PTC cavity and towards the nascent chain tunnel entry port. Furthermore, in numerous previously published studies, the monovalent K^+ and bivalent Mg^{2+} metal ions have been shown to be key ingredients in the structure of the ribosome and of ribonucleic acids rRNAs and tRNAs [Nierhaus 2014; Rozov et al. 2019; Wang et al. 2020].

Efforts in understanding the kinetics of the protein synthesis in vivo and the factors

that affect elongation rate have been conducted for decades by the research community [Rodnina 2016; Simpson et al. 2020]. Experimental evidence supports the fact that enzymatic kinetics is affected by the electrostatic environment around the active sites of enzymes. Recently, vibrational Stark effect spectroscopy techniques were used to obtain direct measurements of electric fields at active sites of enzymes [Fried and Boxer 2017]. To elucidate the mechanisms of action and the reaction kinetics, there is great interest in knowing quantitatively the potential and the electric field prevailing at the active site of the ribosome, at the peptidyl transferase center and along the ribosome exit tunnel. To this date, there is no experimental protocol that would make the large subunit of the ribosome amenable to such direct measurements, although pioneering efforts were undertaken in the ribosome exit tunnel of rabbit reticulocytes [Lu et al. 2007]. Given the lack of direct experimental measurements for the potential and the electric field, the only way to obtain the physical quantities of interest is through the use of a mathematical model. Such mathematical models are broadly used to shed light on structure, function and properties of often complex biomolecular systems and are used in numerous studies [Brooks and al 2009].

In this paper, we start from five publicly available atomic structures of the ribosome across the three domains of life and use the fundamental laws of electrostatics along with the dielectric properties of media inside the ribosome to calculate the potential and the electric field experienced around the center of the peptidyl transferase reaction. We investigate the contribution of the two tRNAs at P and A-sites to the electrostatic potential around the rRNA PTC shell cavity and. The local distribution of the metal ions charges around the PTC and around the A and P site tRNAs was also studied within a distance of the order of the electrostatic screening length away from these RNA molecules. The electric field and the axial forces on test charged amino acids can be quantitatively estimated. The mechanical work exerted by the electrostatic field on the test charged amino acids (E, K, D, E) upon protein elongation displacement can be calculated. The significance of the electrostatic environment around the PTC shell cavity on the peptide bond formation rate is quantitatively studied. We will rely on the statistical queueing time theory to model the ribosome dwelling time per codon during nascent protein elongation and use this theory for the interpretation of `Ribo-Seq` data. The rate calculations will be compared with previously published experimental results from two independent sources.

4.4 Material and Methods

4.4.1 X-ray crystallographic space position of phosphate moieties and charged amino acid residues around the PTC cavity

We analyzed five publicly available structures of the large ribosomal subunit across the three domains of life: the archeon *Haloarcula marismortui* (PDB code: 4V9F downloaded from https://www.rcsb.org/) obtained from x-ray crystallography at 2.4 Å [Gabdulkhakov et al. 2013]; two bacteria *Escherichia coli* (PDB code: 7K00) at 2.0 Å [Watson et al. 2020], *Thermus thermophilus* (PDB code: 4Y4P) at 2.5 Å [Polikanov et al. 2015]; two eukaryotes, the rabbit *Oryctolagus cuniculus* (PDB code 7O7Y) at 2.2 Å [Bhatt et al. 2021] and the human *Homo sapiens* (PDB code: 6EK0) at 2.9 Å. To localize the PTC shell cavity for the five species, we relied on two complementary methods: (i) when available in the PDB deposited x-ray solved structures, we used the tips of the acceptor arms of the tRNAs at the P-site and/or the A site, i.e. the O3' oxygen of the ribose of adenosine 76 at the 3' end of the tRNAs; (ii) alternatively, we used the 23S or 28S rRNA sequence alignment of the nucleotides that are the most universally conserved across the three domains of life as detailed in [Doris et al. 2015; Tirumalai et al. 2021]. Table 4.4.1 and Fig. 4.4 show, for each of the five studied species, the nucleotides in the 23S/28S rRNA P-loop (between secondary structure helices h90 and h93) and A-loop (helix h89) that are closest to the PTC shell cavity where the peptide bond is catalyzed between the tRNAs at the P and A sites. Once the virtual straight line between the N3 atoms of the P-loop and A-loop nucleotides most likely involved in the catalytic activity of the PTC has been determined, all nucleotides and phosphate moieties of the 23S/28S rRNA molecules that are within a 40 Å distance are selected. These selections are used to show the 3D configurations of the PTC in PyMOL and are used to study the electrostatic environment of the PTC shell cavity. To find the ribosome tunnel entry port and extract the atom coordinates around the PTC, we used a tunnel search algorithm developed by Sehnal *et al* [Sehnal et al. 2013], implemented in MOLE 2.0 and the web-based MOLEonline 2.0 tool publicly available online [Berka et al. 2012; Pravda et al. 2018]. We used PyMOL (PyMOL Molecular Graphics System, Version 2.3.2) and exported the relevant selected atom positions' cartesian coordinates to output files. These files were further processed with algorithms coded in Python to select the charged chemical groups on or near the inner surface of the PTC cavity. To localize the PTC, we used the geometric center of the five known 23S rRNA P loop and A loop nucleotides that are known to interact with the 3'-terminal CCA end of the tRNA acylated to the carboxy terminal amino acid of the nascent chain at the P site or the 3'-terminal CCA end of the aminoacylated tRNA at the A site [Beringer and Rodnina 2007; Lang et al. 2008; Polikanov et al. 2014; Rodnina et al. 2006]. These five nucleotides in the P and A loops of 23S rRNA are A2485(2450), A2486(2451)

for the P loop and U2620(U2585), U2621(U2586) and U2622(U2585) for the A loop [Simonovic and Steitz 2009]. The numbering of the 23S `rRNA` residues is based on the *H. marismortui* sequence, while the corresponding position in the *E. coli* ribosome is shown in the brackets. For *Haloarcula marismortui*, we translated the crystallographic

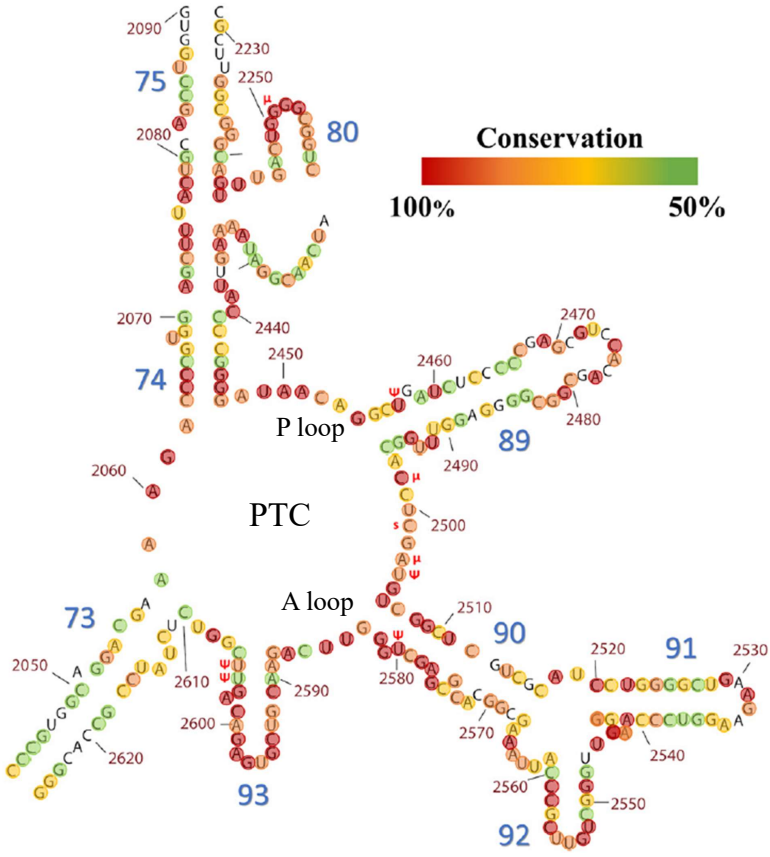


Figure 4.4: Fragment of the secondary structure of 23SrRNA belonging to the large subunit of the ribosome of *Thermus thermophilus* which contains the nucleotides within and around the peptidyl transferase center (PTC). Helix numbers are in large blue font. Nucleotides highlighted in dark red are the most universally conserved across the three domains of life. Nucleotides numbering according to *E.coli*. Figure credits: Tirumalai et al. 2021.

data model space so that the tunnel entry point would be at the origin and we aligned

the direction from the tunnel entry point to the PTC geometric mid-point along the positive z-axis. We selected all charged amino acid residues (NH2 or NZ for arginine or lysine, OE2 or OD2 for aspartate or glutamate) belonging to ribosomal proteins and all charged non-bridging oxygen atoms bound to the phosphate moieties in the 23S rRNA backbone that are closer than 40 Å from the centerline joining the tunnel entry port to the PTC geometrical midpoint. The cavity around the PTC was approximated by fitting a truncated prolate spheroid (ellipsoid of revolution about the major axis) having its semi-major axis aligned with the direction from the tunnel entry port to the PTC. The half prolate spheroid was also scaled in such a way that its semi-major axis spans the distance between the tunnel entry point and the decoding center P site (~ 8.75 nm) and its semi-minor axis spans a half-length of 3 codons ($\sim 4.50/2 = 2.25$ nm). We algorithmically set out the 3D equations of the truncated prolate spheroid in this reference frame to calculate the radial distance of the selected atoms to the surface of the prolate spheroid PTC cavity. We confirmed that there are no charged atoms belonging to the 23S rRNA or to any ribosomal proteins inside the fitted truncated prolate spheroid cavity. Once the PTC cavity and a ribosome exit tunnel is localized, the magnesium, potassium and sodium atoms (and water molecules) around them are counted. The counts are compared to the local charged phosphate moieties count and the positions of all charged groups (phosphate moieties, magnesium, potassium and sodium ions) are extracted to determine the set of the source charges contributing to the electrostatic potential in any given space point around the PTC cavity.

Table 4.4.1: Universally most conserved 23S/28S rRNA sequences around the PTC in the large subunit of the ribosome across the 3 domains of life, adapted from [Doris et al. 2015].

Species (domain)	PDB code	P-loop	A-loop	number of nucleotides within 40 Å around PTC
Haloarcula marismortui (archea)	4V9F	**A2486** 5'-GGAUAAC-3'	**U2620** 5'-GAGCUGGGUUUA-3'	487
Escherichia coli (bacteria)	7K00	**A2451** 5'-GGAUAAC-3'	**U2585** 5'-GAGCUGGGUUUA-3'	484
Thermus thermophilus (bacteria)	4Y4P	**A2463** 5'-GGAUAAC-3'	**U2597** 5'-GAGCUGGGUUUA-3'	484
Oryctolagus cuniculus (eukarya)	7O7Y	**A4143** 5'-GGAUAAC-3'	**U4277** 5'-GAGCUGGGUUUA-3'	488
Homo sapiens (eukarya)	6EK0	**A4397** 5'-GGAUAAC-3'	**U4530** 5'GAGCUGGGUUUA-3'	488

4.4.2 P-site and A-site tRNAs phosphate moieties and metal ions positions

The x-ray or cryo-electron-microscopy solved structures of the ribosomes of *Thermus thermophilus*(PDB code: `4Y4P`), *Escherichia coli* (PDB code: `7K00`) both include two tRNAs at the P and A site [Polikanov et al. 2015]. *Oryctolagus cuniculus* (PDB code: `707Y`) includes one tRNA at the P site and *Homo sapiens* (PDB code: `6EK0`) includes one tRNA at the E site. The O3' oxygen atom of the ribose of the last nucleotide A76 (adenosine 76) in the tRNA at the P-site or the A-site is used to localize the PTC cavity. The space positions of the charged phosphate moieties and of the ions, magnesium, potassium and sodium have been extracted in order to investigate the contribution of both tRNAs and their metal ions to the electrostatic potential around the PTC cavity and along the line towards the tunnel entry port for the ribosome of *Thermus thermophilus*.

4.5 Theory and calculations

4.5.1 Idealized shape model of the ribosomal RNA cavity around the PTC and the Yukawa-Debye-Hückel potential with dielectric screening

After the sphere, the prolate spheroid is the geometrical shape with the smallest surface encompassing the largest volume. An electrostatic model of the cavity around the catalytic center of the large subunit of the ribosome is built using this simplest shape. This fulfills the minimal geometrical constraints which prevail between the ribosome peptide exit tunnel, the mRNA channel and the size of three aminoacylated-tRNAs that are accommodated in the cavity during nascent protein elongation Fig.4.6(a). The prolate spheroid was fitted onto the *Haloarcula marismortui* aforementioned publicly available x-ray solved structure of the large ribosomal cavity around the peptidyl transferase center as explained in Material and Methods.

The electrical scalar potential $\Phi(\vec{\mathbf{r}})$ at the observed position $\vec{\mathbf{r}}$ is expressed, in a homogeneous medium of constant permittivity and in the absence of dielectric screening effects, by the Coulomb law:

$$\Phi(\vec{\mathbf{r}}) = \frac{1}{4\pi\epsilon} \int \int_S \frac{\sigma(\vec{\mathbf{r}'})\, da}{|\vec{\mathbf{r}} - \vec{\mathbf{r}'}|} \tag{4.9}$$

where $\sigma(\vec{\mathbf{r}'})$ is the surface-charge density (measured in coulombs per square meter) at position $\vec{\mathbf{r}'}$ of the source, da is the two dimensional surface element at $\vec{\mathbf{r}'}$ and ϵ is the permittivity of the dielectric medium (Eq. 1.23 in Jackson [Jackson (1998)]) with

$\epsilon = \epsilon_r \epsilon_0$, where ϵ_r is the relative permittivity of the medium and ϵ_0 is the permittivity of free space.

It is worth emphasizing here that the position where the physical values of the potential are calculated is $\vec{\mathbf{r}}$, but the variable over which the surface integration is conducted is $\vec{\mathbf{r}'}$ with its elementary surface elements ($d^{(2)}\vec{\mathbf{r}'} = da$) integrated over the surface domain of interest in the 3D space. We actually sum over all the fixed charges located on the surface. There are two ways to describe the charges and their positions. In the first, the charges are considered to be continuously distributed on the surface. A surface charge density σ must be known which can be a function of $\vec{\mathbf{r}'}$ as expressed in equation (4.9). In the second, the real fixed charges are discrete (with no spatial extensions) but the surface domain on which integration is conducted is still a compact interval in $\mathbb{R}^2$. Implicitly, discrete charges q_i, spatially localized in $\vec{\mathbf{r}'}_i$ are mathematically represented by a generalized function which is the product of the charges q_i by a Dirac distribution [Barton (1989)]: $\sigma(\vec{\mathbf{r}'}) = \sum_{i \in \text{all charges}} q_i \delta^{(2D)}(\vec{\mathbf{r}'} - \vec{\mathbf{r}'}_i)$. This second type of description will be used in the next section. In mathematical terms, in both descriptions, the integration domain is the support of the function to be integrated. The support of the function is the interval upon which the function exists and is not null. $\vec{\mathbf{r}'}$ exists in the support of the charges or the support of the Dirac functions one implicitly uses to define the position of the charges that are sources of the field. $\vec{\mathbf{r}}$ is defined in another space, the set of points where one wants to calculate the field, irrespective of the positions of the source charges. In the surface integral calculation conducted below, $\vec{\mathbf{r}'}$ has coordinates u, v in the support of the charges, which is a surface in 3D (a half prolate spheroid) and $\vec{\mathbf{r}}$ has the cartesian coordinates (0, 0, z) of a straight line in 3D, because for the sake of simplicity, the potential and the electric field will only be calculated along the central axis z.

In the presence of polarizable dielectric material, screening effects occur due to induced dipoles or permanent dipoles reorientations in the medium separating the fixed charged sources of the field and the observation point. In this latter case, the electrical scalar potential $\Phi(\vec{\mathbf{r}})$ at the observed position $\vec{\mathbf{r}}$ is expressed by the Yukawa-Debye-Hückel law:

$$\Phi_{\text{Yuk}}(\vec{\mathbf{r}}) = \int \int_S \frac{\sigma^*(\vec{\mathbf{r}'})}{4\pi\,\epsilon} \frac{e^{-\frac{|\vec{\mathbf{r}}-\vec{\mathbf{r}'}|}{\xi}}}{|\vec{\mathbf{r}} - \vec{\mathbf{r}'}|}\, da \tag{4.10}$$

where $\sigma^*(\vec{\mathbf{r}'})$ is the actual *formal bare* surface charge density. There is a marked exponential damping of the Coulomb interaction where ξ is a characteristic distance of the screening. We assume the screening effect to be homogeneous around the fixed charged surface, and thus we take the assumption that both ϵ and ξ do not depend on $\vec{\mathbf{r}'}$ or $\vec{\mathbf{r}}$ or at least are piecewise constants in a given space domain [Joiret et al. 2022b].

For the prolate spheroid, we can take advantage of the axial symmetry and restrict the observation positions to the spatial points on the z axis, i.e., for $\vec{\mathbf{r}} = (0, 0, z)$. The surface integration is conducted on the support of the fixed source charges. The prolate

spheroid's inner wall is geometrically generated by the $\gamma(u)$ curve moving axially along the z-axis from $z = 0$ to $z = -a$ as drawn in Fig.4.6(a), where a and b are the semi-major and semi-minor axis of the prolate spheroid:

$$\gamma(u) = (b\sqrt{1-\frac{z^2}{a^2}}\cos u,\ b\sqrt{1-\frac{z^2}{a^2}}\sin u, z),\ u \in [0,\ 2\pi]. \tag{4.11}$$

The prolate spheroid's half surface can be written as $S = \phi(K)$ where $K = \{(u, v) \in [0, 2\pi] \times [0, \frac{\pi}{2}]\}$ and where $\phi : \mathbb{R}^2 \to \mathbb{R}^3$:

$$\phi(u, v) = (b\ \sin v\ \cos u,\ b\ \sin v \sin u,\ -a\ \cos v). \tag{4.12}$$

Another equivalent parametrization of the prolate spheroid's half surface can be written as $S = \phi(K)$ where $K = \{(u, v) \in [0, 2\pi] \times [0, 1]\}$ and where $\phi : \mathbb{R}^2 \to \mathbb{R}^3$:

$$\phi(u, v) = (b\sqrt{1-v^2}\cos u,\ b\sqrt{1-v^2}\sin u,\ -a\, v). \tag{4.13}$$

Upon these parametrizations, the distance between any observation point and any point supporting a source charge can be expressed respectively as:

$$|\vec{\mathbf{r}} - \vec{\mathbf{r}'}| = \sqrt{b^2\ \sin^2 v + (z + a\cos v)^2} \tag{4.14}$$

$$= \sqrt{b^2(1-v^2) + (z+av)^2}. \tag{4.15}$$

There is an aperture at the leftmost end of the prolate spheroid where the PTC cavity is connected to the ribosome exit tunnel which is approximated by a cylinder as shown in Fig. 4.6(a). The radius of the aperture is around 5 Å. Hence, the exact support of the resulting truncated spheroid' surface is parametrized as $S = \phi(K)$ where $K = \{(u, v) \in [0, 2\pi] \times [v_{inf}, v_{sup}]\}$ where $v_{sup} < 1$. $D_u\phi$ is the first partial derivative of the parametric equation of the surface $\phi(u, v)$ with respect to u and $D_v\phi$ is the first partial derivative of the parametric equation of the surface $\phi(u, v)$ with respect to v. In the general Eqs (4.9 and 4.10), the surface-charge densities $\sigma(\vec{\mathbf{r}'})$ or $\sigma^*(\vec{\mathbf{r}'})$ are dependent of the position $\vec{\mathbf{r}'}$ on the support of the source charges. Here, we will take the simple approximation that σ or σ^* can be considered a constant parameter over a surface of a given shape, e.g. over the spheroid's surface. This is the surface charge uniform distribution assumption for a given shape. Note however that a space dependence of σ is possible if it is compensated for by a similar space dependence of ϵ ensuring the combined ratio σ/ϵ is constant in a region of interest. This piecewise constant ratio is the strictly necessary assumption for the mathematical surface integration calculations of our models to be analytically tractable. In subsection 4.6.5, we will discuss how reasonable this assumption is by comparing the potential profiles calculated from the exact x-ray crystallographic structural data of the large subunit of the ribosome and

calculated from the idealized spheroidal shape with constant surface charge density and a constant medium permittivity.

The electrostatic scalar Yukawa screened potential (4.10) results from the surface integral calculation:

$$\Phi(z) = \frac{\sigma^*}{4\pi\epsilon} \int\int_{K=\{(u,v)\in[0,2\pi]\times[v_{inf},v_{sup}]\}} \frac{e^{-\frac{\sqrt{b^2(1-v^2)+(z+av)^2}}{\xi}}}{\sqrt{b^2(1-v^2)+(z+av)^2}} |D_u\phi \wedge D_v\phi| \, du \, dv \tag{4.16}$$

$$D_u\phi = (-b\sqrt{(1-v^2)}\,\sin u, b\sqrt{1-v^2}\,\cos u, 0) \tag{4.17}$$

$$D_v\phi = (-\frac{b\,v}{\sqrt{1-v^2}}\,\cos u, -\frac{b\,v}{\sqrt{1-v^2}}\,\sin u, -a) \tag{4.18}$$

$$|D_u\phi \wedge D_v\phi| = \left| det \begin{pmatrix} \vec{e_x} & \vec{e_y} & \vec{e_z} \\ -b\sqrt{1-v^2}\sin u & b\sqrt{1-v^2}\cos u & 0 \\ \frac{-b\,v\,\cos u}{\sqrt{1-v^2}} & \frac{-b\,v\,\sin u}{\sqrt{1-v^2}} & -a \end{pmatrix} \right| \tag{4.19}$$

$$= |(-ab\sqrt{1-v^2}\cos u, -ab\sqrt{1-v^2}\sin u, b^2\,v)| \tag{4.20}$$

$$= b\sqrt{a^2(1-v^2)\,+\,b^2\,v^2} \tag{4.21}$$

$$\Phi(z) = \frac{\sigma^*}{4\pi\epsilon} \int_0^{2\pi} du \int_{v_{inf}}^{v_{sup}} dv \frac{e^{-\frac{\sqrt{b^2(1-v^2)+(z+av)^2}}{\xi}} \cdot b\sqrt{a^2(1-v^2)\,+\,b^2v^2}}{\sqrt{b^2(1-v^2)+(z+av)^2}} \tag{4.22}$$

$$\Phi(z) = \frac{\sigma^*}{2\epsilon} \int_{v_{inf}}^{v_{sup}} dv \frac{e^{-\frac{\sqrt{b^2(1-v^2)+(z+av)^2}}{\xi}} \cdot \sqrt{a^2(1-v^2)\,+\,b^2v^2}}{\sqrt{(1-v^2)+(\frac{z+av}{b})^2}} \tag{4.23}$$

$$\Phi(z) = \frac{\sigma^*}{2\epsilon} \int_{v_{inf}}^{v_{sup}} dv \cdot \sqrt{\frac{a^2\,-\,(a^2-b^2)v^2}{(1-v^2)+(\frac{z+av}{b})^2}} \cdot e^{-\frac{\sqrt{b^2(1-v^2)+(z+av)^2}}{\xi}} \tag{4.24}$$

The integral converges if the denominator in the integrand of Eq.(4.24) does not vanish, i.e., if $v_{sup} < 1$, meaning that the spheroid (ellipsoid of revolution) is truncated. The ribosome cavity around the PTC further goes towards the ribosome exit tunnel which is

approximated by a cylinder surface. By virtue of continuity and physical consistency, it is necessary that the electrostatic potential be a continuous function of space inside the cavity of the catalytic center of the ribosome and inside the ribosome exit tunnel.

According to the x-ray solved structure of the ribosomal large subunit, the length scales for the spheroid are with semi-major axis $a = 8.97\,\text{nm}$, semi-minor axis $b = 2.25\,\text{nm}$, giving $v_{sup} = 0.975$ such that the aperture of the spheroid cavity towards the cylindrical ribosome exit tunnel has a radius $R = 5\,\text{Å}$. The length between the center of the spheroid, at $z = 0$, and the entry point of the ribosome exit tunnel is 8.75 nm. The P loop 23S rRNA nucleotide A2486 (2451) interacting at the P site in the peptide bond formation has its axial position 20.6 Å away from the tunnel entry port. This provides the lower integration limit for the support of the charges on the spheroid's surface $v_{inf} = 0.74536$.

The integral in Eq.(4.24) converges but cannot be solved analytically as it is an elliptical integral. We resorted to two classical numerical approaches using Newton-Cotes method or Gaussian quadrature where the nodes sampling followed the Gauss-Konrod rule [Johansson (2019)]. Both methods gave the same result for the solution of the electrostatic potential $\Phi(z)$ profile along the centerline of the truncated spheroid's cavity.

The axial electric field is obtained from the negative of the first partial derivative with respect to z:

$$\vec{E}_z = -\frac{\partial \Phi(z)}{\partial z}. \tag{4.25}$$

Finally, the axial force on a test charge can be calculated from:

$$\vec{F}_z = q_e \vec{E}_z \tag{4.26}$$

where q_e is the charge of the test charge.

In Eq.(4.24), we require the knowledge of at least two phenomenological parameters: the *formal bare* surface charge density σ^* on the truncated spheroid surface in the vicinity of the PTC and the dielectric response (permittivity) ϵ of the medium around the PTC. In fact, only the ratio σ/ϵ is needed. The phenomenological screening length parameter ξ is also required.

4.5.2 Structural data model of the ribosomal RNA cavity around the PTC and the Yukawa-Debye-Hückel potential with dielectric screening

The Coulomb or Yukawa-Debye-Hückel electrostatic potential can be calculated from the x-ray solved exact distribution of the source charges (phosphate moieties and charged

amino acids) for which the positional 3D map is shown in Fig.4.6(b) to (e). The method to compute the electrostatic potential based on the real observed atom positions involves a discrete sum and complies with the superposition principle due to the linearity of the electrostatics equations. The Yukawa-Debye-Hückel potential is used and the exact positions $\vec{r}_i^{\,'}$ of the sources and their charges q_i are summed over all source charges:

$$\Phi(z) = \sum_{k \in \text{regions}} \sum_{i \in \text{source charges}} \frac{q_{i,k}}{4\pi\epsilon_0\,\epsilon_r(k)} \cdot \frac{e^{-\frac{|\vec{\mathbf{r}}_i^{\,'} - (0,0,z)|}{\xi_k}}}{|\vec{\mathbf{r}}_i^{\,'} - (0,\,0,\,z)|} \tag{4.27}$$

In this formula, two phenomenological parameters are required which are ϵ_r, the relative permittivity of the medium and the screening length ξ. The screening length ξ is a generic placeholder which can be computed from one of three possible different approaches (Bjerrum, Gouy-chapman or Debye screening length). The Coulomb potential is a particular case of the Yukawa potential when the screening length goes to infinity.

The assumption is made that the two phenomenological parameters are constant (homogeneous) in the media where the potential is computed. The standard or default homogeneous values of these parameters are $\epsilon_r = 78$ (water) and $\xi = 10$ Å. Eq.(4.27) neglects surface charge polarization effects at dielectric media discontinuities.

In the above Coulomb-Yukawa electrostatic potential Eq.(4.27), different values of $\epsilon_r(k)$ and different ξ_k screening lengths can be used in the different k-indexed regions (or media). The elementary unit charge value of $+|e|$ or $-|e| = -1.602\,10^{-19}$ C is used for each of the charges q_i associated to the positively or negatively charged atoms at their given $\vec{\mathbf{r}}_i^{\,'}$ positions.

4.5.3 Phenomenological parameters: σ^*/ϵ, media permittivities and dielectric screening lengths ξ

Electrostatic interactions in ribonucleoprotein structures are potentially quite strong, but these interactions are mitigated by the screening effects of water, nucleic acids (both rRNA and tRNAs) or nearby protein atoms [Lockhart and Kim 1993], even in the absence of mobile ions. In addition to the electrostatic screening, surface charge polarization effects also occur due to dielectric response discontinuities at media interfaces [Joiret et al. 2022b].

The screening of electrostatic interactions results primarily from electronic polarization, reorientation of dipolar groups in the vicinity of charges and dipoles. These effects are well understood and can be accurately determined for interactions in isotropic, homogeneous media. However, in complex inhomogeneous environments such as

those near the surface of ribonucleoproteins, dielectric screening is difficult to predict. Two factors are expected to be especially important in the case of the ribosome cavity around the PTC and the exit tunnel: the confined geometry and composition of the inner wall close to the surface, and whether the interactions involve direct charges or dipoles. The x-ray solved atomic space positions in the immediate 12 Å vicinity of the tunnel wall show that water molecules in addition to tRNAs contribute to the screening of the *formal bare* charges carried by the non-bridging oxygen atoms bound to the phosphorus atoms of 23S/28S rRNA. Due to the dissociation (charge regularization) of surface groups, the rRNA phosphate moieties support surface acquires a net surface charge density that we call σ^*. These *bare* charges do not stay unbalanced due to a screening effect involving water solvent. The water molecules dipole moments re-orient so that a layer of positively charged hydrogens oppose the negatively charged phosphate moieties.

Water as a bulk solvent has a relative electric permittivity of 78 (25°C), while experimental and theoretical evidence suggest that proteins (or the nascent polypeptide) have an average dielectric response that can be approximated with a dielectric constant of about 3-4 [Sharp and Honig 1990 and references therein]. The dielectric constant of nucleic acids in bulk solution has been measured to be around 8 [Cuervo et al. 2014]. Thus, depending on the abundance of water molecules in the PTC cavity volume, the presence of tRNAs and the presence of the carboxy terminal end of the growing nascent peptide, the cavity micro-environment cannot be viewed as uniform. The dielectric constants ϵ should be used in a range from $\epsilon = 3 - 4$ (polypeptide) to $\epsilon = 78$ (water).

Selecting the most appropriate screening theory reduces to knowing which length scale parameter ξ to use. Three length scales, i.e., the Bjerrum length (λ_B), the Debye length ($\lambda_D = \kappa^{-1}$) and the Gouy-Chapman length (ξ_{GC}) deserve specific attention as highlighted by Van Roij [Roij 16 July 2009]. The expressions of the Bjerrum, Debye and Gouy-Chapman screening lengths are respectively:

$$\lambda_B = \frac{e^2}{4\pi\,\epsilon\,\epsilon_0\,k_B T}. \tag{4.28}$$

$$\kappa^{-1} = \left(\frac{2\,e^2\,I\,N_A}{\epsilon\,\epsilon_0\,k_B T}\right)^{-1/2}. \tag{4.29}$$

$$\xi_{GC} = \frac{1}{2\pi\,\lambda_B\,\sigma^*}. \tag{4.30}$$

where e is the elementary charge of the electron, ϵ the relative permittivity of the medium, ϵ_0 the permittivity of the vacuum, I the ionic strength, k_B the Boltzmann constant, N_A the Avogadro number, T the absolute temperature and σ^* is the *bare* surface charge density of the wall.

4.5.4 Effects of forces on Gibbs free activation energy for the transition state in the peptide bond formation

The peptide bond formation is conducted at the peptidyl transferase center (PTC) in the large subunit of the ribosome. It has been known for more than thirty years that the peptidyl transferase center is not an enzyme but a ribozyme, composed of ribosomal RNA (rRNA)[Noller et al. 1992]. This ribozyme is a template assisted catalyst using an information rich mRNA single molecule to process two amino acid substrates previously acylated to tRNAs. These two tRNAs were initially accommodated and proofread on the A site for cognate anticodon-codon matching and then transferred to the P site. The detailed mechanisms of the peptide bond formation have largely been studied in the literature and the ribosome catalytic site described both as an entropy water trap and a water trap [Sievers et al. 2004; Wallin and Aqvist 2010]. During the peptidyl transfer reaction, the α-amino group of aminoacyl-tRNA positioned in the A site of the ribosome nucleophilically attacks the carbonyl carbon at the ester bond of the peptidyl-tRNA in the P site, which results in peptidyl-tRNA extended by one amino acid in the A site and deacylated tRNA in the P-site [Simonovic and Steitz 2009]. The peptide bond formation between the two aminoacylated-tRNAs proceeds 10 million times faster when catalyzed by the ribosome than when uncatalyzed in bulk solution [Beringer and Rodnina 2007]. The ribosome catalyzed peptide bond formation kinetics is known to be affected by the particular context of the upstream amino acid sequence and by the amino acid distribution embedded in the nascent chain ribosome exit tunnel [Joiret et al. 2022b; Rodnina et al. 2006].

The theory of kinetics of catalysis relies on the transition state theory that was introduced by Eyring who linked the rate constant of a chemical reaction to the Gibbs free activation energy of the transition state, TS [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983]. In this representation, a catalyzer accelerates a (bio)chemical reaction through a significant reduction in the transition Gibbs free energy barrier that the reactants have to overcome as shown in Figure 4.5.

Here, we hypothesize that the physical forces transmitted mainly through the backbone of the peptidyl-tRNA play a role in the reduction of the Gibbs free energy barrier of the transition state. The mechanical work of these physical forces affects the activation Gibbs free energy of the transition state. The modulation of the Gibbs free energy activation barrier changes the reaction rate constant through Eyring's relation [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983].

The catalyzed reaction scheme for the deacylation and peptide bond formation (transpeptidation) between the two substrates occurs in two steps with the formation of an activation complex (ribozyme complex) in the first step (rate limiting step). This is followed by a fast second step. If the second step is very fast, the overall reaction scheme can be viewed as a nucleophilic substitution (S_N2-like scheme) and there is a

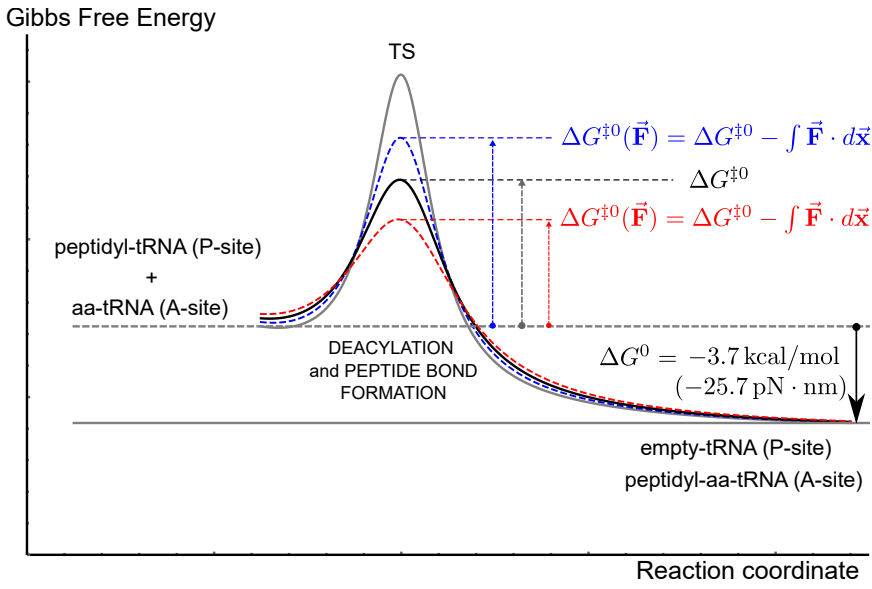


Figure 4.5: Gibbs free energy activation barrier to overcome at the transition state (TS) for the peptide bond formation at the peptidyltransferase center of the ribosome as a function of the reaction coordinate. The activation energy for the catalyzed reaction without mechanical forces (thick black line) is $\Delta G^{\ddagger 0}$ as compared to the uncatalyzed transition state (thin black line). The activation energy for the catalyzed reaction can be higher (dashed blue line) or lower (dashed red line) when the mechanical work exerted by any applied force on the peptidyl-tRNA substrate at the P-site during the reaction is negative or positive respectively.

single transition state as shown in Figure 4.5. The overall reaction can be written:

$$\underbrace{\text{peptidyl tRNA}}_{P-site} + \underbrace{\text{aminoacyl tRNA}}_{A-site} \longrightarrow \underbrace{\text{empty tRNA}}_{P-site} + \underbrace{\text{peptidyl(+1) tRNA}}_{A-site}. \tag{4.31}$$

This reaction is exergonic (thermodynamically favourable) as the change in standard Gibbs free energy is $\Delta G^0 \sim -3.7\,\mathrm{kcal/mol} (= -25.7\,\mathrm{pN \cdot nm})$ [Kaiser and Tinoco 2014]. The rate constant depends on the Gibbs activation energy of the transition state $\Delta G^{\ddagger 0}$. Applying external forces on molecules involved in catalyzed or uncatalyzed chemical reactions affects the kinetics of the reactions. The mechanical work of these applied mechanical forces can quantitatively be incorporated in the calculation of the activation Gibbs free energy of the transition state as already conceptually introduced

by Bell [Bell 1978], Bustamante [Bustamante et al. 2004] and others [Ribas-Arino and Marx 2012]:

$$\Delta G^{\ddagger 0}(\vec{\mathbf{F}}) = \Delta G^{\ddagger 0}(\mathbf{0}) - \int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}} \tag{4.32}$$

where $\Delta G^{\ddagger 0}(\vec{\mathbf{F}})$ is the activation energy for the transition state in the presence of an applied force acting on the system, $\Delta G^{\ddagger 0}(\mathbf{0}) \sim +14\,\mathrm{kcal/mol} = +97.2\ \mathrm{pN \cdot nm}$) is the activation energy for the transition state without any applied force [Bustamante et al. 2004], and $W = \int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}}$ is the mechanical work exerted by the force upon a test body along its curvilinear path. The mechanical work W is algebraically positive if the force and the displacement are parallel or negative if they are antiparallel. In the former case, $\Delta G^{\ddagger 0}(\vec{\mathbf{F}})$ is smaller than $\Delta G^{\ddagger 0}(\mathbf{0})$, whereas it is larger in the latter case as can be seen in Figure 4.5 (dashed red and dashed blue line respectively). In turn, the modulation of the Gibbs free energy activation barrier changes the reaction rate constant through Eyring's relation [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983]:

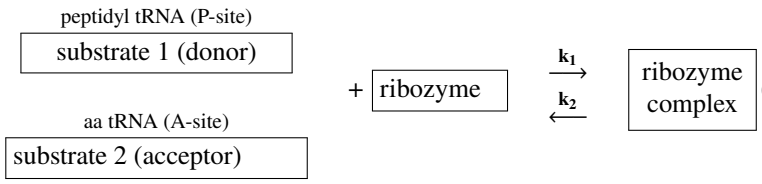
$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot e^{-\Delta G^{\ddagger 0}(\vec{\mathbf{F}})/N\,k_B\,T} \tag{4.33}$$

$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot e^{-\left(\frac{\Delta G^{\ddagger 0}(\mathbf{0})}{N\,k_B\,T} - \frac{\int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}}}{k_B\,T}\right)} = k(\mathbf{0}) \cdot e^{\frac{\int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}}}{k_B\,T}} \tag{4.34}$$

where $k(\vec{\mathbf{F}})$ is the reaction rate constant of the rate limiting step in the presence of an applied force upon the substrate at the P-site, $k(\mathbf{0})$ is the reaction rate constant in the absence of applied force. N, k_B, h and κ are Avogadro's number, Boltzmann's constant, Planck's constant and transmission coefficient respectively [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983].

4.5.5 Modified Michaelis-Menten kinetics of the peptide bond formation

The peptide bond formation kinetics at the catalytic center of the ribosome can be described by the Michaelis-Menten model where the aminoacyl-tRNA (acceptor substrate) is the canonical substrate and where the C-terminal amino acid of the peptidyl-tRNA (donor substrate) behaves as an allosteric substrate.



(4.35)

$$\boxed{\begin{matrix}\text{ribozyme}\\ \text{complex}\end{matrix}} \xrightarrow{k_3} \boxed{\text{ribozyme}} + \begin{matrix}\text{empty tRNA}\\ \text{(P-site)}\end{matrix} + \begin{matrix}\text{peptidyl(+1) tRNA}\\ \text{(A-site)}\end{matrix} \tag{4.36}$$

The rate of peptide bond formation, $k_{\text{pep}} = \frac{dP}{dt}$, is written:

$$\frac{dP}{dt} = v'_{max} \cdot \frac{S}{K_{1/2} + S} \tag{4.37}$$

$$v'_{max} = \frac{v_{max}}{1 + \frac{S_{allo}}{k_{allo}}} \tag{4.38}$$

where $K_{1/2} = \frac{k_2+k_3}{k_1}$ is the Michaelis constant of substrate 2, S, at the A-site; S_{allo} is the C-terminal amino acid at the peptidyl-tRNA (substrate 1); k_{allo} is the reaction rate constant between substrate 1 and the ribozyme (PTC) at the P-site and v_{max} is the maximum rate in the absence of allosteric effect ($v_{max} = k_3 \cdot [\text{ribozyme}]$). Incorporating the right hand side Maxwell-Boltzmann factor of (4.34) into (4.37) to account for the effect of mechanical forces in the kinetics of the ribozyme catalyzed peptide bond formation leads to the final kinetics equation:

$$\frac{dP}{dt} = e^{\frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}} \cdot v'_{max}(\mathbf{0}) \cdot \frac{S}{K_{1/2} + S} \tag{4.39}$$

In single molecule experiments, P, when proper normalized, represents the probability of the formation of the peptide bond over time in equation (4.39).

4.5.5.1 Elongation minimal case with dipeptide and puromycin.

Studying the catalyzed kinetics of peptide bond formation, as described in the reactional scheme (4.35) and (4.36), is possible only when peptidyl transfer is uncoupled from accommodation. The reason is that the accommodation rate of `aa-tRNA` in the A site is in the range $5 - 10\ \text{s}^{-1}$ and peptide bond formation follows instantaneously. Because accommodation precedes peptide bond formation, it limits the rate of product formation because it is much slower than the peptidyl transfer [Pape et al. 1999]. One way to circumvent the accommodation rate limiting step is to use substrate analogs that bind to the A site rapidly and do not require accommodation. If the full length `aa-tRNA` is replaced by the shorter puromycin substrate as the last acceptor substrate (substrate 2), the kinetics of the catalytic chemical step can be monitored experimentally by the quench flow technique and is amenable to quantitative measurements [Beringer and

Rodnina 2007]. These experimental measurements were conducted on prokaryotic ribosomes by Rodnina and coworkers in the elongation minimal case when the donor substrate 1 is the minimal dipeptidyl-tRNA, i.e. fMet – X – tRNA and the acceptor substrate 2 is puromycin, Pmn [Wohlgemuth et al. 2008]. X can be any of the 20 natural amino acids (positively charged, neutral or negatively charged). The global reaction scheme 4.31 becomes:

$$\underbrace{\text{fMet-X-tRNA}}_{P-site} + \underbrace{\text{puromycin(Pmn)}}_{A-site} \longrightarrow \underbrace{\text{empty tRNA}}_{P-site} + \underbrace{\text{fMet-X-Pmn}}_{A-site} \tag{4.40}$$

The product of the catalyzed reaction, dipeptidyl-puromycin fMet – X – Pmn, is released from the ribozyme upon completion of the reaction. The experimental initial condition for the puromycin concentration was 20 mM and the reaction rate constant measurements made at $37°C$ were obtained by single exponential fitting [ibid.].

4.5.5.2 Elongation cycle at codon resolution obtained by ribosome profiling Ribo-seq.

The ribosome profiling technology, or Ribo-seq, introduced by Ingolia *et al.* is based on deep sequencing of ribosome-protected mRNA fragments which produces a detailed account of the ribosome occupancy on specific mRNAs during translation in the sampled cells (bulk Ribo-seq) [Ingolia et al. 2009]. The density of ribosome occupancy on mRNAs at codon resolution is a proxy for the ribosome residence time (RRT) on each codon under endogenous conditions. A number of ribosome normalized footprint counts (NFC) distributions have been obtained for each codon in a number of previous published studies across multiple species [Dana 2014; Dana and Tuller 2012, 2014; Ingolia et al. 2011].

Dana & Tuller analyzed ribosome profiling data for *S.cerevisiae* and *E.coli*, among other species, based on published data in the GEO database, accession number GSE13750 (GSM346111, GSM346114) [Dana 2014; Dana and Tuller 2014]. They produced normalized footprint counts (NFC) of ribosomes on codons of the same type, originating from different genes, controlling for the transcripts copy number and initiation rates. The NFC enables measuring the relative time a ribosome spends translating each codon in a specific transcript relative to other codons in it, while considering the total number of codons in the transcript. Specifically, these authors denoted by T_j the translation time of codon j in transcript J and denoted the mRNA levels (transcript copy number) of transcript J by m, and its initiation rate by B, and calculated that the ribosome protected read counts RC_j for codon j is [Dana and Tuller 2014]:

$$RC_j \sim m \cdot B \cdot T_j \tag{4.41}$$

and thus

$$NFC_j \sim \frac{m \cdot B \cdot T_j}{m \cdot B \cdot \text{mean}(T)} = \frac{T_j}{\text{mean}(T)} \tag{4.42}$$

This last relationship, (4.42), indicates that NFC values represent the time a ribosome spends decoding each codon in a specific transcript relative to other codons in that transcript ibid.

Hence, if we divide the NFC by the mean elongation turnover $1/\text{mean}(T)$, i.e., the number of translated codons per unit of time for the species under study, we have:

$$\frac{NFC_j}{\text{turnover}} = \frac{T_j}{\text{mean}(T)} \cdot \frac{1}{1/\text{mean}(T)} \tag{4.43}$$

$$NFC_j \cdot \text{mean}(T) = T_j \tag{4.44}$$

We conclude that we can get the distribution of the ribosome residence time for codon j relative to the other codons from the NFC distribution. As the relative comparison is conducted on the translation mean time (elongation mean turnover), we have an averaged residence time of the ribosome for that codon in that species. Equivalently, if we have a probability density function for the NFC_j, multiplying this density by the elongation turnover, while scaling the probability space by the mean elongation time, we have the probability density function of the ribosome residence time for the codon of type j. Explicitly, the probability space scaling by $\text{mean}(T)$ means that if $\text{pdf}_{\text{NFC}_j}(x)$ denotes the probability density function of the normalized footprint count for codon j relative to the other codons, then:

$$f_{\text{RRT}_j}(t) = \frac{1}{\text{mean}(T)} \text{pdf}_{NFC_j}(x/\text{mean}(T)) \tag{4.45}$$

is the probability density function for the ribosome residence time on codon of type j. In equation (4.45), the turnover, or its inverse, $1/\text{mean}(T)$, just acts as a scaling of the NFC distribution. Given the way the NFCs were generated, one should refrain to interpret the probability space, the x-axis, as absolute times. Relative time ratio can be obtained from the comparison of two RRT distributions. Drawing conclusions on absolute time differences from such distributions might be misleading. In other words, the comparison of RRTs between codons should be made in time fold change, not in absolute time differences.

The time spent by a ribosome on codons provides information on the elongation cycle dynamics at codon resolution but this information is statistically aggregated translatome wide and statistically aggregated for all sequential sub-steps possibly occurring during the elongation cycle. The ribosome footprint count may help infer ribosome residence time on a given codon in a given condition but does not directly provide the temporal breakdown of the sub-steps occurring during one elongation cycle. How long does take

the accommodation and proofreading of the tRNA on the codon? How long does take the peptide bond formation at the PTC and how long does take the translocation to the next codon? All these individual sub-step times are aggregated together in a global RRT per codon and translatome wide, i.e. for a large number of transcripts collectively.

Dana and Tuller interpreted the distribution of ribosome residence time on codons as the distribution generated by the sum of two independent random variables, i.e., a Gaussian and an exponentially distributed random variable [Dana 2014; Dana and Tuller 2014]. The convolution of the two distributions leads to the exponentially modified Gaussian distribution. Our interpretation differs from theirs in considering that the empirically observed distribution of the ribosome residence time on a given codon is the sum of (at least) three contributing random variables that are exponentially distributed. The convolution of three such exponential distributions leads to the hypo-exponential distribution, see appendix A.1. Our interpretation derives from the queueing theory of a succession of sequential waiting times. The underlying statistical models in both interpretations result in the same topology for the empirical observed residence time distribution of a ribosome on a given codon. Indeed, the exponentially modified Gaussian distribution, the hypo-exponential distribution, the Gamma distribution (but also the log-normal distribution) convey very similar information as detailed in appendix A.1. In a pioneering paper, Tinoco and Wen used the Gamma distribution to simulate the time spent by a single ribosome on a codon during translation [Tinoco and Wen 2009]. In a recent paper, estimation of peptide elongation times from ribosome profiling spectra was derived where the copy number of ribosome on any codon on any open reading frame was assumed to be Poisson distributed and where the frequency densities of the peptide elongation times are well fitted to log-normal distributions [Pavlov et al. 2021].

The Kullback-Leibler (KL) divergence distance between the exponentially modified Gaussian and its least squared fitted hypo-exponential is very close to zero. From the theory of information, a KL value close to zero indicates that both models are statistically equivalent in practice, when it comes to the characterization of the ribosome residence time on a given codon like `GAA` (Glu-E) or `CGG` (Arg-R), see appendix A.1 and below.

Here, we adopt the point of view that the observed empirical distribution of the residence time of a ribosome on a given codon derives from the sum of three random waiting times that are exponentially distributed. These three time steps are (1) `STEP 1`: tRNA accommodation and proofreading at the A-site decoding center; (2) `STEP 2`: peptide bond formation at the PTC; (3) `STEP 3`: translocation of the ribosome to the next codon downstream the mRNA (after the unloaded-tRNA left the E-site). The ribosome elongation cycle rate limiting step was shown to be `STEP 1` as previously reported by others [Rodnina et al. 2006; Rodnina and Wintermeyer 2001; Wohlgemuth et al. 2008].

Through the comparisons of the observed empirical ribosome residence time (RRT)

on specific codons deciphering positively charged amino acid and negatively charged amino acid, respectively, we can gain access to the Maxwell-Boltzmann's factor of equation (4.34) modulating the rate of `STEP 2`, the peptide bond formation. We quantitatively compare the peptide bond formation rate when arginine (or lysine) is incorporated to the nascent chain at the PTC with the situation when glutamate (or aspartate) is incorporated. This allows to question whether the *ratios* between these rates are quantitatively consistent with the Maxwell-Boltzmann's factor modulation caused by the change in mechanical work, when a positive or negative `aa` is incorporated at the PTC, while the ribosome is deciphering the corresponding codon on the mRNA.

In summary, the theoretical workflow is to use the observed distribution of RRTs for a positively charged `aa` (resp. negatively charged `aa`), fit the hypo-exponential distribution, estimate the rate of `STEP 2` by educated deconvolution of the hypo-exponential distribution, compute the *ratio* of the two rates (positively and negatively charged `aa`) and compare with the Maxwell-Boltzmann modulation factor. To our knowledge, this workflow shows for the first time how Ribo-seq data at codon resolution, when properly deconvoluted, can be used to indirectly gain information on peptide bond formation relative rates along with the other sub-steps rates in the elongation cycle. The experimental results for the observed NFCs (and the derived RRTs) were extracted from file #007 (excel), publicly available, in the supplementary data of Dana and Tuller [Dana 2014]. This file provides for each of the 61 sense codons, the exponentially modified Gaussian best fitted parameters to the observed NFC distribution across 4 species.

4.6 Results

The ribosome exit tunnel was modeled by a cylinder concatenated to a cone frustum in reference [Joiret et al. 2022b]. As a result of the scaling exposed in Material and Methods, the peptidyl transferase center (PTC) is approximately surrounded by a minimal surface prolate spheroidal volume accommodating for 3 tRNAs having their anticodon matching the 3 codons at the A, P and E site respectively. The distance between the amino acid residues acylated to their cognate tRNAs and their anticodon loops is 8 nm (the size of the L-shaped tRNA single molecule). The cylinder radius is 5 Å [Joiret et al. 2022b; Lu et al. 2007] and is connected to the PTC cavity by the truncated prolate spheroidal volume as shown in the simplified representation of Fig.4.6(a). The angle between the support of the PTC arrow and the direction from the tunnel entry to the tunnel exit port is 38°.

Figure 4.6: (caption of next page).(a) Ribosome exit tunnel and cavity around the PTC. (b) and (c) Front and right view of the charged groups neighboring the PTC cavity. (d) and (e) 3D scatter plot of the 291 charged atoms around the PTC. Phosphorus atoms bridging the 23S rRNA backbone (orange dots), positively charged moieties of lysine (NH2 atom in PyMol convention) or arginine (NZ atom in PyMol convention) belonging to ribosomal proteins (red dots), negatively charged moieties of aspartate or glutamate (OE2 or OD2 atom in PyMol convention) belonging to ribosomal proteins (blue dots), phosphorus atoms of P loop 23S rRNA nucleotides A2485 and A2486 (black dots), phosphorus atoms of A loop 23S rRNA nucleotides G2588, U2589 and U2590 (gray dots). The atomic positions were retrieved from *H. marismortui* x-ray solved structure of the large subunit of the ribosome.

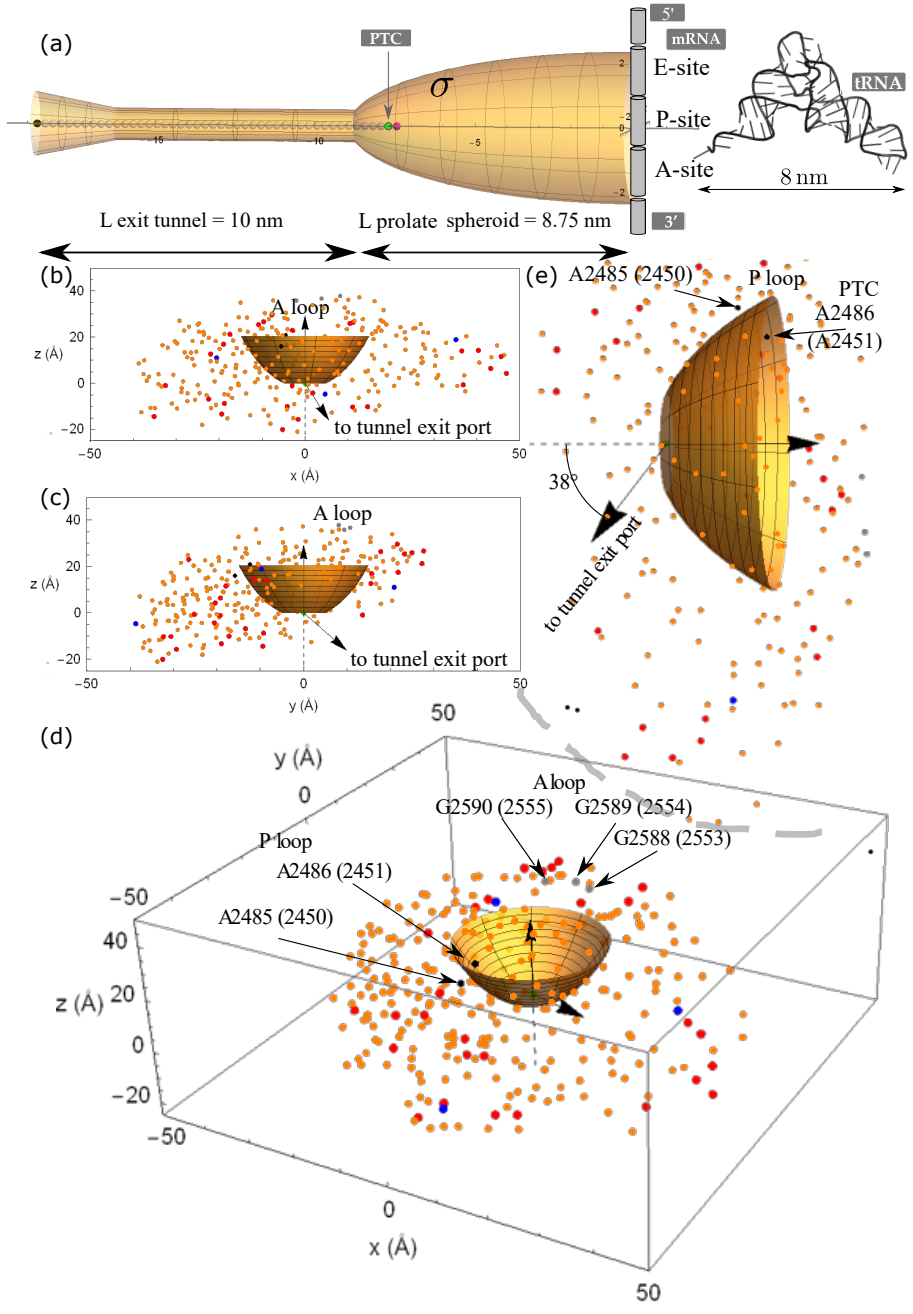
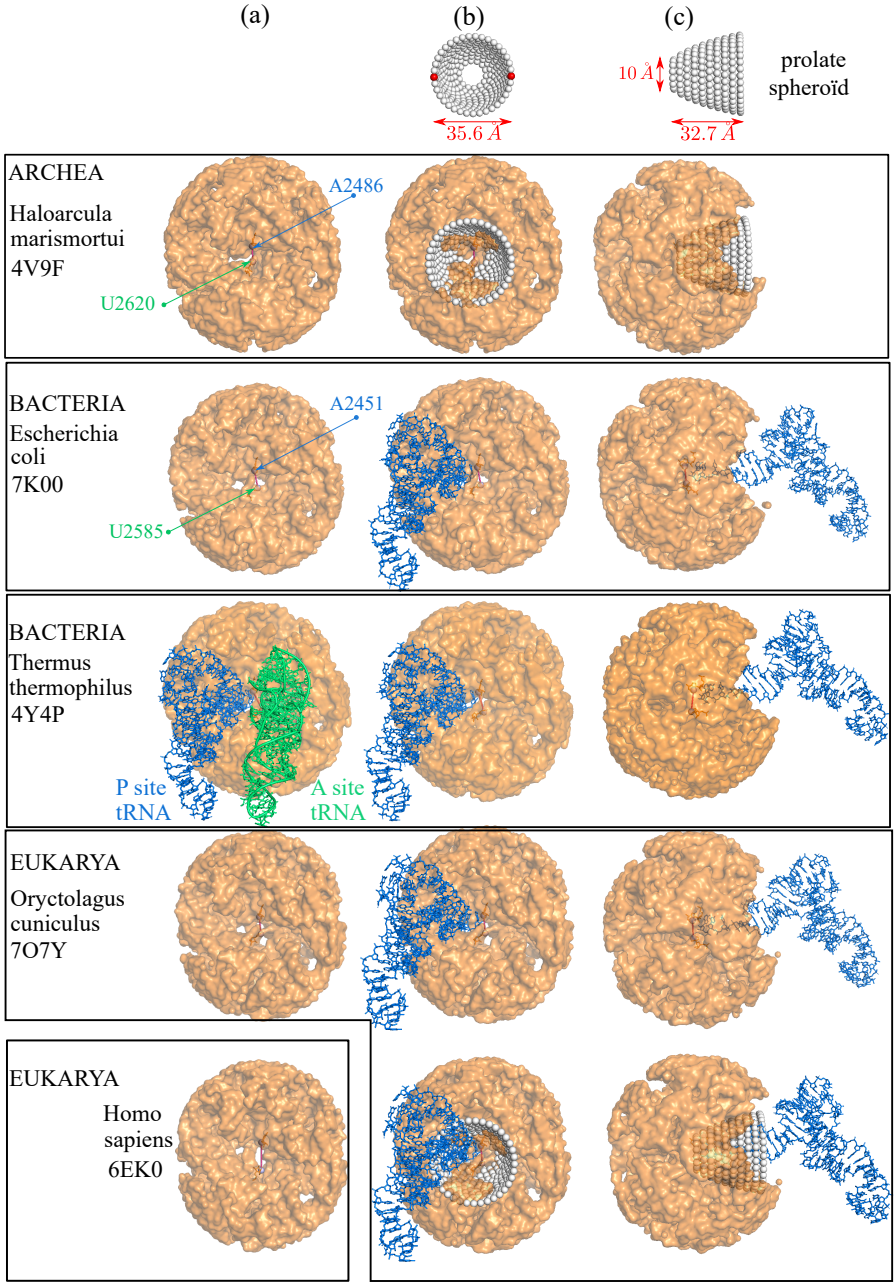


Figure 4.6: Ribosome exit tunnel and cavity around the PTC. Caption: previous page.

4.6.1 Overview of the atlas of the 5 maps of the PTC shell cavity across the 3 domains of life and impact on the electrostatic potential profiles

The nucleotides residues of the 23S/28S rRNA which are taking part in the catalytic activity of the peptide bond formation, i.e. at the PTC, are not contiguous. Instead, as shown in Fig. 4.4, the PTC is formed by a diverse collection of distant residues that, when the rRNA is folded into its tridimensional shape, act together as a catalytic unit [Tirumalai et al. 2021]. The universally most conserved nucleotides sequences of the 23S/28S rRNA across the three domains of life (better than 90% conservation across archea, prokarya and eukarya) were used to localize the homologous nucleotides that are known to take part in the peptide bond formation [Doris et al. 2015]. The key nucleotides belonging to the P-loop and A loop respectively are given in Table 4.4.1. The key nucleotides that are bold typeset in Table 4.4.1 in the P-loop and the A-loop are separated by a distance of 134 nucleotides in the rRNA primary sequence. In the tridimensional 23S/28S rRNA folded structure, these key nucleotides are at a physical distance ranging from 7 to 10 Å only. A number of studies have shown the mutational flexibilities of a number of the nucleotides residues around the PTC. Most mutations do not affect the transpeptidation activity [Tirumalai et al. 2021]. The PTC is also enriched in modified residues but none of these modifications appear to be essential for peptide bond formation [ibid.]. Here we further inquire if the functional importance of the universal conservation of the rRNA sequence can be visualized in the 3D structures of the PTC shell cavity. Figure 4 shows the atlas of five 23S/28SrRNA tridimensional maps representing the PTC shell cavity of five species across the 3 domains of life. This atlas compares the tridimensional structures of the PTC shell cavities and shows strong common geometrical features for the 23S/28S rRNA ribose-phosphate backbone across the 5 species that might shed light on important functional roles. There are between 484 and 488 nucleotides of the 23S or 28S rRNA that are within a 40 Å distance of the key nucleotides of the P and A loop of table 4.4.1. We used the 3D space positions of the phosphate moieties to compare the resulting electrostatic profile around the PTC and more specifically, we calculated the electrostatic profile along an arbitrary virtual path in the PTC shell cavity between the A-loop and the P-loop positions and towards the tunnel entry port.

Figure 4.7: (caption of next page). Atlas of the tri-dimensional PTC shell cavity for 5 species across the 3 domains of life. (a) and (b) Front views of the shell cavity around the PTC. (c) Side views of the shell cavity around the PTC. The atomic positions were retrieved from the 5 atomic solved structures of the large subunit of the ribosome for PDB codes $4V9F$, $7K00$, $4Y4P$, $7O7Y$ and $6EK0$. A synthetic representation of the prolate spheroid was superposed in the front (b) and side views (c) for *Haloarcula marismortui* and *Oryctolagus cuniculus*.


(a)
(b)
(c)
10 Å
prolate
spheroïd
35.6 Å
32.7 Å
ARCHEA
Haloarcula
marismortui
4V9F
A2486
U2620
BACTERIA
Escherichia
coli
7K00
A2451
U2585
BACTERIA
Thermus
thermophilus
4Y4P
P site
tRNA
A site
tRNA
EUKARYA
Oryctolagus
cuniculus
7O7Y
EUKARYA
Homo
sapiens
6EK0

The tridimensional spatial conservation of the 484-488 nucleotides from the 23S or 28S rRNA within a 40 Å distance of the PTC key nucleotides results in the conservation of the electrostatic potential profile along any arbitrary path from the A site to the P site or along the centerline from the PTC towards the ribosome nascent chain tunnel entry point. The common qualitative feature that is shared for the 5 species across the 3 domains of life is the funnel shape of the cavity when moving along an arbitrary path from the A-loop to the P-loop, and then towards the exit tunnel. Any arbitrary path along that direction encounters closer phosphate moieties positions. This causes the electrostatic potential to decrease when moving in the PTC shell cavity from the A-loop to the P-loop and then to the tunnel entry port. A decrease in the electrostatic potential means that the electric field points in the shell cavity from the A-loop to the P-loop and to the entry port of the tunnel. To support the effects of the qualitative observation of the shared common feature across species, Fig. 4.8 shows the calculated electrostatic potential profiles along an arbitrary straight line path starting around the PTC and heading towards the tunnel entry point for the two bacteria species. The potential also decreases when moving from the A-loop to the P-loop (not shown here). In the immediate vicinity of the PTC, the electric field points in the direction from the A-loop to the P-loop and in the direction of the tunnel entry port. The calculated values of the electrostatic potential for *E.coli* and *T. thermophilus* show a decrease in the potential from $-150\,\mathrm{mV}$ to $-200\,\mathrm{mV}$ and $-250\,\mathrm{mV}$. The resulting estimated values for the axial force experienced by a positively charged test particle are in an averaged range of $-20\,\mathrm{pN}$ to $-30\,\mathrm{pN}$. In the chosen z reference frame, the negative sign means the force points towards the tunnel for a positively charged test particle. These results are similar to the ones obtained below for the archeon *H. marismortui*.

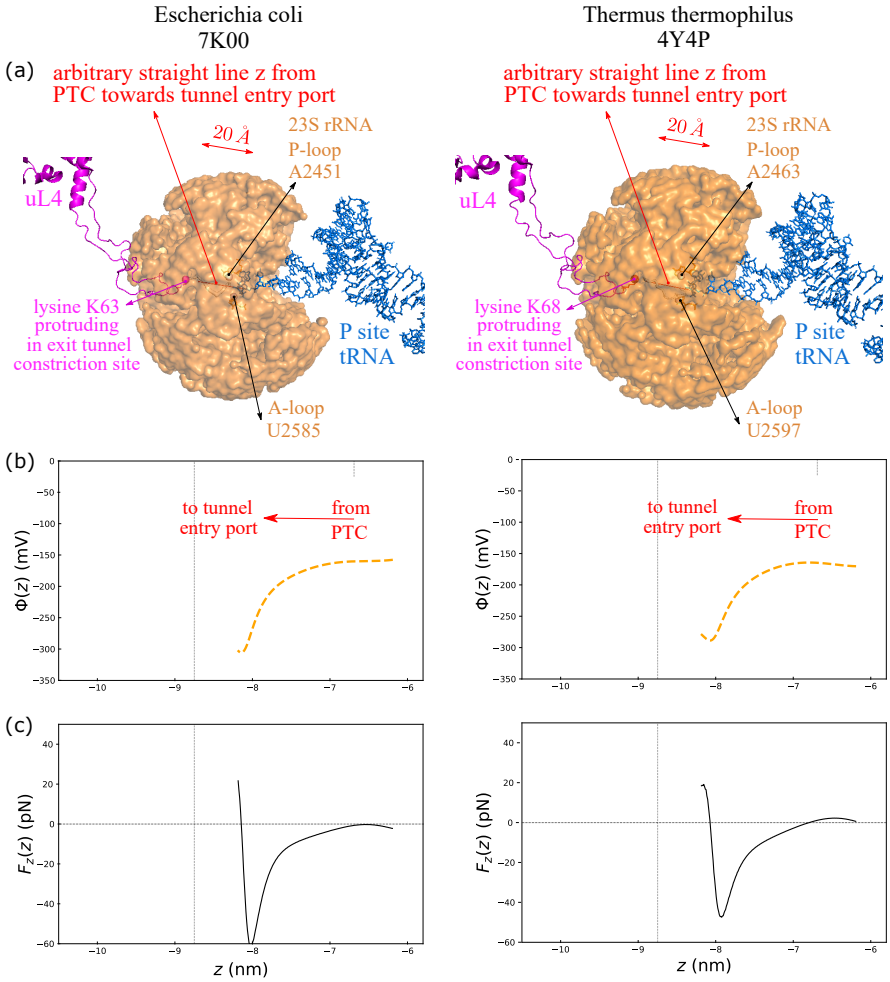


Figure 4.8: Electrostatic potential profiles and electric field caused by the funnel shape of the PTC shell cavity. Comparison between species: *E.coli* (left), *T. thermophilus* (right). The arbitrary paths are taken as straight lines from the mid-point between A-loop and P-loop nucleotides to the lysine K amino acid in uL4 protruding in the tunnel constriction site. (a) Comparison of spatial representation of the x-ray solved structures of the 484 nucleotides of 23S rRNA in the immediate vicinity of the PTC. (b) Comparison of electrostatic potential $\Phi(z)$. (c) Comparison of axial force $F_z(z)$ along z path.

4.6.2 Magnesium, potassium ions counts around the PTC shell cavity

Depending on the presence or not of tRNAs at A and P sites in the x-ray crystallographic solved structure of the ribosomes (depending on the PDB entry codes that are used), it is difficult to discriminate the metal ions (monovalent K^+ or bivalent Mg^{2+}) that are bound to the tRNAs structures from the ones that are bound to the 23S rRNA in the immediate vicinity of the PTC shell cavity or to determine whether these metal ions are free or bound. As shown in Fig.6 (b) and(c), we localized the `CCA` nucleotides at the 3' end of the tRNA molecules at the P and A sites of the *Thermus thermophilus* PDB code `4Y4P` solved structure of the ribosome. We counted and mapped the 3D positions of the metal ions (K^+ or Mg^{2+} indistinctively) that are within a 12 Å distance of the $3'-$oxygen atom at the tip of the ribose of adenosine 76 at the P-site tRNA. Except for a limited number of papers, such as the one by Rozov *et al.*, the bound metal ions counts around the PTC cavity were not of specific concerns. In their paper, Rozov *et al.* located seven metal ions next to the inner shell of the PTC for the ribosome with PDB code `4V6F` [Rozov et al. 2019]. In the x-ray solved structure of the ribosome with PDB code `4Y4P` of *Thermus thermophilus*, our count of the bivalent Mg^{2+} ions would be 8 and there would be only one monovalent K^+ ion (and a total of 24 water molecules) within a 12Å distance of the $3'-$oxygen atom at the tip of the ribose of adenosine 76 at the P-site tRNA, i.e, very close to the PTC shell cavity as shown in Fig.4.9 (a). The metal ions and water molecules that are within a 12Å distance from all the phosphorus atoms belonging to the whole tRNA at the P-site are 34, 1 and 70 for Mg^{2+}, K^+ and water respectively as shown in Fig.4.9 (c). Altogether, the metal ions and water molecules count is consistent with their screening role as free mobile ions of the 76 phosphate moieties of a single tRNA molecule at the P-site. For the tRNA molecule at the A site, there are 36, 0 and 68 counts of Mg^{2+}, K^+ and water respectively within a 12Å distance from all the phosphorus atoms belonging to the whole `tRNA` at the A-site (not shown here).

4.6.3 P-site and A-site tRNAs phosphate moieties and associated screening metal ions positions

The substrate of the peptide bond formation at A-site tRNA that will find itself accommodated in the PTC shell cavity of the ribosome for peptide bond formation is the amino acid residue that was loaded on its cognate tRNA by its specific aminoacyl tRNA synthetase. More precisely, it is the amino group, i.e. $-NH_2$, of the amino acid that will be processed in the PTC shell cavity along with the carboxy terminal part of the peptidyl-tRNA at the P-site. The nitrogen atom of this amino group will be closer to nucleotides of the 23S/28S rRNA (PTC shell cavity) than to any of the 76 nucleotides of the aminoacyl-tRNA. Figure 6 (a) shows that the first closest phosphate moiety (P76)

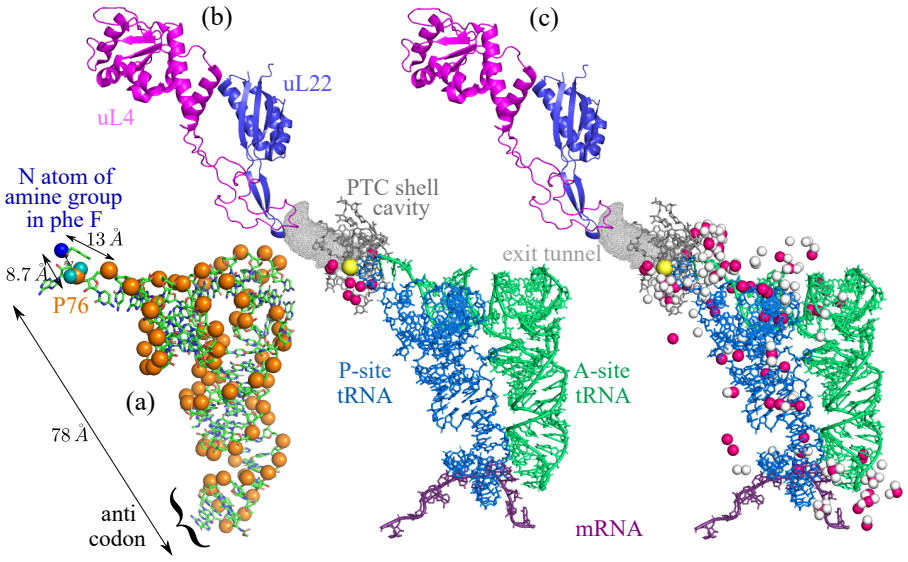


Figure 4.9: (a) *Thermus aquaticus* PDB code 1TTT: tRNA structure of aminoacyl-tRNA^{phe} showing distances from amino group of the loaded amino acid to the 2 closest phosphate moieties of the tRNA; 76 phosphorus atoms in tRNA (orange spheres) (b) Mapping of metal ions around PTC shell cavity and tRNAs. *Thermus thermophilus* ribosome X-ray solved structure PDB code 4Y4P: 9 metal ions Mg^{2+} (pink spheres) or K^+ (yellow sphere) within 12Å of the tip of ribose $3'O$ atom of P-site tRNA adenosine 76 in close vicinity of the PTC shell cavity. (c) PDB code 4Y4P: metal ions (pink and yellow spheres) and water molecules (white spheres) within 12Å of all 76 phosphate moieties of the P-site tRNA. Cautionary note: the distinction between the metal ions Mg^{2+} or K^+ is elusive on these standard x-ray crystallographic solved structures of the ribosome and their tRNAs [Rozov et al. 2019].

of the tRNA molecule to the amino group of the amino acid residue is at a distance of 8.7 Å. The second closest phosphate moiety of the tRNA is at a distance of 13 Å. This last distance is already larger than the screening length in the Yukawa-Debye-Hückel model. Most of the 76 phosphorus atoms and their non-bridging oxygen atoms are much more distant than the electrostatic screening length. The electrostatic effect of all the phosphate moieties belonging to the tRNA molecules is therefore expected to be negligible in the immediate vicinity of the 23S/28S rRNA PTC shell cavity where the peptide bond formation is catalyzed. The number of metal ion equivalents within a 12 Å distance to the tRNAs is approximately the same as the number of charged equivalents in the phosphate groups belonging to the tRNAs. These ions, on average, are screening the phosphate groups for the tRNAs. The screening effect of these metal ions along

with water molecules is such that the tRNAs taken as a whole with their metal ions do not contribute significantly to the electrostatic profile around the PTC or along the line from the PTC to the tunnel entry port. This is expected since the tRNAs shuttle in and out from the PTC cavity to the cytoplasm where they easily equilibrate with the ionic composition of the cytoplasm. These facts suggest that the tRNAs should not experience significant hindrance of electrostatic origin when they are accommodated at the A site and then at the P site.

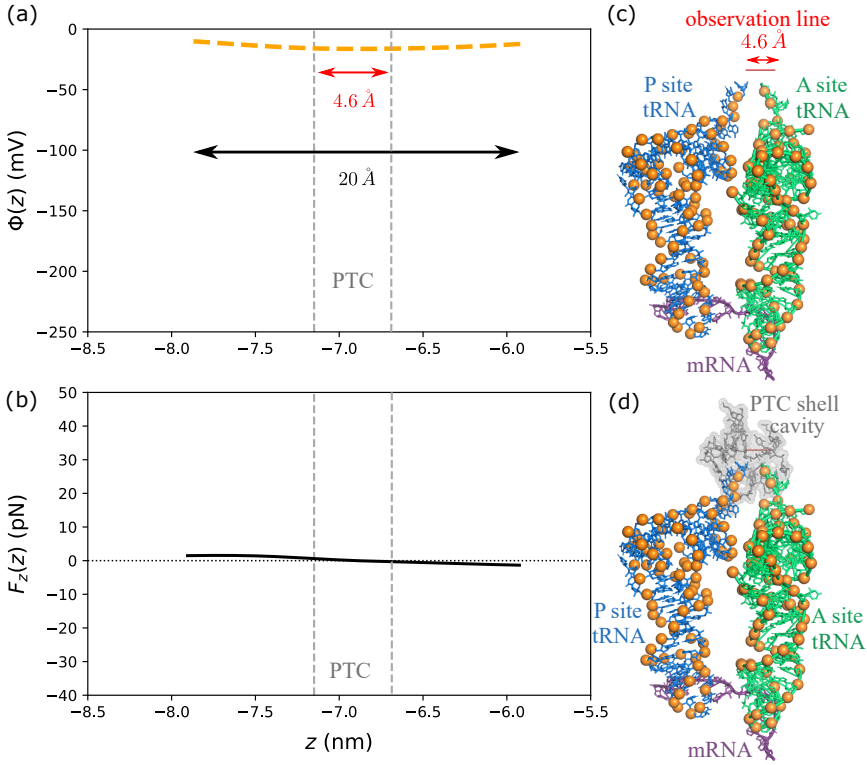


Figure 4.10: Electrostatic potential contributed by *Thermus thermophilus* PDB code 4Y4P tRNA structures as accommodated at P and A sites of the ribosome showing phosphate moieties positions of the tRNAs. The potential and axial forces profiles are plotted with the same y-scales as the ones used for the contribution of the 23S/28S rRNA (see Fig. 4.8 and Fig. 4.12 for comparison) (a) Electrostatic potential profile (dashed orange) along the red straight line (c-d), in the vicinity of the PTC and contributed by the tRNAs. The potential profile was calculated over a 20 Å extended distance along the red straight line. (b) Axial force profile (black) in the vicinity of the PTC and contributed by the tRNAs. (c) Observation straight line (red line) in the vicinity of the PTC and the 152 phosphorus atoms of the tRNAs (orange spheres). (d) PTC shell cavity near the tRNAs' tips. 71 metal ions (Mg^{2+} or K^{+}) and 138 free water molecules (not shown here) contribute to the ionic strength screening the phosphate moieties.

To substantiate quantitatively the fact that the phosphates belonging to the tRNAs both at P-site and A-site have negligible electrostatic effect around the PTC shell cavity (and taking into account the metal ions within screening distance of these tRNAs), we calculated the electrostatic potential along a virtual observation line close the two tRNAs and in the vicinity of the PTC cavity. We replicated the method used in section 4.5.2 and extracted 76 + 76 phosphorus atoms' positions of the tRNA molecules, and 35+36 metal ions from PDB file 4Y4P of the solved structure of ribosome for *Thermus thermophilus*, having the tRNAs at both A and P-sites. We constructed the virtual observation line (red straight line in Fig.4.10. (c-d)) by taking a parallel straight line at a 5Å distance to the line joining the two oxygen 3-prime atoms of the riboses of adenine 76 at both tRNAs (P site and A site) and in the plane including nucleotide 34 of the P site tRNA (decoding center). This red line is approximately located in the PTC shell cavity where we want to estimate the electrostatic potential and electric field. Fig.4.10 (a-b) shows that the contribution to the electrostatic potential and to the electric field caused by these charged moieties of the tRNAs accommodated at P and A sites is negligible when compared to the one caused by the 23S/28S rRNA in the immediate vicinity of the PTC shell cavity (within a distance of $7 - 12$Å). The values, calculated with the Yukawa-Debye-Hückel theory, of the electrostatic potential around the PTC contributed by the 2 tRNAs are about -20 mV (Figure 4.10(a)) while the values contributed by the 484 nucleotides of the 23S rRNA around the PTC are between -150 mV to -200 mV (Figure 4.8(b)). The potential profile of Figure 4.10(a) is flat and the resulting contribution to the electric field or force is negligible: compare Figure 4.10(b) (*T. thermophilus*) to Figure 4.8(c) (*E.coli* and *T. thermophilus*) or Figure 4.11(d) (*H. marismortui*).

At the scale of the tRNA full length (~ 80 Å), the tRNAs are neutral. For electrostatic charges to be observed as unbalanced, the observation scale must be smaller than the Debye, Bjerrum or Gouy-Chapman screening lengths all of which are smaller than 10 Å.

4.6.4 Surface charge density and dielectric screening lengths around the PTC

Fig.4.6(b-e) shows the 3D scatter plot of 291 charged groups in the region around the PTC and the tunnel entry port. The exact number of phosphate moieties around the PTC was counted from the x-ray solved structural data. The *bare* surface charge density σ^* was calculated by dividing the number of phosphate moieties (= 100) closer than 12 Å from the truncated spheroid by its surface area. The surface area was analytically calculated to be $15.75\,\mathrm{nm}^2$ as detailed in the appendix B.1. Hence, the *bare* surface charge density σ^* is estimated to be $\sigma^* \sim -\frac{100}{15.75}\,\frac{|e|}{\mathrm{nm}^2} = -6.35\,\frac{|e|}{\mathrm{nm}^2}$ on the surface of the spheroid cavity around the PTC. This numerical result for the surface charge density

is approximately three times higher than the surface charge density prevailing on the inner surface of the ribosome exit tunnel [Joiret et al. 2022b].

It is unclear whether Bjerrum or Debye theory should be used for the electrostatic screening length in the medium inside the PTC cavity. The Bjerrum screening length as computed from Eq.(4.28) is 0.72 nm when water permittivity $\epsilon = 78$ is used at 298 K inside the truncated spheroid medium around the PTC. The Debye screening length as computed from Eq.(4.29) is 1.1 nm when water permittivity $\epsilon = 78$ and an ionic strength $I = 75$ mM ($= 75\,\mathrm{mol/m^3}$) are used at 298 K inside the truncated spheroid medium around the PTC. Note that a Debye screening length equal to the Bjerrum length corresponds to the ionic strength's value of $I = 76$ mM at 25°C and for water permittivity. The mathematical expressions of both theories are then numerically equivalent and knowing which theory prevails is numerically irrelevant.

4.6.5 Comparison of the potential profiles and fields calculated from the spheroid idealized shape and from the x-ray solved structural data around the PTC

The x-ray solved structure dataset includes the exact 3D coordinate positions of a total of 291 charged atoms that are closest to the ribosome exit tunnel entry point or the PTC. Of these 291 atoms, 113 are at a distance shorter than 12 Å from the truncated spheroid's outer surface. The volume inside the truncated spheroid is empty, i.e., there are no 23S rRNA phosphorus atoms or ribosomal proteins charged amino acid moieties inside the volume. The charges which are at a distance larger than 12 Å from the outer surface of the cavity are so strongly screened by the dielectric medium that they can be neglected as was detailed in [ibid.].

The electrostatic potential profile $\Phi(z)$ for the uniformly charged idealized spheroidal shape, as numerically calculated from Eq.(4.24), is shown in Fig.4.11(c, black solid line). Upon implementing Eq.(4.27) in Python and using the exact positions of the 113 charged atoms within 12 Å of the spheroid's outer surface (as mapped in Fig.4.6), we obtained the electrostatic potential along the spheroid's centerline shown in Fig.4.11. The electrostatic potential contributed by the phosphate moieties (Fig.4.11(a)) is negative while the potential contributed by the charged amino acid residues (Fig.4.11(b)) is positive as arginine or lysine outnumber aspartate or glutamate residues. The net resulting electrostatic potential is negative, orange solid line in Fig.4.11(c,) as there are $\sim$ 100 phosphate groups and only $\sim$ 13 charged amino acid residues at a distance closer than 12 Å from the PTC surface cavity. The charged amino acid residues are located much further away from the catalytic center than the phosphate moieties. The potential profile is due to the dominant presence of negatively charged phosphate moieties harbored by 23S/28S rRNA on the inner surface cavity in the immediate vicinity of the PTC.

The potential profiles calculated from the uniformly charged spheroid idealized shape (Eq.4.24) or calculated from the discrete sum of the Yukawa-Debye-Huckel formula over the charged atoms at their exact positions (Eq.4.27) are compared in Fig.4.11(c). The two potential profiles look similar and provide an estimation of the order of magnitude of the electrostatic potential along the virtual centerline between the PTC and the entry port of the exit tunnel. The uniformly charged assumption of the surface in the region between the PTC and the tunnel entry port appears to be reasonable. The discrepancy between the two potential profiles (idealized spheroidal shape with uniform surface charge density versus exact atomic positions of the charges, shown in Fig.4.11(c)), or axial electrostatic force profiles (Fig.4.11(d)), is sensitive to the two phenomenological parameters of the models i.e., permittivity and screening length and their possible local heterogeneity (not shown here). Actual measurements of the electric field through vibrational Stark effect spectroscopy would provide indirect support and constraints on the ranges of the local values for ϵ and ξ (permittivity and screening length).

The electrostatic potential particular profile results from the funnel shape of the cavity. The impact of the funnel shape on the potential profile is that when moving from the PTC to the tunnel entry port along the $z-$axis, the potential goes from smoothly convex (smaller curvature) to sharply convex (larger curvature) (Fig.4.11(c)). The resulting negative inverse bell shaped peak for the electric field (or force) has a larger width in the region from the PTC to the tunnel entry port than the width of the positive bell shaped peak in the region near the tunnel entry port (Fig.4.11(d)). The respective electric fields profiles and hence the forces along the $z-$axis centerline are also compared in Fig.4.11(d). From these numerically estimated electrostatic force profiles, the maximum force at the center of the cavity neighboring the PTC would be between -21 and -27 pN for a unit positive test charge. The negative sign means that the force would point from the PTC to the tunnel entry port for a positive test charge.

Figure 4.11: Caption continuation of next page. (b) Electrostatic potential (mV) computed from the exact atomic positions of the 13 charged amino acid within 12 Å of the surface cavity. (c) Electrostatic potential (mV) computed from the exact atomic positions of all 113 charged moieties within 12 Å of the surface cavity (orange solid) compared to the potential resulting from a uniformly charged truncated spheroid surface of $\sigma^* = -6.35|e|/\text{nm}^2$ (black solid). (d) Axial electrostatic force (pN) computed from the exact atomic positions of all 113 charged moieties within 12 of the surface cavity (gray dashed) compared to the force resulting from a uniformly charged truncated spheroid surface of $\sigma^* = -6.35|e|/\text{nm}^2$ (black solid). All panels: 3 gray dashed vertical lines, from left to right: tunnel entry port position (radius= 5 Å), P loop A2485 z position, and truncated spheroid position where the cavity radius = 15 Å. A Bjerrum screening length of 0.72 nm was used for the analytical potential calculation and a Debye screening length of 1.1 nm was used in the discrete sum over the exact atomic positions, with ionic strength $I = 75$ mM. The water permittivity was assumed to be $\epsilon = 78$ inside the PTC cavity.

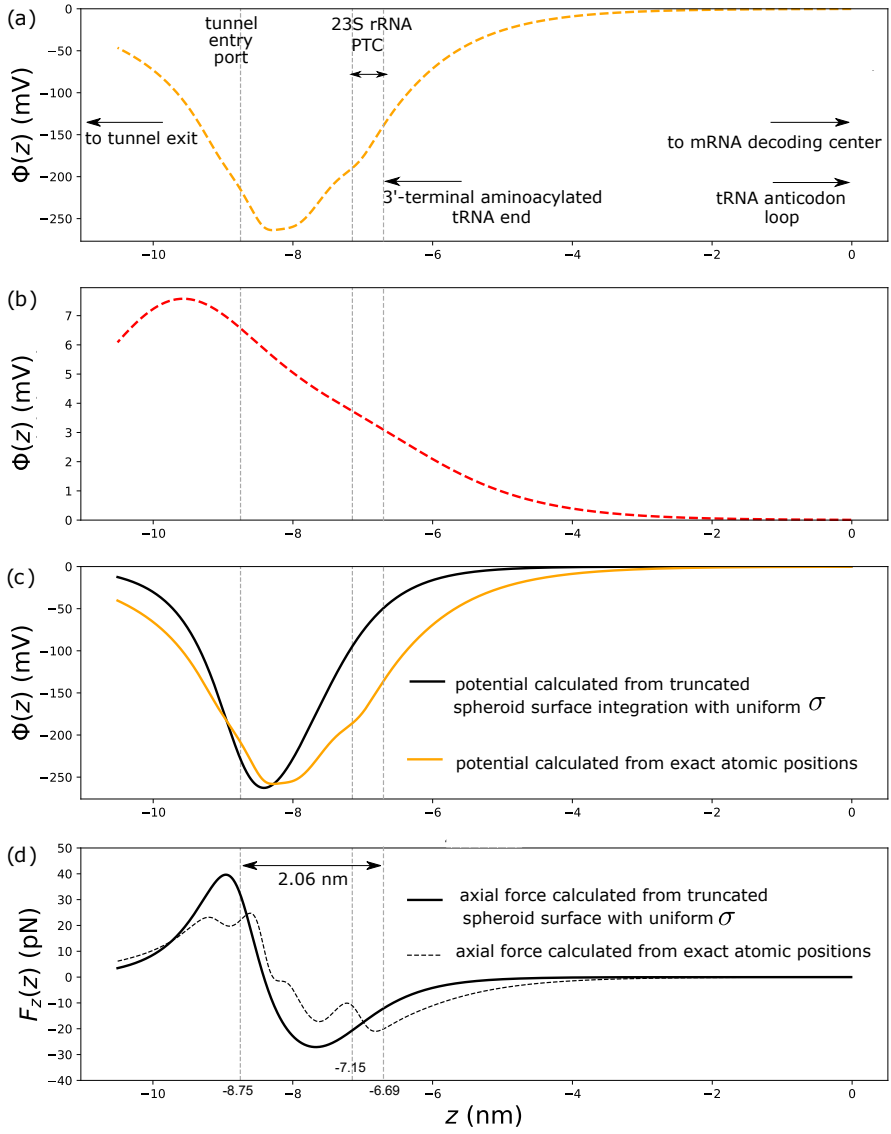


Figure 4.11: Electrostatic potential profiles contributed by the ribosomal large subunit cavity around the PTC. (a) Electrostatic potential (mV) computed from the exact atomic positions of the 100 phosphorus within 12 Å of the surface cavity.

4.6.6 Electric field estimation in the vicinity of the PTC

The ratio of the *bare* surface charge density over the medium permittivity σ^*/ϵ has the dimension of an electric field. In the absence of a screening layer of water molecules, the order of magnitude of this ratio in the immediate vicinity of the PTC cavity surface is $\sigma^*/\epsilon = -\frac{6.35\,|e|}{78\cdot\epsilon_0} \sim 1.473 \cdot 10^9 \mathrm{V/m} = 14.73\,\mathrm{MV/cm}$, at least if the cavity medium is fully filled with water. If the medium in the PTC is only partially filled with water ($\epsilon = 78$), and the main components are ribonucleic acids (tRNAs) ($\epsilon = 8$) and the carboxy terminal end of the growing nascent protein ($\epsilon = 4$), than the resulting medium coarse-grained permittivity would be $\epsilon \sim 8$ (for a medium composed of 5% water and 95% protein). In this latter case, the estimated electric field, or the ratio $\sigma^*/\epsilon = -\frac{6.35\,|e|}{8\cdot\epsilon_0} \sim 14.36 \cdot 10^9 \mathrm{V/m} = 143.6\,\mathrm{MV/cm}$.

Using vibrational Stark effect spectroscopy, Fried, Boxer and coworkers measured the electric field in a typical enzyme-substrate configuration at the catalytic site of the enzyme ketosteroid isomerase (KSI), with a magnitude of $144 \pm 6\,\mathrm{MV/cm}$ [Fried et al. 2014; Fried and Boxer 2017]. Our numerical result shows that the order of magnitude of the electric field in the vicinity of the catalytic surface of the peptidyl transferase center of the ribosome is similar to that of the catalytic sites of known protein enzymes.

This intense electric field is possibly and locally electrostatically screened by constitutive water molecules of the PTC cavity inner wall. The reorientation of the permanent electric dipoles of constitutive water molecules on the inner surface of the PTC cavity would strongly damp the electrostatic potential and give rise to an apparent surface charge density $\sigma = \sigma^* \cdot e^{-\frac{\delta}{\xi_{GC}}}$ at a distance δ from the support of the *bare* charges, where $\xi_{GC} = 0.105\,\mathrm{nm}$ is the Gouy-Chapman length. The Gouy-Chapman length is the screening length used when dealing with charges distributed on a surface wall and when the screening is mainly due to water molecules (not mobile ions) [Joiret et al. 2022b; Roij 16 July 2009].

The apparent electric field along the z-axis as numerically calculated from the negative of the first derivative of Eq.(4.24)(Fig.4.11(d) black solid line) or of Eq.(4.27) (Fig.4.11(d) black dashed line), is approximately equal to $1.3\,\mathrm{MV/cm}$. The corresponding axial force experienced by a positively charged unit test probe would be $21 - 27\,\mathrm{pN}$ around the PTC region (Fig.4.11(d)).

The shape of the prolate spheroid with a uniformly charged inner surface results in a particular profile for the electrostatic potential and for the resultant electric field along the centerline of the cavity towards the tunnel entry point.

4.6.7 Impacts and functional significance of the electrostatic profile around the PTC on the peptide bond formation kinetics and on the complete elongation cycle

In higher eukaryotes, the protein elongation turnover is around 5.6 aa/s, i.e, the average elongation cycle has a time span of $\sim$ 178 ms. In bacteria and lower eukaryotes (e.g. yeasts) the elongation turnover is around 9.6 aa/s ($\sim$ 105 ms). Three sequential steps are involved in the protein elongation cycle by the ribosome: (1) accommodation and proofreading, (2) peptide bond formation and (3) translocation. The insert of Fig. 4.12 shows the relative time spans of each step in the complete elongation cycle in the case of higher eukaryotes.

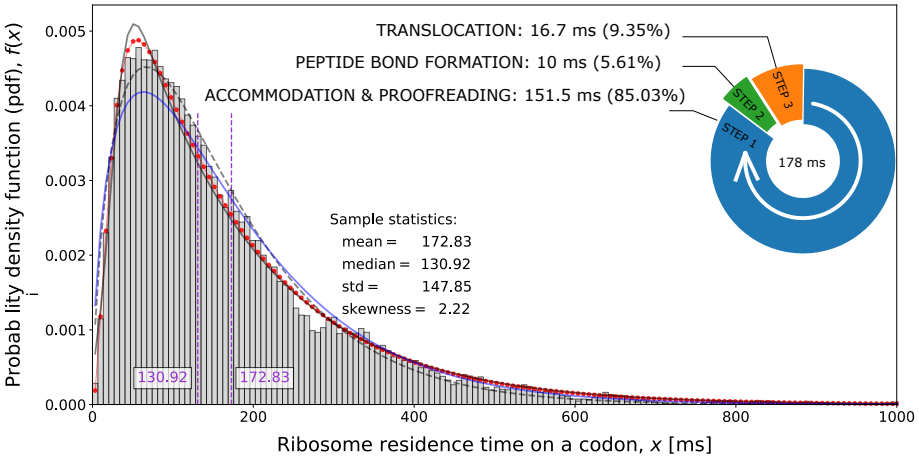


Figure 4.12: Ribosome elongation cycle simulation in three sequential time steps. Hypo-exponential distribution with $1/\lambda_i = (151.5, 10.0, 16.7)$: red dots. Grey bars: frequency distribution sampled from the hypo-exponential distribution. Exponentially modified Gaussian distribution least sq. fit to hypo-exponential: grey line. Gamma distribution fit to frequency distribution, least sq.: dashed line; max. likelihood estimation: blue line.

The total time spent by a ribosome on a given codon results from the queueing time of these three steps. From a theoretical perspective, the ribosome residence time on an arbitrary codon is a stochastic process that can be viewed as a random variable which is the sum of three elementary independent random variables of the time spent in each step. The mathematical developments of this multi-step queueing theory in the applied fields of probability and stochastic modeling are detailed in appendix A.1. It is important to keep in mind this sequence of three time steps in the elongation cycle and the fact that the electrostatic interaction only affects the rate of `STEP 2` directly. Indeed, the global rate limiting step is `STEP 1`, i.e.,the aa-`tRNA` A-site accommodation and proofreading at the `mRNA` decoding center, which occurs at the `16SrRNA-23SrRNA` interface, at a distance ~ 80 Å away from the PTC. This first step is not dependent on the electrostatic potential around the PTC. The last step, `STEP 3`, i.e., the translocation to the next codon, is not dependent on the electrostatic potential around the PTC either. This explains why the impact of electrostatic interaction on the complete elongation cycle should not be overstated.

In the previous sections, we calculated the electrostatic potential profile around the PTC and from there the electric field and axial force acting on a charged test probe. Specializing on the *Escherichia coli* large ribosomal subunit 50S, the axial force at the PTC is estimated to be 21.2 pN, for a charged probe amino acid residue (Fig. 4.8 (c) and 4.11 (d)). This force keeps acting for the short time during the displacement $d\vec{\mathbf{x}}$ along the curvilinear path followed by a charged probe amino acid at transpeptidation. The dot product of the force by the probe small displacement yields the mechanical work upon curvilinear integration, $W = \int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}$. This mechanical work is exerted by the surrounding electric field. The Gibbs free activation energy of the peptide bond formation is modulated by this mechanical work according to equations derived in section 4.5.4. The quantitative values obtained for the axial force are used to estimate the mechanical work and to calculate theoretically the impact on the peptide bond reaction rate. Any axial force acting on the nascent chain backbone will have similar effects. The electrostatic interaction of the nascent chain with the ribosome exit tunnel can also be estimated in the same way [Joiret et al. 2022b]. We compare these theoretical calculations to two independent sets of experimental results published previously by others [Dana 2014; Wohlgemuth et al. 2008]. In the next subsection, we examine the kinetics of the peptide bond formation between a dipeptide-tRNA and puromycin in *E.coli* [Wohlgemuth et al. 2008]. In the subsection after that, we examine the significance of the transpeptidation electrostatic modulation on the dwelling time distribution of the ribosome during the elongation cycle upon specific codons in *Saccharomyces cerevisiae* [Dana 2014].

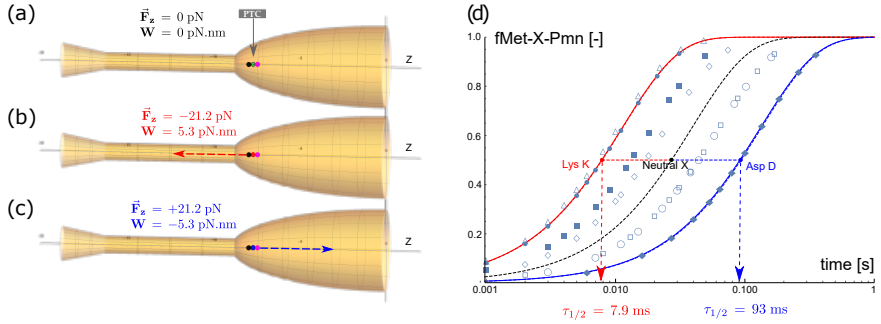


Figure 4.13: Elongation minimal case: effect of force on the rate of peptide bond formation when the P-substrate is a dipeptidyl-tRNA (fMet-X-tRNA). Electrostatic force acting on X at the P site: **(a)** X=Neutral amino acid (green sphere). **(b)** X=Positively charged amino acid K or R (red sphere). **(c)** X=Negatively charged amino acid D or E (blue sphere). Methionine (black sphere). Puromycin (magenta sphere). **(d)** Experimentally measured normalized time courses of the Pmn (20mM) reaction with different dipeptidyl-tRNAs: Lys(△), Arg(●), Ala(■), Ser(◇), Phe(◯), Val(□), Asp(♦) adapted from reference [Wohlgemuth et al. 2008]. Calculated theoretical normalized time courses using Maxwell-Boltzmann factors from equations (4.34) and (4.39). Neutral amino acid ($\tau_{1/2}$ = 27.1 ms), dashed line. Positively charged amino acid C-terminal dipeptidyl-transfer rate: neutral rate × $\exp(\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}/k_B T)$ = neutral rate × $\exp(5.3/4.28)$ = neutral rate × 3.45, red line. Negatively charged amino acid dipeptidyl-transfer rate: neutral rate × $\exp(\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}/k_B T)$ = neutral rate × $\exp(-5.3/4.28)$ = neutral rate × 0.29, blue line.

4.6.7.1 Kinetics experimental results for the transpeptidation minimal case with dipeptide-tRNA and puromycin in *E.coli*

The rate constants of the Pmn reaction with $\mathrm{fMet-X-tRNA^X}$ are tabulated by decreasing order in Table 4.6.1 reproduced and adapted from reference [Wohlgemuth et al. 2008].

In the particular elongation minimal cases studied by Rodnina and coworkers [ibid.], the peptidyl-tRNA is a dipeptidyl-tRNA with a very limited length that does not even reach the ribosome exit tunnel entry point (Fig. 4.13 (a-c)). When the X amino acid is neutral, no force is exerted at the P-site ($\vec{\mathbf{F}}_\mathbf{z}$ = 0 pN). On the contrary, if the X amino acid residue is a positively charged arginine R, or lysine K, a pulling force ($\vec{\mathbf{F}}_\mathbf{z}$ = −21.2 pN) is exerted on the backbone of the dipeptidyl-tRNA at the P site, while if the X amino acid residue is a negatively charged aspartate D, or glutamate E, a pushing force ($\vec{\mathbf{F}}_\mathbf{z}$ = +21.2 pN) is exerted on the backbone of the dipeptidyl-tRNA

during the peptide bond formation when the A-site amino-group of puromycin comes close to the carbonyl carbon of the ester bond in the dipetidyl-tRNA at the P-site. In the former case (X = R or K), the force points toward the exit tunnel (negative $z-$ axis), whereas in the latter case (X = D or E), the force points toward the PTC (positive $z-$ axis). Upon elongation with puromycin as the last acceptor substrate (substrate 2), the displacement upon transpeptidation to the A-site is $d\vec{\mathbf{z}} \sim 0.25$ nm (median distance between two C$-\alpha$ in a peptide bond) and the mechanical work is estimated to be $W = 21.2 \cdot 0.25 = +5.3$ pN $\cdot$ nm for R or K (force and displacement are parallel) and $W = -5.3$ pN $\cdot$ nm for D or E (force and displacement are antiparallel). Using the final equation for the peptide bond formation kinetics as derived in equation (4.39) and using the quantitative values for the mechanical work in the two charged cases as compared to the neutral residue case, we predict the Maxwell-Boltzmann factors and the reaction rate constant values tabulated in Table 4.6.2. The time courses of the dipeptidyl-tRNA reaction with Pmn are shown in Figure 4.13 (d). The red and blue lines are the time course calculated from equation (4.39) for the positively and negatively charged case respectively as compared to the neutral case (dashed line). These theoretical results are qualitatively and quantitatively consistent with the experimental values obtained by Rodnina and coworkers [ibid.]. This provides indirect evidence of the forces caused by the electrostatic interaction of charged amino residues in the peptidyl-tRNA with the negatively charged phosphate moieties lining the inner surface of the ribosomal RNA

Table 4.6.1: Experimentally measured dipeptityl transfer rate constants (k_{pep}), apparent affinities of Pmn binding ($K_{\frac{1}{2}}$) and waiting time ($\tau_{1/2}$) to peptide bond formation event with a probability of 0.5; reproduced and adapted from reference [Wohlgemuth et al. 2008].

P-site dipeptide substrate	k_{pep} (s^{-1})	$K_{\frac{1}{2}}$ (mM)	$\tau_{\frac{1}{2}}$ (ms)	charge
$fMet-Lys-tRNA^{Lys}$	100 ± 7	14 ± 3	7.2	(+)
$fMet-Arg-tRNA^{Arg}$	90 ± 7	6 ± 3	7.8	(+)
$fMet-Ala-tRNA^{Ala}$	57 ± 4	35 ± 4	13.2	neutral
$fMet-Ser-tRNA^{Ser}$	44 ± 2	30 ± 3	17.0	neutral
$fMet-Phe-tRNA^{Phe}$	16 ± 1	4 ± 1	43.8	neutral
$fMet-Val-tRNA^{Val}$	16 ± 1	6 ± 1	44	neutral
$fMet-Asp-tRNA^{Asp}$	8 ± 1	22 ± 2	91.5	(−)
$fMet-Pro-tRNA^{Pro}$	0.14 ± 0.02	12 ± 4	5,102.2	neutral

around the PTC. These results also support the functional impact of these forces on the kinetics of the peptide bond formation.

Table 4.6.2: Maxwell-Boltzmann factors $\exp(\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}/k_B T)$ modulating the dipeptidyl transfer rate constants $k(\vec{\mathbf{F}})$ and the waiting time $\tau_{1/2}$ to peptide bond formation event with a probability of 0.5. Maxwell-Boltzmann factors and reaction rate constants are calculated from equation (4.34) or (4.39) at $T = 310.15\,K$, $k_B T = 4.282$ pN.nm. Acceptor substrate at A-site is puromycin.

P-site dipeptide substrate and C-terminal charge at P-site	Mechanical work (pN.nm)	Maxwell Boltzmann factor (-)	Rate constant $k(\vec{\mathbf{F}})$ (s^{-1})	$\tau_{1/2}$ (ms)
$fMet \odot -tRNA\,(0)$	0.0	1	$k(\mathbf{0})$	27.1
$fMet \oplus -tRNA\,(+1)$	+5.3	3.45	$3.45 \cdot k(\mathbf{0})$	7.9
$fMet \ominus -tRNA\,(-1)$	−5.3	0.29	$0.29 \cdot k(\mathbf{0})$	93

4.6.7.2 Comparison of ribosome residence time empirical distribution results for the complete elongation cycle at specific codons in yeasts and bacteria coding for positively and negatively charged amino acids

In what follows, the elongation turnover for *S. cerevisiae* was taken as 9.5 aa/s. [Shah et al. 2013]. The mean elongation time is $\text{mean}(T) = 105.3$ ms and the probability space for the ribosome residence time on a codon is expressed in millisecond units of time, ms.

The elongation turnover for *E.coli* was taken as 16 aa/s, i.e., the mean elongation time is $\text{mean}(T) = 62.5$ ms at 37°C [Nieß et al. 2019]. Note that the elongation turnover drops to 7.5 aa/s at 28°C in *E.coli* [Farewell and Neidhardt 1998].

It is worth recalling that one should refrain from drawing conclusions on absolute times from Ribo-Seq derived data such as the normalized footprint counts (NFC) distributions, considering the definition equation (4.42) of the NFC. Comparisons of ribosome residence times (RRTs) between codons in terms of time fold changes are meaningful, but not comparisons in terms of absolute times.

The workflow detailed in section 4.5.5.2 provided the following results for the comparison of the deconvoluted elongation cycle for specific pairs of amino acid and

their codons. In the four vertical panels of Fig. 4.14 and Fig. 4.15, pairwise comparison of the RRT distributions for positive versus negative amino acids deciphering codons are conducted. The first row plot in each panel shows the red and blue curve for the positive and negative amino acid respectively: Fig 4.14(a(i), b(i)) and Fig. 4.15(a(i), b(i)). The second and third row plots shows, for the positively (Fig 4.14 and 4.15 (ii)) respectively the negatively (Fig 4.14 and 4.15 (iii)) charged amino acid, the Levenberg-Marquardt least squares fitting of the observed empirical exponentially modified Gaussian (full line) with the hypo-exponential distribution (dotted line). The inserts of the second and third row plots in each panel display the inferred individual rates $\lambda_1, \lambda_2, \lambda_3$ of the three exponentials as deconvoluted components of the hypo-exponentials, plus the time shift Δ. The quality of the fit, or similarity of the two statistical models, is assessed by the Kullback-Leibler (KL) divergence distance (relative entropy). A KL divergence distance close to zero indicates that the two models are very similar and convey the same information. For all eight best least squares fitting, the Kullback-Leibler divergence measures were smaller than 0.013, i.e, $KL(\text{EMG} \,||\, \text{HYPO}) < 0.013$.

A first broad inspection of Fig. 4.14 and Fig. 4.15 immediately shows that the ribosome on average spends significantly more time on average on triplets coding for glutamate or aspartate, i.e., the negatively charged amino acids, than for triplets coding for arginine or lysine *in S.cerevisae*, i.e., the positively charged amino acids.

The educated deconvolution of the hypo-exponential distribution least square fitting into its three independent exponential components, plus time shift, provides an estimated breakdown of the three elongation sub-steps as shown on Fig. 4.14 a(ii), a(iii), b(ii), b(iii) and Fig. 4.15 a(ii), a(iii), b(ii), b(iii) and summed up in table 4.6.3.

The smallest value of the λ_i, or largest of $1/\lambda_i$, is interpreted as the rate limiting step in the elongation cycle, i.e. the rate of accommodation and proofreading, `STEP 1`, with the other two λ values corresponding to `STEP 2` and `STEP 3`. Note that the time shift Δ can be interpreted as a time lag being part of `STEP 3`.

The ratios of λ_2 are calculated for each pair of positively charged versus negatively charged amino acids in *Saccharomyces cerevisiae*.

Overall, the educated deconvolution results show that the ratio of the fast rate (for + amino acid) over the slow rate (for - amino acid) of STEP 2 is ~ 11.90. The inferred fold change as compared to the median rate for a neutral amino acid is $\sqrt{11.90} = 3.45$. The Maxwell-Boltzmann factor is quantitatively close to 3.45 (at 37 Celsius) and is consistent with our derivation of section 4.6.7.1.

Table 4.6.3: Ribosome residence time hypo-exponential distribution deconvolution into its exponential rate components and time steps breakdown of the elongation cycle for codons deciphering positive and negative amino acids (λ, ms^{-1}; $1/\lambda$, Δ and RRT, ms).

				STEP 1 ACCOMMODATION & PROOFREADING		**STEP 2** PEPTIDE BOND FORMATION		**STEP 3** E-SITE EVICTION & TRANSLOCATION			ELONGATION CYCLE
species	charge	amino acid	codon	λ_1	$1/\lambda_1$	λ_2	$1/\lambda_2$	λ_3	$1/\lambda_3$	Δ	RRT
S. cerevisiae	(+)	**Arg R**	**CGC**	0.01619	61.76	0.92	1.09	0.10178	9.82	13.95	86.6
S. cerevisiae	(−)	**Glu E**	**GAG**	0.01076	92.9	0.077	12.90	0.33561	3.0	12.6	121.4
	λ_2 rate fold change for + over − =				$\frac{\lambda_2^+}{\lambda_2^-}$ =	11.90					
S. cerevisiae	(+)	**Arg R**	**CGU**	0.01602	62.41	0.92	1.09	0.08494	11.77	5.79	81.1
S. cerevisiae	(−)	**Asp D**	**GAU**	0.01213	82.5	0.077	12.90	0.23087	4.30	7.30	107.1
	λ_2 rate fold change for + over − =				$\frac{\lambda_2^+}{\lambda_2^-}$ =	11.90					
S. cerevisiae	(+)	**Lys K**	**AAG**	0.01092	91.62	0.92	1.09	0.08171	12.24	6.82	111.8
S. cerevisiae	(−)	**Glu E**	**GAA**	0.01017	98.4	0.077	12.90	0.19955	5.0	6.5	122.9
	λ_2 rate fold change for + over − =				$\frac{\lambda_2^+}{\lambda_2^-}$ =	11.90					
E. coli	(+)	**Lys K**	**AAG**	0.03583	27.91	10.82	0.09	2.51917	0.40	0.31	28.7
E. coli	(−)	**Glu E**	**GAG**	0.03562	28.1	0.91	1.1	21.30965	0.0	0.1	29.3
	λ_2 rate fold change for + over − =				$\frac{\lambda_2^+}{\lambda_2^-}$ =	11.90					

Cautionary note: comparisons between codons in terms of rate fold changes or time fold changes are meaningful but not comparisons in terms of absolute times.

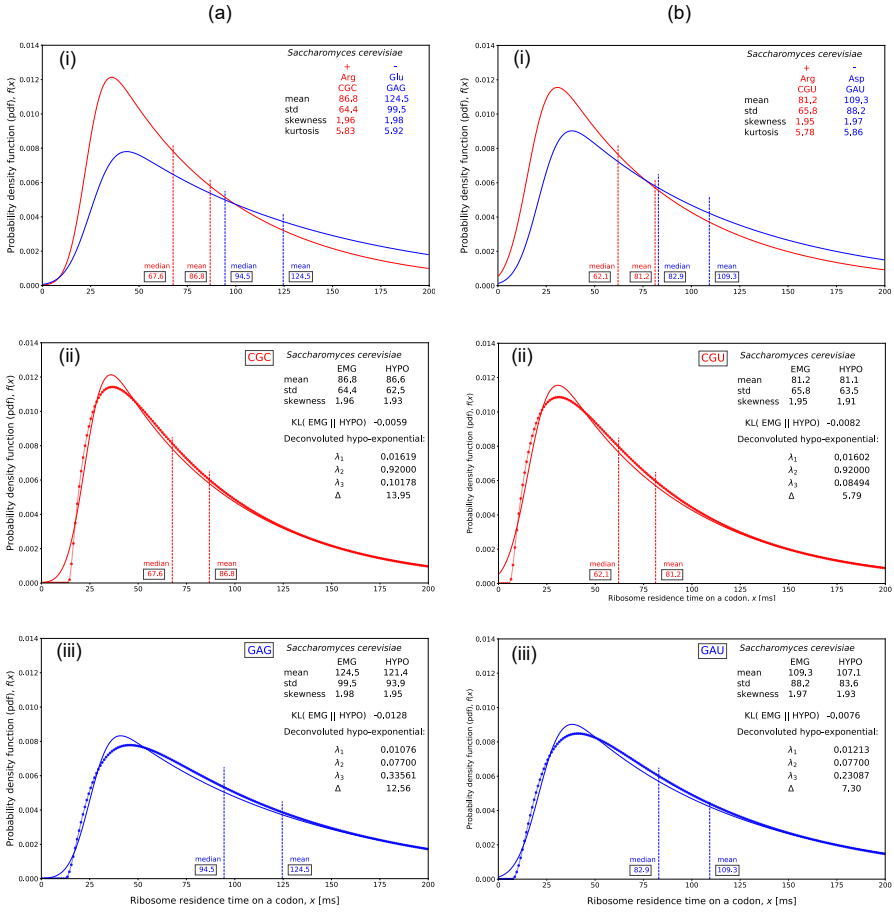


Figure 4.14: Ribosome residence time distribution on specific codons in *Saccharomyces cerivisiae*. a,b-(i) Exponentially modified Gaussian empirical distributions as experimental reference comparing (+) red and (-) blue lines aa encoding codons. Data from [Dana 2014]. a,b-(ii) Least squares fitting with shifted hypo-exponential distribution for the (+) aa, red dotted line. a,b-(iii) Least squares fitting with shifted hypo-exponential distribution for the (-) aa, blue dotted line.

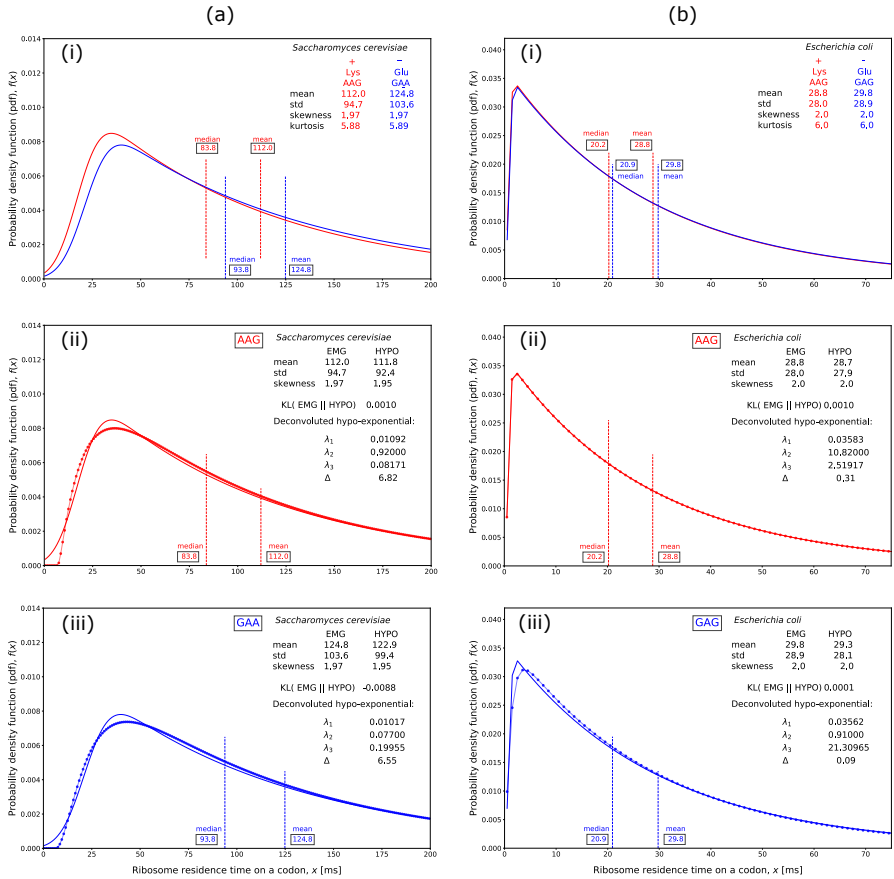


Figure 4.15: Ribosome residence time distribution on specific codons in *Saccharomyces cerivisiae* and *Esherichia coli*. a,b-(i) Exponentially modified Gaussian empirical distributions as experimental reference comparing (+) red and (-) blue lines aa encoding codons. Data from [Dana 2014]. a,b-(ii) Least squares fitting with shifted hypo-exponential distribution for the (+) aa, red dotted line. a,b-(iii) Least squares fitting with shifted hypo-exponential distribution for the (-) aa, blue dotted line.

4.7 Discussion

We studied the electrostatic environment around the catalytic center of the ribosome by using the Yukawa-Debye-Hückel theory applied to structural data from 5 publicly available x-ray solved structures of the ribosome across the three domains of life. The salient feature of the catalytic center of the ribosomal large subunit is that it is made only of nucleic acids (ribozymes). This is uncommon as the catalytic activity in biochemical processes are carried out mostly by proteins (enzymes). By contrast to proteins, nucleic acids have a molecular backbone harboring a very large number of phosphate moieties. Our study results suggest that the exact tridimensional distribution of these phosphate moieties has a functional role in the peptide bond formation and can affect its rate. The comparison of 5 species across the 3 domains of life of the ribosome catalytic center shell cavity shows that a common feature lies in the 3D shape of the distribution of non-adjacent 484 phosphate moieties within the polynucleotide backbone of 23S/28S rRNA at a 40 Å distance from the peptidyl transfer center (PTC). The tridimensional phosphate moieties distribution in the immediate vicinity of the PTC appears to be the main contributing structure that determines the electrostatic profile around the catalytic center. A simple truncated prolate spheroid shape fits the catalytic center of the ribosome. This funnel shape feature and its surface charge density generated by the fixed phosphate moieties explains the decrease in the electrostatic potential when a test charged probe moves from the A-loop to the P-loop and towards the tunnel entry port. To quantitatively determine the electrostatic potential profile and the electric field, the exact path followed by a newly incorporated amino acid should be known. To this date, we do not have this information. For simplicity, the charged amino acid residues participating in the reaction were approximated by immobile point charges, apparently equivalent to the positions of the α-atoms of the corresponding residues. However, in reality, these charges sit on the tip of rather large and highly mobile aliphatic chains. Lysine (K) and arginine (R) are notoriously large entities ($\sim$ 11 Å), and even aspartate (D) and glutamate (E) are by no means small (6.5 and 8 Å, respectively, in a fully extended conformation). Their aliphatic segments are largely unconstrained, and rotation around the C-C bonds occurs with sub-nanosecond rates, which means that during the relatively slow transpeptidation, the charges carried on the tip of these amino acids can in principle sample a large volume around them. This volume is actually comparable to that of the entire PTC cavity (the distance between the PTC and the tunnel entry port is $\sim$ 2nm, and the amino acid displacement upon transpeptidation is only 0.25nm). If these charges are indeed so mobile (i.e. their exact position is subject to high uncertainty), then, at this spatiotemporal scale, it might be difficult, by using the proposed framework, to measure the exact mechanical forces they experience. This mobility may even dampen the effect of the electrostatic field on the reaction energy barrier. In the best case scenario, since the conformational changes of the side chains must be quite rapid, one could envisage that arginine and lysine, directed by the field, would be relatively immobile and stretched in the direction of

the tunnel entry port, bringing their positive charges 1 nm further than expected from the α-carbon approximation. Based on Fig. 4.11(d), this would mean that they would experience an even stronger axial electrostatic force (closer to 30 pN). By contrast, aspartate and glutamate, stretched in the opposite direction, would suffer a lesser drag than expected by the theory. This side chain mobility is difficult to incorporate in the current model and is a limitation of the proposed coarse-grained framework.

The relative permittivity prevailing in the medium around the PTC is not known with accuracy and is not homogeneous. The medium inside the PTC cavity is more aqueous than the medium in the more confined micro-environment of the ribosome exit tunnel. It was hypothesized in this study that a coarse-grained permittivity should be taken in the range corresponding to a mixture of protein, nucleic acid and water, i.e at least between $\epsilon_r = 8$ and $\epsilon_r = 78$. The uncertainty on the numerical value of the permittivity on an extended spatial domain mainly affects the y-axis scale of the electrostatic potential. This is also modulated by the screening length. Some authors only provide a so-called arbitrary scale for the electrostatic potential to avoid these uncertainties on the empirical parameters [Dao Duc et al. 2019; Wang et al. 2020].

Metal ions, Mg^{2+} and K^{+} specifically, have long been described in the literature to be key players in stabilizing nucleic acid secondary and tertiary structures and also in stabilizing the ribosomal subunits [Nierhaus 2014; Tirumalai et al. 2021]. For this reason, it is relevant to incorporate the positions of the metal ions when available in the structural data of the ribosome and in the vicinity of the PTC. Unfortunately, methodological shortcomings in the metal ions assignments and counts in the large ribosome deposited x-ray solved structures put the research community in a quandary [Auffinger et al. 2021]. Due to the elusive distinction between bivalent Mg^{2+}, monovalent K^{+} or even oxygen of water molecules, the charge equivalents (valence) attributed to the metal ions is uncertain if the exact positions of these metal ions are used in the Yukawa-Debye-Hückel theory as source charges of the electrostatic potential and field [Rozov et al. 2019]. The convenient way to deal with this metal ions assignment uncertainty in the field is to keep the metal ions, Mg^{2+} and K^{+} free mobile ions indistinctively, as part of the empirical screening length parameter ξ generally adopted in the literature for the Yukawa-Debye-Hückel formula to calculate the electrostatic potential, without using the exact positions of these metal ions. The comparison of solved structures of ribosomes, around the PTC cavity for different species, shows that there is heterogeneity in the distribution of the metal ions across different x-ray or cryo-EM structural data from different species or PDB entries [Rozov et al. 2019; Wang et al. 2020]. Given that, the assumption that the metal ions can be considered as free mobile ions screening the phosphate moieties fixed charges is reliable, at least around the PTC or around the tRNA molecules at the P and A sites.

Our study is based on x-ray or cryo-EM solved structures of the ribosome. These high resolution solved structures inherently are static data. In this static framework, the conformational rearrangements and rotation motions that occur during elongation,

such as tRNAs transitioning to hybrid states and ribosome ratcheting were not taken into account. Dynamic behavior of rRNA nucleotides was also ignored. How these dynamic rearrangements could alter the electrostatic environment around the PTC is left for later molecular dynamics studies.

The present work does not explain the peptide bond formation rate dependence on the nature of the substrate at the A-site. Both pair of amino acids at the P and A-sites in the PTC shell cavity must play a role in the peptide bond formation rate. This has been partially covered by other studies. Organic chemistry steric hindrance, pH-sensitivity, water trap and entropy trap explain why different A site substrates incorporate at different rates in the PTC [Johansson et al. 2011; Sievers et al. 2004; Wallin and Aqvist 2010]. The A-site substrates context was also investigated in references [Melnikov et al. 2016; Pavlov et al. 2009; Pavlov et al. 2021]. Some amino acids like proline even require the help of specific elongation factors, e.g., EF-P or eIF5A [Doerfel et al. 2012; Doerfel et al. 2015; Peil et al. 2013; Starosta et al. 2014]. Our model implication focuses on how the electrostatic environment around the immediate vicinity of the PTC (within a distance less than $5-10$Å) has its most significant impact in the energy barrier to overcome the deacylation of the peptidyl-tRNA at the P-site and this particularly appears when charged amino acids are at the carboxy-terminal end of the peptidyl-tRNA at the P-site.

The contribution of the tRNA molecules themselves at the P and A sites to the electrostatic environment in the catalytic shell cavity of the PTC was also considered. The analysis of structural data of the x-ray solved large ribosomal subunit including tRNA molecules showed that except for the last nucleotide at the $3'-$end of the tRNAs, all 76 phosphate moieties belonging to each of the tRNA at P and A sites are much further away from the PTC cavity than the Debye screening length. The electrostatic potential calculated contribution from the two tRNAs is much smaller than the calculated contribution from the 484-488 nucleotides of the 23S/28S rRNA in the vicinity of the PTC. The electrostatic field contributed by the two tRNAs fixed phosphate moieties is negligible in the vicinity of the PTC.

The significance and functional consequences of the electrostatic potential profile around the catalytic center are affecting the kinetics of the protein elongation rate. Following the literature in mechano-bio-chemistry, we have applied the classical Eyring theory of catalysis and incorporated a modulation of the Gibbs free energy activation energy barrier by the mechanical work of physical forces acting on at least one substrate. We hypothesized that the physical forces transmitted mainly through the backbone of the peptidyl-tRNA play a role in the reduction of the Gibbs free energy barrier of the transition state. During the peptidyl transfer reaction, the α-amino group of aminoacyl-tRNA positioned in the A site of the ribosome nucleophilically attacks the carbonyl carbon at the ester bond of the peptidyl-tRNA in the P site, which results in peptidyl-tRNA extended by one amino acid in the A site and deacylated tRNA in the P-site. Our analysis advocates that the biochemical deacylation of the peptidyl-tRNA at

the P site appears to be facilitated (or hampered) by a pulling (pushing) force exerted on the backbone of oligopeptide attached at the $3'-$tip of tRNA at the P-site. If the carboxy-terminal end of the peptidyl-tRNA is a positively charged amino acid, a pulling force of electrostatic origin reduces the activation energy while if the carboxy-terminal end of the peptidyl-RNA is a negatively charged amino acid, a pushing force increases the activation energy. Overall, the 23S/28S rRNA catalytic shell cavity of the PTC functionally works like some kind of electrostatic bottle cap opener where the bottle part is the tRNA at the P-site and the cap is the carboxy-terminal end of the nascent peptide chain. The opening itself would be the deacylation of the peptidyl-tRNA at the P-site during the peptide bond formation. The development of the Michaelis-Menten kinetics incorporating the mechanochemical effect of the forces acting at the peptidyl transfer center led to theoretical quantitative predictions in the relative rates of peptide bond formation when comparing lysine or arginine to glutamate or aspartate as substrates at the carboxy-terminal end of the P-site tRNA. The consistency of these predictions with previously published experimental data sheds light on the significance and functional consequence of the electrostatic profile around the PTC, on the kinetics of the peptide bond formation and on its dependence on the nature of the second last incorporated amino acid residues. The peptide bond formation is the second step of the protein elongation cycle in the ribosome and the only one affected by the electrostatic interaction around the PTC shell cavity. We claimed that the global residence time of the ribosome on a given codon is a stochastic process resulting from the sum of three independent elementary queueing times, each of which being exponentially distributed. We relied on the queueing time theory in probability and statistics and used the convolution product to show that the ribosome residence time on a given codon should be hypo-exponentially distributed. The hypo-exponential distribution is the distribution best describing the sequential sub-steps involved in the elongation process. In the previous literature in the field, the exponentially modified Gaussian distribution and the Gamma distribution have been used as stochastic models for the RRT. In practice, we showed that the information conveyed by these distributions, HYPO, EMG or GAMMA, is very similar. These three distributions are statistically equivalent from the perspective of the information theory. However, the HYPO, when deconvoluted into its three exponential components has the advantage to provide the temporal breakdown of the sub-steps occurring during the elongation cycle. The educated deconvolution of the hypo-exponential distribution allows to use Ribo-Seq data or normalized footprint counts to infer kinetics information on the three sub-steps of elongation. To the knowledge of the authors, this is the first time that Ribo-Seq data were interpreted using a deconvolution of the hypo-exponential distributions that were previously fit to the ribosome normalized footprint count on codons.

4.8 Concluding remarks and future perspectives

One of the expected future developments in the field of protein synthesis and translational control will be in biochemical kinetics and dynamics. In this study, the structural data from x-ray crystallography was used to provide an inherently static picture of the peptide bond catalytic center. Molecular dynamics studies are expected to provide a better understanding of the dynamical interactions between key nucleotides of the P-loop and A-loop and the amino acid substrates loaded on the tRNAs at the P and A sites. Experimental studies using Stark effect spectroscopy, Förster resonance energy transfer and optical tweezers have been used and will continue to be used to probe the details of the elongation cycle and especially of the peptide bond formation dynamics in the ribosome in different contexts. To date, Ribo-Seq experimental results are difficult to interpret. The research community will greatly benefit from the development of computational biology models that will be able to generate simulated synthetic data under controlled settings. The comparison of data patterns between the simulated ribosome footprints on arbitrary transcripts and their real experimentally observed ribosome footprints (Ribo-Seq) will eventually help to disentangle the complex factors that are modulating the elongation cycle rates. The statistical queueing theory and the convolution product of the probability distribution for the queueing times of the sequential steps involved in the elongation cycle developed in this study provide fundamental insights in the stochastic behavior of the ribosome when a mRNA is translated *in singulo*. Our proposed interpretation of the normalized ribosome footprint profiles at codon resolution as an hypo-exponential distribution will improve data mining and parameter learning of existing Ribo-Seq big repositories like the Sequence Read Archive (SRA) and the European Read Archive (ENA). Future bioinformatics and machine learning studies trained on these repositories will help to gain better quantitative knowledge on the elongation cycle sub-steps and on the time spent by ribosomes on individual codons in different experimental settings and for species with different codon usage.

Chapter 5

tRNAs pool and tRNA modifications

This chapter deals with the key factor which is of central interest in our agent-based model: tRNAs and tRNA modifications. The chapter explains how codon usage and tRNA modifications impact on protein elongation and how the tRNA modifications are incorporated in the ABM. There are hundreds of known tRNA modifications but we will focus here in only two generic enzymatic modifications targeting nucleotide 34 in the anticodon stem loop of tRNAs. From a modeling perspective, tRNA modifications can be used to perturb protein elongation, adjust parameters of TASEP agent-based models, and qualitatively test the model's `Ribo-Seq` density pattern predictions. Predefined modulating factors, calibrated from published metadata analyses, influence tRNA accommodation rates (step 1 of the elongation cycle) for specific targeted codons. At the end of a simulation, the algorithm generates ribosome density maps for all transcripts of interest, allowing for comparisons between a control (no tRNA modification) and a 'treatment' case (with tRNA modification). Predictions of protein abundance and translation efficiency can be compared to experimental results, as can changes in polysome fragmentation and ribosome footprint profiles, when corresponding experimental data are available. When such data are unavailable, the model serves as a valuable tool for synthetic or systems biology research. Additionally, we introduce the experimental datasets used for comparison with our model predictions, including data of interest to cancer research.

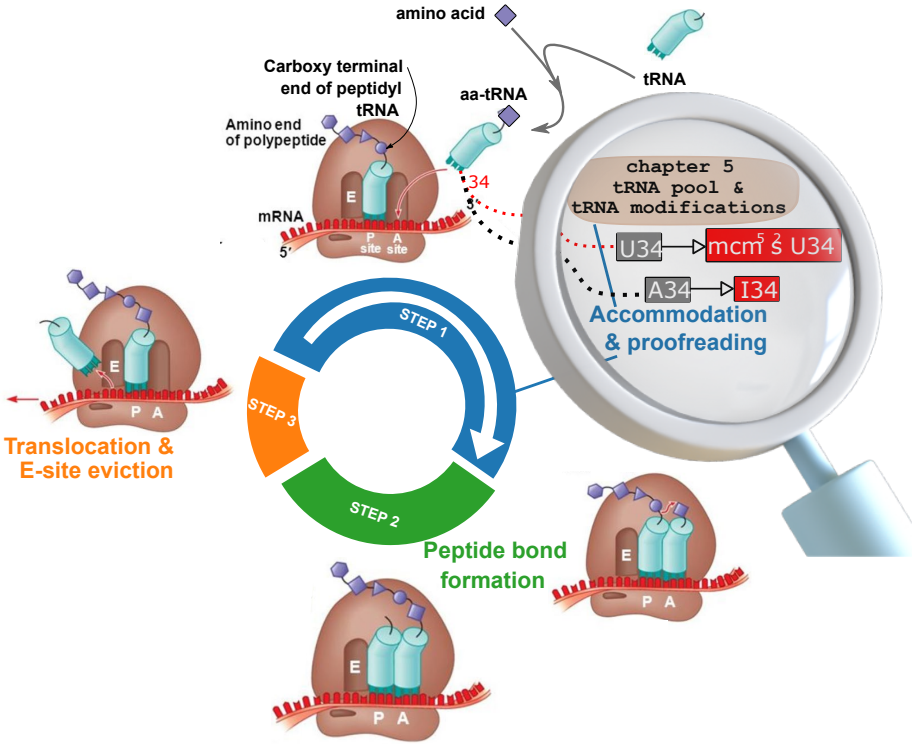


Figure 5.1: Graphical abstract of the chapter. This chapter depicts the impact of tRNA modifications at anticodon position 34 on the accommodation rates at the codon level. The effects of two tRNA modifications are presented: (i) the modification of U34 to 5-methoxycarbonyl-methyl-2-thio-uridine (mcm5s2U34), which directly affects three anti-codons complementary to AAA, GAA and CAA and, indirectly the decoding of 6 sensitive codons corresponding to 3 amino acids (lysine (K), glutamic acid (E), and glutamine (Q)); (ii) the modification of A34 to I34 by the enzyme adenosine deaminase acting on tRNA (ADAT), which impacts the accommodation and decoding rates of 37 codons corresponding to 8 amino acids (threonine (T), alanine (A), proline (P), serine (S), leucine (L), isoleucine (I), valine (V), and arginine (R)). The probability distribution of queueing times for STEP 1 is altered at each codon sensitive to these modifications, as is the total queueing time for the entire elongation cycle. We calibrated our agent-based model with these changes in the hypo-exponential distributions using data from published meta-analyses, representing a novel contribution to the field.

5.1 Transfer RNAs, tRNAs, key adaptor intermediates in protein synthesis

The transfer RNAs, **tRNAs**, are key players in the protein elongation by ribosomes. They are the adaptor molecules between the transcriptome and the proteome. In 1958, Francis Crick made this highly innovative hypothesis that eventually became an established fact: the adaptor in protein synthesis is **tRNA**. Holley, Khorana and Nirenberg later revealed the detailed compositions and structures of these adaptor molecules and cracked the genetic code, for which they were awarded the 1968 Nobel Prize in Physiology and Medicine. In a side publication to this PhD thesis, we showed that, in the proteomic and transcriptomic era, with the huge omics data currently publicly available, the genetic code can be cracked by a completely data driven approach. This approach implemented machine learning algorithms and artificial intelligence, using the generic capabilities of neural networks. The neural networks were trained on a dataset having thousands of pairs of transcripts and their proteins associated sequences [Joiret et al. 2023b].

In this first section, we recall the structure and chemical features of these remarkable adaptor molecules. These properties help comprehend the rules of the agent-based model we implemented to represent the interactions of codons with tRNAs and ribosomes during a protein elongation cycle.

A **tRNA** molecule contains an *amino acid attachment site* and a *template recognition site*, Figure 5.2 (a). A **tRNA** carries a specific amino acid in an activated form to the catalytic site of protein synthesis in the ribosome. The carboxyl group of this amino acid is esterified to the 3'- (or 2') hydroxyl group of the ribose unit at the 3' end of the tRNA chain, Figure 5.2 (b). The esterified amino acid may migrate between the 2' and the 3' hydroxyl group during the protein synthesis reaction. The joining of an amino acid to a tRNA to form an aminoacyl-tRNA is catalyzed by a specific enzyme called an *aminoacyl-tRNA synthetase* (or activating enzyme). This esterification is driven by ATP. Two high-energy phosphoryl groups of ATP molecules are consumed in the synthesis of an aminoacyl-tRNA. The ester bond between the amino acid and the tRNA is energy rich. It is important to note that this activation and enrichment in biochemical energy made in the aminoacylated-tRNA is conducted in the cytoplasm, outside of the ribosomes. But this biochemical energy will be released during the peptide bond formation at the catalytic site of the ribosome and will contribute to the energy budget required by the ribosome to proceed an elongation cycle. Indeed, the Gibbs free energy content in the ester bond is larger than the Gibbs free energy content of the peptide bond between two amino acids (the elongating amino-acid and the carboxy-terminal amino aid of the peptidyl-tRNA being elongated). There is at least one specific synthetase for each of the twenty amino acids.

The template recognition site on tRNA is a sequence of three bases called the *anticodon*. The anticodon on tRNA recognizes a complementary sequence of three bases on

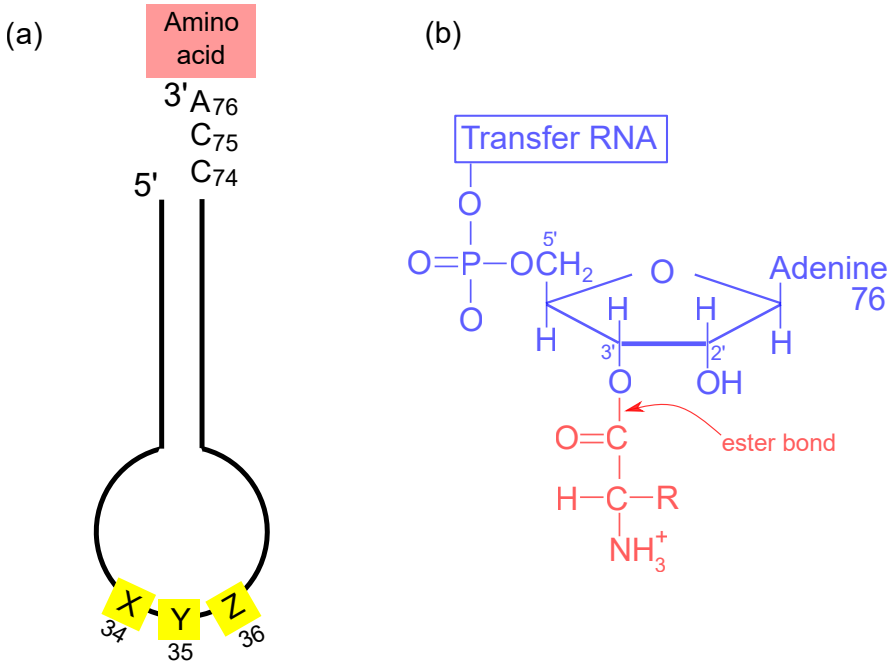


Figure 5.2: (a) Symbolic diagram of an aminoacyl-tRNA showing the amino acid attachment site (red) and the template recognition site (yellow), which is the anticodon. (b) Mode of attachment of an amino acid (red) to a tRNA molecule. The amino acid is esterified to the 3'-hydroxyl group of the terminal adenosine of tRNA. A tRNA having an attached amino acid is an aminoacyl-tRNA or a "charged" tRNA, whereas a tRNA without an attached amino acid is "uncharged" (adapted from [Berg et al. 2019]).

mRNA, called the *codon*. The fidelity and accuracy of the genetic code rely both on the specificity of the aminoacyl-tRNA synthetase (aa-synthetase) and on the template recognition (pairing) between the anticodon and the codon.

The tRNA molecules have a common design. tRNAs are obviously RNA molecules. The nucleotide sequence (primary sequence) of a transfer molecule was first determined by Holley in 1965. It was the first nucleic molecule ever sequenced and it took Holley seven years to do it. The sequence of yeast alanine tRNA is shown in Figure 5.3. Most tRNA molecules across the three domains of life have the same structure and entail a single chain of seventy-six ribonucleotides. The 5'-terminus is phosphorylated (pG), whereas the 3' terminus has a free hydroxyl group. A distinctive feature of tRNA molecules as compared to other nucleic acid molecules is in their high content of bases other than A, U, G, C. There are several unusual nucleosides. In the case of

alanine tRNA, the unusual nucleosides are inosine, pseudo-uridine, dihydrouridine, ribothymidine and methylated derivatives of guanosine and inosine. The amino acid attachment site is always located in the 3' terminus of the molecule. The 3'terminus of all tRNAs share a common sequence of CCA as the three last nucleotides of the primary sequence. In the middle of all tRNA molecules and almost always at nucleotide positions 34, 35, 36 lies the anticodon sequence. In the case of alanine tRNA in Figure 5.3, the anticodon sequence is ${}^{5'}_{34}\text{IGC}^{3'}_{36}$. It is complementary to the 5'GCC3' sequence, one of the codons for alanine. It is actually also complementary to two other synonymous codons for alanine, namely, 5'GCU3' and 5'GCG3'. So, this single tRNA molecule is able to recognize, on its own, three of the four synonymous codons of alanine. The rules that this template recognition comply with will be explained shortly.

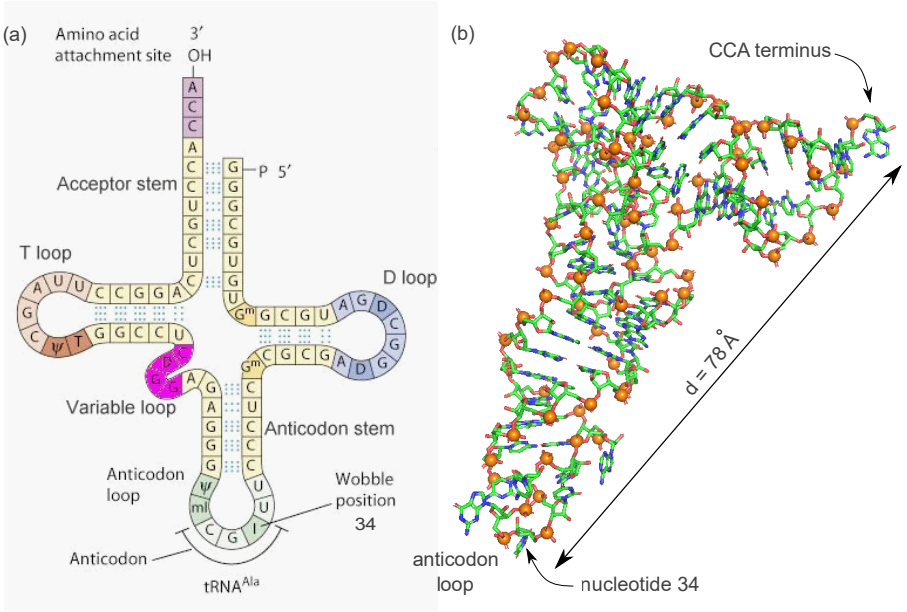


Figure 5.3: (a) Base sequence of yeast alanine tRNA. Modified nucleosides are abbreviated as follows: inosine (I), methylinosine (mI), dihydrouridine (D), ribothymidine (T), pseudouridine (ψ), methylguanosine (G^m), adapted from [Berg et al. 2019] (b) PyMol representation of a X-ray solved structure of yeast alanine tRNA at 4 Å resolution, from PDB entry code 8ASW downloaded from Protein Data Base repository. The representation shows the L-shape and helices in the anticodon stem and acceptor stem. The orange spheres represent the phosphorus atoms of all nucleotides.

All transfer RNA molecules share these features in common, see Figure 5.3:

1. tRNA molecules are single chains containing between 73 and 93 ribonucleotides (most observed 76) and about 25 kDal each.

2. mature tRNA molecules are stable, with half-lives around 2 to 3 days in eukaryotes [Berg and Brandl 2021].

3. tRNA molecules contain many unusual bases, typically between 7 and 15 per molecule. Many of these unusual bases are formed by post-transcriptional enzymatic modifications (see next section below). The roles of these unusual bases are diverse and uncertain. Possibilities among others are: helping the stability or proper folding of the molecule, taking part in specific interaction with the right aminoacyl-tRNA synthetase to achieve fidelity of the genetic code, taking part in specific interaction with the ribosome accommodation site, A-site, or peptidyl-tRNA, P-site.

4. The 5' end of tRNAs is phosphorylated. The 5'terminal residue is usually pG.

5. The base sequence at the 3' end of tRNAs is CCA. The activated amino acid is attached to the 3'-hydroxyl group of the terminal adenosine.

6. About half of the nucleotides in tRNAs are based paired to form double helices and build secondary structures within the molecule. Five groups of bases are not based paired: (i) the 3' CCA terminal region; (ii) the TψC loop, which acquired its name from the sequence ribothymine-pseudouracyl-cytosine; (iii) the extra arm which contains a variable number of residues; (iv) the DHU loop, which contains several dihydouracil residues; and (v) the anticodon loop.

7. The anticodon loop consists of seven bases, with the following well conserved sequence:

$$^{5'}\text{Pyrimidine} - \text{Pyrimidine} -_{34}\text{X} - \text{Y} -_{36}\text{Z} - \text{modified purine} - \text{variable base}^{3'}$$

8. The molecule is L-shaped.

9. The molecule has a clover-leaf secondary structure due to internal base pairing between about 10 base pairs in each of two segments of double helix. The helical segments are perpendicular to each other, which gives the molecule its L-shape.

10. The CCA terminus containing the amino acid attachment site is at one end of the L. The other end of the L, most distal from the CCA terminus, is the anticodon loop. Thus, the amino acid in aminoacyl-tRNA is far from the anticodon (about 78 Å). The DHU and TψC loops form the corner of the L. The fact that the distance between the CCA-terminus and the anticodon loop is 78 Å tells us that, within the large subunit of the ribosome, the mRNA decoding site and the catalytic site of the peptide bond formation are mutually separated by at

least the same distance. This was illustrated in chapter 4, in the x-ray solved structure of the LSU along with tRNA molecules both at the P- and A-sites [Joiret et al. 2023a]. Moreover, synthetic hybrid aminoacyl-tRNA experiments convincingly showed that the codon recognition does not depend on the amino acid that is attached to tRNA. The processes of codon recognition and peptide bond formation are physically separated and are also chronologically separated.

What are the rules that govern the recognition of a codon by the anticodon of a tRNA? The first simple hypothesis was that each of the bases of the codon forms a Watson-Crick type of base pair with a complementary base of the anticodon. The codon and anticodon would then be lined up in an antiparallel fashion. A specific prediction of this rule is that a particular anticodon can recognize only one codon. The facts are otherwise. Some pure tRNA molecules can recognize more than one codon (but these codons should be synonymous to ensure a consistent and unequivocal genetic code). We saw that the alanine **tRNA** studied by Holley with its anticodon ${}^{5'}_{34}\mathrm{IGC}^{3'}_{36}$ binds to three synonymous codons **GCU**, **GCC**, **GCA**. The first two bases of these codons are the same, whereas the third is different. Crick proposed that the third base of a codon is sometimes less discriminating than that of the first other two. The pattern of degeneracy of the genetic code indicates that it could be so. **XYU** and **XYC** always code for the same amino acid, whereas **XYA** and **XYG** usually (but not always) do. Crick surmised from these data that the steric criteria for pairing of the third base might be less stringent than for the other two. Inosine was included in the set of nucleotides whose pairing properties were to be determined because inosine appears in at least 8 anticodons of eukarya tRNAs. Assuming some steric freedom ("wobble") in the pairing of the third base of the codon, the combination shown in table 5.1.1 seemed a plausible rule for the allowed pairing between the first anticodon base and the third codon base.

The *wobble hypothesis* is now firmly established. The anticodons of tRNAs of known sequence bind to the codons predicted by this wobble base rule.

Table 5.1.1: Wobble base rule: allowed pairings at the third codon base with the first anticodon base.

First anticodon base	**Third codon base**
C	G
A	U
U	A or G
G	U or C
I	U, C or A

We saw in the case of alanine tRNA how inosine can pair with U, C or A. Phenylalanine tRNA which has the anticodon **GAA**, recognizes the codons **UUU** and **UUC** but not **UUA** or **UUG**. Thus G pairs with either U or C in the third position of the codon as predicted by the wobble rule hypothesis. Two consequences result from the wobble rule:

1. The first two bases of a codon pair in the standard Watson and Crick way. Recognition is precise. Hence, codons that differ in either of their first two bases must be recognized by different tRNAs. This is the case for the aminoacids with six synonymous codons. For example, both **UUA** and **CUA** code for leucine but are read by different tRNAs. The tRNAs which differ by their anticodon but carry the same amino acid at the 3' terminus attachment site are called tRNA *iso-acceptors* (isoacceptor). There are tRNAs which both have the same anticodon (and thus also carry the same amino acid) but which differ in the nucleotide sequence except the anticodon. These tRNAs are called iso-*decoders* (isodecoder).

2. The first base of an anticodon determines whether a particular tRNA molecule reads one, two or three kinds of codons: C or A (1 codon), U or G (2 codons), or I (3 codons). Thus, part of the degeneracy of the genetic code arises from imprecision (wobble) in the pairing of the third base of the codon. We see here, that inosine, mainly encountered in tRNAs of eukarya species, maximizes the number of codons that can be read by a particular tRNA molecule.

The wobble base pairing explains why a transfer RNA molecule may recognize more than one synonymous codon. The consequence is that it is possible for a species to have a pool of different tRNAs for which the total number of the different anticodons is much less than the 61 sense codons used in this species. For instance, in yeasts, there are 41 different tRNA iso-acceptors recognizing the 61 sense codons of the 20 standard amino acids. In Humans, there are 46 different tRNA iso-acceptors recognizing the 61 sense codons of the 20 standard amino acids.

Finally, we recall here that the easy way to assess the relative abundance of the different tRNA iso-acceptors or estimate the relative abundance of the different tRNA families in the cells of a given species is through the gene copy number (GCN) of these respective tRNAs. In eukarya, the tRNA genes are transcribed by RNA polymerase III in the nucleus. In yeasts, the 41 different tRNAs are coded by 275 genes. For instance, there are 16 tRNA genes out of the 275 which code alanine tRNAs. Of these 16 iso-acceptors tRNAs' genes for alanine, 11 have the anticodon **IGC** (able to decode 3 of the 4 synonymous codons of alanine) and 5 have the anticodon **UGC** (able to decode 2 out of the 4 synonymous codons of alanine). In humans, the 46 different tRNAs are coded by about 428 genes. For instance, there are 38 tRNA genes out of the 428 coding for alanine tRNAs. Of these 38 alanine isoacceptor-tRNAs'genes, 28 have the anticodon **IGC** (able to decode 3 out of the 4 synonymous codons), 4 have the anticodon **CGC** (able to decode 1 out of the 4 synonymous codons) and 8 have the anticodon **UGC**

(able to decode 2 out of the 4 synonymous codons). We see that the tRNAs relative abundances are species specific and that the tRNAs iso-acceptors decoding capacities have overlapping sets of synonymous codons that are species-dependent as well.

A significant correlation has been observed between the codon usage frequency and the gene copy number for any given species. This indicates a general balance between supply (the abundance of cognate or near-cognate tRNAs) and demand (the codon usage frequency) in the protein elongation processes. The strength of this correlation supported the introduction of useful bioinformatics scoring metrics such as the codon adaptation index CAI, and the tRNA adaptation index, tAI. These scoring metrics can be used to infer the elongation efficiency of a given transcript in any given expression vector. This is used in biotechnology applications to optimize gene expression of heterologous inserts in cloned vectors.

5.2 Introduction to tRNA modifications

Like ribosome biogenesis, tRNA biogenesis is a highly coordinated process that is spatially and temporally organized. The precursors of tRNA molecules are transcribed from their genes by RNA Polymerase III in the nucleus and go through a series of maturation steps and post-transcriptional modifications to become fully active [Berg and Brandl 2021]. A complete mature tRNA molecule may require shuttling between the nucleus and the cytoplasm in eukaryotic species. Figure 5.4 illustrates the life cycle of a tRNA molecule. Over 170 different RNA base modifications have been described in the literature and a majority is found in tRNA molecules of all three domains of life. The functional consequences of tRNA modifications can be separated into three main categories depending on the modified positions:

(i) stabilizing the structural integrity of the core tRNA fold;

(ii) contributing to the correct amino-acylation of respective tRNAs at the acceptor stem loop;

(iii) enhancing the decoding potential, improving the translation fidelity at the ribosome or changing the local elongation rate depending of the codon usage of specific transcripts.

The third category of tRNA modifications is mostly found around the anticodon stem loop (ASL), particularly at the 'hot spot' positions 34 and 37. As modifications in this region can provide additional chemical bonds between the ASL and its cognate-tRNA or near cognate (near- and non- cognate tRNA) codons during the ribosomal mRNA sequence deciphering process, they are crucial for fine tuning translation elongation and co-translational folding dynamics.

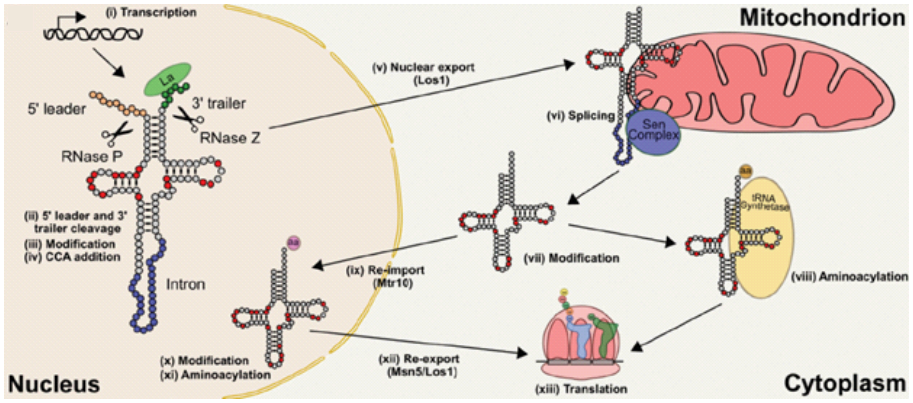


Figure 5.4: tRNAs are transcribed from RNA polymerase III (i) with 5' leader and 3' trailer sequences that are cleaved by RNase P and RNase Z respectively (ii). After trimming, partial modification (iii) indicated by red circles, and addition of the terminal CCA nucleotides (iv) occur in the nucleus. In yeast, the tRNA is exported out of the nucleus to the cytoplasm by Los1 (v) and if the tRNA contains an intron, it is spliced on the mitochondrial surface (vi). In mammals, splicing occur before nuclear export. In the cytoplasm, additional modifications may be added (vii). tRNAs can be aminoacylated by their cognate aminoacyl-tRNA synthetase (viii) or be re-imported (ix) by Mtr10 into the nucleus for further modification (x) and quality control through nuclear aminoacylation (xi). Re-export of tRNAs is facilitated by both Los1 and Msn5 dependent pathway that specifically recognizes mature tRNAs (xii). Once aminoacylated, tRNAs are used in translation by ribosomes (xiii). tRNAs can be degraded both in the cytoplasm and the nucleus by 5'-3' exonuclease Xrn1 and Rat1, respectively (not shown here). The typical half-life of a mature tRNA molecule is 2-3 days in eukaryotic cells. Figure reproduced with permission from [Berg and Brandl 2021]).

The number of target substrates of enzymatic tRNA post-transcriptional modifications in the anticodon stem loop of different tRNA isoacceptors can differ. Cases where a limited group of different tRNA iso-acceptors are targeted are those for which one, two or three tRNAs belonging to synonymous codons of one, two or three amino acid residues, are modified. This first case is described in the next section. In summary, for three **tRNAs**, namely $_{\mathrm{UUC}}\mathrm{tRNA}^{\mathrm{Glu}}$, $_{\mathrm{UUG}}\mathrm{tRNA}^{\mathrm{Gln}}$ and $_{\mathrm{UUU}}\mathrm{tRNA}^{\mathrm{Lys}}$, decoding the **mRNA** codons **GAA**, **CAA** and **AAA** respectively, a primal tRNA modification is succeeded by an additional thiolation leading to $\mathrm{mcm}^5\mathrm{s}^2\mathrm{U34}$-**tRNA** (mcm5s2-tRNA).

For cases where a larger number of different tRNA iso-acceptors are targeted, e.g., when up to eight distinct iso-acceptors are modified by a single enzyme, the accommodation and decoding rates of up to 37 codons for eight different amino acid residues will

be affected. This modification is catalyzed by a heterodimeric enzyme adenosine deaminase acting on tRNA (ADAT), which deaminates tRNAs coding for the amino acids T, A, P, S, L, I, V, and R (TAPSLIVR) and the eight targeted tRNAs of this ADAT enzyme are $_{3'-\mathrm{TGA}-5'}\mathrm{tRNA}^{\mathrm{Thr}}$, $_{\mathrm{CGA}}\mathrm{tRNA}^{\mathrm{Ala}}$, $_{\mathrm{GGA}}\mathrm{tRNA}^{\mathrm{Pro}}$, $_{\mathrm{AGA}}\mathrm{tRNA}^{\mathrm{Ser}}$, $_{\mathrm{GAA}}\mathrm{tRNA}^{\mathrm{Leu}}$, $_{\mathrm{TAA}}\mathrm{tRNA}^{\mathrm{Ile}}$, $_{\mathrm{CAA}}\mathrm{tRNA}^{\mathrm{Val}}$, and $_{\mathrm{GCA}}\mathrm{tRNA}^{\mathrm{Arg}}$ respectively. The eight targeted tRNA share in common that nucleotide 34 at the anticodon stem loop is an adenine. After enzymatic deamination of A34, these eight modified tRNAs are able to decode 24 different codons. This significantly expands the decoding capacity of these tRNAs and hence the translation efficiency.

5.3 ELP3-TRM9 and URM1: tRNA modifying enzymes of U34

Figure 5.5 shows different post-transcriptionnaly modified structures of the wobble U34 nucleoside.

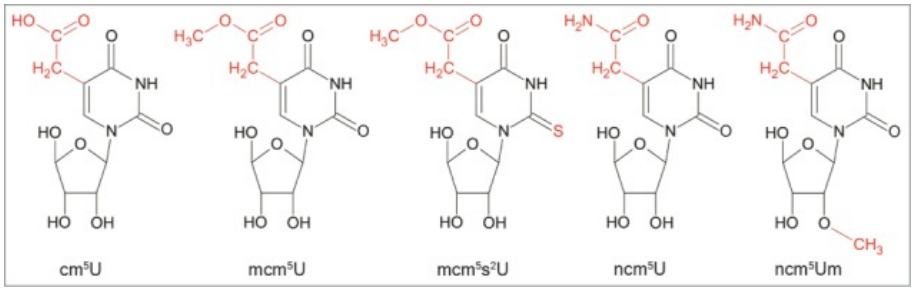


Figure 5.5: Side groups on uridines at wobble position in tRNA. Structures of 5-carboxymethyluridine ($\mathrm{cm}^5\mathrm{U}$), 5-methoxycarbonylmethyluridine ($\mathrm{mcm}^5\mathrm{U}$), 5-methoxycarbonylmethyl-2-thiouridine ($\mathrm{mcm}^5\mathrm{s}^2\mathrm{U}$), 5-carbamoylmethyluridine ($\mathrm{ncm}^5\mathrm{U}$) and 5-carbamoylmethyl-2'-O-methyluridine ($\mathrm{ncm}^5\mathrm{U}_\mathrm{m}$). The uridine side groups are highlighted in red (reproduced with permission from [Karlsborn et al. 2015]).

The eukaryotic Elongator complex **ELP1-6** and its acetyl transferase activity in the **ELP3** subunit is a tRNA modification enzyme, which catalyzes the cm^5U34 modification, representing the first step in a cascade leading to different types of U34 modifications. The 5-carboxy-methyl (cm^5) moiety can be subsequently methylated by the methyl transferase **Trm9** resulting in the 5-methoxycarbonyl-methyluridine at the wobble position (mcm^5U34). An additional thiolation can lead to mcm^5s^2U34-**tRNA**. The thiolation (s^2) is achieved by the **URM1** pathway, through the action of the ubiquitin-related modifier 1 (**Urm1p**), its activating enzyme **Uba4p**, thiouridine modification protein 1 (**Tum1p**), and proteins called 'Needs Cla 4 to survive 2 and 6' (**Ncs2p**, **Ncs6p**). **Uba4p** first activates **URM1** as acyl-adenylate and then transfers the sulfur on the catalytic cysteine in **Urm1p**. Subsequently, **Ncs2p/Ncs6p** mediate the transfer of sulfur from **Urm1p** to tRNA in an ATP-dependent manner [Ranjan and Rodnina 2017]. These post-transcriptional enzymatic wobble U34 tRNA modifications and the enzymes catalysing these pathways are illustrated in Figure 5.6 (a) and (b).

In yeasts and higher eukaryotes, U34 in $tRNA^{Lys}$ (anticodon UUU), $tRNA^{Gln}$ (anticodon UUG) and $tRNA^{Glu}$ (anticodon UUC) carries the mcm^5s^2 which is introduced by the enzymes of the **ELP** and **URM1** pathways (Figure 5.6). Deletions or silencing of these **URM1** or **ELP3** genes in yeasts impede translation of mRNAs with repeats of **AAA**, **CAA** and **GAA** codons read by these tRNAs, whereas the global protein synthesis is hardly affected [ibid.]. The lack of the s^2 modification (lack of thiolation) results in a higher ribosome density at these codons in ribosome profiling experiments, suggesting a direct involvement of the modification at some step of translation elongation, most probably accommodation and proofreading (A site initial binding and codon reading). The lack of the modification increases the dissociation rate of the codon-anticodon complex which is involved both at the initial-selection and proofreading phase on the A-site [ibid.]. Using a bacterial translation system (*E. coli*), Ranjan *et al.* [ibid.] estimated that the lack of s^2 modification increased the time required to decode 50% of the **Lys** codons from 1 to 3 s. The total time for the elongation cycle increases from 2.5 s to about 3.5 s. Hence, the residence time of the ribosomes on **Lys** codons is expected to increase by 40%. This nicely accounts for the modest (20 – 40%) increase in the ribosome occupancy of the **AAA** codons in the ribosome profiling experiments [Nedialkova and Leidel 2015; Zinshteyn and Gilbert 2013]. Ranjan et al. [Ranjan and Rodnina 2017] concluded that U34 tRNA modifications appear to play an important role in maintaining the exact rhythm of translation.

The three lists of the tRNAs iso-acceptors for the K, E, Q amino acid synonymous codons in three different species, their tRNA gene copy numbers and the associated synonymous codon frequencies statistics are provided in table 5.3.1. The tRNA copy number-related data reference is from GtRNAdb database [http://gtrnadb.ucsc.edu/GtRNAdb2/]. The codon usage frequencies database for the collected species is from [https://kasuza.or.jp/codon/].

The rate fold changes tabulated in 5.3.1 for each of the three amino acids were

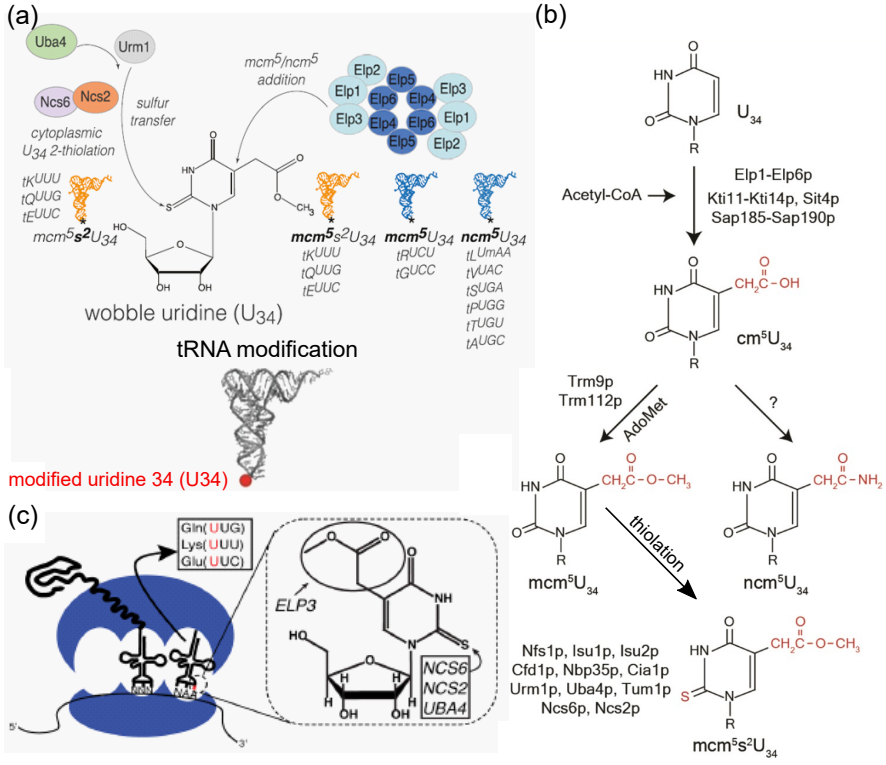


Figure 5.6: Simplified representation of the pathways and enzymatic modifications of the wobble U34 tRNA. (a) Elongator complex ELP proteins involved in U34 acetylation and pathways for wobble U34 modification in the eukaryotic cytoplasm (reproduced with permission from [Nedialkova and Leidel 2015]); (b) tRNA modification pathways modifying uridine 34 to mcm^5s^2U34 (reproduced from [Karlsborn et al. 2015]) and (c) mcm^5s^2U34 is found at the nucleotide 34 of the anticodon in three yeasts tRNAs, i.e., $5' - UUU - 3'$, $5' - UUG - 3'$ and $5' - UUC - 3'$ of lysine, glutamine and glutamate respectively (reproduced from [Zinshteyn and Gilbert 2013]).

calculated based on the fold changes in codon occupancy upon U34 tRNA enzymatic hypomodifiction, as experimentally measured by Nedialkova and Leidel through ribosome profiling [Nedialkova and Leidel 2015]. Their published paper, particularly Fig. 1 (B), reports the $\log_2 -$fold change in codon occupancy for each of the U34 sensitive codons in yeasts. The average total fold change in the ribosome residence time on the corresponding codons can be estimated from these $\log_2 -$fold changes. Assuming that only the accommodation step (Step 1) of elongation is affected, the fold

change in the rate of accommodation can be inferred, while the rates of Step 2 and Step 3 remain unchanged, as discussed by Rodnina and co-workers [Ranjan and Rodnina 2017].

For instance, the calculation of the rate fold change in accommodation for the **AAA** codon (lysine) is as follows.

The average total ribosome residence time (RRT) for the **AAA** codon (in wild-type yeast) is $\tau_{\text{tot.}} = 111.56\,\text{ms}$, with substep durations $\tau_1 = 82.49\,\text{ms}$, $\tau_2 = 14.53\,\text{ms}$, and $\tau_3 = 14.53\,\text{ms}$ (see Chapter 4). For **AAA**, the $\log_2$-fold change in ribosome occupancy upon U34 tRNA hypomodification, as reported by Nedialkova and Leidel, ranges from 1.20 to 1.29 [Nedialkova and Leidel 2015]. This corresponds to a fold change in ribosome occupancy, or equivalently RRT, of $\exp[1.20 \times \ln 2] = 2.30$ to $\exp[1.29 \times \ln 2] = 2.44$. Accordingly, the total RRT upon U34 tRNA hypomodification for **AAA** is in the range $111.56 \times 2.30 = 256\,\text{ms}$ to $111.56 \times 2.44 = 272\,\text{ms}$.

The required net fold change for the accommodation substep, τ_1, is thus calculated as $\frac{(256-111.56)+82.49}{82.49} = 2.75$ to $\frac{(272-111.56)+82.49}{82.49} = 2.94$. Since the accommodation rate is the inverse of τ_1, the corresponding fold change in the rate of step 1, inferred from these data, is in the range 0.33 to 0.36, as summarized in Table 5.3.1. All fold change factors listed in this table were calculated in a similar manner based on the published experimental results of Nedialkova and Leidel [ibid.].

Table 5.3.1: Codon-anticodon relation for U34-modified tRNA of lysine, glutamine and glutamate amino acid. tRNA gene copy number and codon usage frequencies are listed for 3 species (*Homo sapiens*, *Saccharomyces cerevisiae* and *Escherichia coli*). Step 1 Rate fold change rules upon ELP3-URM1 depletion estimated from *S. cerevisiae* from references [Nedialkova and Leidel 2015; Ranjan and Rodnina 2017].

Lysine (K)	species		
Gene copy number	H. sap.	15	12
	S. cer.	21	7
	E. coli	0	6
tRNA anticodon		3'-UUC-5'	3'-UUU-5'
mRNA codon		5'-AAG-3'	5'-AAA-3'
codon usage [in ‰]	H. sap.	31.86	24.44
	S. cer.	30.82	41.87
	E. coli	12.10	33.19
step 1 rate fold change		1.10	0.33
Glutamine (Q)	species		
Gene copy number	H. sap.	13	6
	S. cer.	1	9
	E. coli	2	2
tRNA anticodon		3'-GUC-5'	3'-GUU-5'
mRNA codon		5'-CAG-3'	5'-CAA-3'
codon usage [in ‰]	H. sap.	34.23	12.34
	S. cer.	12.11	27.28
	E. coli	27.72	12.10
step 1 rate fold change		1.0	0.55
Glutamate (E)	species		
Gene copy number	H. sap.	8	8
	S. cer.	2	14
	E. coli	0	4
tRNA anticodon		3'-CUC-5'	3'-CUU-5'
mRNA codon		5'-GAG-3'	5'-GAA-3'
codon usage [in ‰]	H. sap.	39.59	28.96
	S. cer.	19.24	45.60
	E. coli	18.35	43.73
step 1 rate fold change		1.18	0.62

The tRNA with anticodon listed in red does not exist in *E. coli*. The step 1 rate fold change values are the factors applied to multiply the first parameter of the corresponding hypo-exponential distribution for the queueing times on the codon upon ELP3-URM1 depletion activation in **`Ribosomer`**.

It is an intriguing fact that the enzymatic U34 modifications affect the deciphering of codons whose codon usage is strongly biased in most species and address amino acids (lysine, glutamine and glutamate) for which there are only two synonymous codons. This would suggest a functional role for a codon usage bias in genomes. The interpretation of this codon usage bias is that the tRNA modifications could reprogram the translation elongation efficiency and favor the protein synthesis of the selected transcripts for which the codon usage is enriched in the targeted codons, supporting the hypothetical existence of a so-called 'second genetic code'.

The lack of certain uridine modifications in the wobble position (U34), such as 5-methoxy-carbonyl-methyl-2-thio (mcm^5s^2), was shown to induce cellular stress, increase protein aggregation [Nedialkova and Leidel 2015], disturb proteome homeostasis, or to be associated with perturbations in plant immunity [Ramírez et al. 2015], developmental dysfunctions, neurological disorders, type 2 diabetes [Torres et al. 2014], tumors in promoting oncoproteomes and resistance to targeted cancer therapies [Rapino et al. 2021; Rapino et al. 2017].

5.4 Adenosine Deaminase Acting on tRNAs (ADAT): a tRNA modification enzyme expanding the decoding capacity of multiple tRNAs and the translation efficiency

Most transcriptomes from eukaryotes species exhibit a bias for C at the third nucleotide position (wobble position) of codons. Eukaryotic transcriptomes are significantly enriched in $5' - NNC - 3'$ codons. This is at odds with the fact that tRNAs with $3' - NNG - 5'$ anticodons are completely absent in these species [Lyu et al. 2020]. This dichotomy was a riddle that was solved by recognizing the role of a special category of tRNA modifications involving A34-to-I editing by adenosine deaminases targeting tRNAs, known as ADATs ADAT2. The unmodified A34 base (almost) only pairs with uridine (U) but when adenine is converted to inosine (I), I34 is able to form wobble pairs with C, U and A (but not G). Thus, the $3' - NNI - 5'$ anticodons allow tRNAs to recognize more synonymous codons, Fig. 5.7. Hence, ADAT enzymes expand the decoding capacity of tRNAs [Gerber and Keller 2001; Gerber and Keller 1999; Rafels-Ybern et al. 2018; Wolf et al. 2002]. A34-to-I editing by ADATs convincingly reinforced the explanation of the strong correlation between codon usage, tRNA abundance and tRNA gene copy number, in both bacteria and eukaryotes [Novoa et al. 2012]. Figure 5.7 shows the hydrolytic deamination of adenine to inosine by ADAT and how the A-to-I edited tRNA expands the decoding capacity to three codons in the example of the 4 synonymous codons for the amino acid alanine [Rafels-Ybern et al. 2015].

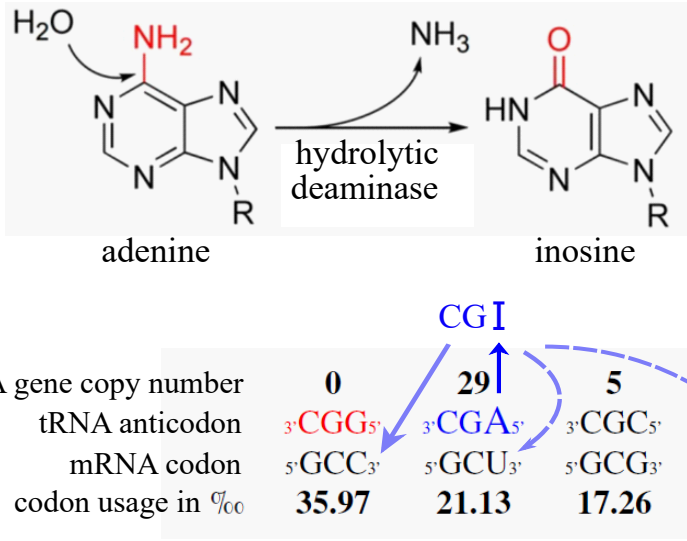


Figure 5.7: Adenine is converted into an inosine by a hydrolytic deamination catalyzed by the ADAT enzyme. Codon-anticodon relations for ADAT-associated alanine amino acid. tRNA gene copy number and codon usage frequencies is shown for each pair for the fungus *Neurospora crassa*. Note that ${}_{3'-CGG-5'}tRNA^{Ala}$ does not exist (red anticodon nucleotides).

Figure 5.8 shows the wobble pairing by hydrogen bonds of inosine (I) with cytosine (C), adenine (A), and uridine (U) respectively.

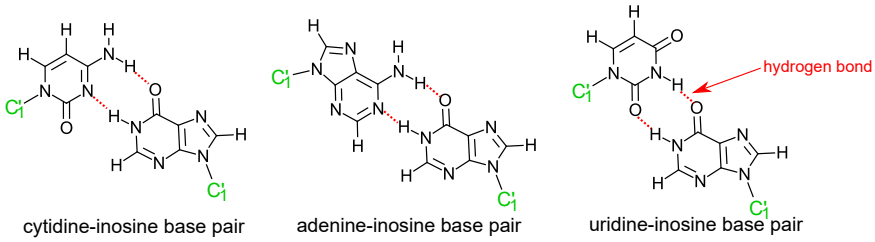


Figure 5.8: Inosine expands the decoding capacity of adenine after deamination. Inosine can wobble pair with cytosine (left), adenine (center) and uridine (right) through 2 hydrogen bonds.

5.5 Implementation and calibration of the tRNA modifications factor in the Agent-Based model

The tRNA modifications in the anticodon stem loop are believed to impact the most on the kinetics of what was called in chapter 4 (page 95), the first step in the elongation cycle, i.e., accommodation and proofreading. Most of the time, this step 1 is the rate limiting step of the elongation cycle. Genetic engineering techniques allow to silence the expression of the tRNA modification enzymes such as U34 modification enzymes or the ADAT enzymes.

Upon U34 modification enzyme depletion (**ELP3** or **6** and or thiolation by **URM1** or **NCS2**), the consequences for the engineered cells is a change in the binding affinity of the unmodified or hypo-modified tRNA at the A-site (accommodation and proofreading). Ribosome profiling experiments in yeasts, show a slowdown in the elongation rate for the cognate codons **AAA**, **GAA** and **CAA**. A slight increase (or no increase) is observed for the more rare codons **AAG**, **GAG** and **CAG** (these last three codons are the rare codons in *Saccharomyces cerevisiae*).

One of our contribution to study the tRNAs modification consequence on protein elongation rates was to implement a so-called "ELP3-URM1 depletion" factor in the ABM model. When this "ELP3 depleted" factor is activated, a modified rule is enforced on the ribosome agent to determine the hypo-exponential density distribution to sample from for the ribosome queueing time on the relevant codon. Explicitly, the new rule is that the rate parameter λ_1 for step 1 is decreased for the A ending codons of lysine, glutamine and glutamate (5'NNA3' codons) but (slightly) increased for the three rare codons (5'NAG3' codons). The fold change factors affecting the step 1 rates are listed in the last rows of table 5.3.1 for each of the three targeted amino acids. This rule can be applied to yeasts but may not strictly hold for humans. The codon usage for **Lys**, **Gln** and **Glu** is biased in the opposite direction in the case of *Homo sapiens* as listed in table 5.3.1.

Upon ADAT silencing, the consequence for the engineered cells is that their tRNAs $3' - \mathrm{NNA} - 5'$ anticodons are not edited to $3' - \mathrm{NNI} - 5'$ anymore. These unmodified tRNAs are not able to wobble decode C and A (but can better decode the U) in the codon at the third position. A slowdown in the elongation rate (slowdown in accommodation) is predicted to occur for all the near-cognate codons of these unmodified tRNAs. Ribo-Seq profiling experiments in *Neurospora crassa* provided strong evidence that it is indeed the case [Lyu et al. 2020]. As observed by Ribo-Seq analysis, upon ADAT silencing, the ribosome residence time on all eight 5'NNC3' codons are increased, i.e., the elongation rate is decreased for these codons. The ribosome residence time on seven out of eight 5'NNU3' codons are decreased, i.e., the elongation rate is increased for these codons, with the exception of the 5'AUU3' codon for isoleucine.

Another contribution to study the tRNAs modification consequence on protein elongation rates was to implement a so-called "ADAT-silencing" factor in the ABM model. When this "ADAT-silencing" factor is activated, a modified rule is enforced on the ribosome agent to determine the hypo-exponential density distribution to sample from for the ribosome queueing time on the relevant codon. The values of step 1 kinetics parameters of the subset of codons able to wobble pair with $3' - \text{NNI} - 5'$ are changed. Specifically, the new rule is that the rate parameter λ_1 for step 1 is decreased for eight 5′NNC3′ codons but increased for seven out of the eight 5′NNU3′ codons with the exception of isoleucine (the rate of step 1 for codon AUU is decreased). The hypo-exponential statistical distribution has 3 parameters but only the first one is changed upon ADAT-depletion activation. The factors applied to change the λ_1 rates of step 1 only affect the subset of codons in the following key-value pairs with the empirically predefined factor values. These predefined values for the new rule were based on Ribo-Seq profiling observations comparing control cases and ADAT-silenced cases for *Neurospora crassa* [ibid.].

The rate fold change factors tabulated in Table 5.5.1 were calculated in a similar way as detailed above for the U34 sensitive codons but this time using the codon occupancy fold change for the ADAT sensitive codons reported in Fig.4 (C) of reference [ibid.].

Table 5.5.1: Dictionary of the 37 keys (codons): λ_1 fold change values (step 1 rate fold change factor) upon ADAT silencing.

AUC:	0.57	**AUU:**	0.4	**AUA:**	2.5	**ACC:**	0.85	**ACU:**	1.5
ACG:	1.0	**ACA:**	0.67	**GCC:**	0.57	**GCU:**	2.0	**GCG:**	0.71
GCA:	0.9	**CCC:**	0.57	**CCU:**	2.0	**CCG:**	0.71	**CCA:**	0.9,
GUC:	0.85	**GUU:**	1.5	**GUA:**	0.67	**GUG:**	1.05	**UCC:**	0.5
UCU:	1.05	**UCG:**	1.1	**UCA:**	0.8	**AGC:**	1.15	**AGU:**	1.05
CUC:	0.9	**CUU:**	1.35	**CUG:**	1.4	**CUA:**	0.5	**UUG:**	0.8
UUA:	1.2	**CGC:**	0.4	**CGU:**	2.0	**CGG:**	1.0	**CGA:**	1.2
AGG:	0.85	**AGA:**	0.95						

Note that all NNU codons will have their λ_1 increased while all NNC codons will have their λ_1 decreased, upon ADAT silencing. Exception for AUU coding for isoleucine. Also note that for serine, λ_1 is significantly decreasing for UCC but slightly increasing for AGC.

The complete lists of all the tRNAs iso-acceptors for the TAPSLIVR amino acid synonymous codons in three different eukarya species, their tRNA gene copy number and the associated synonymous codon frequencies statistics are provided in the following tables. The tRNA copy number-related data reference is from GtRNAdb database [http://gtrnadb.ucsc.edu/GtRNAdb2/]. The codon usage frequencies database for the collected species is from [https://kasuza.or.jp/codon/].

Table 5.5.2: Codon-anticodon relation for ADAT-related threonine, alanine and proline amino acid. tRNA gene copy number and codon usage are listed for 3 species (*Homo sapiens*, *Saccharomyces cerevisiae* and *Neurospora crassa*). Step 1 Rate fold change rules upon ADAT silencing estimated from *N. crassa* from reference [Lyu et al. 2020].

Threonine (**T**)	species				
Gene copy number	H. sap.	0	9	5	6
	S. cer.	0	11	1	4
	N. cra.	**0**	**12**	**3**	**3**
tRNA anticodon		3'-UGG-5'	3'-UGA-5'	3'-UGC-5'	3'-UGU-5'
mRNA codon		5'-ACC-3'	5'-ACU-3'	5'-ACG-3'	5'-ACA-3'
codon usage [in ‰]	H. sap.	18.89	13.12	6.05	15.11
	S. cer.	12.73	20.28	7.96	17.76
	N. cra.	**24.71**	**11.16**	**13.54**	**10.75**
step 1 rate fold change		0.85	1.5	1.0	0.67
Alanine (**A**)	species				
Gene copy number	H. sap.	0	26	4	8
	S. cer.	0	11	0	5
	N. cra.	**0**	**29**	**5**	**3**
tRNA anticodon		3'-CGG-5'	3'-CGA-5'	3'-CGC-5'	3'-CGU-5'
mRNA codon		5'-GCC-3'	5'-GCU-3'	5'-GCG-3'	5'-GCA-3'
codon usage [in ‰]	H. sap.	27.73	18.45	7.37	15.82
	S. cer.	12.60	21.17	6.18	16.21
	N. cra.	**35.97**	**21.13**	**17.26**	**12.56**
step 1 rate fold change		0.57	2.0	0.71	0.9
Proline (**P**)	species				
Gene copy number	H. sap.	0	9	4	7
	S. cer.	0	2	0	10
	N. cra.	**0**	**12**	**4**	**3**
tRNA anticodon		3'-GGG-5'	3'-GGA-5'	3'-GGC-5'	3'-GGU-5'
mRNA codon		5'-CCC-3'	5'-CCU-3'	5'-CCG-3'	5'-CCA-3'
codon usage [in ‰]	H. sap.	19.79	17.54	6.92	16.92
	S. cer.	6.78	13.51	5.29	18.31
	N. cra.	**22.42**	**15.09**	**14.56**	**12.36**
step 1 rate fold change		0.57	2.0	0.71	0.9

The tRNA with anticodon listed in red does not exist in eukarya. The step 1 rate fold change values are the factors applied to the first parameter of the corresponding codon hypo-exponential distribution queueing times upon ADAT silencing activation in `Ribosomer`.

Table 5.5.3: Codon-anticodon relation for ADAT-related isoleucine and valine amino acid. tRNA gene copy number and codon usage are listed for 3 species (*Homo sapiens*, *Saccharomyces cerevisiae* and *Neurospora crassa*). Step 1 Rate fold change factors to apply upon ADAT silencing, as estimated from *N. crassa* in reference [Lyu et al. 2020].

Isoleucine (I)	species				
Gene copy number	H. sap.	3	14	5	
	S. cer.	0	13	2	
	N. cra.	**0**	**17**	**1**	
tRNA anticodon		3'-UAG-5'	3'-UAA-5'	3'-UAU-5'	
mRNA codon		5'-AUC-3'	5'-AUU-3'	5'-AUA-3'	
codon usage [in ‰]	H. sap.	20.82	16.0	7.49	
	S. cer.	17.17	30.13	17.79	
	N. cra.	**26.48**	**14.0**	**4.09**	
step 1 rate fold change		0.57	0.4	2.5	

Valine (V)	species				
Gene copy number	H. sap.	0	9	5	13
	S. cer.	0	14	2	2
	N. cra.	**0**	**19**	**2**	**4**
tRNA anticodon		3'-CAG-5'	3'-CAA-5'	3'-CAU-5'	3'-CAC-5'
mRNA codon		5'-GUC-3'	5'-GUU-3'	5'-GUA-3'	5'-GUG-3'
codon usage [in ‰]	H. sap.	14.46	11.03	7.08	28.12
	S. cer.	11.78	22.07	11.77	10.76
	N. cra.	**24.83**	**13.84**	**5.4**	**15.51**
step 1 rate fold change		0.85	1.5	0.67	1.05

The tRNA with anticodon listed in red (most often) does not exist in eukarya. The step 1 rate fold change values are the factors applied to the first parameter of the corresponding codon hypo-exponential distribution queueing times upon ADAT silencing activation in `Ribosomer`.

Table 5.5.4: Codon-anticodon relation for ADAT-related serine, leucine and arginine amino acid. tRNA gene copy number and codon usage are listed for 3 species (*Homo sapiens*, *Saccharomyces cerevisiae* and *Neurospora crassa*). Step 1 Rate fold change factors applied upon ADAT silencing, as based on *N. crassa* in reference [Lyu et al. 2020].

Serine (S)	species						
Gene copy number	H. sap.	0	9	4	4	8	0
	S. cer.	0	11	1	3	6	0
	N. cra.	**1**	**14**	**6**	**3**	**8**	**0**
3'-tRNA anticodon-5'		AGG	AGA	AGC	AGU	UCG	UCA
5'-mRNA codon-3'		UCC	UCU	UCG	UCA	AGC	AGU
codon usage [in ‰]	H. sap.	17.68	15.22	4.41	12.21	19.46	12.13
	S. cer.	14.22	23.50	8.56	18.67	9.75	14.15
	N. cra.	**19.99**	**11.95**	**14.51**	**9.22**	**17.43**	**8.66**
step 1 rate fold change		0.5	1.05	1.1	0.8	1.15	1.05
Leucine (L)	species						
Gene copy number	H. sap.	0	9	9	3	6	4
	S. cer.	1	0	0	3	10	7
	N. cra.	**1**	**16**	**6**	**2**	**6**	**1**
3'-tRNA anticodon-5'		GAG	GAA	GAC	GAU	AAC	AAU
5'-mRNA codon-3'		CUC	CUU	CUG	CUA	UUG	UUA
codon usage [in ‰]	H. sap.	19.59	13.19	39.64	7.15	12.93	7.67
	S. cer.	5.44	12.25	10.48	13.41	27.17	26.15
	N. cra.	**26.79**	**14.25**	**18.26**	**5.95**	**14.95**	**2.73**
step 1 rate fold change		0.9	1.35	1.4	0.5	0.8	1.2
Arginine (R)	species						
Gene copy number	H. sap.	0	7	4	6	5	6
	S. cer.	0	6	1	0	1	11
	N. cra.	**0**	**21**	**3**	**3**	**5**	**6**
3'-tRNA anticodon-5'		GCG	GCA	GCC	GCU	UCC	UCU
5'-mRNA codon-3'		CGC	CGU	CGG	CGA	AGG	AGA
codon usage [in ‰]	H. sap.	10.42	4.54	11.42	6.17	11.96	12.17
	S. cer.	2.60	6.40	1.74	2.99	9.23	21.28
	N. cra.	**17.64**	**8.88**	**8.54**	**7.05**	**11.84**	**7.91**
step 1 rate fold change		0.4	2.0	1.0	1.2	0.85	0.95

The tRNA with anticodon listed in red does (most often) not exist in eukarya. The step 1 rate fold change values are the factors applied to the first parameter of the corresponding codon hypo-exponential distribution queueing times upon ADAT silencing activation in `Ribosomer`.

The values indicated in the above dictionary of 37 key-values pairs are consistent with a thorough analysis of the tables 5.5.2, 5.5.3 and 5.5.4, describing for each TAPSLIVR amino acid, the synonymous codons statistics of the tRNAs associated iso-acceptors gene copy numbers. Figure 5.9 shows how the hypo-exponential distributions for the queueing time of a ribosome on the four synonymous codons are changed in the case of alanine (A) amino acid upon ADAT-silencing of the targeted tRNA iso-acceptor
.

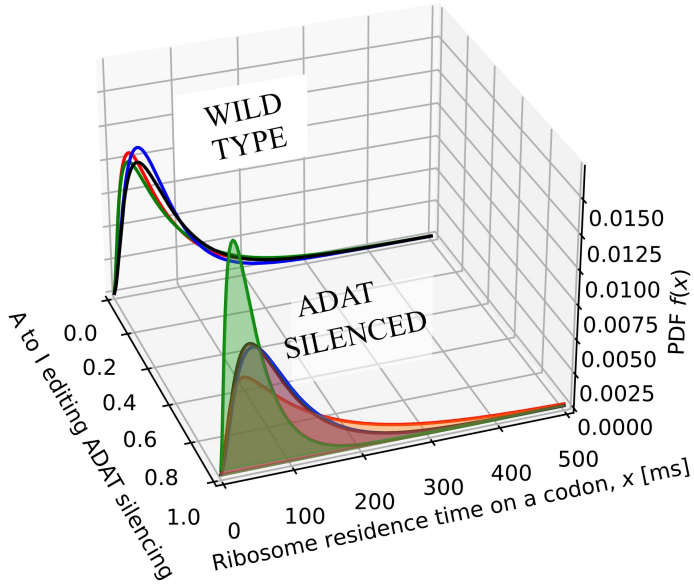


Figure 5.9: Probability density function (PDF) of the queueing time for each of the 4 alanine codons. Effects of ADAT enzyme silencing on the alanine synonymous codons queueing time sampling distributions. Rear plane: wild type. Front plane: ADAT silenced. Alanine synonymous codons: GCC (red), GCU (green), GCG (blue), GCA (black).

The pairwise comparison between the wild type and the ADAT silenced cases, for each of the four alanine individual codons are detailed in Figures 5.10 and 5.11. Due to the fact that the mean and the variance are not independent for a hypo-exponential sampling distribution, an increase in the averaged queueing time on a codon is also expected to be associated to an increase in the variance (increased heterogeneity). The standard deviation (std), i.e., the dispersion, increases when the queueing time mean is increased, e.g., as is the case for the alanine `GCC` codon upon ADAT silencing (Fig. 5.10 a).

The opposite is also expected: a decreased queueing time mean is associated to a decreased dispersion in the hypo-exponential sampling distribution, e.g., as is the case for the alanine GCU codon upon ADAT silencing (Fig. 5.10 b).
These ADAT silencing effects are codon-dependent. Figures 5.10 and 5.11 illustrate these expected effects. The inherent stochasticity described through the hypo-exponential sampling distributions of the queueing time of a ribosome on a codon results in changes in the variance composition in studies comparing a wild type control with an ADAT silenced case. The variance composition is codon-dependent and it

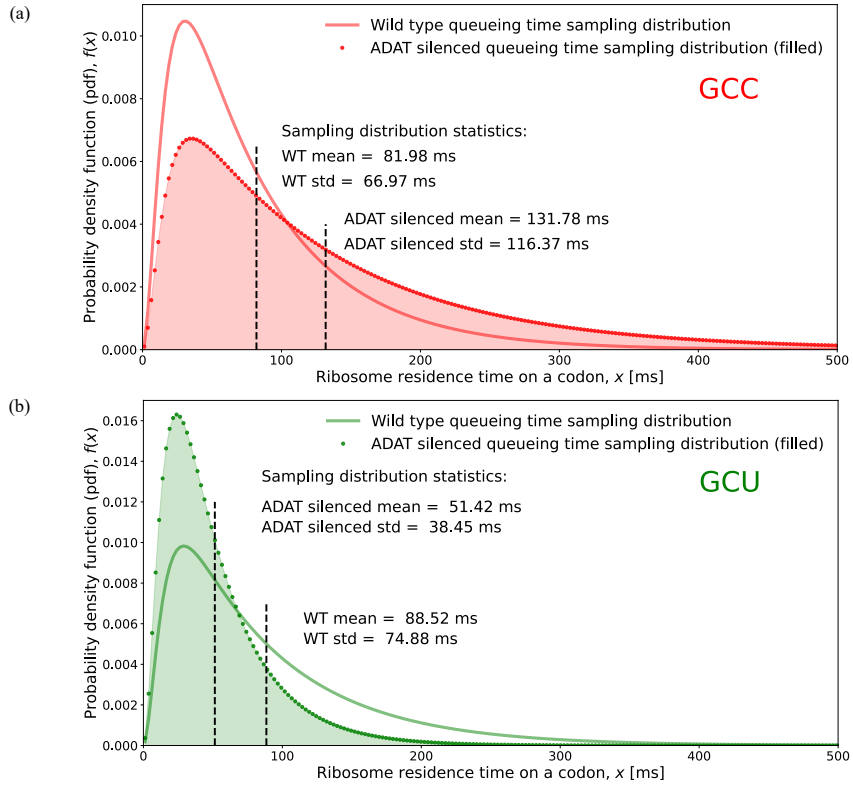


Figure 5.10: Probability density function (PDF) of the queueing time for each of the 4 alanine codons. Effects of ADAT enzyme silencing on the alanine individual codons queueing time sampling distributions. Pairwise comparison per codon. Wild type: simple line; ADAT silenced: dot line filled area under the curve. Alanine synonymous codons: (a) GCC (red), (b) GCU (green). See next figure for GCG (blue) and GCA (black).

may not be the same sets of codons contributing most to the aggregated heterogeneity when comparing a case with a control. This suggests why the experimental Ribo-seq profiling between samples datasets are so noisy and why classical analysis of variance statistical methods could be inappropriate.

Similar changes in the queueing time sampling distributions occur for the seven other amino acids (TPSLIVR) and their 33 codons whose elongation kinetics are affected by

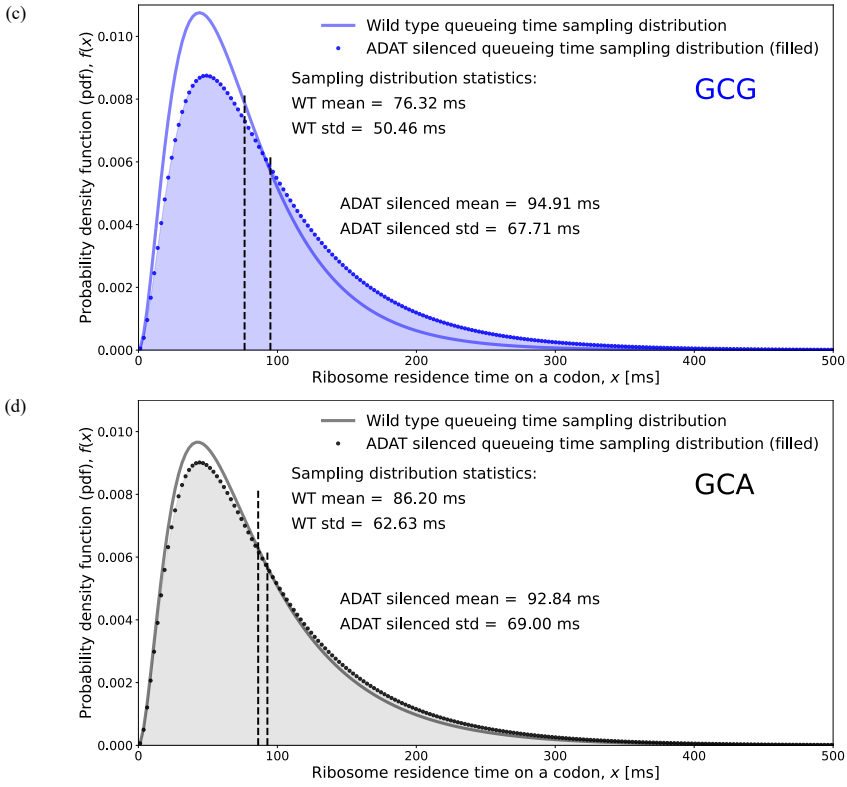


Figure 5.11: Probability density function (PDF) of the queueing time for each of the 4 alanine codons. Effects of ADAT enzyme silencing on the alanine individual codons queueing time sampling distributions. Pairwise comparison per codon. Wild type: simple line; ADAT silenced: dot line filled area under the curve. Alanine synonymous codons: (c) GCG (blue), (d) GCA (black). See previous figure for GCC (red) and GCU (green).

the A to I editing ADAT enzyme silencing (similar plots to Fig. 5.9 not shown here).

All the affected elongation kinetics for all the 37 codons have been implemented in the ABM model conditionally on the ADAT silencing factor. The effects of an ADAT silencing on the proteome output, translational efficiency and on the ribosome density footprint patterns can be studied through extensive computer simulations presented in the preliminary results in chapter 11. These first simulations will be continued in future research extending the work presented in this thesis.

5.6 Summary of main findings and insights

This chapter 5 of the thesis focused on the role of tRNA pools and tRNA modifications in translation elongation. It highlighted how codon usage and specific tRNA modifications influence elongation kinetics and translation efficiency. The chapter centered on two key enzymatic modifications at position 34 of the anticodon loop: (i) the conversion of U34 to mcm5s2U34, which affects codons for lysine, glutamic acid, and glutamine, and (ii) the modification of A34 to I34 by adenosine deaminase (ADAT), which significantly expands the decoding capacity of tRNAs for eight amino acids. These modifications modulate ribosome queueing times, altering elongation rates at specific codons. The chapter also discussed how these factors were integrated into our agent-based model (ABM) to simulate their effects on translation and ribosome density mapping. The model was calibrated with meta-analysis data and provides a predictive tool for synthetic biology and disease research, including applications in cancer studies.

Chapter 6

Ribosome exit tunnel electrostatic interaction

This chapter was previously published in two original contributions:

1. M. Joiret et al. (2022b). "Ribosome exit tunnel electrostatics". In: *Physical Review E* 105.1, p. 014409. DOI: 10.1103/PhysRevE.105.014409
2. M. Joiret et al. (2024). "Reversing the relative time courses of the peptide bond reaction with oligopeptides of different lengths and charged amino acid distributions in the ribosome exit tunnel". In: *Computational and Structural Biotechnology Journal* 23, pp. 2453–2464. DOI: 10.1016/j.csbj.2024.05.045

These two papers are referred to as PRE and CSBJ02, respectively.

The core overview section highlights the key results of these two published articles that are directly aligned with the central framework of the thesis. This introduction provides sufficient context for understanding the Ribosomer model and maintaining continuity with the thesis's core narrative.

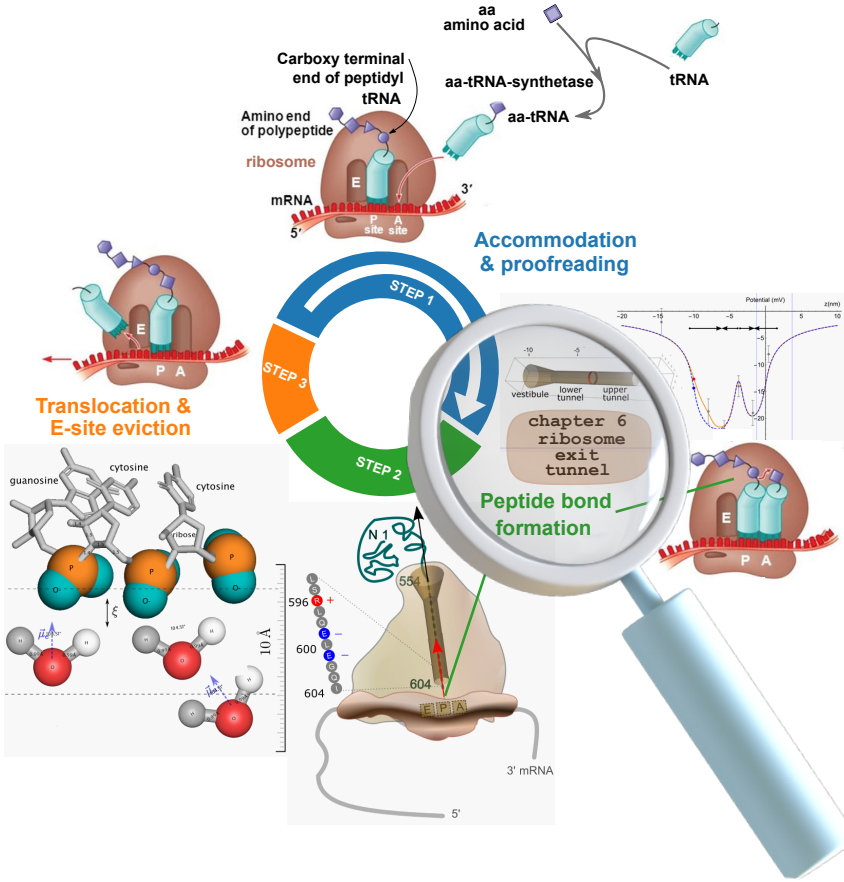


Figure 6.1: Graphical abstract of the impact of the ribosome exit tunnel electrostatic interaction on the elongation cycle. The inner wall surface of the ribosome exit tunnel is lined with a large number of phosphate moieties bridging the ribonucleotides of the single-stranded 23S/28S rRNA. These phosphates carry formal negative charges, creating a negative electrostatic potential profile along the entire centro-axial length of the exit tunnel. Charged amino acids within the tunnel are sensitive to this electrostatic interaction and transmit the probed forces through the nascent chain backbone down to the peptidyl-tRNA at the P-site. These forces modulate the activation energy of peptide bond formation, which, in turn, affects the rate of the peptide bond formation.

6.1 Core overview and connection to the thesis backbone

The central idea, common to both papers, emphasizes the role of the ribosome exit tunnel as a critical factor influencing the kinetics of the protein elongation cycle. The first paper (PRE) covers an extended study of the ribosome exit tunnel electrostatics. The second paper (CSBJ02) builds on the theoretical model of PRE (but also of CSBJ01 in Chapter 4) to make predictions that could be experimentally tested. The key results are as follows: a model of the electrostatic potential along the central axis of the ribosome exit tunnel was developed, leading to a fully analytical expression for both the electrostatic potential and the corresponding electric field.

One of the key factors influencing the non-uniform rate of protein elongation by the ribosome during mRNA translation is the interaction between the nascent peptide chain, embedded within the ribosome exit tunnel, and the tunnel's internal wall. This interaction is electrostatic in nature, arising from the negatively charged surface of the tunnel wall and the charged amino acids incorporated into the growing peptide chain. The charge distribution of the nascent chain depends on the encoded mRNA sequence.

The permanent negative charge of the tunnel's inner wall is primarily due to the phosphate moieties of the 23S/28S rRNA phosphodiester backbone, each carrying a formal negative charge on one of its two non-bridging oxygen atoms. Given that rRNA constitutes the majority of the ribosomal large subunit's mass, this electrostatic environment plays a significant role in modulating translation kinetics.

The model was calibrated using previously published measurements of electrostatic potential in ribosomes from rabbit reticulocytes. The quantitative theoretical predictions presented in CSBJ02 could be experimentally tested in the future, providing validation for this aspect of the model and reinforcing confidence in the mechanistic elucidation offered by this factor.

The key contribution addresses how this factor is incorporated into the agent-based model.

6.1.1 Key contributions

In chapter 4, a quantitative link was established between the rate constant of a (bio)chemical reaction and an (external) mechanical work through the Maxwell-

Boltzmann factor in the rightmost part of the equation:

$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot e^{-\left(\frac{\Delta G^{\ddagger 0(\mathbf{0})}}{N k_B T} - \frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}\right)} = k(\mathbf{0}) \cdot \underbrace{e^{\frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}}}_{\text{Maxwell-Boltzmann factor}} \tag{6.1}$$

where $k(\vec{\mathbf{F}})$ is the reaction rate constant of the rate limiting step in the presence of an applied force acting on the transition state, $k(\mathbf{0})$ is the reaction rate constant in the absence of applied force. N, k_B, h and κ are Avogadro's number, Boltzmann's constant, Planck's constant and transmission coefficient respectively [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983].

The Maxwell-Boltzmann factor can be larger than 1 and increase the rate, or smaller than 1 and decrease the rate, depending on the algebraic sign of the mechanical work.

It is hypothesized that step 2 of the elongation cycle—peptide bond formation—is modulated by the electrostatic interaction between the nascent chain and the ribosome exit tunnel. A plausible mechanistic explanation is that a pulling force exerted on the peptidyl side of the peptidyl-tRNA at the P-site facilitates deacylation of the peptide from its ester bond with the ribose of A76, following the nucleophilic attack of the amino group from the aa-tRNA at the A-site. Additionally, this pulling force may promote the entropically favorable positioning of the tRNA substrates at the P-site and A-site, further influencing the reaction dynamics.

The key results of this chapter follow from the following basic biophysical reasoning:

1. The quantitative estimation of the mechanical work exerted by the force on the transition state—necessary for calculating the Maxwell-Boltzmann factor—requires precise knowledge of the force acting on the backbone of the nascent chain.

2. The force arises from the permanent electric field within the tunnel environment. As the field is static, it is estimated from the gradient of the electrostatic potential, specifically the negative gradient of the potential.

3. Knowledge of the electrostatic potential profile along the centro-axial line of the exit tunnel, or a calibrated version of this profile, is required. The calibration utilizes experimental data previously published by the pioneering group of Carol Deutsch on the ribosome exit tunnel of rabbit reticulocytes [Lu et al. 2007]. A fully analytical, piecewise continuous expression of the electrostatic potential profile was derived using two converging approaches:

   (a) A geometrical theoretical model of the tunnel was conceptualized as the concatenation of a straight hollow cylinder, a conical frustum, and a localized constriction site located in the first third of the cylinder's length.

(b) A model based on structural data from x-ray crystallographic or cryo-EM solved structures of real ribosomes at atomic scale resolution was used.

Both approaches (a) and (b) utilized the Poisson-Boltzmann equation for electrostatics in dielectric media, or its linearized Yukawa-Debye-Hückel version. A thorough literature review was conducted to estimate the physical values of empirical parameters, such as the dielectric constants (relative permittivities) and screening lengths of the media, in the complex heterogeneous environment where nucleic acids, proteins, water, and ions coexist.

Using structural data, both approaches determined the surface charge densities by counting the phosphorus atoms in the phosphate moieties and their negatively charged non-bridging oxygen atoms (the sources of the electrostatic field), incorporating their exact 3D spatial distributions. The induced polarization and screening effects of water molecules were considered, with particular attention to the contributions of a small number of positively charged amino acids from ribosomal proteins protruding into the tunnel.

The physical values of the electrostatic potential (in millivolts, mV) and the local electric field (in megavolts per centimeter, MV/cm) were estimated and compared with published literature values.

4. The electric field profile was derived from the analytical expression of the electrostatic potential, by taking its first derivative along the centro-axial line of the tunnel. The local forces acting on the nascent chain backbone were obtained by multiplying the charges (positive or negative) of the individual amino acids with the local electric field strength. The total force is the sum of all local forces, or equivalently, the mechanical work was calculated as the curvilinear integral of the local forces along the infinitesimal displacements of the nascent chain during each elongation cycle.

6.1.2 Key outcomes included in the ribosomer model framework

A simple bioinformatics algorithm, based on the aforementioned line of reasoning, enables the computation of the Maxwell-Boltzmann factor from Equation 6.1 at each elongation step, as a new amino acid is added to the carboxy-terminal end of the peptidyl-tRNA at the P-site. The algorithm operates stepwise, with single-codon resolution, updating at each ribosomal translocation event. Figure 6.2 reproduced from the first paper, illustrates the algorithm's workflow. At each step, the algorithm scans a moving window whose length corresponds to the number of amino acids embedded within the ribosome exit tunnel. This window spans a minimum of one amino acid and a maximum of fifty. During execution, the algorithm maintains a memory of the

spatial sequence distribution of the last 50 amino acids elongated by each ribosome agent as it processes a transcript. The primary sequence is depicted in Figure 6.2, as a linear thread of multicolored beads, where red and blue beads represent positively and negatively charged amino acids, respectively. The algorithm's output modulates, at codon resolution, the rate constant of the peptide bond formation reaction k_2 for each ribosome on a given transcript. Consequently, the mean queueing time setpoint of step 2 in the elongation cycle, τ_2, is dynamically adjusted. Each ribosome agent resets its queueing timer at the end of an elongation cycle and adopts the newly computed τ_2 setpoint. This occurs as part of the updating loop of the three key setpoints τ_1, τ_2 and τ_3 which are all calculated in a context-dependent manner, as previously described in the agent-based model of the ribosome[1].

The second paper (CSBJ02) introduced a methodology to test the theoretical quantitative predictions of our electrostatic model using synthetically designed transcript sequences of varying lengths and distinct charge distributions. The inverse bell-shaped profile of the electrostatic potential within the tunnel plays a crucial role in explaining how the relative timing of peptide bond formation can be reversed, depending on both the charge (positive or negative) and the spatial positioning of amino acids within the ribosome exit tunnel.

[1]The τ_3 queueing time set point for step 3 – the translocation step – is the topic of chapter 8.

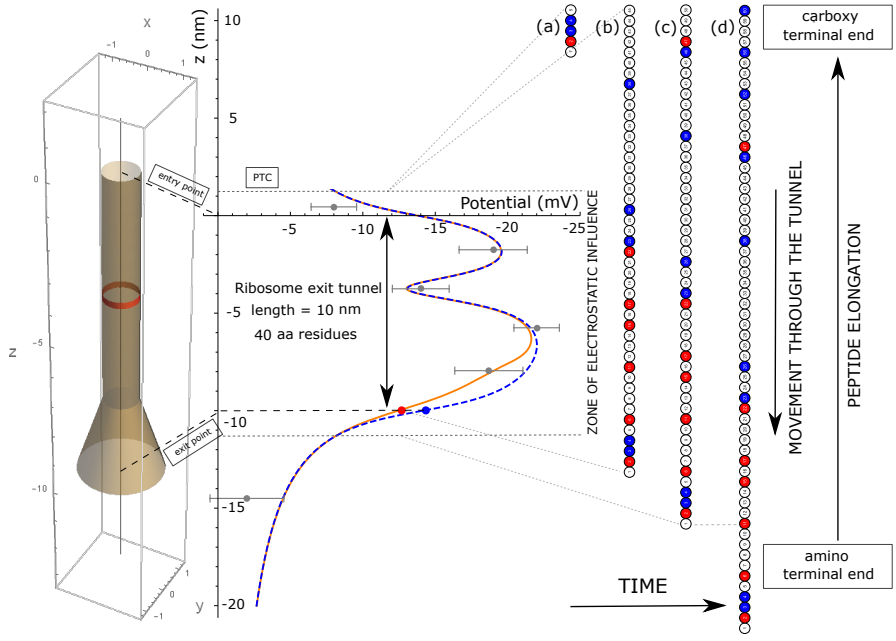


Figure 6.2: Simple bioinformatics algorithm for the computation of electrostatic axial forces acting on the nascent chain in the ribosome exit tunnel. The axial forces acting on each individual charged amino acid depend on the nature and arrangement of the amino acids within the segment of the nascent polypeptide chain embedded in the tunnel. The computed axial forces are updated at each new translocation while the chain is elongated further during the translation by the ribosomes from the carboxy-terminal end. Four snapshots showing in (a), (b), (c), (d) the peptide nascent chain growing towards the tunnel exit with its amino-terminal end exiting first. Full explanation in the text and references [Joiret et al. 2022b] and [Joiret et al. 2024].

6.2 Ribosome exit tunnel electrostatics (PRE)

Sections `6.4.1`, `6.4.6`, `6.5.1`, `6.5.4`, `6.6.1` as well as appendix `C.5` (respectively sections `II.A`, `II.F`, `III.A`, `III.D`, `IV.A`, `V.A.3` as well as Appendix 5 of this published article) are the key elements required for the context of the `Ribosomer` model.

PRE abstract

The impact of the ribosome exit tunnel electrostatics on the protein elongation rate or on the forces acting upon the nascent polypeptide chain are currently not fully elucidated. In the past, researchers have measured the electrostatic potential inside the ribosome polypeptide exit tunnel at a limited number of spatial points, at least in rabbit reticulocytes. Here, we present a basic electrostatic model of the exit tunnel of the ribosome, providing a quantitative physical description of the tunnel interaction with the nascent proteins at all centro-axial points inside the tunnel. We show that a strong electrostatic screening is due to water molecules (not mobile ions) attracted to the ribosomal nucleic acid phosphate moieties buried in the immediate vicinity of the tunnel wall. We also show how the tunnel wall components and local ribosomal protein protrusions impact on the electrostatic potential profile and impede charged amino acid residues from progressing through the tunnel, affecting the elongation rate in a range of minus 40% to plus 85% when compared to the average elongation rate. The time spent by the ribosome to decode the genetic encrypted message is constrained accordingly. We quantitatively derived, at single residue resolution, the axial forces acting on the nascent peptide from its particular sequence embedded in the tunnel. The model sheds light on how the experimental data point measurements of the potential are linked to the local structural chemistry of the inner wall, the shape and the size of the tunnel. The model consistently connects experimental observations coming from different fields in molecular biology, X-ray crystallography, physical chemistry, biomechanics, synthetic and multi-omics biology. Our model should be a valuable tool to gain insight into protein synthesis dynamics, translational control and into the role of the ribosome's mechanochemistry in the co-translational protein folding.

6.3 PRE introduction

Ribosomes are the cells' manufacturing tools building up proteins. They decode the 61 sense codons from a primary message encrypted in a messenger RNA (mRNA) single molecule, translate it with the help of a set of fewer than 61 transfer RNAs

(tRNAs) into 20 amino acids to be sequentially polymerized in a nascent polypeptide that will eventually fold into its final structure. At each elongation cycle, the ribosome incorporates a new amino acid into the nascent protein and translocates to the next codon – shifting along the single stranded mRNA by three nucleotides (triplet). Ribosomes have three binding sites for tRNAs: the aminoacyl (A), the peptidyl (P), and exit (E) sites, each located between the small and the large subunit of the ribosome. The elongation cycle starts with recognition, accommodation by induced fit and proofreading of an aminoacylated tRNA on the A site of the ribosome if the cognate anticodon pairs the codon being read on the mRNA [Beringer and Rodnina 2007; Rodnina and Wintermeyer 2001]. Elongation proceeds with the binding of the carboxyl terminal end of the peptide acylated to the previous tRNA at the P site to the amino moiety of the amino acid acylated on the tRNA at the A site. The formation of the new peptide bond between the nascent chain and the new amino acid is catalyzed at the peptidyl transferase center (PTC), Fig. 6.3, by a ribozyme belonging to the large subunit of the ribosome [Nissen et al. 2000]. Two energy rich guanosine triphosphate molecules (GTP) are used and two elongation factors with GTPase activity assist the ribosome during each elongation cycle. For more than five decades, attempts to model protein synthesis and mRNA translation from first principles have been pursued extensively [Haar 2008; MacDonald and Gibbs 1969; MacDonald et al. 1968; Zur and Tuller 2016]. Although the average codon translation rate is rather constant transcriptome wide, estimated at 5.6 amino acid residues per second in eukaryotes, codon translation rates have been shown to vary up to 100-fold across a single transcript [Ingolia et al. 2011; Morisaki et al. 2016]. Many factors influence translation speeds across a single transcript (mRNA), including differences in cognate, near-cognate and non-cognate tRNA relative abundance, nascent-chain charged residues inside the ribosome exit tunnel, mRNA secondary structure, proline residues at either A or P site of the ribosome, steric hindrance between contiguous ribosomes translating the same mRNA molecule, and the finite resource of the ribosome pool available in the cell [Dana 2014; Dao Duc and Song 2018; Fluitt et al. 2007; Lu and Deutsch 2008; Pavlov et al. 2009; Raveh et al. 2016; Riba et al. 2019; Rodnina 2016; Shah et al. 2013; Sharma et al. 2018; Simpson et al. 2020; Tuller et al. 2011; Yang et al. 2014]. The individual contributions of each of the previous factors to the rate of the translation are difficult to assess quantitatively and separately. Sometimes, depending on the local sequence in the mRNA encrypted message, all these factors interfere and may either antagonize each other or, on the contrary, add up to increase or decrease the rate of translation significantly [Dao Duc and Song 2018; Gorochowski et al. 2015b; Rodnina 2016; Sharma et al. 2018; Tuller et al. 2011; Yang et al. 2014]. This hampers our understanding of the dynamics of protein synthesis and specifically of the elongation rate.

Although this has been disputed and it remains a debated question [Artieri and Fraser 2014; Requião et al. 2016; Sabi and Tuller 2015], some studies have argued that the charged residues are the major determinants of ribosomal velocity [Charneski and Hurst 2013]. The nascent protein gets out of the ribosome through a tight tunnel

approximately 8.5-10 nm long and 1-2 nm wide [Voss et al. 2006]. The inner wall of the ribosomal exit tunnel is lined with fixed negative charges causing a local negative electrostatic potential inside the tunnel as shown in Fig.6.3 [Lu et al. 2007]. Among the 20 amino acids, two of them are positively charged in physiological conditions, namely arginine and lysine [Pace et al. 2009]. A third one, histidine, is only weakly positively charged. When the ribosome incorporates a local increased number of such positively charged amino acid residues in the nascent protein, a local variation in the elongation rate is often reported. This is also true for the negatively charged amino acid residues, i.e. glutamate and aspartate. Ribosome profiling experiments results (Ribo-Seq) are difficult to interpret and to reconcile with RNA sequencing (RNA-Seq) profiles and proteome expression results in any given biological condition [Dao Duc and Song 2018]. *In vitro* laser optical tweezers experiments [Desai et al. 2019; Kaiser and Tinoco 2014; Liu et al. 2014a; Wen et al. 2008], with high resolution dual traps, involving ribosome specifically, are now being conducted to probe the forces acting upon the mRNA at each translocation step or upon the nascent protein emerging from the exit tunnel [Kaiser and Tinoco 2014].

The research community would benefit from highly predictive and quantitatively accurate computational models of translation dynamics and specifically of elongation rates. A fully realistic model of the electrostatics inside the ribosome exit tunnel is lacking, despite experimental point measurements of the electrostatic scalar potential in the ribosomal exit tunnel being available, at least in one eukaryote species cell type, i.e. rabbit reticulocytes [Lu et al. 2007]. Stochastic models for protein synthesis have been developed for more than fifty years [Haar 2008; MacDonald and Gibbs 1969; MacDonald et al. 1968; Shah et al. 2013; Sharma et al. 2018; Zur and Tuller 2016]. The extended totally asymmetric simple exclusion process (TASEP) is a widely used stochastic model family dedicated to dynamically simulate the translation rate of a set of transcripts in various conditions [Greulich et al. 2012; Sharma et al. 2018; Zia et al. 2011; Zur and Tuller 2016]. In the parametrization of TASEP models, most researchers impose an empirical penalty factor to account for the influence of the electrostatic molecular interaction of the ribosome exit tunnel with newly incorporated charged amino residues at the peptidyl transferase center (PTC). For example, a fixed 20% decrease in the translation rate is imposed for those codons that are within five positions downstream of a codon encoding positively charged amino acid residues (lysine, arginine or histidine)[Sharma et al. 2018]. The negatively charged residues (aspartate and glutamate) are ignored in most studies. This approach is considered inconsistent or too naive if more accurate predictions are expected from TASEP models and to be compared to specific ribosome profiling experimental data [Dana 2014] or to real-time specific single RNA molecule translation dynamics experiments *in vivo* [Morisaki et al. 2016] or *in vitro* [Kaiser and Tinoco 2014; Wen et al. 2008].

In this study, we focus on one of these specific factors which affects the local speed of elongation during protein synthesis, namely the electrostatic interaction between

the charged amino acid residues embedded in the nascent polypeptide chain and the ribosome exit tunnel. To model this electrostatic interaction, we developed a full analytical expression of the electrostatic potential inside the tunnel, starting from two very basic and idealized theoretical geometries for the tunnel. The assumptions on the phenomenological parameters of the model are confronted to the atomic structure of the large ribosomal subunit determined by X-ray crystallography. Together, the model and the experimental structural constraints allow us to investigate the origin of the electrostatic screening that prevails in the very confined environment of the ribosome exit tunnel. The model is used to explore the physical consequences of a possible dynamically variable geometry of the tunnel from a theoretical perspective. The model is used to quantitatively estimate the profile of the axial forces and requires knowledge of the primary sequence of a significant length of the nascent polypeptide chain or its encrypted mRNA to compute the local axial forces acting at the PTC center during elongation. An algorithm is proposed to compute the axial forces acting locally at the PTC and due to a spatially extended electrostatic interaction inside the tunnel. The model is used to conduct comparative analyses of the axial force profiles for different synthetic or real protein sequences. Knowing the axial forces quantitatively allow to estimate the mechanical work and the biochemical energy required at each elongation step to overcome the electrostatic potential barrier inside the ribosome exit tunnel. These estimations are compared to the energy sources and uptakes involved in the mechanochemistry of the ribosome at each elongation cycle. The ribosome exit tunnel electrostatic model we describe can stand as a building block for computational tools that should be beneficial for the analysis of different experimental techniques like the probing of force by laser optical tweezers, the study of conformational changes with fluorescence resonance energy transfer at the ribosome subunits, the measure of the longitudinal electric field along the ribosome exit tunnel axis by vibrational Stark spectroscopy and for bioinformatic processing of multi-omics data [Fried and Boxer 2017].

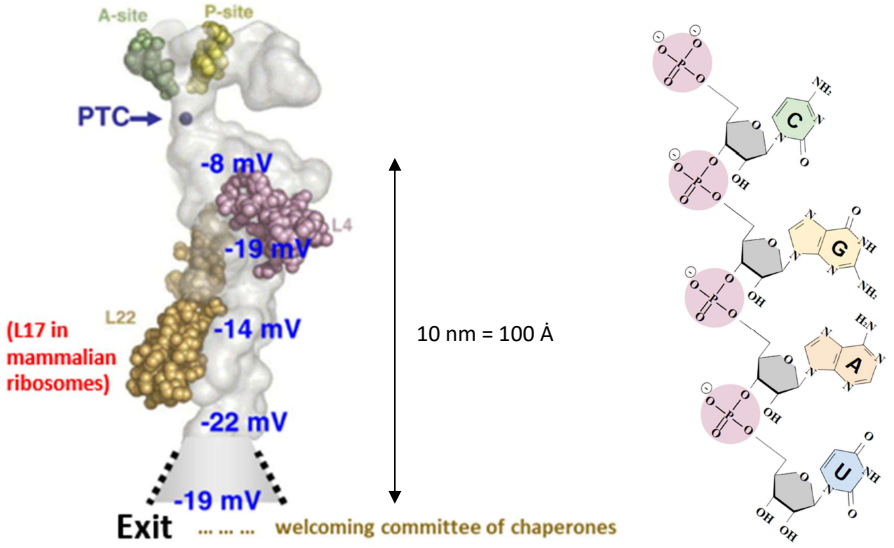


Figure 6.3: Left panel: Ribosomal exit tunnel structure. The light grey shape is made up of rRNAs. The peptidyl transfer center (PTC) is where a new amino acid residue is bound to the nascent peptide. The figure is taken with permission from Lu *et al* [Lu et al. 2007]. Right panel: RNA molecular structure showing the ribose-phosphate alternating units. The minus signs represent formal negative charges.

6.4 Geometrically idealized electrostatic models of the ribosome exit tunnel

The local negative electrostatic potential inside the ribosome exit tunnel, shown in Fig.6.3 (left panel), from which the nascent proteins emerge, originates from the ribosome composition. Ribosomes are composed of two subunits: 50S and 30S in prokaryotes, 60S and 40S in eukaryotes, identified by their sedimentation coefficients, measured in Svedberg unit *S*; the whole prokaryotic and eukaryotic ribosomes are 70S and 80S, respectively. The ribosome exit tunnel is found in the larger (50S or 60S) of the two subunits. Each of the subunits entails proteins and ribosomal ribonucleic acids (rRNAs). The essential feature of interest of rRNAs, shown in Fig. 6.3 (right panel), is that, like all RNAs, they are single stranded polymerized molecules with a backbone made up of alternating ribose sugars and phosphate groups all esterified alternatively together. In this long strand, each phosphate group harbors a formal negative charge. The inner wall of the ribosome exit tunnel is mainly lined up with rRNAs (more than 80% w/w in eukaryotic ribosome exit tunnels), though in some locations specific proteins are also present.

6.4.1 General modeling approach and assumptions

Starting with idealized shapes, the ribosome exit tunnel is successively considered as one of the exact three dimensional surfaces: hollow straight cylinder Fig. 6.4 left panel (a), cone frustum (b) and a cone frustum concatenated to a cylinder (c). The wall material is not of the conductor type with mobile free charges but is rather a dielectric material harboring fixed charges – the fixed phosphate moieties lining the inner wall. As a first reasonable assumption, the fixed charges are supposed to be uniformly distributed on the surface of the inner wall. The size of the hollow cylinder closest to the shape of the ribosome exit tunnel documented in the literature would be $85-100$ $(8.5-10\,\text{nm})$ in length and $10-20$ $(1-2\,\text{nm})$ in diameter Dao Duc et al. 2019; Voss et al. 2006. The precise length for the ribosomal exit tunnel as measured by cryo-electron microscopy is 9.2 nm on average in prokaryotes and 8.3 nm on average in eukaryotes Dao Duc et al. 2019. The *in vivo* lengths are believed to be a bit larger due to thermal dilatation at the higher temperatures prevailing in living organisms as compared to the cryogenic conditions. For a given uniformly distributed charge density σ on the inner surface wall of the tunnel, the determination of the electrostatic scalar potential $\Phi(\vec{\mathbf{r}})$ and of the electric field $\vec{\mathbf{E}}(\vec{\mathbf{r}})$, at any spatial point close to or far away from the surface, are well stated problems in classical electromagnetism Jackson (1998). For the sake of simplicity, we restrict ourselves here on spatial points located on the axis of the tunnel, lying anywhere inside or outside of the tunnel. In this schematic pictorial description, a new amino acid is incorporated into the nascent protein which gets into the tunnel from

one side, conventionally from the right of Fig. 6.4 left panels. The nascent oligopeptide is then pushed by the multi-tasking ribosomal enzymatic functions inside the tunnel and out of the tunnel at the other side (left side of Fig. 6.4 left panels). The movement is strictly asymmetric as the nascent protein always enters the tunnel from the same side with the amino terminal end of the protein getting in first and the carboxyl terminal end of the protein getting in last.

The electrical scalar potential $\Phi(\vec{\mathbf{r}})$ at the observed position $\vec{\mathbf{r}}$ is expressed by:

$$\Phi(\vec{\mathbf{r}}) = \frac{1}{4\pi\epsilon} \int \int_S \frac{\sigma(\vec{\mathbf{r}'})\, da}{|\vec{\mathbf{r}} - \vec{\mathbf{r}'}|} \tag{6.2}$$

where $\sigma(\vec{\mathbf{r}'})$ is the surface-charge density (measured in coulombs per square meter) at position $\vec{\mathbf{r}'}$ of the source, *da* is the two dimensional surface element at $\vec{\mathbf{r}'}$ and ϵ is the permittivity of the dielectric medium (formula 1.23 in Jackson Jackson (1998)) with $\epsilon = \epsilon_r \cdot \epsilon_0$, where ϵ_r is the relative permittivity of the medium and ϵ_0 is the permittivity of free space. We can take advantage of the axial symmetry and restrict to the spatial points on the z-axis, i.e. for $\vec{\mathbf{r}} = (0, 0, z)$. The surface integration is conducted on the support of the source charges. The complete detailed derivations of the electrostatic potential and the axial electric field on the tunnel axis are given in supplementary material for three idealized geometries, i.e the hollow straight cylinder, the cone frustum and the cone frustum concatenated to a hollow cylinder.

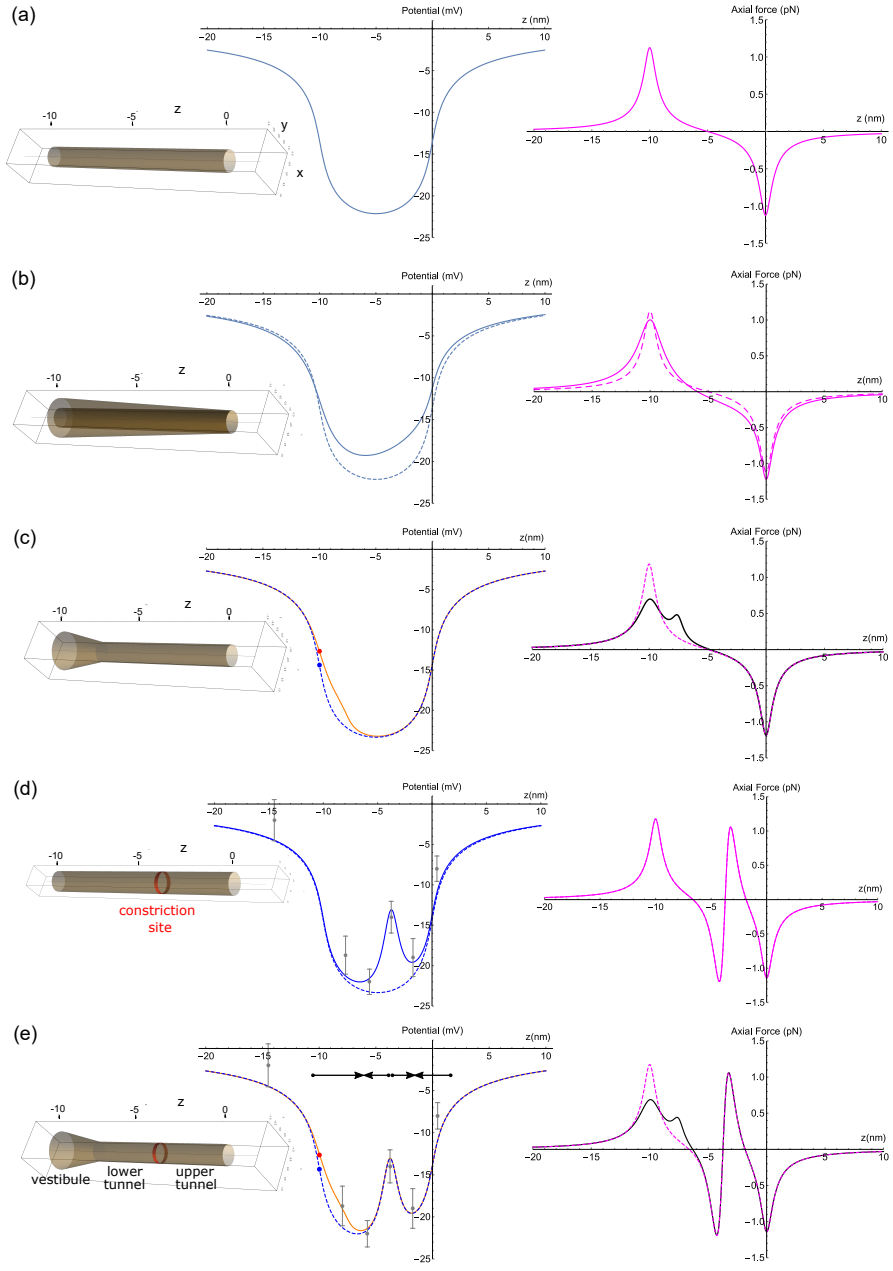


Figure 6.4: caption on next page

Figure 6.4 (previous page): Idealized and realistic ribosome exit tunnel models. (a) Left panel: hollow cylinder of length L = 10 nm and R = 0.5 nm with uniformly charged inner wall; (b) Left panel: normally truncated cone of length L = 10 nm, R_{in} = 0.5 nm and R_{out} = 1.0 nm with uniformly charged inner wall. The entry point is at $z = 0$ and exit point at $z = -L$; (c) Left panel: hollow cylinder of length L_1 = 8 nm and R = 0.5 nm concatenated to a truncated cone of length L_2 = 2 nm with a uniformly charged inner wall. Transition between cylinder and truncated cone at 80% of the tunnel total length: $\lambda = 0.80$. Electrostatic scalar potential (center panel) and axial force (right panel) profiles for a hollow cylinder (a), a cone (b) and cylinder concatenated to a truncated cone (c) dashed lines: cylinder only; full lines: cylinder concatenated to cone. Red and blue points: the potential at the tunnel exit point is higher than the potential at the entry point. The axial forces in the tunnel exit region are smoother and more dispersed in the combined geometry (c) and (e). (d) and (e) Best fitted models for the electrostatic scalar potential of the ribosomal exit tunnel. Center panel: a Lorentzian peak was added locally to the idealized models potential which was fitted to the experimental data points obtained by Lu *et al.* [Lu et al. 2007]. The protein protrusion's position is indicated by the red ring (left panel). Dashed blue line: potential resulting from the idealized uniformly charged hollow cylinder. Orange line: potential resulting from the superposition of a cylinder and a truncated cone. The transition between the cylinder and the truncated cone is at $\lambda = 0.75$ (e). 95% confidence intervals error bars computed from the experimental data. Right panel: the axial forces for a positively unit charged test amino acid residue on the tunnel axis as a function of axial position in the tunnel for the best fitted models. Black line: axial forces profile resulting from the superposition of a cylinder and a truncated cone. Right panel: dashed magenta idealized uniformly charged hollow cylinder. Black arrows (e) indicate the direction of the axial forces upon a positively charged test residue.

6.4.2 Hollow straight cylinder

In a first simplified approach, the ribosome exit tunnel is considered a hollow straight cylinder (Fig. 6.4 (a) left panel). The electrostatic potential is given by the expression:

$$\Phi(z) = \frac{\sigma R}{2\epsilon} \log \frac{\left|\frac{z+L}{R} + \sqrt{(\frac{z+L}{R})^2 + 1}\right|}{\left|\frac{z}{R} + \sqrt{(\frac{z}{R})^2 + 1}\right|} \tag{6.3}$$

and the axial electric field can be written as:

$$E_z(z) = -\frac{\sigma R}{2\epsilon}\Big(\frac{1}{\sqrt{R^2 + (z+L)^2}} - \frac{1}{\sqrt{R^2 + z^2}}\Big). \tag{6.4}$$

The axial force applied on a test particle is the product of the axial electric field with the charge of the test particle:

$$F_z = q \cdot E_z. \tag{6.5}$$

The plots of electrostatic scalar potential $\Phi(z)$ and of the axial force F_z acting on a unit test charge located on the tunnel axis at any point of coordinate z are displayed in Fig. 6.4 (a), with the medium permittivity prevailing inside the ribosome exit tunnel (see below). A negative force means that the test particle is forced to move towards negative z values whereas a positive force means that the test particle is forced to move towards positive z values. In these plots, the ratio σ/ϵ is adjusted so that the potential fits the range of the experimentally measured values given for instance in Lu *et al.* [Lu et al. 2007].

6.4.3 Normally truncated straight cone model

An alternative approach would depict the tunnel as a hollow cone normally truncated at both ends (Fig. 6.4 (b) left panel). The section radius at the entry point is equal to R = 0.5 nm but with a section radius twice that value at the tunnel exit point, and equal to R = 1 nm. With the total axial length kept at L = 10 nm, the half opening angle along the axis is $\alpha \sim 0.05$ radian (2.86 arc degrees) and exactly such that $\tan\alpha = \mathrm{R/L}$ complying with the observation that the diameter at the exit point is around twice the diameter at the entry point of the tunnel. To alleviate the notations, the two following substitutions are adopted:

$$f_1(z) = R\cos\alpha - z\sin\alpha \tag{6.6}$$

$$f_2(z) = R\sin\alpha + z\cos\alpha \tag{6.7}$$

$f_1(z)$ is always positive for $z \leq 0$ (and even for $z < R/\tan\alpha$, i.e. the virtual z position of the cone summit), which is the domain we are interested in. The z position values are negative in the tunnel and beyond its exit point.

The electrostatic potential can be written as:

The electrostatic potential can be written as:

$$\Phi_{cone}(z) = \frac{\sigma}{2\epsilon}\left\{ f_1(z)\cos\alpha \cdot \log\left[\frac{\left| \frac{L}{\cos\alpha} + f_2(z) + \sqrt{z^2 + 2L(z+R) + R^2 + \frac{L^2}{\cos^2\alpha}} \right|}{\left| f_2(z) + \sqrt{R^2+z^2} \right|} \right] \right.$$
$$\left. + \sin\alpha \cdot \left[\sqrt{(z+L)^2 + (R + L\tan\alpha)^2} - \sqrt{R^2+z^2} \right] \right\} \tag{6.8}$$

This last equation (6.8), valid for any conical geometry with entry section of radius R and any cone angle α, replaces equation (6.3) of the cylindrical geometry. Note that the electrostatic potential vanishes at $z = \pm\infty$ as physically expected. It is also worth noticing that equation (6.8) for the truncated cone restores, as a special case, equation (6.3) for the cylinder when $\alpha = 0$, as expected as well.

The axial electric field is given by the expression:

$$E_{z\,cone}(z) = \frac{\sigma}{2\epsilon}\left\{ \sin\alpha\cos\alpha \log \frac{L/\cos\alpha + f_2(z) + \sqrt{z^2 + 2L(z+R) + R^2 + L^2/\cos^2\alpha}}{f_2(z) + \sqrt{R^2+z^2}} + \right.$$
$$\frac{f_1(z)\cos\alpha(\cos\alpha + \frac{z}{\sqrt{R^2+z^2}})}{f_2(z) + \sqrt{R^2+z^2}} -$$
$$f_1(z)\cos\alpha \frac{\cos\alpha + \frac{z+L}{\sqrt{z^2+2L(z+R)+R^2+L^2/\cos^2\alpha}}}{L/\cos\alpha + f_2(z) + \sqrt{z^2+2L(z+R)+R^2+L^2/\cos^2\alpha}}$$
$$\left. - \sin\alpha\left(\frac{z+L}{(z+L)^2 + (R+L\tan\alpha)^2} - \frac{z}{\sqrt{R^2+z^2}} \right) \right\} \tag{6.9}$$

Multiplying eq. (6.9) by a positive unit test charge yields the axial forces acting on a positive unit test charge. The plot of the axial forces as a function of the position in the tunnel is displayed in Fig. 6.4 (b) rightmost panel for the truncated cone geometry and

compared to the cylinder case. Experimental measurements made on ribosome exit tunnels show that the tunnel exit section radius is around 1 nm, i.e. twice the radius of the innermost part of the tunnel. If the ribosome tunnel were of the cone type, the cone opening angle would be around $\alpha \sim 0.05$ radian (2.86 arc degrees). The consequence on the electrostatic potential profile is of importance because, with this conical geometry, and if the total charges are kept the same for the two surfaces, the electrostatic potential inside the tunnel will necessarily be algebraically higher than the potential profile in the case of the cylinder as displayed in Fig. 6.4 (b) (central panel) where the analytical equation for the electrostatic potential for the truncated cone was plotted and compared to the cylinder case. A simple geometrical calculation shows that if the two surfaces support the same total charges $Q_1 = Q_2$, then $\sigma_2 = \mathrm{S}_{\mathrm{cylinder}}/\mathrm{S}_{\mathrm{cone}} \times \sigma_1 = \frac{2}{3} \times \sigma_1$, for a geometry where both tunnels have the same radius at the entry point, the same total lengths L, but where the cone exit section has a radius twice as large as the cylindrical radius. The surface charge density σ_2 on the lateral truncated cone inner surface would be two third of the surface charge density σ_1 prevailing on the lateral inner surface of the cylinder. Moreover, the potential profile in the conical geometry is skewed to the left as compared to the potential profile for the cylindrical geometry. An asymmetry in the potential profile appears due to the change in radius along the z-axis of the cone. The minimal value of the potential is shifted to the left. The slope of the cylindrical potential profile is steeper than the conical potential at the tunnel exit point, meaning that the electric field intensity will be a bit weaker in that region for the conical geometry as can be seen in Fig. 6.4 (b) of the axial forces curves. The axial forces vary more smoothly and are more dispersed in the conical geometry than in the cylindrical geometry.

6.4.4 Normally truncated cone concatenated to a cylinder

The question to know whether or not the tunnel is geometrically exactly more like a cylinder or like a truncated cone is less important than the consequence on the electrostatic potential profile. The salient feature of the real ribosome exit tunnel is that there is indeed a widening in the tunnel section at the exit. Therefore, a better simple geometrical model that fits most of the experimental observations to date is a model combining the cylinder and the normally truncated cone as shown in Fig. 6.4 (c). The transition from a cylindrical shape to a conical shape results in an electrostatic potential rise along the tunnel axis when moving from the entry point to the exit point. This is a fundamental difference between the two geometries (truncated cone combined with cylinder versus cylinder alone) that has both energetic and biological consequences. The electrostatic potential resulting from such a configuration results from the linear combination of the straight cylinder model part and the normally truncated cone model part (supplemental material, linear superposition of integrands in equations (C.3)

and(C.24)):

$$\Phi(z) = \frac{1}{4\pi} \int\int_{K=\{(u,v)\in[0,2\pi]\times[-1,0]\}} \Big(\chi_{[-1,\,-\lambda[}(v) \cdot \frac{\sigma_{\text{cone}}}{\epsilon_{\text{cone}}} \cdot \text{Integrand}_{\text{cone}}(u, v, z + \Delta z_{\text{shift}}) + \chi_{[-\lambda,\,0]}(v) \cdot \frac{\sigma_{\text{cylinder}}}{\epsilon_{\text{cylinder}}} \cdot \text{Integrand}_{\text{cylinder}}(u, v, z) \Big)\, du\, dv \quad (6.10)$$

where the characteristic function $\chi_{[a,\,b]}(v)$, used here for the correct setting of the charged sources distribution, is defined by

$$\chi_{[a,\,b]}(v) = \begin{cases} 1 & \text{if } v \in [a, b] \\ 0 & \text{if } v \notin [a, b] \end{cases} \quad (6.11)$$

A z-shift was also incorporated in equation (6.10) to account for the shift in axial position of the truncated cone, as can be noticed by comparing the left panels in Fig. 6.4 (b) and (c). The z-shift must match the λ value adopted as the interval limit in both characteristic functions in equation (6.10), delineating the limit between the start of the truncated cone and the end of the cylinder when moving axially to the left from $z = 0$. For Fig. 6.4 (c), $z_{\text{shift}} = \lambda \cdot L = 0.8\,L$ (L is the tunnel total axial length). The α angle for the cone may be a free parameter to be determined. Different surface charge densities can be incorporated as well, using two different values for σ in the two integrands, providing an extra degree of freedom to fit the model to the experimental observations. Different permittivities can be incorporated as well, considering some heterogeneity in the tunnel medium dielectric response (see next section). In summary, four parameters can be fitted to the experimental data: σ_{cylinder}, σ_{cone} (if the permittivity is hypothesized to be uniform along the tunnel medium length) or $\sigma_{\text{cylinder}}/\epsilon_{\text{cylinder}}$, $\sigma_{\text{cone}}/\epsilon_{\text{cone}}$ (if a piecewise discontinuity is incorporated for the ratio of the surface charge density over permittivity along the tunnel medium length), λ, and α which are the surface charge density on the cylindrical surface, the surface charge density on the cone frustum, or their quotients over permittivities, the fraction of the ribosome length occupied by the cylinder, and the cone frustum half opening angle respectively. Each of these phenomenological parameters can influence the electrostatic potential profile of the tunnel.

The axial electrical field resulting from the combination of the cylindrical geometry for 75% of the tunnel length ($0.75\,L$) from its entry point and of the truncated cone geometry for the remaining 25% ($0.25\,L$) of the length in the ribosome exit tunnel, with or without an added Lorentzian peak (see 6.4.5), is the superposition of equations (6.4) and (6.9). The parameter settings have to be consistent with the chosen geometry, with the surface charge densities (σ_1 and σ_2) and with the dielectric responses (ϵ_1 and ϵ_2). More specifically, for a given λ, one would have $L_1 = \lambda \cdot L$ and $L_2 = (1 - \lambda) \cdot L$. The surface charge density σ_2 of the truncated cone that we adopted was such that

$\sigma_2/\sigma_1 = S_{1\,\text{lateral}}/S_{2\,\text{lateral}}$ because it best fits the observational data that were gained from the rabbit reticulocytes ribosomes, given a fixed uniform dielectric response ϵ throughout the tunnel. This condition is consistent with the possibility that the conical part of the tunnel end could result from an elastic deformation of an initially cylindrical shaped tunnel with a uniform surface charge density (conservation of total initial charge before and after this hypothetical elastic deformation of the inner surface turning the cylinder into a truncated cone at the exit side of the tunnel). Alternatively, if the dielectric response along the tunnel medium entails piecewise heterogeneity when going from the cylinder to the cone frustum, the previous assumptions of charge conservation and elastic deformation can be relaxed. In the latter case, the fitted parameters are the ratios $(\sigma/\epsilon)_{\text{cylinder}}$ and $(\sigma/\epsilon)_{\text{cone}}$ (or σ_1/ϵ_1 and σ_2/ϵ_2). Under the dielectric response piecewise heterogeneity assumption, the individual numerical values of σ and ϵ cannot be determined separately; only their ratio σ/ϵ is numerically accessible upon fitting to the measured electrostatic potential data points.

The area under the curve of the axial forces profile yields the mechanical energy required for a unit charge to move between two axial points inside the tunnel. Equivalently, the required mechanical energy can easily be computed by multiplying the unit charge with the electrostatic potential difference between the tunnel exit point and the tunnel entry point.

$$W \quad = \quad q \cdot \left(\Phi_{z=-L} - \Phi_{z=0}\right) \tag{6.12}$$

where W is the mechanical work required for a unit test charge q to move across an electrostatic potential difference (in Volt units) from the tunnel entry point ($z = 0$) to the tunnel exit point ($z = -L$). The result is expressed in J/mol units by multiplying equation (6.12) with the Avogadro number.

The geometrical asymmetry induced by the widening in the tunnel radius at the exit of the tunnel is important because it introduces a permanent difference in the electrostatic potential between the exit and the entry points of the tunnel as shown in Fig. 6.4 (b) and (c) center panel.

This is unfavorable to the positively charged amino acid residues that will have to traverse the tunnel and will require more mechanical energy to overcome this electrostatic potential difference than their negatively charged amino acid counterparts, when moving from the entry point $z = 0$ to the exit point $z = -L$. In an adopted geometry that best fits the experimental observations (see section 6.4.5), with $\lambda = 0.75, \alpha = 0.198$ and with $\sigma_2 = 2/3\,\sigma_1$, the potential difference is $-12.65 - (-14.35) = 1.70\,\text{mV}$ and the required mechanical work for traversing the tunnel is $0.164\,\text{kJ/mol} \sim 0.039\,\text{kcal/mol}$ for a positive unit charge embedded in an otherwise neutral nascent peptide stretch. This provides a rough estimate of the energy required for a single positively charged amino acid residue to traverse the ribosomal exit tunnel if this positive residue is embedded in a completely neutral peptide sequence. The mechanical energy requirement for real

sequences depends on the particular distribution of the charged amino acid residues along the primary sequence. There might be particular sequence contexts for which the local mechanical energy requirements could be much higher than the estimated values given above. For a straight cylinder, the mechanical energy is equal to zero (symmetry in the potential between exit and entry points), whereas for the truncated cone concatenated to the cylinder (asymmetry in the potential between the exit and entry points), the mechanical energy uptake when moving a single positive unit charge from the entry point ($z = 0$) to the exit point ($z = -L$) is estimated to be around 164 Nm/mol for a stretch of 40 residues in the tunnel, according to our electrostatic interaction model. The estimated mechanical energy uptake per residue incorporation would be around 4.1 Nm/mol per residue incorporation(~ 0.001 kcal/mol per residue incorporation). This is due to the fact that the axial forces profile is not symmetrical in the cone concatenated to the cylinder geometry. However, in this truncated cone geometry (asymmetrical potential), the axial forces amplitudes are reduced and are more spatially dispersed than in the cylindrical geometry as displayed in Fig. 6.4 (c) right panel.

6.4.5 Empirically improved electrostatic potential profile of cylindrical shape best fitted to experimental point measurements

The real electrostatic potential profile inside the ribosome exit tunnel was experimentally measured with an ingenious biochemical technique of molecular tape at least in rabbit reticulocytes ribosomes by Carol Deutsch and co-workers [Lu et al. 2007]. The measurements were mapped on the X-ray resolved spatial structure of the ribosomal tunnel of the archeon *Haloarcula marismortui*. To the authors' knowledge, the potential profiles in other eukaryotic or prokaryotic ribosomes exit tunnels have not been measured experimentally ever since. The real electrostatic potential profile is actually not symmetric. We further need to build an improved and more realistic potential profile by adding to the previous idealized models a small Lorentzian peak function. The motivation for this comes from the experimental data showing that the electrostatic potential locally increases at a distance one third of the length of the tunnel away from the PTC center (approximately at least 15-17 amino acid residues in the nascent protein upstream from the amino acid residue incorporation site). This local increase in the potential is located near the position of the ribosomal constriction, where specific ribosomal conserved constitutive proteins protrude inward the tunnel, i.e. L4 both in bacteria and eukaryotes, L22 in bacteria and L17 in eukaryotes, see Fig. 6.3 (left). Dao Duc *et al.* Dao Duc et al. 2019 confirmed, with multiple sequence alignments of uL22 and L4 proteins across 20 species in the three domain of life, the presence of a highly conserved sequence enriched in arginine (R) and or lysine (K). In uL22, there are up to 7 R or Ks conserved between position 154 and 176. In L4, there are 5 Rs (or Ks) conserved between position 71 and 92 across eukaryotic species and up to 6 Ks or Rs

conserved between position 69 and 82 across prokaryotic species. Similar conservation has been shown for uL23 (bacteria) and eL39 (eukaryotes) [ibid.]. These positively charged residues protrude near the tunnel constriction and explain the local rise of the potential. The Lorentzian local peak potential as expressed in equation (6.13) that we added was fitted to the experimental data obtained by Lu *et al.* [Lu et al. 2007]

$$\Phi_{\text{Lorentzian}} = \frac{f_{\text{scale}}}{\Gamma/2} \cdot \frac{\Gamma/2}{(z - z_0)^2 + (\Gamma/2)^2} \tag{6.13}$$

The fitted parameters values are $f_{\text{scale}}/(\Gamma/2) = 9\,10^{-9}\,\text{Volt} \cdot \text{m}$ for the scale factor divided by $\Gamma/2$, $\Gamma = 18 \cdot 10^{-10}$ m for the Lorentzian peak full width at half maximum and $z_0 = -3.75\,10^{-9}$ m for the peak center location, i.e. 37.5Å measured from the entry side point towards the protein tunnel exit. The experimental data points taken from Lu *et al.* [ibid.] and the fitted adapted function for the ribosome exit tunnel electrostatic potential are displayed in the center panel of Fig. 6.4(d) when the simple straight cylinder geometry is adopted. The extended expression for the total electrostatic potential in this improved model version is

$$\begin{aligned} \Phi_{\text{total}} &= \Phi_{\text{cylinder}} + \Phi_{\text{Lorentzian}} \\ &= \frac{\sigma R}{2\epsilon} \log \frac{\left|\frac{z+L}{R} + \sqrt{(\frac{z+L}{R})^2 + 1}\right|}{\left|\frac{z}{R} + \sqrt{(\frac{z}{R})^2 + 1}\right|} + \frac{f_{\text{scale}}}{\Gamma/2} \cdot \frac{\Gamma/2}{(z - z_0)^2 + (\Gamma/2)^2} \end{aligned} \tag{6.14}$$

An important characteristic of the Lorentzian function $\Phi_{\text{Lorentzian}}$ that is shared with the potential Φ_{cylinder} is its vanishing at infinity in both directions. At this point, it is important to underline the physical interpretation of the f_{scale} parameter in formula (6.13) and (6.14). Indeed, f_{scale} has units of $\text{Volt} \cdot \text{m}^2$ and can be related to a local σ/ϵ physical value in the constriction site inner volume. A dimensional analysis shows that the f_{scale} units are equivalent to the units of σ/ϵ (Volt/m) times the units of a volume (m^3). Conventionally naming σ_3/ϵ_3, the local ratio of charge density over the dielectric response at the constriction site, we have the following dimensional relation:

$$f_{\text{scale}} \equiv \frac{\sigma_3}{\epsilon_3} \cdot \pi R^2 \Gamma \tag{6.15}$$

where R is the radius of the cylinder and Γ, the full width at half maximum of the Lorentzian peak, is an approximation of the length of the constriction (electrostatic functional length at the constriction site). From this dimensional relation (6.15), we can infer the ratio $\frac{\sigma 3}{\epsilon 3}$ at the constriction site from the above fitted values of f_{scale} and Γ. With $R = 5$ Å, $\sigma_3/\epsilon_3 = 5.73 \cdot 10^9$ Volt/m (or ~ 57.3 MV/cm).

The total axial electric field (and the axial force) is obtained by

$$E_{z,\text{total}} = E_{z,\text{cylinder}} + E_{z,\text{Lorentzian}}$$
$$= -\frac{\sigma R}{2\epsilon}\left(\frac{1}{\sqrt{R^2+(z+L)^2}} - \frac{1}{\sqrt{R^2+z^2}}\right) + \frac{f_{\text{scale}}}{\Gamma/2} \cdot \frac{2\,(\Gamma/2)\,(z-z_0)}{\left((z-z_0)^2+(\Gamma/2)^2\right)^2} \quad (6.16)$$

from which the axial force results immediately by $F_z = q \cdot E_z$ and is displayed in Fig. 6.4 (d) right panel in the case of a single positively unit charged amino acid residue.

The charge surface density σ can be made dependent on the z variable in a stepwise manner to account for local heterogeneity on the ribosome wall and to account for the experimentally observed potential profile. The dielectric constant ϵ or the ratio σ/ϵ can also be made dependent on the z variable in a stepwise manner.

6.4.6 Empirically improved electrostatic potential profile of truncated cone combined with cylindrical shape best fitted to experimental point measurements

A still better fit of the experimental data of Deutsch and co-workers [Lu et al. 2007] is obtained with the truncated cone concatenated to the cylinder geometry. Keeping the same Lorentzian peak, the best extended expression for the total electrostatic potential

in this last improved version of the model is

$$\Phi_{\text{total}} = \Phi_{\text{cone}}(z+\lambda L, L_2, \sigma_2) + \Phi_{\text{cylinder}}(z, L_1, \sigma_1) + \Phi_{\text{Lorentzian}}(z)$$

$$= \frac{\sigma_2}{2\epsilon_2}\Bigg\{ f_1\big(z+L_1\big)\cos\alpha\cdot$$

$$\log\Bigg[\frac{|L_2/\cos\alpha + f_2(z+L_1) + \sqrt{(z+L_1)^2 + 2L_2\Big((z+L_1)+R\Big) + R^2 + \frac{L_2^2}{\cos^2\alpha}}|}{|f_2(z+L_1) + \sqrt{R^2+(z+L_1)^2}|}\Bigg] +$$

$$\sin\alpha\cdot\Big[\sqrt{\Big((z+L_1)+L_2\Big)^2 + (R+L_2\tan\alpha)^2} - \sqrt{R^2+(z+L_1)^2}\Big]\Bigg\} +$$

$$\frac{\sigma_1 R}{2\epsilon_1}\log\frac{\Big|\frac{z+L_1}{R} + \sqrt{(\frac{z+L_1}{R})^2+1}\Big|}{\Big|\frac{z}{R} + \sqrt{(\frac{z}{R})^2+1}\Big|} + \frac{f_{\text{scale}}}{\Gamma/2}\cdot\frac{\Gamma/2}{(z-z_0)^2+(\Gamma/2)^2} \quad (6.17)$$

where $L_1 = \lambda\cdot L$, $L_2 = (1-\lambda)\cdot L$, $L = L_1 + L_2$.

Similarly for the axial electric field (and the axial force) along the $z-$axis

$$E_{z,\text{total}} = E_{z,\text{cone}}(z+\lambda L, \sigma_2, L_2) + E_{z,\text{cylinder}}(z, \sigma_1, L_1) + E_{z,\text{Lorentzian}}(z)$$

$$= E_{z,\text{cone}}(z+L_1, \sigma_2, L_2) - \frac{\sigma_1 R}{2\epsilon_1}\Big(\frac{1}{\sqrt{R^2+(z+L_1)^2}} - \frac{1}{\sqrt{R^2+z^2}}\Big)$$

$$+ \frac{f_{\text{scale}}}{\Gamma/2}\cdot\frac{2\,(\Gamma/2)\,(z-z_0)}{\Big((z-z_0)^2+(\Gamma/2)^2\Big)^2} \quad (6.18)$$

where the detailed expression of the first term in the last right hand side is easily obtained from equation (6.9) by substituting $z+\lambda L$ to L, σ_2 to σ and L_2 to L. The plots of the electrostatic potential and of the total axial force profiles in the ribosome exit tunnel under this last improvement of the model are displayed in Fig. 6.4 (e) central and right panels. The central panel shows the goodness of the fit with the experimental data of Lu *et al.* [ibid.]. The improvement of the fit due to the cone geometry concatenated to the hollow cylinder is worth noticing (orange line in the central panel of Fig. 6.4(e)). This last version of the model perfectly fits the four experimental points located inside the tunnel. The model is a very good fit of all the

6 experimental data points if the tunnel length is taken in a range from $L = 8.5\,\text{nm}$ to $L = 9.5\,\text{nm}$, keeping all the other parameters constant. Fig. 6.5 shows the plots for the potential curves for these two tunnel lengths boundaries and shows that the 6 experimental measurements are correctly captured within the 95% confidence intervals of the potential measurements between these two length boundaries. In their study, Lu *et al.* mapped their 6 experimental points on the ribosomal crystal structure of *Haloarcula marismortui* (archae) for which the X-ray or cryo-electron microscopy resolved ribosome structure gives a tunnel length of $\sim 9.5\,\text{nm}$ [Dao Duc et al. 2019]. It is recognized that there might be some deviations in mapping distance and with respect to the actual length of the ribosome exit tunnel of the biological material that they used. Also, the actual *in vivo* lengths might slightly differ from the lengths determined in the cryogenic conditions prevailing in cryo-electron microscopy and the functional length might also slightly differ to the geometrical length. The X-ray and cryo-electron microscopy ribosome structure resolution conducted on 23 species across the three domain of life are supporting the fact that the ribosome exit tunnel in bacteria is a bit longer than the one in eukaryotes while archea have intermediate lengths between bacteria and eukaryotes [ibid.]. Hence, a tunnel length of $L \approx 8.5\,\text{nm}$, should be adopted if we aim at building a model for the eukaryotic/mammalian ribosome exit tunnel, that could be eventually used for computational biology and bioinformatics purposes on eukaryotic/mammalian omics data. The plot of the electrostatic potential in a ribosome exit tunnel with a length of 8.5 nm and with a variable angle of the cone frustum at the tunnel exit is shown in the animated figure Fig. AF1, provided in the supplemental material of the published paper.

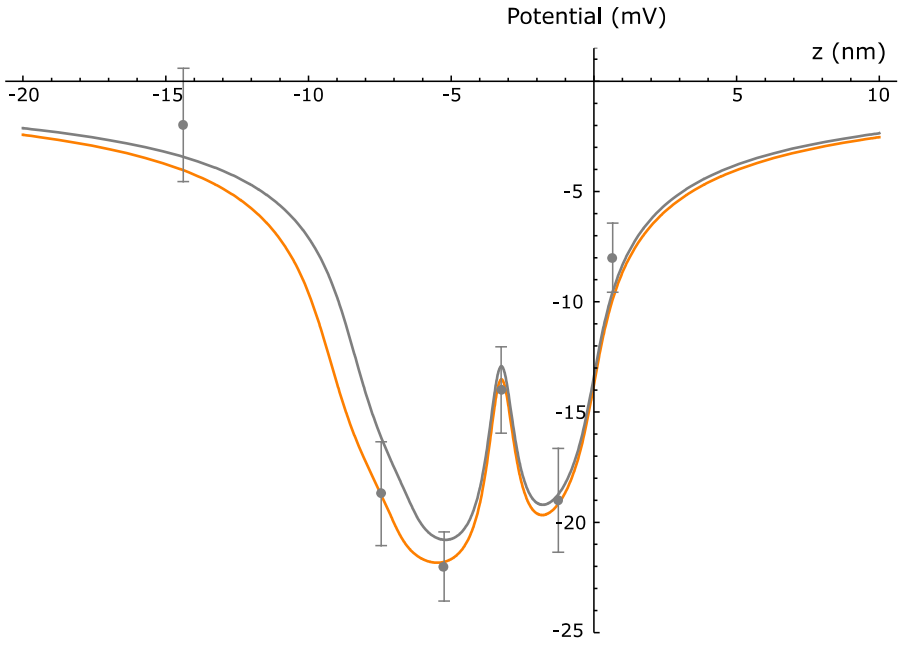


Figure 6.5: Electrostatic scalar potential curves for two tunnel length boundaries. Tunnel length = 9.5 nm: orange curve. Tunnel length = 8.5 nm: gray curve. Both curves capture the 6 experimental points within the 95% confidence interval of the potential measurements that were mapped on the ribosomal crystal structure of *H. marismortui* by Lu *et al* [Lu et al. 2007].

6.5 PRE models assumptions

The main simplifying assumptions related to the physical properties of the tunnel lumen and inner wall on which the models are based are:

- Uniform distribution of the charges with possible different values for the surface charge densities (σ) in the different tunnel inner wall parts (upper tunnel, constriction site, lower tunnel and tunnel vestibule). This is the piecewise uniform surface charge density of the tunnel inner wall assumption.
- Piecewise continuous constant dielectric responses (ϵ) in the medium prevailing inside the tunnel (upper tunnel, constriction site, lower tunnel and tunnel vestibule). This assumption combined with the previous one is equivalent to the piecewise constant assumption for the ratio σ/ϵ. A space dependence of σ is possible if it is compensated for by a similar space dependence of ϵ provided the combined ratio σ/ϵ is constant in a region of interest. This piecewise constant ratio is the strictly necessary assumption for the mathematical surface integration calculations of our models to be analytically tractable.
- Polarization surface charge density effects due to media discontinuities which are implicitly incorporated as a net apparent surface charge density σ.
- Existence of a strong water screening effect inside the tunnel wall (and not due to mobile ions) which results in the substitution of a *renormalized apparent* surface charge density (σ) to the *bare* surface charge density (σ^*).

These assumptions are studied in this section. In order to show how reasonable or realistic these assumptions are, X-ray crystallographic data of a real ribosomal large subunit were analysed and a 2D map at atomic resolution of the tunnel inner wall was built. The model assumptions were confronted to the crystallographic observations. We first present hereafter the tunnel mapping of the X-ray crystallographic solved atomic structure of the large ribosomal subunit.

6.5.1 Mapping the negatively charged phosphate groups close to the tunnel inner wall and counting the number of charged groups contributing to the formal surface charge density of the ribosome tunnel inner wall from a X-ray solved structure of the large ribosomal subunit

We analyzed the publicly available structure of the large ribosomal subunit of the archeon *Haloarcula marismortui* (PDB code: `4V9F` downloaded from

`https://www.rcsb.org/`) obtained from X-ray crystallography at 2.4 Å [Gabdulkhakov et al. 2013]. Other ribosomal large subunit structures are available across the three domains of life [Dao Duc et al. 2019], but we chose **`4V9F`** because it was on *Haloarcula marismortui* ribosomal structure that Carol Deutsch and co-workers mapped on their data points measurement of the potential and also because this PDB entry is one of the best resolution of the ribosomal large subunit to date. The structure has an atom count of 103, 831 (non-hydrogen atoms) and includes 2, 808 nucleotides of the 2, 922 nucleotides in the **`23S RNA`** , all of the 122 nucleotides in the **`5S RNA`** and 4, 311 amino-acid residues from 32 ribosomal proteins, as well as 261 ions (138 Mg^{2+}, 30 Cl^{-1}, 85 Na^{+}, three K^{+}, five Cd^{2+}) and 7, 745 water molecules. The structure has been refined to an R factor of 0.166 and an R free of 0.206. The refinement statistics indicate root mean square deviations of bond lengths of 0.006 Å and bond angles of 1.08°. It is estimated, that with this 2.4 Å resolution and a R-factor smaller than 0.2, the upper limit precision on the atoms position is around 0.2 Å for **`4V9F`**. We can therefore consider that atom positions are precise to within one-twelfth of the stated resolution [Rhodes (2006)]. The atom positions are precise to within one-eighth (12.5%) of the length of a phosphorus-oxygen bond in the phosphodiester bond between a phosphorus atom and the bridging oxygen of the `O3'`-ribose or `O5'`-ribose in the nucleotides.

To find the ribosome exit tunnel and extract the tunnel centerline coordinates, we used a tunnel search algorithm developed by Sehnal *et al* [Sehnal et al. 2013], implemented in MOLE 2.0 and the web-based MOLEonline 2.0 tool publicly available online [Berka et al. 2012; Pravda et al. 2018]. We used PyMOL (The PyMOL Molecular Graphics System, Version 2.3.2) and exported the relevant selected atom positions cartesian coordinates to output files. These files were further processed with mapping algorithms coded in Python to build 2D positional maps of the charged chemical groups on or near the inner surface of the ribosome exit tunnel, as viewed from the tunnel centerline. The protocol to build the 2D maps is detailed in supplemental material. We algorithmically set out the 3D equations of the cylinder and the cone frustum in a reference frame aligned with the centerline of the real X-ray solved structure, to calculate the closest distance of the selected atoms to the surface of the tunnel model. The Δ values shown in blue in Fig.6.6 were obtained as the closest (orthonormal) distance of the PyMOL selected atoms either from the model cylinder surface or from the model cone surface, depending on where the atoms are located. The 2D maps of the selected atoms within a 12 Å of the tunnel centerline, projected from the 3D space surfaces of the tunnel inner wall model as shown in Fig.6.6 allow to determine the *formal bare* charge surface densities σ^{*}. The total phosphorus atom counts is 94 on or near the tunnel model surface, 56 of which are close to the cylinder surface and 38 close to the vestibule surface. Among the 56 phosphorus atoms in the immediate vicinity of the model cylinder surface, 23 are in the upper tunnel, 10 near the constriction site and 23 in the lower tunnel. We tested the null hypothesis of a uniform distribution of the phosphorus atoms across the cylinder surface. We calculated the p-value by 100, 000 Monte-Carlo (or bootstrap) simulations of sets of 56 atoms that were uniformly distributed across

the cylinder surface and counted how many of these sets entailed atom counts in the three regions that were equal or more extreme than the actual observed combination of 23, 10, 23. The p-value to observe the mapped distribution or a more extreme one across the three cylindrical developed areas, i.e. upper tunnel, constriction site and lower tunnel, is larger than 0.083. The assumption of a uniform 2D-joint distribution of the negatively charged phosphate moieties in the immediate vicinity of the tunnel cylinder wall is reasonable and can not be ruled out. We also tested the null hypothesis of uniformity of the 1D-marginal distributions along the z-axis or along the ϕ- arc angle with the chi-squared and the Kolmogorov-Smirnov tests (1D marginal along ϕ: χ^2-test p-value= 0.549; 1D marginal along z: χ^2-test p-value= 0.487, KS-test p-value= 0.41). Therefore, we concluded that the formal bare charge distribution contributed by the phosphate moieties on the cylinder is $\sigma_1^* = -2.1\,|\mathrm{e}|/\mathrm{nm}^2$ and that this value can be retained for the upper cylinder, the constriction site and the lower tunnel altogether. Fig.6.6(b) shows the 56 phosphorus atoms with an orthogonal distance to the cylinder surface of less than 12 Å. The average distance of a phosphorus atom to the model tunnel inner wall is 6.68 Å. The surface charge density on the $7.45\,\mathrm{nm}^2$ of the model tunnel surface of the vestibule is $\sigma_2^* = -5.1\,|\mathrm{e}|/\mathrm{nm}^2$. The vestibule surface appears to be more enriched in phosphate moieties near the end edges of the tunnel. The 2D-joint distribution of the phosphorus atoms across the cone surface cannot be considered uniform, p-value <= 0.009. This p-value was calculated using two methods, i.e. the rejection-sampling method and the method of root squared random variable to take into account the increase of the surface elements on a cone surface when moving from the small to the large radius of the cone. We also tested the null hypothesis of uniformity of the 1D-marginal distributions along the z-axis or along the ϕ- arc angle for the cone with the chi-squared and the Kolmogorov-Smirnov tests (1D marginal along ϕ: χ^2-test p-value= 0.0847; 1D marginal along z: χ^2-test p-value= 0.00012, KS-test p-value= 0.008), where we also included the non-linearity of the surface metric element of the cone. Altogether, these results show that in the truncated cone, the surface charge density $\sigma_{\mathrm{cone}} = \sigma_{\mathrm{cone}}(z)$ depends on z. Around $50 - 60\%$ of the phosphate moieties of the cone surface are at the edge of the exit port, in the last 3 interval $z = [-101.80, -98.80]$ of the cone, showing a highly significant enrichment of the phosphates at the exit port where the tunnel radius is the largest.

The potential created by the observed structural distribution of charges as mapped on Fig.6.6 along the tunnel centerline is computed in supplemental material to see how it compares with the inverse bell shape results from a uniformly charged tunnel. The constraints of the observed potential data point measurements on the heterogeneity of the phenomenological parameters (dielectric response and electrostatic screening lengths) are also studied and compared in supplemental material subsection C.5.

Fig.6.6(c) and (d) show the maps of the positions of the charged atoms carried by the charged amino acid residues R, K in red and D, E in blue within a distance of 8 Å from the model wall surface. The markers' shapes determine which ribosomal protein these

amino acids belong to: squared markers for ribosomal protein uL22, stared markers for uL4, filled circled markers for eL39 and L24. The threshold distance of 8 Å was adopted for the amino acid residue selection because it is the largest atomic Euclidean length between the side chain distal charged nitrogen (called NH2 in the atomic model PDB format) and the proximal carboxylic carbon in the peptide bond for the arginine amino acid residue. Extending the threshold distance to 12 Å for the amino acid does not affect the maps significantly (not shown here). Fig.6.9 (a) and (b) highlights the relative positions of these 7 positively charged residues with respect to the tunnel surface. There are no charged amino acid closer than 8 Å to the tunnel surface in the upper tunnel. The salient feature of the constriction site is that there are exactly 7 positively charged residues belonging to uL4 (Arg65, Lys72, Arg76, Arg78) and uL22 (Arg125, Arg128, Arg132), with their center of charges at z position = −40.53 Å which is only 3 Å away of $z_0 = -37.5$ Å which was the independently fitted location of the Lorentzian peak for the electrostatic potential (see 6.4.5). There is only a single Arg35(eL39) in the lower tunnel. The average closest distance of the charged atoms from the cylinder tunnel wall is 4.56 Å.

All the 5 positively charged amino acid residues within close distance of the vestibule tunnel wall are located near the end edge of the cone, with two of them in an enriched hydrophilic region, i.e. Arg31(eL39) and Lys81(L24).

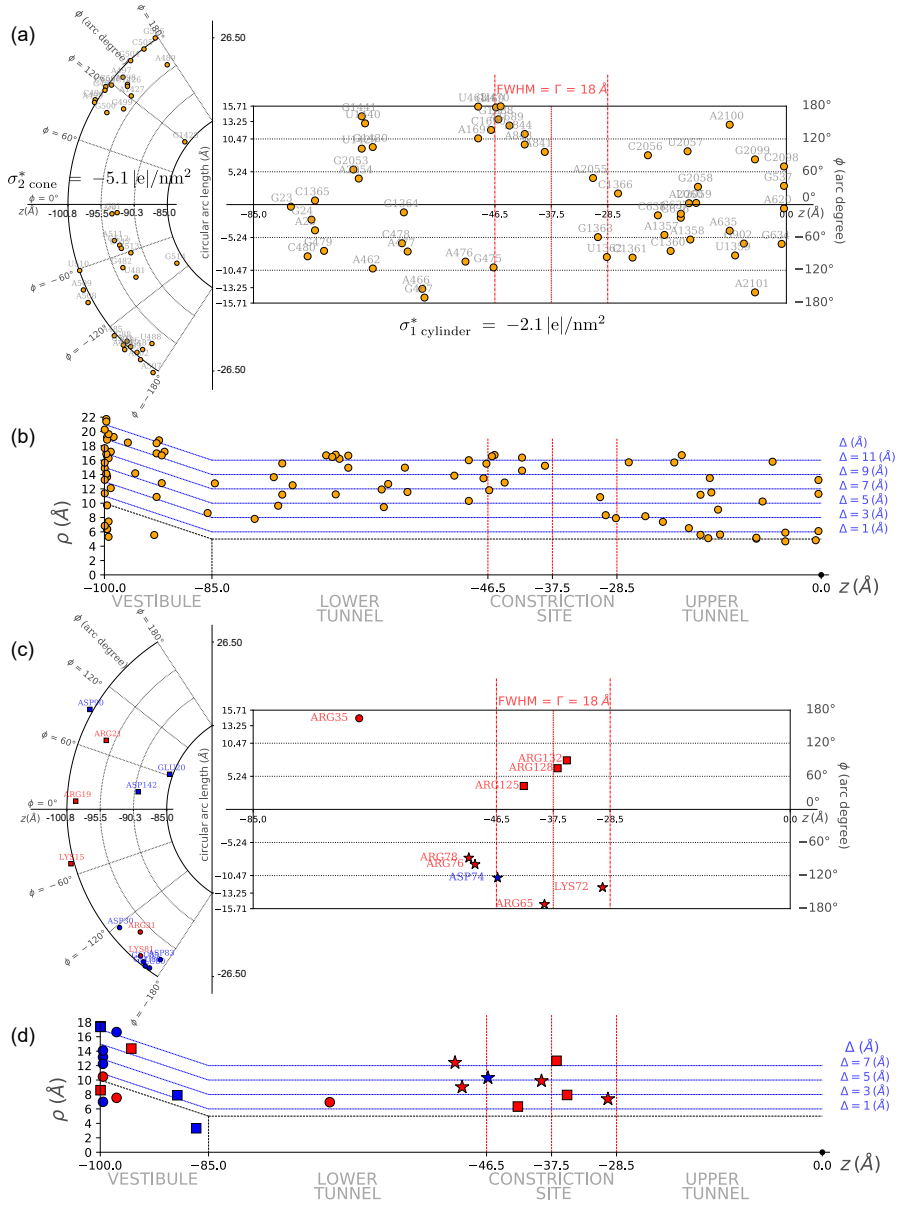


Figure 6.6: caption on next page

Figure 6.6 (previous page): 2D Mapping of phosphorus atoms and charged amino acid residues close to the ribosome exit tunnel wall surface. Calculated wall surfaces: 8.95 nm^2 (upper tunnel), 5.65 nm^2 (constriction site), 12.09 nm^2 (lower tunnel) and 7.45 nm^2 (vestibule). Position of the constriction site with full width at half maximum of the Lorentzian peak (dashed red vertical lines). (a) 2D mapping of phosphorus atoms in the phosphate moieties (orange disks) on the tunnel cylindrical surface (right) and tunnel vestibule cone frustum surface (left). Bare surface charge density on cylindrical surface $\sigma_1^* = -2.1|e|/nm^2$. Bare surface charge density on cone frustum surface $\sigma_{2(cone)}^* = -5.1|e|/nm^2$. Horizontal axis: z distance along tunnel centerline, vertical axis and arc axis: angle to atom line of sight as viewed from centerline (see text for orientation reference frame). (b) Mapping of the phosphorus atoms (orange disks) closest to the tunnel centerline. Horizontal axis: z distance along tunnel centerline, vertical axis: radial distance from tunnel centerline. (c) 2D mapping of positively charged (red squares) and negatively charged (blue stars) amino acid residues on the tunnel cylindrical surface (right) and tunnel vestibule cone frustum surface (left). (d) Mapping of the positively charged (red squares) and negatively charged (blue stars) amino acid residues closest to the tunnel centerline. Note the enrichment of 7 positively charged amino acids at the constriction site.

6.5.2 Dielectric response of the medium prevailing in the tunnel and mobile ions

The medium inside the ribosome wall is of course a dielectric and not a conductor and not free space (vacuum). In formula (6.3), we see that the potential on the cylinder axis depends on the geometry and surface charge density. The ϵ parameter is the medium dielectric permittivity with $\epsilon = \epsilon_r \cdot \epsilon_0$. The ϵ_0 parameter is the vacuum's permittivity: $8.854\ 10^{-12}$ Farad/m. The formula applies to the real dielectric medium prevailing inside the ribosome exit tunnel. The ionic strength inside the tunnel and the electric polarizability of all the molecules inside the tunnel are also expected to play a role that is further discussed below.

The volume of a cylinder of length 85 Å and radius 5 Å is $V_{cyl} = 6,676$ Å^3. The volume of the cone frustum (tunnel vestibule) is $V_{cone} = 3,312$ Å^3. The total volume inside the tunnel is $V_{tot} = 9,988$ Å^3. The volume of an individual water molecule is estimated to be $v_W = 29.9$ Å^3 [Sansom et al. 1997]. Hence, the estimated maximum number of water molecules that could be accommodated in an otherwise empty ribosome exit tunnel is around 333 (223 water molecules in the cylinder and 110 in the cone frustum). In their paper, Dao Duc *et al.* [Dao Duc et al. 2019] estimated the volume of the exit tunnel in eukarya to be on average $27.8\ 10^4$Å^3, which is 2.8 time as much as our calculation. This would amount to a maximum number of water molecules of ~ 930 in an otherwise empty exit tunnel (including the large cavity around the PTC). In their paper, Dao Duc *et al.* included the volume of the large cavity between the PTC and the tunnel entry point and this might be the origin of the discrepancy with our estimation. Besides, the tunnel searching algorithm being used may or may not remove solvent water molecules and does not discriminate between the mobile water molecules and the wall constitutive water molecules, increasing the tunnel apparent measured volume. Otherwise the tunnel length and radius averaged values in Kao Duc *et al.* paper are very similar as the ones used in our calculation. When a nascent polypeptide is progressing through the exit tunnel, the volume left for mobile water molecules is significantly reduced. So, the real number of water molecules present in the tunnel during polypeptide elongation is believed to be of the order of ~ 100 and most of them probably present in the vestibule region (tunnel exit end). Water as a bulk solvent has a relative electric permittivity of 78 (25°C) or 74 (37°C), while experimental and theoretical evidence suggest that proteins (or the nascent polypeptide) have an average dielectric response that can be approximated with a dielectric constant of about 3-4 [Sharp and Honig 1990 and references therein]. The dielectric constant of nucleic acids like DNA in bulk solution has been measured to be around 8 [Cuervo et al. 2014]. Thus, depending on the abundance of water molecules in the tunnel volume accommodating the nascent chain, the tunnel micro-environment cannot be viewed as uniform. At least two dielectric constants ϵ should be used in a range from $\epsilon = 3 - 4$ (polypeptide) to $\epsilon = 78$ (water). It follows that the dielectric medium permittivity is

not uniform along the tunnel axis nor across the tunnel wall and is not necessarily equal to the one of water in bulk solution: ϵ could be higher in the vestibule region (more water molecules) than in the narrower section of the tunnel (less water molecules). In our model settings, two different values can be used for ϵ: one in the cylinder section (upper and lower tunnel) and another one in the vestibule region. A third permittivity value can also be implicitly used as well in the tunnel constriction through the scale factor f_{scale} (see 6.4.5).

Ionic strength and mobile ions effects are very important to study the behavior of biological macromolecules in aqueous *bulk solution at equilibrium* [Sharp and Honig 1990]. The inside of the ribosome exit tunnel is however too narrow to be considered as a plain aqueous bulk solution. The inside of the ribosome exit tunnel is also far from equilibrium during the dynamical process of protein elongation. The hydrated radius of magnesium Mg^{2+} (naked ionic radius of 0.65Å but with octahedral coordination geometry in $Mg(H_2O)_6^{2+}$), calcium Ca^{2+} potassium K^+ and sodium Na^+ ion in bulk aqueous solution at equilibrium are 0.7 – 0.8 nm, 0.6 nm, 0.3 nm, 0.4 – 0.45 nm respectively [Harris (1988)]. The ribosome exit tunnel radius is $\sim$ 0.5 nm in the cylinder but the tunnel vestibule exit opening is 1 nm in radius. Only monovalent ions and possibly magnesium that has a great affinity for phosphate oxygens could be accommodated in the tunnel cylinder region and probably only in the absence of a nascent polypeptide growing inside the tunnel. If there were mobile ions, they would be expected only in the vestibule region of the tunnel. A number of publications have addressed the issue of the presence of metal ions inside the ribosome and particularly underscored the importance and roles of magnesium and potassium as phosphates counter ions stabilizing secondary and tertiary structures of both rRNA and tRNAs respectively [Hsiao and Williams 2009; Klein et al. 2004; Nierhaus 2014; Petrov et al. 2012]. Klein *et al.* numbered 116 Mg^{2+} ions in total inside the 50S large ribosomal subunit of *Haloarcula marismortui* and 88 monovalent ions in total [Klein et al. 2004]. Most of these metal ions usually bind to the nonbridging phosphate oxygen atoms of the RNA phosphodiester backbone or to specific parts of RNA secondary structure motifs or to highly idiosyncratic sites in 23S RNA. This number of metal ions is at least one order of magnitude lower than the number of phosphate moieties in the rRNAs encountered in the 50/60S ribosomal subunit. We could not find supporting published evidence of the presence of (mobile or free) magnesium or potassium ions in abundance inside the upper or lower tunnel during protein synthesis *in vivo*. Furthermore, the very notion of ionic concentration or ionic strength is less clear in the tunnel confined environment where the total number of water molecules is limited and of magnitude order of a hundred molecules.

Moreover, as we analyzed above in 6.5.1 with PyMOL from the PDB entry id `4V9F` of the X-ray solved structure of the 50S subunit of *Haloarcula marismortui* at 2.4 Å [Gabdulkhakov et al. 2013], the total number of water molecules was 7, 745 in the 50S structure. The number of water molecules found inside the tunnel was 44. The number

of Mg atoms in the 50S subunit was 138, the number of Mg atoms inside the tunnel was 3 at a distance of 12 Å from the tunnel centerline. The total number of Na atoms was 85, 3 of which were found at a distance of 12 Å from the tunnel centerline. The total number of potassium atoms was 3, none of which in or near the tunnel wall. Altogether, the crystallographic data do not support evidence of the presence of free mobile ions in number inside the ribosome exit tunnel.

During the peptide bond formation in the catalytic center of the large ribosomal subunit at the peptidyltransferase center (PTC), the precise positioning [Rodnina et al. 2006; Simonovic and Steitz 2009] of the local rRNA, peptidyl-tRNA at the P-site, aminoacyl-tRNA at the A-site and of a possible single solvatation water molecule involved in the proton shuttle occurring during the catalytic deacylation and transpeptidation, the introduction of water molecules and or (large) ions such as Mg^{2+} or K^{+} or even Na^{+} is sterically hindered (if not impossible). Dynamically, during the peptide elongation process and the progress of the nascent chain into the ribosome exit tunnel from the PTC, it is expected that only aminoacid residues can enter the tunnel and there is hardly space left for third party water molecules or large ions. It is believed that only small H^{+} or OH^{-} (having the highest electrochemical mobility) could be the counter ions accompanying charged amino acid residues. In any case, the elongation process itself prevents the environment inside the ribosome exit tunnel to be in equilibrium with external aqueous ions. Due to the dynamics of elongation, the ribosome exit tunnel interior is a medium far from equilibrium. For all these reasons, our model neglected such effects as mobile ions and ionic strength. It is possible however that mobile ions and ionic strength effects could be present in the vestibule (exit end of the tunnel). The Poisson-Boltzmann theory of charged macromolecules in bulk solution with strong electrolytes or its Debye-Hückel linearized version are not fully appropriate in the context of the ribosomal exit tunnel as their basic assumptions are not met in the very confined micro-environment of the tunnel.

However, a strong screening effect due to the permanent dipole moments of the water molecules that are buried inside the tunnel wall medium and close to the phosphate moieties has to be taken into consideration as will be studied under the Gouy-Chapman approach and shown in Fig. 6.8 (d) below.

6.5.3 Polarization surface charge densities on the tunnel wall due to discontinuity between dielectric media: image charge effect with dielectrics

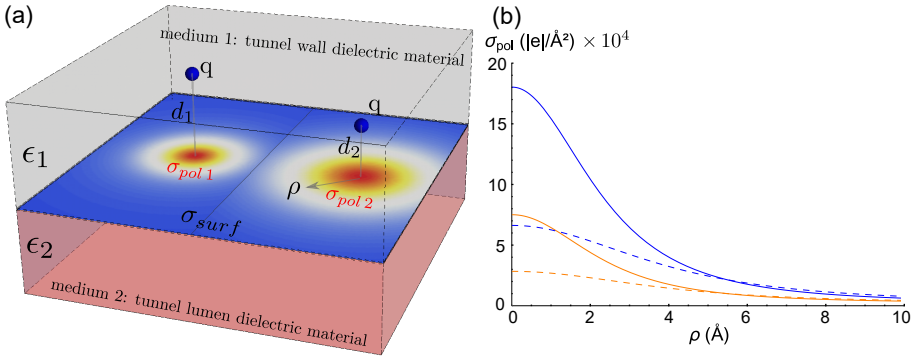


Figure 6.7: Polarization surface charge densities on tunnel wall due to discontinuity between dielectric media. (a) Two charges (blue spheres) buried in dielectric medium 1 and their polarization effects at the boundary with dielectric medium 2. (b) Surface polarization charge density σ_{pol} as a function of the charge burial depth d in medium 1 and the tangential distance ρ at the interface. Significant dielectric discontinuity: $\epsilon_1 = 8$ (rRNA) and $\epsilon_2 = 78$ (water) for $d_2 = 3$ Å (blue line), $d_1 = 5$ Å (dashed blue). Weak dielectric discontinuity: $\epsilon_1 = 8$ (rRNA) and $\epsilon_2 = 4$ (polypeptide), $d_2 = 3$ Å (orange line), $d_1 = 5$ Å (dashed orange).

Polarization charge surface densities σ_{pol} appear at media discontinuities between two material with different dielectric constants. This is a well known classical effect in electrostatics that can be treated in the framework of boundary-value problems with multiple dielectrics, see Jackson's textbook chap.4.4 p.154 [Jackson (1998)]. The polarization charge density appearing on the surface boundary separating the two dielectric media is expressed in formula 4.47 in reference [ibid.]:

$$\sigma_{pol} = -\frac{q}{2\pi}\frac{\epsilon_0(\epsilon_2-\epsilon_1)}{\epsilon_1(\epsilon_2+\epsilon_1)}\frac{d}{(\rho^2+d^2)^{3/2}} \tag{6.19}$$

We consider here a radial discontinuity at the tunnel's inner surface, with dielectric constant ϵ_2 inside the lumen of the tunnel and dielectric constant ϵ_1 for the tunnel wall material as shown in Fig.6.7(a). When applying formula (6.19), a fixed phosphate charge moiety is supposed to be in medium 1 at distance d from the boundary surface separating the media as shown in Fig.6.7(a). If the fixed charged phosphates moieties are deeply buried in the tunnel inner wall, two situations should be discussed depending

on the dielectric response differences between the media and the importance of the discontinuity.

1. Case one: $\epsilon_2 = \epsilon_{water} = 78$ and $\epsilon_1 = \epsilon_{rRNA} \approx 8$. The discontinuity is significant. $\epsilon_2 \gg \epsilon_1$, the dielectric ϵ_2 (tunnel lumen) behaves much like a conductor in that the electric field inside it becomes very small and the surface charge density approaches the value appropriate to a conducting surface, apart from a factor of ϵ_0/ϵ_1. The polarization surface charge density has a sign opposed to the buried source fixed charge. The polarization charge partially screens the buried fixed charge.
$$\sigma_{pol}(d, \rho = 0) = -\frac{q}{2\pi}\frac{\epsilon_0}{\epsilon_1}\left(\frac{70}{86}\right)\frac{1}{d^2}. \tag{6.20}$$
2. Case two: $\epsilon_2 = \epsilon_{peptide\ nascent\ chain} \approx 4$ and $\epsilon_1 = \epsilon_{rRNA} \approx 8$. The discontinuity is smoother than in the previous case. The polarization surface charge density that appears at the boundary is 2.44 times smaller in absolute value than in the previous case.
$$\sigma_{pol}(d, \rho = 0) = -\frac{q}{2\pi}\frac{\epsilon_0}{\epsilon_1}\left(\frac{-4}{12}\right)\frac{1}{d^2} \tag{6.21}$$

Finally, if the charged phosphate moieties are localized exactly on the surface of the inner wall ($d = 0$), no polarization appears and $\sigma_{pol} = 0$. Our model relies on the assumption that the negative charges of the phosphate moieties are indeed exposed in numbers at the immediate vicinity of the surface of the wall and are carried by the non-bridging oxygen atoms in the phosphodiester bonds between riboses. A fraction of the inner negative charges buried deeper inside the rRNA material are neutralized by constitutive metal ions like magnesium and sodium as discussed previously or by charged residues of ribosomal proteins. The rest of the inner negative charges buried inside the rRNA that are possibly fixed and not screened directly by counter cations (positively charged amino acid residues from ribosomal proteins or metal cations) will generate local polarization charge densities on the boundary surface between the two dielectric media as calculated above. This polarization due to dielectric discontinuities contributes to the screening of the phosphate moieties.

Media discontinuity surface polarization effects do occur in the ribosome exit tunnel. They are believed to be local and would not affect, on a large scale, the profile of the electrostatic potential along the tunnel length. Note that when moving tangentially along the boundary surface, the polarization surface charge density decreases according to the third power of the tangential distance ρ. Polarization due to dielectric discontinuities is local and fades away rapidly: σ_{pol} is 7.34 times smaller at $\rho = 5$ Å than at $\rho = 0$ Å, see Fig. 6.7(b).

We did not explicitly incorporate in our simplified electrostatic model such medium discontinuity surface polarization effects. We took the approach of estimating an

apparent or *renormalized* fixed negative charge surface density on the tunnel wall (see the Gouy-Chapman screening length approach in next subsection). It is not established whether this net surface charge density σ results from the direct contribution of fixed negative charges directly lying on the surface, i.e. $\sigma_{\text{surf.}}$ or from surface polarization effects due to dielectric media discontinuities and due to the charges buried deeper inside the wall material, i.e. $\sigma_{\text{pol.}}$. The charged surface density σ that is explicitly used in all our formulas should be considered as a resulting net *apparent* surface charge density $\sigma_{\text{app.}}$ where

$$\sigma = \sigma_{\text{app.}} = \sigma_{\text{surf.}} + \sigma_{\text{pol.}} \tag{6.22}$$

Interestingly, a decrease in the ratio σ/ϵ in the tunnel vestibule (cone frustum) as compared to the ratio σ/ϵ in the lower or upper tunnel (cylinder) could result from a stronger discontinuity in the dielectric constants between the tunnel inner wall material and the tunnel lumen material. Indeed, if the dielectric constant ϵ in the lumen of the vestibule is higher due to the presence of more water molecules there than in the upper tunnel cylinder, the polarization surface charge density on the vestibule wall would be higher and would better screen out the buried negative source charges. The apparent surface charge density would be smaller. Besides, a larger ϵ value in a more aqueous medium would also contribute to decrease the ratio σ/ϵ as compared to a medium containing a nascent polypeptide without water molecules.

6.5.4 Water screening inside the tunnel wall medium

Electrostatic interactions in ribonucleic structures are potentially quite strong, but these interactions are mitigated by the screening effects of water or nearby protein atoms, even in the absence of mobile ions [Lockhart and Kim 1993]. The screening of electrostatic interactions results primarily from electronic polarization, reorientation of dipolar groups in the vicinity of charges and dipoles. These effects are well understood and can be accurately determined for interactions in isotropic, homogeneous media. However, in complex inhomogeneous environments such as those near the surface of ribonucleoproteins, dielectric screening is difficult to predict. In the case of the ribosome exit tunnel, the confined geometry and composition of the inner wall close to the tunnel surface and whether the interactions involve direct charges or dipoles are expected to be especially important. The X-ray solved atomic space positions in the immediate 8 to 12 Å vicinity of the tunnel wall show that water molecules are indeed good candidates to explain the screening of the *formal bare* charges carried by the non-bridging oxygens bound to the phosphorus atoms as shown on Fig.6.8 (a) and (b). We investigated with PyMOL scripts the number of water molecules (only oxygen atoms of solvent are seen in X-ray crystallography) that are within less than 4, 5 and 6 Å respectively of the 94 phosphorus atoms around the tunnel wall and obtained the following water molecule counts: 110, 208 and 304. It should be noted here that the total number of water molecules sufficiently well organized to be observed

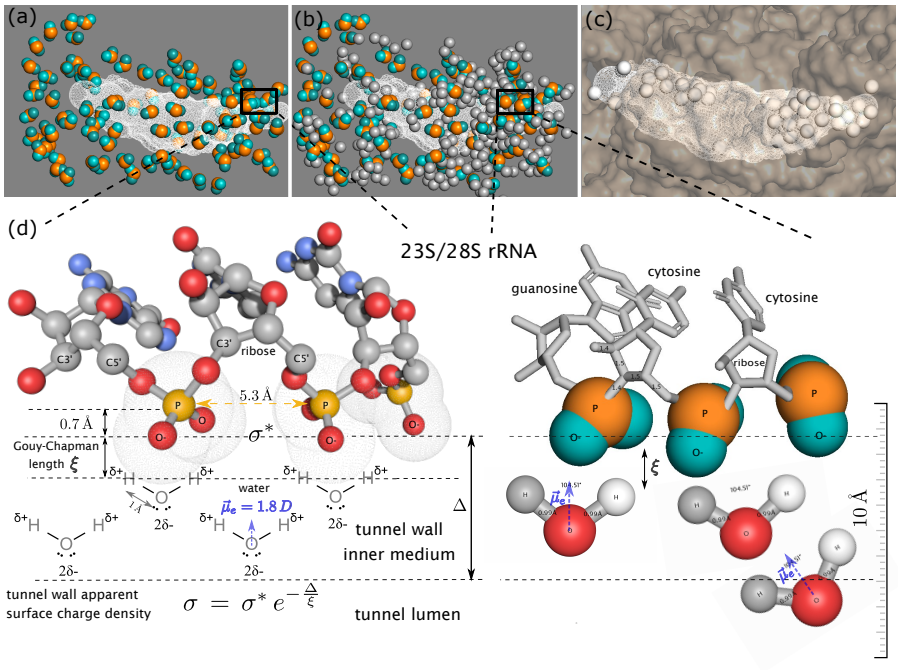


Figure 6.8: (a) Representation of the 94 phosphorus atoms (orange spheres) within 12Å of the tunnel surface (white mesh) with their non-bridging oxygen atoms (teal spheres);(b) 304 oxygen atoms (gray spheres) of water molecules within 6Å of the 94 phosphorus atoms; (c) 44 oxygen atoms (white spheres) of water molecules within 5Å of the tunnel centerline; (d) Electrostatic screening effect due to water molecules. Phosphate moieties carry bare negative charges re-orienting the dipole moment of water molecules $\vec{\mu}_e$, shown in blue. The *bare formal* charge surface density σ^* is screened beyond the Gouy-Chapman length ξ resulting in the *apparent* or *renormalized* charge surface density σ on the tunnel wall. The cartoon was produced with PyMOL extracted data from PDB entry code '4V9F' and ChemDoodle 3D [LLC 2021].

by X-ray crystallography inside the LSU is $7,745$ and that the total number of water molecules observed inside the tunnel is 44 as shown on Fig.6.8 (c). Fig.6.8 (b) shows the abundance of water molecules in the immediate vicinity of the phosphorus atoms in 23S rRNA near the ribosome tunnel wall. Almost no mobile ions were observed in this region in the X-ray solved atomic structure: only 3 Mg^{2+} were found inside the tunnel, no K^+ and only 3 Na^+ within a 12 Å distance of the 94 selected phosphorus atoms. These results support the assumption that the dominant electrostatic screening is due to constitutive water molecules that are buried in the inner wall medium of the tunnel; each phosphate moiety being surrounded by 1 to 3 water molecules in the immediate

vicinity of the tunnel wall surface.

Gouy-Chapman screening length and charge renormalization incorporating electrostatic screening of buried phosphate moieties due to solvent water molecules inside the tunnel wall medium

Due to the dissociation (charge regularization) of surface groups, the rRNA phosphate moieties support surface acquires a net surface charge density that we call σ^* as shown on Fig.6.8 (d). These *bare* charges do not stay unbalanced due to a screening effect involving water solvent. The water molecules dipole moments re-orient so that a layer of positively charged hydrogens oppose the negatively charged phosphate moieties. The solvent water oxygen atoms also form the next layer and so on. The Poisson-Boltzmann-like derivation in the case of a uniformly charged single planar surface provides a far field solution for the screened electrostatic potential which is equivalent to a Yukawa-Debye-Hückel potential:

$$\Phi_{Yuk}(\vec{\mathbf{r}}) = \int \int_S \frac{\sigma^*(\vec{\mathbf{r}'})\, da}{4\pi\, \epsilon\, \epsilon_0} \cdot \frac{e^{-\frac{|\vec{\mathbf{r}}-\vec{\mathbf{r}'}|}{\xi}}}{|\vec{\mathbf{r}} - \vec{\mathbf{r}'}|} \tag{6.23}$$

This is equivalent to a marked exponential damping of the Coulomb interaction where ξ is a characteristic distance of the exponential screening. If we assume the screening effect to be uniform along the charged surface, we can recast the far field potential as the result of an *effective* or *renormalized* or *apparent* surface charge density σ that is different from the actual *formal bare* charge σ^*. The "mapping" between the *bare* and the *renormalized apparent* charge density is simply:

$$\sigma = \sigma^* \cdot e^{-\Delta/\xi} \tag{6.24}$$

where Δ is the distance between the original bare charged wall and the support wall of the electrostatic surface as obtained within non-linear screening theory, see Fig.6.8 (d). The actual *bare formal* surface charge σ^*, when screened by water molecules, results in an *apparent* or *renormalized* surface charge density σ [Roij 16 July 2009]. Unless otherwise stated in this paper, σ always refers to the *apparent* charge surface density obtained within non-linear screening of water molecules present in the inner wall of the tunnel. Selecting the most appropriate screening theory reduces to knowing which length scale parameter ξ to use. Three length scales, i.e. the Bjerrum length (λ_B), the Debye length ($\lambda_D = \kappa^{-1}$) and the Gouy-Chapman length (ξ_{GC}) deserve specific attention as highlighted by Van Roij [ibid.]. The Gouy-Chapman length results from the fact that the Bjerrum length multiplied by the number of charges per surface unit in a uniformly charged wall has the dimension of an inversed length and comes from a boundary condition in the Poisson-Boltzmann equation applied to a uniformly charged planar wall complying with the charge neutrality condition [ibid.]. By convention, this

so-called Gouy-Chapman length is given by

$$\xi = \frac{1}{2\pi\, \lambda_B\, \sigma^*} \tag{6.25}$$

and is a property of the uniformly charged wall, as ξ denotes the length scale over which the potential due to the charged wall, in the absence of screening ions but in the presence of the solvent (characterized by λ_B) equals the agitation thermal energy $k_B T$. The atomic X-ray solved structure of the region around the ribosome exit tunnel showed that the wall buried phosphate moieties are surrounded by one to three water molecules and no mobile ions as shown in Fig.6.8 (b). These experimental facts support the Gouy-Chapman approach for the electrostatic screening in the particular confined micro-environment of the ribosome exit tunnel. The Gouy-Chapman length as shown on Fig.6.8 (d) can physically be interpreted as the separation between two charged layers. The first wall layer is made of the non-bridging oxygens that carry the negative *formal bare* charges in the phosphodiester backbone of the 23S rRNA close to the tunnel wall as shown on Fig.6.8 (d). The second layer is made of the hydrogens belonging to the oriented water molecules having their dipole moment pointing towards the first layer. Table 6.5.1 shows how the screening length scales compare in the three screening theories.

Table 6.5.1: Screening lengths for three screening theories for different dielectric response values at standard temperature $T = 298.15\,\mathrm{K}$

		Bjerrum length λ_B	Debye length[a] $\kappa^{-1} = \lambda_D$	Gouy-Chapman length per unit charge [b] ξ
Medium		$\lambda_B = \frac{e^2}{4\pi\,\epsilon\,\epsilon_0\,k_B T}$	$\kappa^{-1} = \left(\frac{2e^2 I N_A}{\epsilon\,\epsilon_0\,k_B T}\right)^{-1/2}$	$\xi = \frac{1}{2\pi\,\lambda_B\,\sigma^*}$
Vacuum	$\epsilon = 1(\epsilon_0)$	56 nm	–	–
Protein	$\epsilon = 4$	14 nm	0.18 nm	–
rRNA	$\epsilon = 8$	7 nm	0.25 nm	–
Water	$\epsilon = 78$	0.72 nm	0.78 nm	0.105 nm

The atomic positions mapped on the tunnel surfaces, built from high precision X-ray solved structures of real ribosomes set experimental constraints on the phenomenological constants σ, ϵ, or their ratio σ/ϵ, that are complementary to the experimental electrostatic measurements in the ribosome exit tunnel. Altogether, these experimental constraints allow to assess the model assumptions on the piecewise heterogeneity

[a] Debye length calculated for an ionic strength of $I = 0.15\,\mathrm{M}\,(= 150\,\mathrm{mole/m^3})$. N_A is the Avogadro number.

[b] Gouy-Chapman length calculated for a *bare formal* surface charge density of $\sigma^* = 2.1|\mathrm{e}|/\mathrm{nm}^2$.

for σ, ϵ along the tunnel length as well as on the particular origin of electrostatic screening effects inside the ribosome exit tunnel. Indeed, taking as screening length scale, the Gouy-Chapman screening length value, i.e. $\xi = 1.05$, and taking for $\Delta = 5.97 = 6.67 - 0.7$, the difference of the average distance between the closest phosphorus atoms to the model cylinder wall surface (6.67 Å), and the distance (0.7 Å) between the phosphorus atoms and the non-bridging oxygen atoms in the phosphodiester bonds with the riboses, the *apparent normalized* surface charge density is

$$\begin{aligned}
\sigma_1 &= \sigma_1^* \, e^{-\frac{\Delta}{\xi_{GC}}} \\
&= -2.1 \, e^{-\frac{5.97}{1.05}} \\
&= -2.1 \cdot 0.00339 \\
&= -0.00712 \, |\mathrm{e}|/\mathrm{nm}^2 \\
&= -1.141 \mathrm{mC}/\mathrm{m}^2
\end{aligned} \tag{6.26}$$

Substituting this numerical value in the formula (6.3) of the potential for the cylinder, with $z = -50$ Å, $R = 5$ Å and taking the two extreme assumptions on the dielectric response in the tunnel lumen $\epsilon = 4$ (nascent protein occupying the tunnel and no water molecule) or $\epsilon = 78$ (water molecules only occupying the tunnel), we can get an estimate for the range of the electrostatic potential $\Phi(z)$ expected values at the centerline of the tunnel at position $z = -50$ Å where the potential reaches its maximal negative value.

With $\epsilon = 4$ (nascent chain only occupies the tunnel):

$$\begin{aligned}
\Phi(z) &= \frac{\sigma_1 \, R}{2 \, \epsilon_1 \, \epsilon_0} \cdot \log \Big| \frac{7 + \sqrt{7^2 + 1}}{-10 + \sqrt{101}} \Big| \\
&= \frac{\sigma_1 \, R}{2 \, \epsilon_1 \, \epsilon_0} \log \, 282.12 \\
&= \frac{-1.141 \, 10^{-3} \, \cdot \, 5 \cdot 10^{-10}}{2 \cdot 4 \cdot 8.85 \cdot 10^{-12}} \cdot 5.6423 \\
&= -0.0455 \, \mathrm{Volt}
\end{aligned} \tag{6.27}$$

The expected value of the electrostatic potential would be -45.5 mV at the tunnel center if the dielectric response of the medium in the lumen is at the lower limit $\epsilon_1 = 4$ (nascent protein).

With $\epsilon = 78$ (water molecules only in the tunnel):

$$\begin{aligned}\Phi(z) &= \frac{-1.141\,10^{-3}\,\cdot\,5\cdot10^{-10}}{2\cdot78\cdot8.85\cdot10^{-12}}\cdot5.6423\\ &= -0.0023\,\text{Volt}\end{aligned}\tag{6.28}$$

The expected value of the electrostatic potential would be −2.3 mV at the tunnel center if the dielectric response of the medium in the lumen is at the upper limit $\epsilon_1 = 78$ (fully filled with water).

The observed value at the tunnel center is −22 mV. Keeping the apparent charge density on the cylinder wall at $\sigma_1 = -1.141\ 10^{-3}\,\text{C/m}^2$ (corresponding to a bare charge density $\sigma_1^* = 2.1\ |e|/nm^2$ for the cylinder as counted on the X-ray solved structure), the mean dielectric response of the medium in the cylinder lumen of the tunnel would be $\epsilon_1 = 8.3$. This would be consistent with a medium composition of 94% of nascent protein and 6% of water in the tunnel lumen.

To support the dimensional relation (6.15) and the stepwise values for the ratio σ/ϵ in the different tunnel parts, we can estimate the local bare surface charge density σ_3^* from the ratio $\frac{\sigma_3}{\epsilon_3}$ at the constriction site. In section 6.4.5, we inferred the ratio $\frac{\sigma_3}{\epsilon_3}$ at the constriction site from the values of f_{scale} and Γ, experimentally fitted to the measured electrostatic potential. With $R = 5\,\text{Å}$, $\sigma_3/\epsilon_3 = 5.73\cdot10^9$ Volt/m. For a dielectric constant prevailing inside the constriction site, i.e. the dielectric constant of the polypeptide nascent chain ($\epsilon_r \sim 4$) and in the *absence of screening* water molecules in the confined region of the constriction site, we can estimate $\sigma_3 = \sigma_3^*$:

$$\sigma_3^* = \epsilon_{3\,r}\cdot\epsilon_0\cdot5.73\cdot10^9 \tag{6.29}$$

$$= 4\cdot8.85\,10^{-12}\cdot5.73\cdot10^9 \tag{6.30}$$

$$= 0.2029\ \text{C/m}^2 \tag{6.31}$$

The lateral surface being $S = 2\,\pi R\cdot\Gamma = 5.655\,\text{nm}^2$, the bare formal charge in this local surface is estimated to be $q_3 \sim 0.2029\cdot5.655\,10^{-18} = 1.147\,10^{-18}$ Coulomb or approximately 7 elementary charge units ($|e| = 1.602\,10^{-19}$ C). The fitted values and the dimensional interpretation of section 6.4.5 indicate that the number of apparent positive charges above the phosphate moieties baseline would be around 7 which is consistent with the number of the positively charged residues at the uL22 and uL4 protrusion (R or K residues in uL22 and uL4 near the tunnel wall) lining in the vicinity of the constriction site surface and within a distance of 8 Å of the tunnel surface as shown in Fig.6.6 (b) and (c) and in Fig.6.9 (a), (b) and (c).

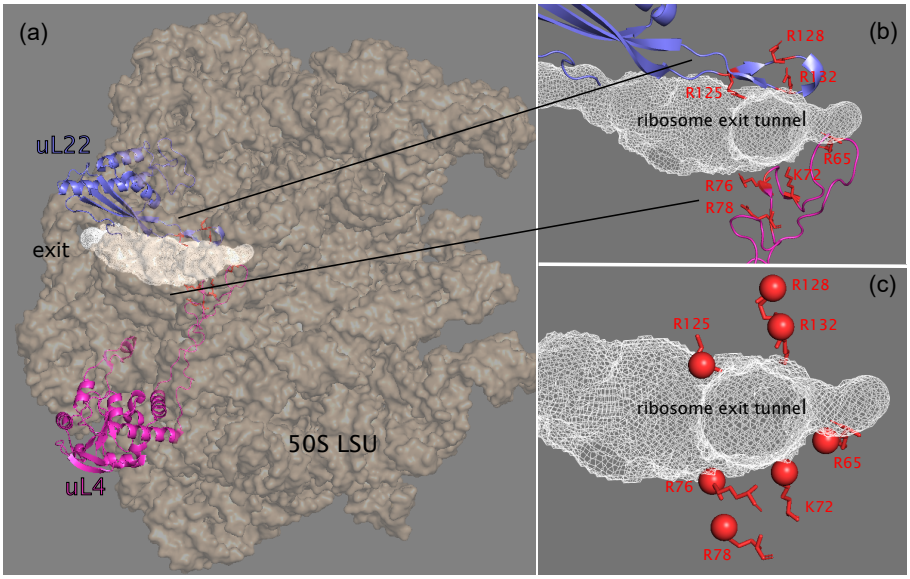


Figure 6.9: (a) Ribosome exit tunnel in the 50S large subunit of *Haloarcula marismortui* showing 2 ribosomal proteins uL22 and uL4 and (b) their local protrusion with 7 positively charged amino acid residues close to the constriction site; (c) the red spheres represent the nitrogen charged atoms bonded to the ϵ-carbon of arginine (NH1 or NH2) or of lysine (NZ).

6.6 Application of the ribosome exit tunnel model

6.6.1 Computing the electrostatic interaction variations of the ribosome exit tunnel for different amino acid sequences in nascent polypeptides

Our established models expressed by equation (6.3) or (6.14) or (6.17) can be used to quantitatively compare how difficult it is for the ribosome to push a nascent polypeptide chain inside and eventually out of the exit tunnel, depending on the amino acid primary sequence. If a peptide sequence is locally enriched in positively charged amino acids residues inside the tunnel and in negatively charged amino acid residues close to tunnel entry point, the axial forces required to push the nascent protein through the tunnel will be higher than for a peptide composed of neutral amino acid residues or carrying only a single cluster of positively charged amino acid residues. The electrostatic potential well, locally trapping charged amino acid enriched peptides, needs to be overcome by

other forces. These compensation mechanisms are exerted either by the ribosome itself or by third party proteins with motor domains from specialized chaperone proteins exerting tugging forces outside of the ribosome. The elongation speed also has to be compatible with the decoding speed of the mRNA encrypted message which depends on the codon usage and on codon position autocorrelation, i.e. codon ordering allowing tRNA recycling (reusage of the same tRNA at successive encodings of the same amino acid can speed up translation or favor fidelity [Cannarozzi et al. 2010; Friberg et al. 2006)]. The elongation speed may also independently be impeded by downstream mRNA secondary structures [Desai et al. 2019; Liu et al. 2014a; Simpson et al. 2020; Yang et al. 2014].

The eukaryote ribosome exit tunnel can accommodate at least 40 amino acid residues and up to more than 70. It is known that the nascent polypeptide can start folding, i.e. finding its final secondary structure, inside the tunnel (and eventually tertiary 3D structure outside). Alpha helices have been shown to be present inside the tunnel close to its exit point. So, a variable number of amino acids larger than 40 can actually be hosted inside the tunnel. Again, for the sake of simplicity, here, we consider that the number of amino acid residues hosted inside the tunnel is exactly 40 and that the maximum number of amino acid residues that are under the electrostatic influence of the tunnel is exactly 50: 5 between the PTC and the tunnel entry, 40 in the tunnel and 5 out of the tunnel.

The incorporation of a single amino acid to the nascent polypeptide chain takes place at the peptidyl transfer center (PTC), at the so called P site of a translating ribosome. This PTC center is located around 5 amino acid residues away from the ribosome tunnel entry point. Stated otherwise, this means that the currently decoded codon, for which the cognate or semi-cognate aminoacylated tRNA, is 5 codons downstream the codon for which the amino acid is currently in the entry point of the tunnel. There are 5 amino acids bound in the oligopeptide part ready to enter the tunnel. Let us also assume that there are 5 bound amino acids out of the tunnel at the exit side that can feel the electrostatic influence of the tunnel. So, from the start codon (AUG coding for methionine), a nascent peptide starts with a 5 amino acid residues stretch elongating to the tunnel entry point, building up progressively to a 45 amino acids sequence fully accommodating the whole length of the tunnel, eventually extending to 50 amino acid residues being under a direct influence of the tunnel, see Fig. 6.10. For this 50−mer stretch to be out of the tunnel influence, another extra 50 amino acids have to be added to the carboxy terminal end of the nascent polypeptide.

Our aim is to compute the force profile and the mechanical power to be applied continuously on peptide stretches to overcome the electrostatic trapping interaction in the tunnel and to exit the ribosome tunnel. The easiest case scenario for computing would be when the downstream sequence is completely neutral. This is of course not always the case and the occurrence of charges in the downstream sequence plays a role that should not be neglected.

Multiplying the axial force acting on the stretches with the stretch axial displacement, i.e. the elongation distance towards the ribosome exit tunnel, yields the mechanical work that was delivered. Multiplying the axial forces acting on the stretches with the protein elongation rate, i.e. the speed of the ribosome along the transcript (mRNA) being translated, yields the required instantaneous net mechanical power.

The electrostatic axial force on the tunnel axis felt by an amino acid residue is the product of its net charge by the axial electric field, see equation (6.5), the latter being the gradient of the electrostatic scalar potential, i.e. the first derivative of the potential with respect to the axial coordinate.

An important simplifying assumption is that all amino acid residues building up the nascent polypeptide are all rigidly bound together and that the resulting nascent protein can be considered a single linear solid rigid body. This peptide, at least in the tunnel, is considered non deformable. With this strong assumption, the axial forces individually computed for each charged amino acids act jointly and apply additively on the resulting rigid peptide body.

The local pH along the tunnel is unlikely to be out of the range 6-8 [Deutsch 2014]. In this pH range, among the 20 amino acid residues, only three are positively charged and two are negatively charged. Arginine (R), lysine (K) and histidine (H) carry a partial positive charge on the amino moiety in the side chain. The intrinsic pK value, referred to as pK_{int}, is the pK value of an ionizable side chain when it is present in pentapeptides [Pace et al. 2009]. Only arginine, $pK_{int} = 12.3$, and lysine, $pK_{int} = 10.4$, are truly positive in physiological conditions at neutral pH whereas histidine, $pK_{int} = 6.5$, would be very weakly positive at a pH in the range $6 - 6.5$. For this reason $+1, +1, +0.05$ net formal charges are arbitrarily adopted for arginine, lysine and histidine respectively. For glutamate (E), $pK_{int} = 4.3$ and aspartate (D), $pK_{int} = 3.9$, both carrying a carboxylic moiety on the side chain, the arbitrarily adopted net formal charges are both -1 in physiological conditions. All other amino acid residues are considered neutral. The positively charged residues are represented in red whereas negatively charged residues are represented in blue on the test sequences to be analyzed under our model as displayed in Fig 6.10. The neutral residues are unsensitive to the electrostatic potential or the axial electric field.

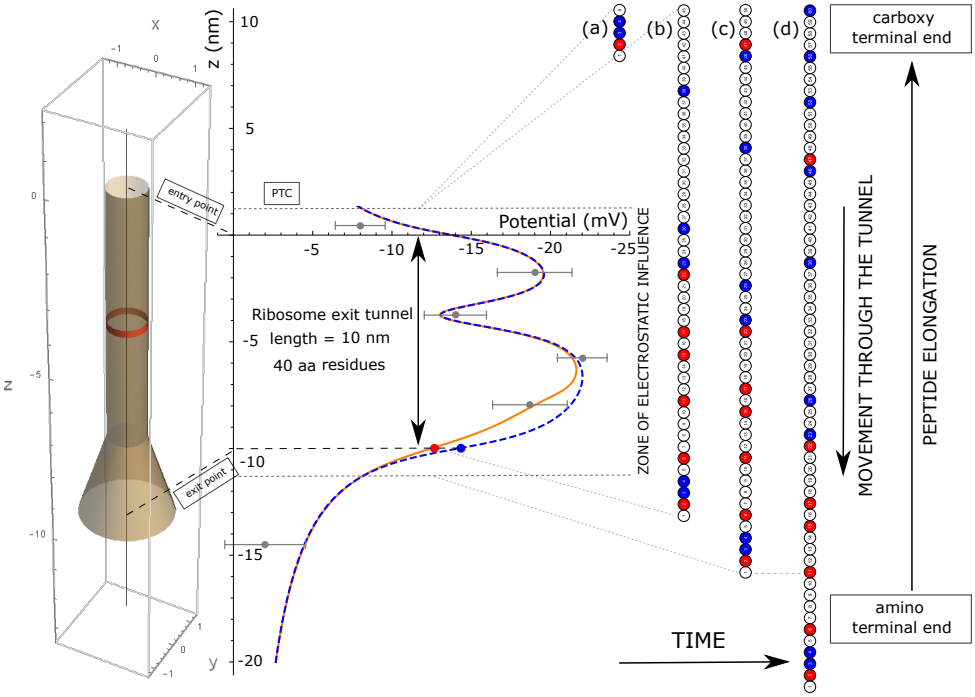


Figure 6.10: Algorithm for computing the axial forces acting on a nascent peptide. Only the last 50 residues, at most, from the PTC are under the electrostatic influence of the ribosome exit tunnel. In this figure, the nascent polypeptide is moving from the top to the bottom through the ribosome exit tunnel during elongation.(a) 5 amino acid residues peptide with residue 5 at the PTC and residue 1 at the tunnel entry point. (b) 45 amino acid residues with residue 45 at the PTC and residue 1 at the tunnel exit point. (c) 50 amino acid residues with residue 50 at the PTC and residue 1 emerging out of the electrostatic influence zone. (d) 60 amino acid residues with residue 60 at the PTC and residues 1 to 10 out of the electrostatic influence zone.

Algorithm and program pseudo-code for computing the axial force on the nascent peptide due to the ribosome exit tunnel interaction

The algorithm for computing the axial force on a given nascent peptide due to the ribosome exit tunnel electrostatic interaction as a function of the amino acid sequence is schematically depicted in Fig. 6.10.

Reading in the given input peptide sequence

(Step a) Read in the peptide sequence from the amino terminal end to the carboxy terminal end.

(Step b) Determine the length of the peptide (number of amino acid residues in the given peptide).

(Step c) Convert the sequence of amino acid residues into an ordered list of formal charges using the following charge coding rule: $K \rightarrow +1, R \rightarrow +1, H \rightarrow +0.05, E \rightarrow -1, D \rightarrow -1$. All other residues are converted to a neutral charge $X \rightarrow 0$.

Computing the axial position of each amino acid in the sequence, compute the axial force acting on the residue at that position and sum the contributions of all charged residues

(Step a) Start with the first 5 residues from the amino terminal end of the peptide (the first five elements in the ordered list) to build the stretch currently computed.

(Step b) Map the axial positions of the residues in the stretch, each separated by a distance $0.25\,10^{-9}$ m. Position $z = 0$ corresponds to the residue located at the ribosome exit tunnel entry point, position $z = 5 \times 0.25\,10^{-9}$ corresponds to the residue located at the PTC, 5 residues downstream in the sequence. All algebraic negative z positions correspond to residues that have entered the tunnel.

(Step c) Compute the ordered list of axial electric fields for each of the previous axial positions using formula (6.4) for the idealized cylindrical model, formula (6.16) or formula (6.18) for the realistic model, incorporating the Lorentzian peak and the truncated cone geometry at the end side of the tunnel, respectively.

(Step d) Multiply element by element, the ordered list of the axial electric fields by the ordered list of formal charges, to obtain the list of the contributing axial forces acting on the peptide stretch currently computed.

(Step e) Sum all the contributing axial forces in the peptide stretch currently computed and store the result in an ordered list of the total axial forces acting on the stretch upon the carboxy-terminal end of the nascent chain at the PTC site.

(Step f) Repeat *Step b* to *Step e* for all iterated stretches by one residue towards the carboxy terminal end, conditionally on a length of 50 residues, and while the last residue has not reached the end of the given input peptide. The 50 residues condition ensures there are at most 40 residues inside the tunnel, 5 residues between the PTC site and the tunnel entry point and at most 5 outside the ribosome exit tunnel, still under the electrostatic influence of the tunnel.

Plot the total axial force acting on the nascent peptide as a function of the last amino acid residue occupying the ribosomal PTC position

Positive axial forces are believed to slow down the elongation rate while negative axial forces are believed to speed up the elongation rate of the ribosome.

6.6.2 Comparing the electrostatic interaction profiles when passing through the ribosome exit tunnel for different amino acid sequences

6.6.2.1 Simulated synthetic oligopeptide sequences

It should be emphasized that due to the symmetry of the potential barrier in the idealized cylindrical model and its finite length, a clustered local enrichment in positive (negative) charge in a polypeptide sequence will first be attracted (repelled) when entering into the tunnel and will then be pulled inside (pushed outside) the tunnel when emerging at the tunnel exit point. Hence an inversion in the sign of the force profile should always be observed for locally clustered net charges that are followed by a neutral tail sequence. This inversion spreads over a distance covering the ribosome exit tunnel length which is 40 amino acid residue in length in the adopted simplified model and with equal areas under the curve, see Fig. 6.11 upper panel (A).

The situation is more complicated when the tail sequence also includes local charges distribution within a range of $20 - 40$ amino acid residues in the tail sequence (Fig. 6.11 lower panel) or if the electrostatic potential well barrier is not symmetric as with the truncated cone concatenated to the cylinder geometry, see Fig. 6.11 upper panel (B).

To highlight the differences between a symmetric potential (idealized cylindrical model) and an asymmetric potential (cylinder plus truncated cone with Lorentzian peak realistic model), we compared the axial force profiles applied for the same synthetic sequences in both cases with typical clustered net charge distributions.

In Fig. 6.11, the axial forces acting after each amino acid incorporation at the PTC are displayed for a peptide of 90 residues in length. Fig. 6.11 upper panel (A) shows

the symmetric potential (idealized cylindrical model) effect on 5 contiguous positively charged residues between position 5 and 9 (net positive charge centered at position 7). The nascent peptide is attracted into the tunnel until amino acid residue number 32 (= 7+25) is incorporated at the PTC. From position 32 to 59 = 9 + 50 (position 59 corresponds to the moment when the last positively charged residue is out of the influence zone), the axial forces acting on the peptide tend to pull it back into the tunnel and these forces tend to prevent the peptide from traversing the tunnel easily.

Equivalently, it is hypothesized that the elongation proceeds at a faster rate when residues 5 to 32 are incorporated at the PTC, and is slower when residues 33 to 59 are incorporated at the PTC. The impact on the elongation speed will be quantitatively assessed with the use of a Maxwell-Boltzmann factor. This Maxwell-Boltzmann factor provides a quantitative modulation of the average elongation speed (see supplemental material).

It is also hypothesized that the ribosome requires more mechanical power to push the nascent chain out of the tunnel when it is repelled due to the electrostatic interactions, when residues 33 to 59, in our example, are incorporated at the PTC. How the extra mechanical power is mobilized is currently unknown. An increased turnover in the biochemical reactions providing Gibbs free energy to the ribosome would probably help. Equivalently, this would require an increased rate in amino acid incorporation because more Gibbs free energy would then be available as there are two energy rich GTPs hydrolyzed per residue incorporation. Fig. 6.11 upper panel (B) shows the asymmetric potential (realistic model) effect on 5 contiguous positively charged residues between position 5 and 9 (net positive charge centered at position 7). The nascent peptide is attracted into the tunnel until amino acid residue number 19 (= 7+12) is incorporated at the PTC. From position 20 to 26 = 9 + 17, the axial forces acting on the peptide tend to pull it back into the tunnel and these forces tend to prevent the peptide from moving out of the tunnel. Then again, from position 27 to 37, the axial forces acting on the peptide tend to move it out of the tunnel. Finally, from position 38 to 58, the axial forces acting on the peptide tend to pull it back into the tunnel and these forces tend to prevent the peptide from traversing the tunnel easily. Compared with Fig. 6.11 upper panel (A), there are two fast moves separated by a short slower move, before residue 38, instead of one single fast move in the symmetric potential case. Equivalently, it is hypothesized that the elongation proceeds at a faster rate when residues 5 to 19 then 27 to 37 are incorporated at the PTC, and is slower when residues 20 to 26, then 38 to 58 are incorporated at the PTC. Note the amplitude of the axial forces are smaller but more dispersed in the positive region, for the asymmetric potential (realistic model) (B), than for the symmetric potential (A). Fig. 6.11 lower panel (A) shows the symmetric potential (idealized model) effect on a peptide with 5 contiguous positively charged residues between position 5 and 9 (net positive charge centered at position 7) and 5 contiguous negatively charged residues between position 45 and 49 (net negative charge centered at position 47) exactly 40 residues away from the first charge cluster. Fig. 6.11

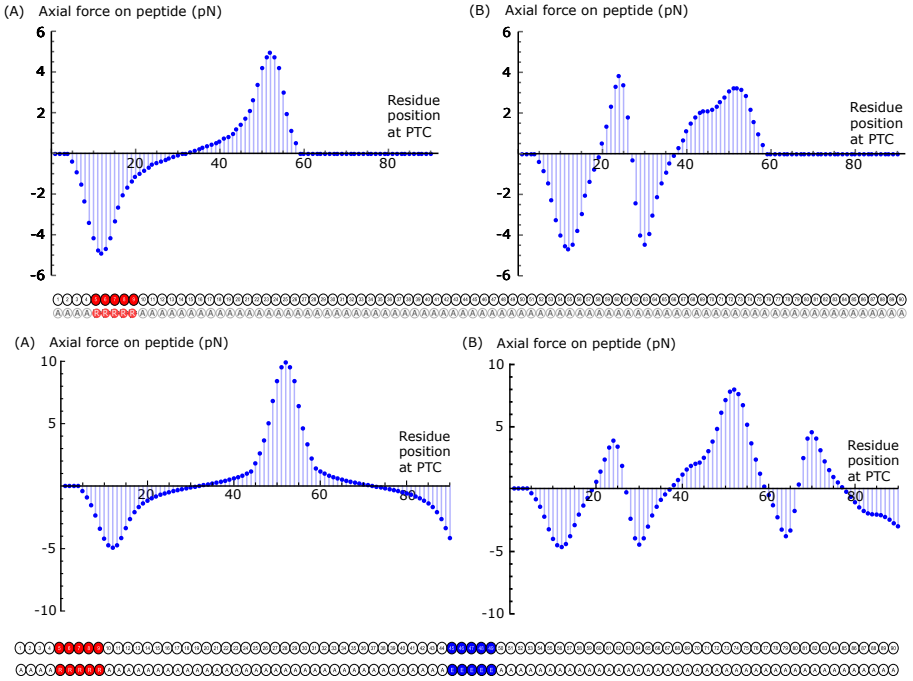


Figure 6.11: Axial forces (pN) ordered values acting on the nascent chain at each residue position incorporated in the primary sequence at the PTC center. Upper panels: symmetric electrostatic potential idealized model (A) with 5 contiguous arginine. Asymmetric electrostatic potential realistic model (B) with 5 contiguous arginine, centered at position 7. Positively charged arginine residues are colored in red in the displayed primary sequence (upper inset). Lower panels: symmetric electrostatic potential idealized model (A) with 5 contiguous arginine and 5 glutamate residues clustered as displayed in the figure lower inset. (B) Asymmetric electrostatic potential realistic model. Opposite charges (+ residues position 5-9: red; - residues position 45-49: blue).

lower panel (B) shows the asymmetric potential (realistic model) effect on the same peptide sequence. When the first plus cluster is emerging at the tunnel exit point, the second minus cluster is at the tunnel entry point. This situation results in high axial forces making difficult for the peptide to get out of the tunnel. As for Fig. 6.11 upper panel (B), there are two fast moves separated by a short slower move, before residue 38, instead of one single fast move in the symmetric potential case. The estimated maximal axial force is more than 8 pN and is reached when residue 52 is at the PTC. The axial

forces tend to prevent the peptide to get out of the tunnel when residues 38 to 59 are at the PTC.

6.6.2.2 Global and local electrostatical work and energy balance

In the two symmetric electrostatic models (cylinder and cylinder with Lorentzian peak), it is important to note that there is no difference in electrostatic potential between the entry point and the exit point of the ribosome exit tunnel, whatever the form of the potential inside of the tunnel. Because electrostatic interaction is conservative, the total work spent or harnessed by a charged residue when moved from the entry point to the exit point of the tunnel will always be equal to zero. Hence, the global net mechanical work for a full sequence to be moved completely through the ribosome exit tunnel should always be equal to zero. In the two asymmetric electrostatic models (cylinder plus truncated cone with or without the Lorentzian peak), there is a net difference in electrostatic potential between the entry point and the exit point of the ribosome exit tunnel. The total work spent or harnessed by a charged residue when moved from the entry point to the exit point of the tunnel will not be equal to zero in general. With a potential difference of 1.7 mV between the tunnel exit and entry points, the required mechanical energy is $-0.164\,\mathrm{kJ/mol}$ ($0.039\,\mathrm{kcal/mol}$), or $\sim 0.3\,\mathrm{pN \cdot nm}$ on a single molecule, to traverse a single positively charged amino acid residue through the tunnel. Moreover, in any case, transiently or locally, the work to overcome positive axial electrostatic forces or the work harnessed in case of negative axial electrostatic forces acting upon any unit charged test residue may not be equal to zero. To illustrate this, the local mechanical work is computed in the case of the simulated synthetic peptide of Fig. 6.11 lower panel (B) with 5 contiguous arginines (+) and 5 contiguous glutamates (-), separated by 40 neutral residues. When the oligopeptide stretch ranging from residue 5 to 19 is incorporated, the sign of the work is positive (work = $+0.67\,\mathrm{kcal/mol}$), according to our adopted conventions in Fig. 6.4, meaning that the stretch is freely benefitting electrostatic energy to traverse the tunnel during the incorporation of those amino acid residues. On the contrary, when amino acid residues 38 to 59 are incorporated in the nascent chain, the sign for the work (work = $-1.42\,\mathrm{kcal/mol}$) is negative, meaning that mechanical energy has to be provided in some way to the nascent chain to help the stretch progressing through the tunnel. It is interesting to compare the computed values for the aforementioned mechanical work that are transiently either harnessed (0.67 kcal for the first stretch of 12 residues), or to be delivered ($-1.42\,\mathrm{kcal} = 9.9\,\mathrm{pN \cdot nm}$ for the second stretch of 21 residues), to the Gibbs free energy released from biochemical reactions at each residue incorporation, i.e. $\Delta G^\circ \sim -18.3\,\mathrm{kcal/mol}$ (per amino acid incorporation) as detailed in the supplemental material. If the chemical energy to mechanical work conversion yield is of the order of $\sim 50\%$, an estimate of the local required chemical energy to push the nascent chain in the case of the second stretch would be around $-1.42/0.5 = -2.84\,\mathrm{kcal/mol}$. This amount of biochemical energy is

about $\sim 15\%$ of the Gibbs free energy released from the biochemical reactions by a single new residue incorporation associated to the ribosome elongation cycle. These simple rough comparisons show that, energetically, the ribosome has enough energy resources to overcome the local electrostatic barrier easily. However, situations may occur for which a nascent peptide will pose difficulties to the ribosome, considering that as much as $\sim 15\,\%$, or possibly more than $\sim 30\,\%$ of the Gibbs free energy normally available to the ribosome per elongation cycle could be required to push the nascent chain out of the tunnel, depending on the charged amino acid distribution content of the nascent chain, and depending on the section widening in the region close to the exit point of the ribosomal tunnel.

6.6.2.3 Real protein sequences

The purpose of the ribosome exit tunnel electrostatic realistic model is to apply it to real protein sequences, to compare them and to quantitatively determine where are the critical spots for the ribosome elongation process, or what are the axial force profiles acting on proteins during the co-translational folding process. To illustrate the application of our model to compute the forces acting on real protein sequences, we use it here in the context of neurodegenerative diseases like Huntington's, Creutzfeldt-Jakob, or Alzheimer's diseases. These diseases share a common pathology in the deposition of misfolded and aggregated conformations of a particular protein in the central nervous system at sites of neuronal degeneration [Hatters 2008]. The mechanisms of misfolding, aggregation and their functional consequences are not yet fully elucidated. Huntington's disease is caused by mutations that expand the number of glutamine codons within an existing poly-glutamine (poly-Q) repeat sequence of the gene coding for the huntingtin protein [Arrasate and Finkbeiner 2011; Bonfanti et al. 2019; Chen and Wolynes 2017; Hatters 2008]. The N-terminus end of a normal huntingtin protein is composed of a N-terminus sequence of 17 residues (NT_{17}), a poly-Q sequence with a number of contiguous glutamines anywhere between 6 and 34 (e.g. Q_{21}), and a polyproline sequence of around 11 proline residues (e.g. P_{11}); see Fig. 6.12. A mutant allele coding for a number of glutamine repeats exceding 36 (e.g. Q_{36}) will inevitably lead to Huntington's disease if the person carrying this allele lives long enough. Huntingtin has a very long sequence with a total length of $3,144$ residues in the normal wild type sequence but the mutated huntingtin is only expanded in the very beginning of the sequence. Here, we do not pretend to solve the mechanism or the detailed molecular steps causing the misfolding of the huntingtin mutant protein but provide an analysis of a possible role of the forces acting on the huntingtin growing sequence while it is biosynthesized by the ribosome and investigate a possible co-translational misfolding situation. We compare, in Fig. 6.13 (A) and (B), the axial forces profiles for the human wild type huntingtin `HTT` and a mutant huntingtin `mHTT` for the first 150 N-terminus residues when their respective transcripts are being translated. The

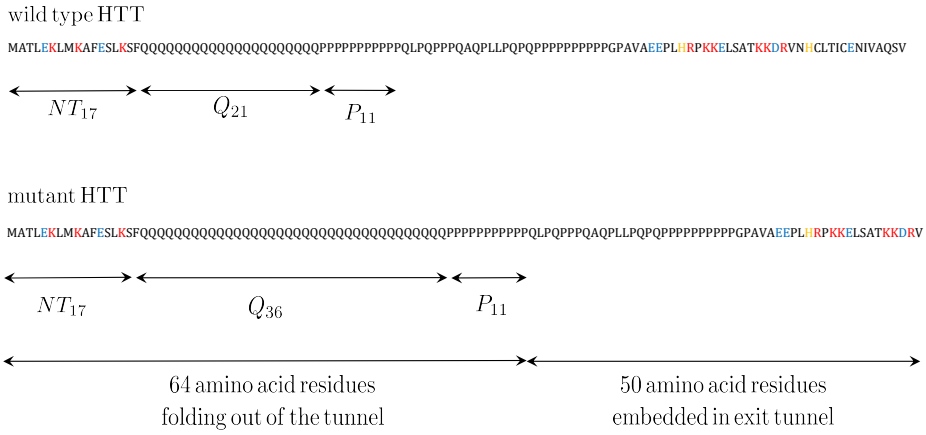


Figure 6.12: Wild type HTT and mutant mHTT huntingtin protein N-terminus starting sequence, showing the lengths of their poly-Q sequences. Positive residues: red. Negative residues: blue. Histidine residues: orange. Neutral residues: black

folding conditions and environments are different as the axial forces acted by the exit tunnel of the ribosome on these two growing nascent huntingtins are different. The two sequences embedded in the tunnel are different and cause the two very different net resulting axial forces. The mutant huntingtin has a length of the N-terminus sequence equal to 64 (= NT_{17} + Q_{36} + P_{11}).

Figure 6.13: (next page) (A) and (B) Axial forces profiles for wild type and mutant protein. Blue and red arrows show the values of forces acting on the nascent chain at the PTC when residue 82 and 114 respectively are incorporated in the peptide at the PTC. (C) Wild type protein. A 2 pN pulling force due to the tunnel interaction with residues 32-82 opposed to the spontaneous folding force when residues 1-31 are out of the tunnel. (D) Mutant protein. No force opposed to the spontaneous folding force when residues 1-31 are out of the tunnel. (E) Wild type protein. A 2 pN pulling force due to the tunnel interaction with residues 64-114 opposed to the spontaneous folding force when residues 1-63 are out of the tunnel. (F) Mutant protein. A pushing force of 4 pN adds to the spontaneous folding force.

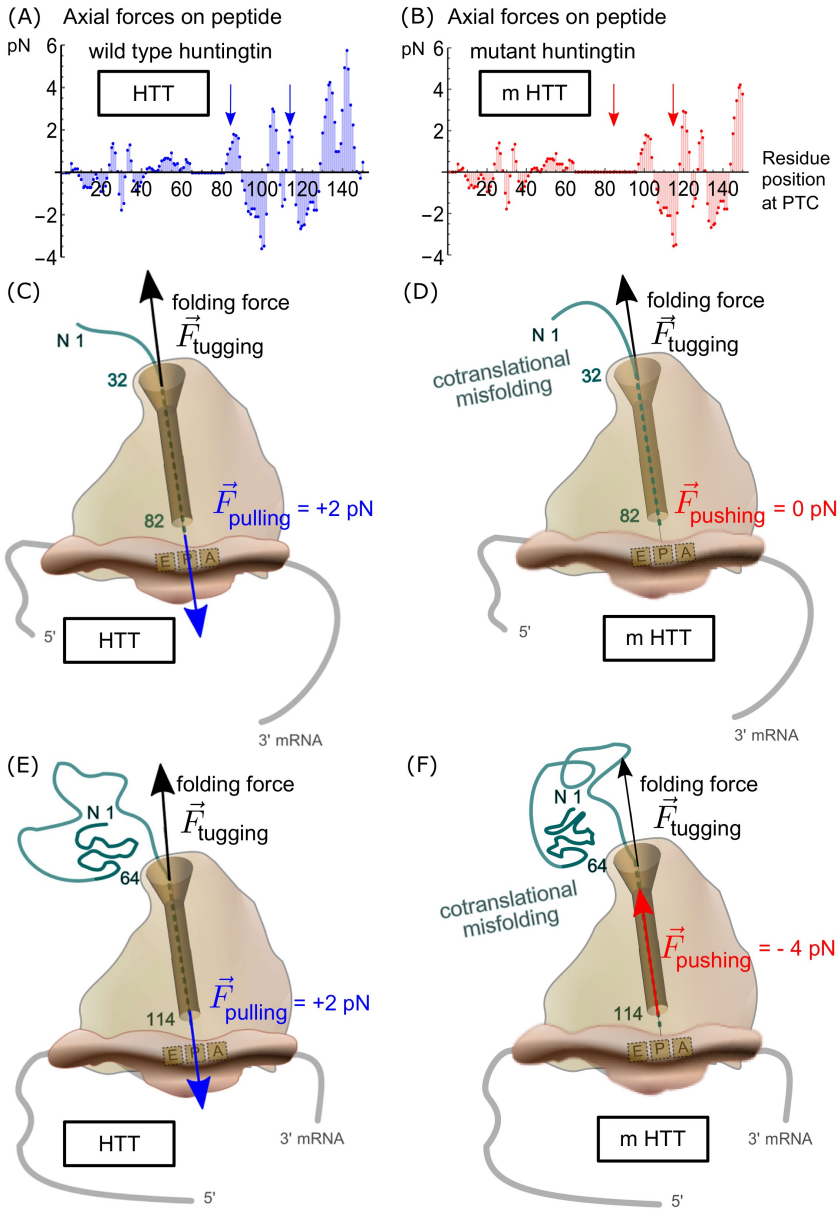


Figure 6.13 (caption continued): Comparison of vectorial co-translational folding for huntingtin wild type (HTT) and mutant protein (mHTT).

The exit tunnel exerts axial forces of electrostatic origin that are either pulling forces or pushing forces. These forces oppose (or not) to the forces generated by the spontaneous folding of the unstructured segments of the nascent polypeptide chain upon lengthening of the chain out of the tunnel. In their computational simulations, Fritch *et al.* estimated that the force difference experienced at the P-site residue upon doubling the length of the chain out the tunnel was of the order of piconewtons [Fritch et al. 2018]. In our Fig. 6.13 (C) to (F), these folding forces are called tugging forces (black arrows). In the extreme case of SecM mediated ribosomal arrest known in bacteria, Goldman *et al.* provided evidence that the minimal tugging force must be $\sim$ 10 pN to relieve the stalled ribosome [Goldman et al. 2015]. These results show that tugging forces in the range $1 - 10$ pN can possibly be generated by emerging folding nascent chains to reach their native conformation.

As seen from the comparison of the axial forces profiles, the wild type nascent protein is experiencing a pulling force ($\sim$ +2pN) of electrostatic origin from the ribosome, when incorporating residue 81 to 90 at the peptidyl transferase center PTC, Fig. 6.13 (A) first arrow, and Fig. 6.13 (C), while the length of the nascent chain out of the tunnel is 31 to 40, i.e. when the critical poly-Q segment is fully emerging from the ribosome. In the mutant **mHTT**, there are no pulling forces from the ribosome at this moment; the axial forces are null at this moment, Fig. 6.13 (B) first arrow, and Fig. 6.13 (D).

When the first 64 amino-terminal residues have just emerged out of the exit tunnel of the ribosome and are exploring the folding space, the downstream 50 residues are under electrostatic interaction with the ribosome exit tunnel. The total axial resulting force acting upon residue 114 (= 64 + 50), and indirectly on the whole growing chain outside the tunnel, is different for the mutant **mHTT** than for the wild type **HTT**. For the wild type **HTT**, the nascent chain out of the tunnel, 64 residues in length, folds while the ribosome is pulling the chain toward the interior of the tunnel (+2 pN), Fig. 6.13 (A) second arrow, and Fig. 6.13 (E), whereas for the mutant protein **mHTT**, the nascent chain out of the tunnel, 64 residues in length, folds without opposing force from the ribosome. On the contrary, for the latter, there is a pushing force on the nascent chain (-4 pN) from the ribosome exit tunnel, Fig. 6.13 (B) second arrow, and Fig. 6.13 (F). This analysis would suggest that proper folding of the poly-Q containing segment of huntingtin protein would require a pulling force from the ribosome. If there is no pulling force, as for the mutant huntingtin, the poly-Q segment would be much more prone to co-translational misfolding. Interestingly, the effect of expanding the length within the poly-Q segment would just cause a shift in the axial forces profile that reverse the forces acted by the ribosome, upon the folding segment, between the wild type and the mutant protein.

Overall, these results suggest that a change in the local distribution of charged residues or an insertion or a replacement by neutral residues has impact on the axial forces profile over a spatially extended region of the nascent protein which is in a range corresponding to the length of the ribosome exit tunnel. The proper co-translational folding of a

nascent polypeptide chain out of the ribosome calls for specific patterns in the charged amino acid distribution in the sequence downstream, embedded in the exit tunnel, down to the peptidyl transferase center PTC. The encrypted sequence indirectly dictates, in a spatially extended way, the electrostatic interaction of the charged amino acid residues in the exit tunnel to generate axial forces profiles acted by the ribosome upon the growing chain. These forces would play a key role in the correct co-translational folding process. The co-translational folding is vectorial, that is, it involves elements that emerge successively from the N-terminus to the C-terminus [Thommen et al. 2017]. The landscape of co-translational folding may differ depending on the charged residue distribution which is embedded in the tunnel. Our model sheds light on how the ribosome could affect the folding trajectory.

6.7 PRE discussions and future perspectives

6.7.1 Summary of results

6.7.1.1 A closed form analytical expression of the electrostatic potential with piecewise phenomenological constant parameters

In this study we set out to model the electrostatics of the ribosome exit tunnel to explore quantitatively the impact of the distribution of the charged amino acid residues embedded in the tunnel on the forces acting on the nascent peptide chain during translation. Our approach was to develop a full analytical expression of the electrostatic potential inside the tunnel, starting from two idealized theoretical geometries for the tunnel, i.e. a cylinder and a cone. We eventually concatenated the cylindrical geometry with the conical geometry, and finally added an empirical Lorentzian function motivated by the known experimental observations of local and highly conserved ribosomal protein protrusions inside the tunnel. The precise geometry of the tunnel is important for quantifying the resulting electrostatic potential profile along its centerline. It shows what part of the electrostatic profile is contributed by the shapes and by the sizes of the tunnel and what part is inherent to the physicochemical properties such as the surface charge density contributed by the large number of phosphates moieties lining the tunnel inner wall, or the dielectric responses of the tunnel lumen and inner wall. In our simplified approach, we incorporated partial piecewise heterogeneity for the phenomenological charge densities and dielectric response properties and used three different ratios σ/ϵ: one in the upper-lower tunnel (σ_1/ϵ_1), one in the tunnel vestibule (σ_2/ϵ_2) and one in the tunnel constriction site (σ_3/ϵ_3).

6.7.1.2 Goodness of the fit of the model to experimental electrostatic potential observation

The main results derived from the theoretical analysis of the electrostatics of the ribosome exit tunnel, as displayed in Fig. 6.4(e) and Fig. 6.5, is the goodness of the fit of the physical model with the measured data points for the electrostatic potential in the ribosome exit tunnel earlier published by Carol Deutsch and co-workers [Lu et al. 2007]. The geometry, crystallographic data and physicochemical properties of the tunnel inner walls of the exit tunnel consistently explain the experimentally measured values for the potential from simple first physical principles. The model draws the attention on the main geometrical and physical features as determinants of the electrostatic potential profile and the derived electric field projected along the tunnel axis. Specifically, the geometrical variation induced by widening the tunnel radius at the exit of the tunnel (cone frustum) introduces a permanent difference in the electrostatic potential between the exit and the entry points of the tunnel. The rise in the potential when moving towards the vestibule can also be contributed by an increase in the dielectric response ϵ in the vestibule medium. This is energetically unfavorable to the positively charged amino acid residues as compared to their negatively charged amino acid counterparts. This provides a simple bioenergetic explanation to the observation that, proteome wide and across species, the protein sequences are slightly but significantly more enriched in negatively charged amino acid as compared to the positively charged amino acid residues [Requião et al. 2017]. This observation would be the consequence of a selection pressure in favor of the negatively charged residues as compared to the positively charged residues; the latter requiring more mechanical energy to traverse the ribosome exit tunnel.

The derived axial forces acting upon the nascent polypeptide stretch, within 50 residues upstream of the amino acid residue that is incorporated at the PTC, stand as a valuable quantitative model. The variation in the axial forces due to electrostatic interaction of the charged nascent chain with the ribosome exit tunnel has been estimated in a range from -10 pN to $+10$ pN in order of magnitude. More importantly, different profiles for these axial forces have been quantitatively related to synthetic polypeptides with arbitrarily charged residue distribution. Arbitrary synthetically engineered transcripts could, in principle, be used in high resolution optical tweezers multiple traps experiments to test experimentally the theoretical profiles of the axial forces acting upon such nascent polypeptides. The electrostatic model best fitted to the experimental data of Lu *et al.* [Lu et al. 2007] is the one combining a cone frustum section concatenated to a cylindrical section with a Lorentzian peak roughly located one third of the tunnel length away from the tunnel entry point. This particular model is used to derive, more accurately, the axial forces acting upon any nascent chain in the tunnel.

6.7.1.3 Model applications to study the contribution of the electrostatic interaction on elongation rate and for ribosomal mechanochemistry

The ordered list of axial forces at single residue resolution also allows to calculate the mechanical work required to overcome the electrostatic potential real profile in the exit tunnel at each residue elongation. From this, a Maxwell-Boltzmann correcting factor can be defined following similar developments as the ones exposed in Bustamante et al. 2004; Ribas-Arino and Marx 2012 and introduced in a seminal article by Bell in the context of cell to cell adhesion [Bell 1978]. These factors can correct, at single amino acid residue resolution, and in a sequence specific way, the elongation rate in TASEP-like modeling tools. For a given transcript, the specific contribution of the tunnel electrostatic interaction locally modulates the elongation rate in a range from minus 40% to plus 85% when compared to the average elongation rate; see supplemental material (E). An interesting advantage of these Maxwell-Boltzmann correcting factors lies with the way they are calculated. The exact local memory of the distribution of the charged amino acid residues is conserved for a sliding window of 50 residues that are upstream the site of incorporation of a new residue at the peptidyl transferase center PTC. This extended stretch of 50 residues is expected to be under the influence of the electrostatic interaction caused by the inner wall of the ribosome exit tunnel, along its whole axial length. All charged residues, positive and negative, embedded in the tunnel, additively contribute to the pace of the elongation process. The route of force transmission to the P site residue is through the nascent polypeptide's backbone as it is also the case for the tugging force generated by the spontaneous folding of the lengthening nascent chain out of the ribosome exit tunnel [Fritch et al. 2018]. Mechanical forces can alter the activation energy barriers that reactants have to overcome in the course of a chemical reaction to be converted into products. Intermediate transition states may be more easily attainable from the reactants when the system is experiencing an external force [Bustamante et al. 2004]. An effect of the external applied force is to provide mechanical work that will linearly decrease the activation energy even without changing the reactants' configurations or the transition state configuration [Bell 1978; Ribas-Arino and Marx 2012]. When the axial forces upon the nascent chain buried in the tunnel are exerted toward the tunnel exit, the Gibbs free energy barrier at the PTC is presumed to be decreased, the rate of the peptidyl-tRNA deacylation step at the P site and the global rate of the peptide bond formation are both expected to be increased. To our knowledge, the model presented here is the first one to take into account the whole size and shape of the ribosome exit tunnel and updates, at single residue resolution, the mobile 50−mer polypeptide window which is embedded in the tunnel. The position dependent precise value of the Maxwell-Boltzmann factor is determined by this spatially extended stretch of 50 residues with a specific charge distribution that is encrypted in the transcript being decoded. These elongation rate correcting factors are at codon resolution and keep the memory of the spatially extended

stretch of amino acid residues embedded in the tunnel. This is a clear improvement over the current state of the art in terms of realism and consistency of the elongation speed calculation. This can be contrasted, for instance, with studies where only positively charged residues within a limited number of residues upstream the incorporation site are considered and where arbitrarily fixed valued correcting factors are used to adjust for the electrostatic interaction in the tunnel [Sharma et al. 2018].

6.7.1.4 Model applications to study co-translational folding

The comparison of the axial force profiles of wild type protein sequences with mutant sequences as illustrated in the case of huntingtin (Fig. 6.12 and Fig. 6.13) may help to study the dynamical folding of a nascent protein that is still in contact with the ribosome. The tugging forces generated by the spontaneous folding of the unstructured segments during the peptide lengthening out of the tunnel were estimated by computer simulations at piconewtons order of magnitude by Fritch *et al.* [Fritch et al. 2018]. These spontaneous co-translational folding tugging forces acting on the nascent chain can be compensated for (or not) by pulling forces from the ribosome due to the electrostatic interaction in the ribosome exit tunnel. The landscape of co-translational folding of the wild type and mutant huntingtin nascent proteins may differ by a sheer difference in the distribution of the charged amino acid residues that are embedded in the full length of the tunnel. This would shed light on how the ribosome takes part in configuring folding intermediates [Thommen et al. 2017]. The specific pattern of the axial forces acting on the residues that are incorporated successively from the N-terminus to the C-terminus could prevent the emerging nascent chain from falling in kinetic traps, or in stable misfolded conformations, eventually resulting in protein aggregation. Our model allows a quantitative analysis of these axial forces profiles and a comparison of such profiles between correctly folded and misfolded protein conformations.

6.7.2 Reliability of model assumptions and model limitations

6.7.2.1 X-ray solved spatial structure data provide insights in phenomenological constants of the model

Publicly available X-ray crystallographic data of the large ribosome subunit were used in the study to assess the values of our model main phenomenological parameters for the charge surface densities in the different regions of the tunnel, namely the upper tunnel, the constriction site, the lower tunnel and the vestibule. The 2D maps that we produced of the phosphate moieties of the 23S rRNA and of the charged amino acid residues belonging to ribosomal proteins in the immediate vicinity of the tunnel wall were used to estimate the *bare formal* charge densities σ_1^* in the cylinder and σ_2^*

in the cone frustum (vestibule). The uniform 2D joint distribution assumption of the phosphate moieties on the cylinder surface of the tunnel wall was not rejected upon Monte Carlo simulations conducted under the null hypothesis of uniformity (p-value > 0.083). Chi-squared and Kolmogorov-Smirnov tests conducted on the 1D marginal distributions did not reject the uniformity hypothesis for the cylinder. The tunnel inner wall bare surface charge density on the cylinder surface is $\sigma_1^* = -2.1\,|e|/\mathrm{nm}^2$. The surface charge density on the tunnel vestibule is significantly higher at the exit port of the cone than at the entry port of the cone (KS test p-value = 0.008). The Kolmogorov-Smirnov test for the 1D-marginal distribution along z shows that more than 50% of the phosphate groups on the tunnel vestibule are located at the edge of the exit port of the cone. Structural data, observed potential data point measurements and coarse grained calculations of the electrostatic potential in the tunnel also provide different lines of evidence that the dielectric response (permittivity) also increases along the tunnel centerline when moving from the lower tunnel exit point towards the cone exit port, keeping the ratio $\sigma_2/\epsilon_2 = \sigma_{\mathrm{cone}}/\epsilon_{\mathrm{cone}}$ at least approximately constant. The average surface charge density on the truncated cone surface is $\sigma_2^* = -5.1\,|e|/\mathrm{nm}^2$ and is higher than the surface charge density on the cylinder. The X-ray crystallographic data showed that water molecules buried in the wall material in the immediate vicinity of the phosphate moieties are good candidates to explain the strong screening of the electrostatic potential. The Gouy-Chapman screening length ξ links the *formal bare* charge density directly contributed by the non-bridging oxygen atoms in the phosphate groups of the 23S rRNA to the *apparent* or *renormalized* charge density on the tunnel inner wall electrostatic surface within the non-linear screening theory. The electrostatic screening exponential factor was estimated under the Gouy-Chapman approach to be $e^{-\Delta/\xi} = e^{-5.97/1.05} = 0.0034$. The electrostatic potential in the centerline of the tunnel is reduced by a factor ~ 300 due to this water screening. The enrichment in 7 positively charged amino acid residues of ribosomal proteins at the constriction site is quantitatively consistent with the local rise of the electrostatic potential in this region. The parameters of the model fitted on the potential measurements provide an independent estimate of the ratio $\sigma_3/\epsilon_3 = 57.3\mathrm{MV/cm}$ in the constriction site. Using the minimal dielectric response $\epsilon_r = 4$ for a nascent protein occupying the tunnel medium, and using the constriction site estimated surface, the number of apparent positive charges is calculated to be ~ 7.

6.7.2.2 An alternative approach to the Poisson-Boltzmann theory for a confined micro-environment that is not in equilibrium with strong electrolytes in bulk solution

The Poisson-Boltzmann theory or its linearized Debye-Huckel version are not appropriate in the particular micro-environment inside of the ribosome exit tunnel. Essential assumptions are not met in the context of the ribosome exit tunnel. First, as

already noted, it is not a bulk solution with strong electrolytes present. Water molecules are not even abundant and the diameter of the tunnel too small to accommodate monovalent or bivalent ions in numbers. Second, the concept of ionic strength is hard to define inside the tunnel, given both the ill-defined concentration and homogeneity in this particularly confined environment. Third, the exit tunnel is not in equilibrium with any bulk solution during the progress of nascent protein elongation. For all these reasons, we relied directly on the simpler Poisson model only (or equivalently on the Coulomb law) which is believed to be more relevant in the (non-equilibrium) ribosome exit tunnel than a Poisson-Boltzmann model for which both the equilibrium assumption and the presence of abundant strong mobile electrolytes are not fully supported either experimentally or theoretically. The Poisson-Boltzmann theory does apply though on the outer surface of the whole ribosome which is in equilibrium with the electrolytes in the bulk solution of the cytoplasm [Baker et al. 2001]. The detailed chemical composition of the large ribosomal subunit and of the direct micro-environment of the ribosomal exit tunnel is known. Overall, the phosphate-ribose repeats building up the homopolymeric backbone of the 23S/28S-rRNA are the dominant molecular constituents of the tunnel inner wall materials. The fixed (negatively charged) phosphates moieties outnumber by at least one order of magnitudes any other charged chemical group (be it from the locally protruding ribosomal proteins origin or from the nascent polypeptide itself embedded in the tunnel): the single 23S/28S rRNA molecule entails more than $3,000/5,000$ phosphate moieties depending on the domain of life. This is primarily the origin of the observation of a net negative electrostatic potential inside the ribosome exit tunnel and even around the external surface of the complete ribosome. There may be polarization and induced dipole effects affecting locally the apparent negative charges harbored by the phosphate moieties inside the tunnel but, from the perspective of the full length of the nascent chain backbone, the curvilinear axis centered in the ribosome exit tunnel experiences the electrostatic effect of these numerous phosphate moieties lining up the tunnel inner wall. The spatial extension of this electrostatic interaction spans throughout the length of the ribosome exit tunnel over a functional distance of around 10 nm. This chemical environment is typical and specific of the ribosome in general and of the ribosomal exit tunnel in particular. With dielectric response of proteins (nascent chain) being $\epsilon \approx 4$ and of nucleic acid (tunnel wall main component) being $\epsilon \approx 8$, both much smaller than the dielectric response of aqueous medium in bulk solution ($\epsilon = 78$ for water), polarization effects in the exit tunnel due to dielectric discontinuities, in general, are not expected to be as important as the ones that would occur in bulk solution at equilibrium.

6.7.2.3 Scope, assumptions and model limitations

Our model of the electrostatic interaction of the ribosome exit tunnel with the nascent chain polypeptide relies on a number of critical assumptions which prevent to consider

the model as a completely realistic representation. As advocated by Lucent *et al.* [Lucent et al. 2010], the understanding of the complexity of molecular behavior in the ribosome exit tunnel should require an atomistic molecular dynamical description including the solvent confined to the tunnel as the medium inside the tunnel does not behave as a continuous isotropic dielectric medium. In our simplified approach, we incorporated partial stepwise heterogeneity for the phenomenological charge densities and dielectric response properties and used three different ratios σ/ϵ: one in the upper-lower tunnel, one in the tunnel vestibule and one in the tunnel constriction site. The exact size of the exit tunnel and the number of accommodated residues inside the tunnel are species (life domain) dependent and polypeptides dependent. The Euclidean distance between two consecutive $\alpha-$carbons is 3.5 Å in a peptide bond in trans configuration (fully stretched polypeptide). In the most compacted form of a polypeptide, i.e. the $\alpha-$helix, the distance between two consecutive $\alpha-$carbons, projected along the helix axis, is 1.5 Å. The median distance between these two upper and lower limit cases is 2.5 Å. The number of residues that could be accommodated in a 100 Å long tunnel would be 28, 40 and 66 depending on the polypeptide configuration. We adopted the median length between two $\alpha-$carbons (0.25 nm) and hypothesized as a general rule that the number of accommodated residues is 40 for a tunnel length of 10 nm. In electrostatics, the electric polarization is defined macroscopically by the volume density of the sum of all microscopic electric dipole moments. The individual dipole moments are either induced dipole moments (electronic cloud polarizability) or permanent dipole moments (of intrinsically polar molecules, e.g. water, $\alpha-$helix) reoriented by the imposed electric field. The electric polarisation is a function of the total electrical field. These physical effects are of fundamental importance in the context of the ribosomal exit tunnel. Unstructured nascent polypeptide sequences do not entail privileged permanent dipoles. On the contrary, it is known that $\alpha-$helices secondary structures entail permanent dipoles with values increasing with their length. The permanent electric dipole moment of an alpha helix secondary structure is oriented from the C-terminal to the N-terminal end of the sequence and is of the order $\vec{\mu}_{\mathrm{e},\,\alpha-\mathrm{helix}} = 3.5\,N\;D$ where N is the number of residues in the helix and D the Debye unit of the electric dipole moment. Such a $\alpha-$helix dipole moment is larger than the one of a single water molecule. The upper and lower tunnel cannot in general accommodate secondary structures other than $\alpha-$helices. When $\alpha-$helices secondary structures appear in the ribosome exit tunnel, they tend to align with the electric field along the $z-$axis. In the upper tunnel or in the lower tunnel just after the tunnel constriction, $\alpha-$helices orient in a parallel direction to the tunnel axis with the $\alpha-$helix dipole moment pointing toward the tunnel exit. This is favorable as it corresponds to the direction of the natural vectorial elongation, resulting in the global movement of the nascent peptide directed toward the tunnel exit, namely from the C-terminal end to the N-terminal end of the peptide embedded in the tunnel. On the contrary, in the end of the lower tunnel or vestibule end, the dipole moment of an $\alpha-$helix would tend to point in the opposite direction, namely toward the PTC. In both cases, this would reinforce locally the electric displacement. These events would locally

modulate the effects of our simplified electrostatic potential profile on the electric fields and axial forces experienced by a charged residue. Although, $\alpha-$helices secondary structures might occasionally form in the tunnel during the nascent chain elongation, for the sake of simplicity of our model, and because such events are transients and context dependent, we did not incorporate these events into our electrostatic model at this stage. Our simplified model applies better to elongation situations of unstructured (random coil or extended uncompacted) polypeptides. It should be noted however, that the profile of the electrostatic potential and the axial electric field derived from our model provide insights on the orientation trends of $\alpha-$helices nascent secondary structures that are consistent with observations of early compaction and early co-translational folding events occurring within the lower, upper tunnel and vestibule. Indeed, reversion of $\alpha-$helices in the lower tunnel end or in the tunnel vestibule have been observed for instance in [Liutkute et al. 2020; Mercier and Rodnina 2018]. Reversion of an $\alpha-$helix in the tunnel is the observed fact that the N-terminal end of the $\alpha-$helix peptide gets back inside the tunnel toward the PTC instead of progressing toward the exit end. These kind of events are the starting point of tertiary structure sometimes initiating at the lower end of tunnel and or the tunnel vestibule. The reversion propensity of $\alpha-$helices in the end of the tunnel might be favored by the orientation of the axial electric field pointing toward the PTC in this zone. Reversion of $\alpha-$helices would not be favorable in the very beginning of the upper tunnel near the PTC. The possible secondary structures that can start to form in the tunnel and the impact of their dipole moments on the local electric field in the tunnel have not been explicitly incorporated into our simplified model. It is also not known with full certainty whether or not the shape and geometry of the ribosome exit tunnel in the large subunit LSU of the ribosome stay the same during the translation process *in vivo* or if reversible continuous elastic deformations occur *in vivo*. We showed that the mechanical energy required to push the growing nascent chain through the LSU exit tunnel, even in difficult scenarios, would be smaller than the Gibbs free energy released from the transpeptidation and the hydrolysis of a single GTP. Overall, the widening of the radius along its central axis toward the exit of the tunnel is however known and contributes to the asymmetric electrostatic potential profile that we estimated. Alternative explanations of the rise in electrostatic potential could also be the rise in the dielectric response when moving from the lower tunnel to the vestibule. This could be due to an increase in abundance of free water molecules. The estimated electrostatic potential profile fits the available observed data rather well at least for the rabbit reticulocytes ribosomes. We must recognize, however, that we only relied on a small sample size of 4 to 6 point measurements. Complementary wider experimental studies on both prokaryotes and eukaryotes ribosomes would be beneficial.

To quantitatively estimate the axial forces applied on the nascent chain, we made a rigid body assumption or assumed the non-deformability of the nascent chain inside the tunnel. This assumption is most certainly not valid for all polypeptide chains and most probably not valid locally. However, this assumption could be legitimate on average and proteome wide. Indeed, the ribosome exit tunnel is universal, meaning

that all the polypeptides that are naturally occurring in the biosphere did traverse the tunnel at the time of their biosynthesis. All the amino acids have progressed through the entire length of the tunnel after they were incorporated in the nascent chain at the peptidyl transferase site. On average, as a first approximation, we can consider that these amino acids followed a centro-axial trajectory in the tunnel and experienced the effect of the electrostatic interaction upon the charged residues with which they are directly or indirectly bound. Fritch *et al.* recently showed that the spontaneous folding force was transmitted directly from the outside of the tunnel to the PTC center through the backbone of the nascent chain [Fritch et al. 2018]. This direct transmission route supports the rigid body assumption we made for the peptide buried in the tunnel.

6.7.3 Comparison to state of the art and literature in the field

The importance of electrostatic interactions should not be overstated and has to be quantitatively incorporated according to its weight to all other key determinants of the ribosome elongation rate. At least 5 determinants in the protein elongation rate are known and have been investigated for decades by the scientific community: tRNAs relative abundance and adaptation to mRNA codon usage (local/global tRNA adaptation index); exit tunnel electrostatic interaction; proline residues at the P and/or A site; downstream mRNA secondary structures hampering the movement of the ribosome toward the mRNA 3' terminus; ribosome interference (traffic jam) and ribosomes pool limited resources. Although Charneski *et al.* [Charneski and Hurst 2013] argued that the electrostatic interaction was the major determinant, the electrostatic interaction is indeed only one of these 5 determinants. To dissect the variance in the elongation rate, all these 5 factors must be taken into account altogether, especially if these factors are meant to be used as predictors for Ribosome profiling (Ribo-Seq) data or are aimed at improving TASEP models. Our study focused only on the electrostatic interaction. Current models of the ribosome exit tunnel electrostatic interaction are incomplete and a correct model is lacking. Our study is aimed to contribute to the improvement of the predictive power of such models. The electrostatic interaction is not the only force acting on the nascent chain during elongation, the entropy driving force and the folding forces acting on the chain also contribute, especially as soon as the N-terminal end of the nascent chain has made its pass through the tunnel (after the ribosome translated the first $\sim 45-50$ codons). Our study did not address the calculation of the entropy driven forces. Position specific biases in the distribution of the 5 determinants contributing to the elongation rate are antagonizing each other and blur our understanding of the elongation rate. Two references [Dao Duc and Song 2018; Tuller et al. 2010] show instances of such position specific biases. Tuller *et al.* [Tuller et al. 2010] initially inferred from Ribo-Seq data that the first 30-50 codons at the 5'-end in yeasts transcripts were low efficiency codons (rare or poorly matched codons to the tRNAs 'pool) and termed this part the 'low efficiency ramp' or the $5'-$ramp. This initial $5'-$ramp (due

to slow accommodation and proofreading of tRNAs at the A site at the beginning of mRNA sequences) was considered to be the effect of a selection pressure enabling to limit the occurrence of later ribosome traffic jams downstream in the course of mRNA translation. This 5'ramp explanation was later argued to be insufficient [Dao Duc and Song 2018]. In contrast, the results of Dao Duc *et al.* [ibid.] suggested that while the N-terminus of the nascent polypeptide has not exited from the tunnel, positively charged amino acids in specific parts of the polypeptide actually facilitate the elongation speed, while the opposite is true for negatively charged amino-acids. The statistical analysis by Dao Duc [ibid.] showed that the presence of positive and negative charges in the upper tunnel may respectively facilitate and inhibit elongation as the nascent chain makes its initial pass through the tunnel. These authors found that the number of positively charged residues in the window [1:11] and the number of negatively charged residues in the window [6:14] upstream the A-site are important features with opposite effects; the former facilitates elongation, while the latter slows down elongation. Overall these results are fully consistent with the shape of the electrostatic potential profile of our model: the longitudinal (axial) direction of the force that a positively charged particle would experience along the tunnel points from the P site toward the tunnel exit end at least when the first 15 residues of the nascent chain enter the tunnel. This is the opposite for negatively charged particles. The electrostatic potential profile also shows that the longitudinal (axial) direction of the force that a positively charged particle would experience along the tunnel when the last 15-20 C-terminal residues are still in the tunnel exit points from the exit toward the P-site (upper tunnel). This is also consistent with the results of Nissley *et al.* [Nissley et al. 2020]. Dao Duc *et al.* [Dao Duc and Song 2018] made the case that once the N-terminus has exited the tunnel, the hydropathy of the part of the nascent polypeptide within the ribosome plays a major role in governing the elongation rate variation. These authors concluded that the movement of the polypeptide inside the tunnel is driven by two distinct biophysical mechanisms. First, when the peptide chain has not yet exited the tunnel, electrostatic interactions in the tunnel play a major role in regulating the movement of the chain down the exit tunnel. Second, when the peptide chain has reached a certain length and its N-terminus has exited the tunnel, it is the structure of the chain itself (captured through the hydropathy) that determines its movement through the tunnel. It should be added that the entropy driving forces upon protein folding outside the tunnel or forces exerted from chaperone proteins also contribute and should be quantified. The electrostatic potential mathematical model that we proposed provides insights into the real measurements that were made in the pioneering experimental studies and that could still be made in the future. It should be emphasized that electrostatic potential measurements should always be conducted in association with precise measurements of size and shape of the tunnel and accurate positional mapping along the tunnel axis.

6.7.4 Future perspectives

As future perspectives in X-ray crystallographic data mining, we suggest to compare the 2D maps of charged moieties in the immediate vicinity of the tunnel wall for ribosomes from different species across the three domains of life in order to investigate the common shared patterns.

We expect the *in silico* research community to assign itself the task of using our suggested electrostatic model, and the ordered list of Maxwell-Boltzmann factors derived from it, to modulate the elongation rate for a better quantitative account of the effect of the tunnel on the charged amino acid residues. On average, it is believed that the comparison of the electrostatic interaction with the exit tunnel of any two different nascent chains, by applying our model, can provide quantitative insights on the effects of the difference in charged amino acid distribution across their primary sequences. This paves the way to a variety of bioinformatic studies on transcriptomic and proteomic data to shed light on translational control. Immediate perspectives and objectives will address (a) accurate predictions of ribosome footprints in Ribo-Seq profiling ensemble experiments; (b) precise dynamical predictions of the speed of elongation in single mRNA molecules experiments; (c) quantitative predictions of the measured tugging force profiles on nascent polypeptide chain emerging from the ribosome exit tunnel in high resolution multiple traps optical tweezers experiments to be conducted on tethered ribosomes *in vitro*; (d) experimental measurement of the strength of the electric fields through vibrational Stark spectroscopy; (e) comparison of axial forces profiles associated to correctly folded or misfolded proteins for the study of co-translational folding and protein aggregation mechanisms.

The model presented in this study consistently connects different results and experimental observations coming from different fields in molecular biology, X-ray crystallography, structural and physical chemistry, synthetic and multi-omics biology and provides a clear picture of the electrostatic interactions in the ribosome exit tunnel and their effects on the protein elongation rate.

6.8 Reversing the relative time courses of the peptide bond reaction with oligopeptides of different lengths and charged amino acid distributions in the ribosome exit tunnel (CSBJ02)

Sections `6.11.2` and `6.12` below (respectively `2.2` and `3` of the third published paper CSBJ02), are the key elements in the context of **`Ribosomer`**. A visual comparison of Fig.6.15 and Fig.6.16 highlights the key findings.

CSBJ02 abstract

The kinetics of the protein elongation cycle by the ribosome depends on intertwined factors. One of these factors is the electrostatic interaction of the nascent protein with the ribosome exit tunnel. In this computational biology theoretical study, we focus on the rate of the peptide bond formation and its dependence on the ribosome exit tunnel electrostatic potential profile. We quantitatively predict how oligopeptides of variable lengths can affect the peptide bond formation rate. We applied the Michaelis-Menten model as previously extended to incorporate the mechano-biochemical effects of forces on the rate of reaction at the catalytic site of the ribosome. For a given pair of carboxy-terminal amino acid substrate at the P- and an aminoacyl-tRNA at the A- sites, the relative time courses of the peptide bond formation reaction can be reversed depending on the oligopeptide sequence embedded in the tunnel and their variable lengths from the P-site. The reversal is predicted to occur from a shift in positions of charged amino acids upstream in the oligopeptidyl-tRNA at the P-site. The position shift must be adjusted by clever design of the oligopeptide probes using the electrostatic potential profile along the exit tunnel axial path. These predicted quantitative results bring strong evidence of the importance and relative contribution of the electrostatic interaction of the ribosome exit tunnel with the nascent peptide chain during elongation.

6.9 CSBJ02 introduction

For more than five decades, attempts to model protein synthesis and mRNA translation from first principles have been pursued extensively [Haar 2012; MacDonald and Gibbs 1969; MacDonald et al. 1968; Zur and Tuller 2016]. Many factors influence translation speeds across a single transcript (mRNA). Several important factors have been identified in the literature. These include differences in cognate, near-cognate, and non-cognate tRNA relative abundance [Dana and Tuller 2014; Dao Duc and Song

2018; Gorochowski et al. 2015b], tRNA enzymatic modifications [Lyu et al. 2020; Novoa et al. 2012], amino acid residues distribution within the nascent-chain embedded inside the ribosome exit tunnel [Charneski and Hurst 2013; Requião et al. 2016; Sabi and Tuller 2015], mRNA secondary structure [Desai et al. 2019; Liu et al. 2014b; Wen et al. 2008], proline residues at either the A or P site of the ribosome [Artieri and Fraser 2014; Pavlov et al. 2009], steric hindrance between contiguous ribosomes translating the same mRNA molecule [Shaw et al. 2003], and the finite resource of the ribosome pool available in the cell [Greulich et al. 2012; Raveh et al. 2016; Shah et al. 2013]. The influences of these many contributing elements are intertwined, which complicates a full understanding of the individual factors. A number of knowledge gaps still prevail and controversies are still unresolved [Artieri and Fraser 2014; Requião et al. 2016; Sabi and Tuller 2015]. For instance, one study has argued that the charged residues are the major determinants of ribosomal velocity [Charneski and Hurst 2013]. Another study described the ribosome exit tunnel as a protein-sensitive channel with gate-latch action. Their authors argued that side chain specific recognition in the ribosome exit tunnel plays an active role in protein elongation regulation and translational folding [Petrone et al. 2008].

At the catalytic center of the ribosome, the enhancement of the rate of peptide bond formation is due to a precise positioning of the two substrates within the active site (peptidyl-tRNA at the P-site and amino-acylated-tRNA at the A-site) and to the very specific electrostatic local environment [Sharma et al. 2005]. The catalytic environment results from the dynamic configuration occurring within the large subunit of the ribosome that can accommodate a variety of substrates pair combination. The C-terminal amino acid of the peptidyl-tRNA at the P-site and the amino acid acylated on the tRNA at the A-site (aa-tRNA) can be any of the twenty amino acids. The number of different pairs of substrates is thus 400, which shows a versatility rarely met for a classical protein enzyme. This variety in substrate pairs only explains part of the variance of the peptide formation rates. The distribution of charged amino acids upstream of the nascent chain also affects the kinetics of the peptide bond formation [Rodnina et al. 2006; Wohlgemuth et al. 2008]. The peptide bond formation at the peptidyl transferase center (PTC) is an entropically driven process achieved by reducing the mobility of the substrates. This happens through precise positioning of the two substrates with respect to each other. As thoroughly reviewed by Rodnina for prokaryotes [Rodnina 2018], there are currently two models for the movement of protons in the active site of the ribosome, both involving a single isolated water molecule, and describing the reaction scheme of the peptide bond formation. According to the first model, a concerted 'eight-membered proton shuttle' involving 2' and 3'−OH on the tRNA ribose sugar of adenine 76 at the P-site explains the proton movement. According to the second model, a proton wire mechanism involving $2' - OH$ of the ribose in adenosine 2451 of the 23S rRNA explains the proton movement. A solvation effect with an isolated single water molecule, near the P-site tRNA A76 ribose sugar at the PTC is involved in the proton shuttle model; or near the orthophosphate between C75 and A76 of the A-site tRNA

at the PTC, in the proton wire model [Beringer and Rodnina 2007; Pape et al. 1999; Polikanov et al. 2014; Rodnina et al. 2006; Schmeing et al. 2005; Sharma et al. 2005; Sievers et al. 2004; Simonovic and Steitz 2009; Trobro and Åqvist 2005, 2006; Wallin and Aqvist 2010; Wohlgemuth et al. 2008]. In both models, a nucleophic attack of the amino group of aa-tRNA from the A-site on the carbonyl carbon of the ester bond in the peptidyl-tRNA in the P-site is key in this reaction scheme.

Steric hindrance near the C-terminal end of the peptidyl-tRNA hampers access to the nucleophilic attack of the $\alpha-$ amino group from the aa-tRNA at the A-site to the carbonyl carbon atom in the ester bond with the tRNA at the P-site. During the peptidyl transfer reaction, the α-amino group of aminoacyl-tRNA positioned in the A site of the ribosome nucleophilically attacks the carbonyl carbon at the ester bond of the peptidyl-tRNA in the P site. This transpeptidation results in a peptidyl-tRNA extended by one amino acid in the A site and a deacylated tRNA in the P site [Simonovic and Steitz 2009]. The nucleophilic attack is believed to be facilitated if the peptidyl-group at the P-site is pulled away from the carbonyl group. This would reduce steric hindrance and open access for the nucleophilic attack from the aa-tRNA A-site amino group. The peptide bond formation between the two aminoacylated-tRNAs proceeds 10 million times faster when catalyzed by the ribosome than when uncatalyzed in bulk solution [Beringer and Rodnina 2007]. The ribosome-catalyzed peptide bond formation kinetics has been extensively studied for decades and is known to be affected by the particular context of the upstream amino acid sequence [Rodnina et al. 2006].

The theory of kinetics of catalysis relies on the transition state theory that was introduced by Henry Eyring who linked the rate constant of a chemical reaction to the Gibbs free activation energy of the transition state (TS) [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983]. In this representation, a catalyzer accelerates a (bio)chemical reaction through a significant reduction in the transition Gibbs free energy barrier that the reactants have to overcome [Joiret et al. 2023a]. In our previously published work ibid., we hypothesized that the physical forces transmitted mainly through the backbone of the peptidyl-tRNA play a role in the reduction of the Gibbs free energy barrier of the transition state. The mechanical work of these physical forces affects the activation Gibbs free energy of the transition state. The modulation of the Gibbs free energy activation barrier changes the reaction rate constant through Eyring's relation [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983]. Computing the mechanical work requires the knowledge of the profile of the axial forces acting on the backbone of the peptidyl-tRNA and of the displacement that results from the peptide elongation. One of the main contributing forces acting on the nascent peptide chain is due to the electrostatic interaction arising from the negative electrostatic potential [Lu and Deutsch 2008; Lu et al. 2007]. The negative electrostatic potential is caused by the presence of the large number of phosphates moieties lining the inner wall of the ribosome exit tunnel and also the inner surface cavity around the PTC, mainly composed of rRNA [Joiret et al. 2022b]. The electrostatic environment exerts forces

on the 4 amino acids that are naturally charged in physiological conditions, namely arginine R, lysine K, glutamate E, and aspartate D. These charged amino acids are occasionally incorporated into the peptide nascent chain as determined by their DNA sequence.

Measuring the kinetics of the elongation cycle of the ribosome at codon resolution is a difficult task. Even more experimentally difficult is to split the time course measurement of the cycle into its main substeps: (i) tRNA accommodation and codon-anticodon proofreading at the A site ($k \simeq 0.010\,\mathrm{ms}^{-1}$ in *E.coli*); (ii) peptide bond formation between the carboxy-terminal amino acid at the P-site and the aa-tRNA at the A-site ($k \simeq 10\,\mathrm{ms}^{-1}$ in *E.coli*); (iii) ribosome translocation to the next codon and unloaded-tRNA eviction at the E site ($k \simeq 2.5\,\mathrm{ms}^{-1}$ in *E.coli*). The peptide bond formation is normally not the rate limiting step in elongation. Hence, its effect on translational control is limited. This is true in normal condition in most of the cases. However, if, by chance, the primary sequence of the polypeptide chain upstream the P-site entails a 'difficult' distribution in charged amino acid residues embedded in the tunnel, the modulation of the peptide bond formation rate due to the electrostatic interaction with the tunnel could turn this step into a limiting step. Other well-known situations where the peptide bond formation is the limiting step are when consecutive prolines, as donor and acceptor substrates, occupy the P-site and A-site. In practice, studying the catalyzed kinetics of peptide bond formation is possible only when peptidyl transfer is uncoupled from the very first step of elongation, i.e., accommodation of the aminoacyl-tRNA at the A site [Beringer and Rodnina 2007; Pape et al. 1999]. The reason is that the accommodation rate of aa-tRNA in the A site is in the range $5 - 10\ \mathrm{s}^{-1}$ and peptide bond formation follows instantaneously. Because accommodation precedes peptide bond formation, it limits the rate of product formation since it is much slower than the peptidyl transfer [Pape et al. 1999].

One way to circumvent the accommodation rate limiting step is to use substrate analogs that bind to the A site rapidly and do not require accommodation. If the full length aa-tRNA is replaced by the shorter puromycin substrate as the last acceptor substrate, the kinetics of the catalytic chemical step can be monitored experimentally by the quench flow technique and is amenable to quantitative measurements [Beringer and Rodnina 2007]. These experimental measurements were conducted on prokaryotic ribosomes by Rodnina and coworkers [Wohlgemuth et al. 2008] in the elongation minimal case when the donor substrate is the minimal dipeptidyl-tRNA, i.e. fMet – X – tRNA and the acceptor substrate is puromycin, Pmn. X can be any of the 20 natural amino acids. The product of the catalyzed reaction, peptidyl-puromycin, is released from the ribozyme upon completion of the reaction. The experimental initial condition for the puromycin concentration was 20 mM and the reaction rate constant measurements made at $37°C$ were obtained by single exponential fitting [ibid.]. The use of puromycin as the last A-site substrate acceptor allows easier experimental measurements of the time courses of peptide bond formation in a quench-flow apparatus measuring the

reaction kinetics. In our earlier works, we made use of the electrostatic local profile in the immediate vicinity of the catalytic center of the ribosome to quantitatively interpret these pioneering experimental results [Joiret et al. 2023a; Wohlgemuth et al. 2008]. From the electrostatic potential profile, we calculated the axial electric field and axial force acting locally on the charged amino acids. These forces modulate the activation energy of the peptide bond formation reaction. The modulation of the reaction rate constant and the time course of the peptide bond formation were calculated accordingly and compared to the experimental kinetics measurements that were made *in vitro* in cell-free extracts [Wohlgemuth et al. 2008]. In summary, the fold change in the median rate of the peptide bond formation between a charged amino acid donor substrate and puromycin as the acceptor substrate was calculated to be 3.45 as compared to a neutral amino acid donor. Following [ibid.], we call the median time course, the time to achieve the peptide bond formation of 50% of the amino acid when at the donor site. The experimentally observed values for the median time courses of peptide bond formation with lysine or arginine as donor substrate were 7.2 ms and 7.8 ms, respectively, to be compared with a median time course of 27.1 ms for neutral amino acids as donor substrate (median time course for alanine, serine, phenylalanine and valine) [Joiret et al. 2023a; Wohlgemuth et al. 2008]. Interestingly, when aspartate is the donor substrate, the calculated fold change is $\frac{1}{3.45} = 0.29$ to be compared with an experimentally measured median time course of 91.5 ms for aspartate. Overall, aspartate (negatively charged aa) is 3.45 times slower than a neutral amino acid, whereas lysine or arginine (positively charged aa) is 3.45 times faster than a neutral amino acid when acting as donor substrate in the minimal case of a dipeptidyl-tRNA at the P-site with puromycin as the acceptor substrate at the A-site. The experimentally observed order of magnitude for the fold change on these median time courses is qualitatively and quantitatively in good agreement with the theoretical calculations [Joiret et al. 2023a].

Here, we expand our previous work to the cases considering the electrostatic interaction on the nascent oligopeptide chain deeper in the exit tunnel. We investigate the effects of the axial forces originating from a much further region than the catalytic center vicinity. A longer oligopeptide acylated to the tRNA at the P-site (much longer than a dipeptide) would probe the electrostatic interaction deeper in the ribosome exit tunnel. The resulting axial forces would still be transmitted to the carboxy-terminal end of the oligopeptidyl-tRNA through the nascent chain backbone and affect the transition state energy barrier of the peptide bond formation. The profiles of the electrostatic potential and the electric field were determined in the full extent of the ribosome exit tunnel in previous works by ourselves and others [Joiret et al. 2022b; Lu et al. 2007]. Those results are used to compute the axial forces acting on charged amino acids arbitrarily distributed in oligopeptides embedded in the ribosome exit tunnel.

One of the big challenges in the field of translational control and protein elongation is how to address the kinetics of the processes involved. Our approach focused on a well identified step in the protein elongation cycle and provides a quantitative tool to

understand the impact of mechanochemical factors on the peptide bond formation rate. The interest of the incorporation of a term accounting for the effect of mechanical forces on the ribosomal catalytic activity and on the equation describing the kinetics is that it allows quantitative estimations of the peptide bond formation rate in different contexts. These quantitative estimations can be confronted with experimental measurements in different settings where some of the above intertwined factors may be adjusted or not. Altogether, these quantitative kinetics approaches will help determine the direct causal links between the factors allegedly affecting, or not, the protein elongation cycle.

6.10 CSBJ02 materials and methods

This study builds on the model developed in a previous contribution [Joiret et al. 2023a]. We briefly recall the key concepts below, before applying them to the case of an elongating nascent polypeptide extending inside the ribosome exit tunnel that is the focus of this study.

6.10.1 Effect of mechanical forces on chemical reaction kinetics

Applying external forces on molecules involved in catalyzed or uncatalyzed chemical reactions affects the kinetics of the reactions. The mechanical work of these applied mechanical forces can quantitatively be incorporated in the calculation of the activation Gibbs free energy of the transition state as conceptually introduced by Bell [Bell 1978], Bustamante [Bustamante et al. 2004] and others [Ribas-Arino and Marx 2012]:

$$\Delta G^{\ddagger 0}(\vec{\mathbf{F}}) = \Delta G^{\ddagger 0}(\mathbf{0}) - \int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}} \tag{6.32}$$

where $\Delta G^{\ddagger 0}(\vec{\mathbf{F}})$ is the activation energy for the transition state in the presence of an applied force acting on the system, $\Delta G^{\ddagger 0}(\mathbf{0}) \sim +14\,\mathrm{kcal/mol} = +97.2\ \mathrm{pN \cdot nm}$ is here the activation energy of the deacylation and peptidyl transfer for the transition state without any applied force [Bustamante et al. 2004; Rodnina et al. 2006; Sievers et al. 2004], and $W = \int \vec{\mathbf{F}} \cdot \mathbf{d}\vec{\mathbf{x}}$ is the mechanical work exerted by the force upon a test body along its curvilinear path. The mechanical work W is algebraically positive if the force and the displacement are parallel or negative if they are antiparallel. In the former case, $\Delta G^{\ddagger 0}(\vec{\mathbf{F}})$ is smaller than $\Delta G^{\ddagger 0}(\mathbf{0})$, whereas it is larger in the latter case. In turn, the modulation of the Gibbs free energy activation barrier changes the reaction rate constant through Eyring's relation [Anslyn and Dougherty 2005; Eyring 1935; Laidler

and King 1983]:

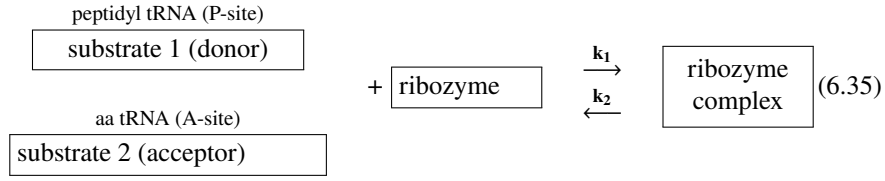
$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot e^{-\Delta G^{\ddagger 0(\vec{\mathbf{F}})}/k_B T} \tag{6.33}$$

$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot e^{-\left(\frac{\Delta G^{\ddagger 0(\mathbf{0})}}{k_B T} - \frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}\right)} = k(\mathbf{0}) \cdot e^{\frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}} \tag{6.34}$$

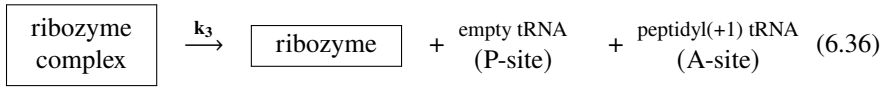
where $k(\vec{\mathbf{F}})$ is the reaction rate constant of the rate limiting step in the presence of an applied force upon the substrate at the P-site, $k(\mathbf{0})$ is the reaction rate constant in the absence of applied force. k_B, h and κ are Boltzmann's constant, Planck's constant and transmission coefficient respectively [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983].

6.10.2 Modified Michaelis-Menten Kinetics

The peptide bond formation kinetics at the catalytic center of the ribosome can be described by the Michaelis-Menten model where the aminoacyl-tRNA (acceptor substrate) is the canonical substrate and where the C-terminal amino acid of the peptidyl-tRNA (donor substrate) behaves as an allosteric substrate [Joiret et al. 2023a].



(6.35)



(6.36)

The rate of peptide bond formation, $k_{\text{pep}} = \frac{dP}{dt}$, is written as follows:

$$\frac{dP}{dt} = v'_{max} \cdot \frac{S}{K_{1/2} + S} \tag{6.37}$$

$$v'_{max} = \frac{v_{max}}{1 + \frac{S_{allo}}{k_{allo}}} \tag{6.38}$$

where $K_{1/2} = \frac{k_2+k_3}{k_1}$ is the Michaelis constant of substrate 2, S, at the A-site; S_{allo} is the C-terminal amino acid at the peptidyl-tRNA (substrate 1); k_{allo} is the reaction rate constant between substrate 1 and the ribozyme (PTC) at the P-site and v_{max} is the maximum rate in the absence of allosteric effect ($v_{max} = k_3 \cdot [\text{ribozyme}]$). Incorporating the right hand side Maxwell-Boltzmann factor of (6.34) into (4.37) to account for the effect of mechanical forces in the kinetics of the ribozyme catalyzed peptide bond formation leads to the final kinetics equation:

$$\frac{dP}{dt} = e^{\frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}} \cdot v'_{max}(\mathbf{0}) \cdot \frac{S}{K_{1/2} + S} \tag{6.39}$$

where $v'_{max}(\mathbf{0})$ is the maximum reaction rate in the absence of external force. In single molecule experiments, P, when proper normalized, represents the probability of the formation of the peptide bond over time in equation (6.39).

6.10.2.1 Electrostatic Potential and Axial Forces Profile Contributed by the Catalytic Center Cavity and by the Ribosome Exit Tunnel.

A full closed analytical expression for the electrostatic potential inside the ribosome exit tunnel was proposed earlier which was fitted to the experimental data point measurements of the electrostatic potential obtained by Deutsch and coworkers on ribosomes collected from rabbit reticulocytes [Lu and Deutsch 2008; Lu et al. 2007]. More recently, an electrostatic model of the cavity around the catalytic center considering the most simple shape fulfilling the minimal geometrical constraints existing between the ribosome peptide exit tunnel, the mRNA channel and the size of the aminoacylated-tRNAs was developed in Joiret et al. 2023a leading to the potential profile shown in Figure 6.14. The axial force profile allows to determine the mechanical work exerted upon a charged amino acid residue during the peptide elongation process. The method to derive the calculation of the axial forces acting upon any charged residue distribution from the electrostatic potential profile inside the exit tunnel is developed in Joiret *et al.* [Joiret et al. 2022b].

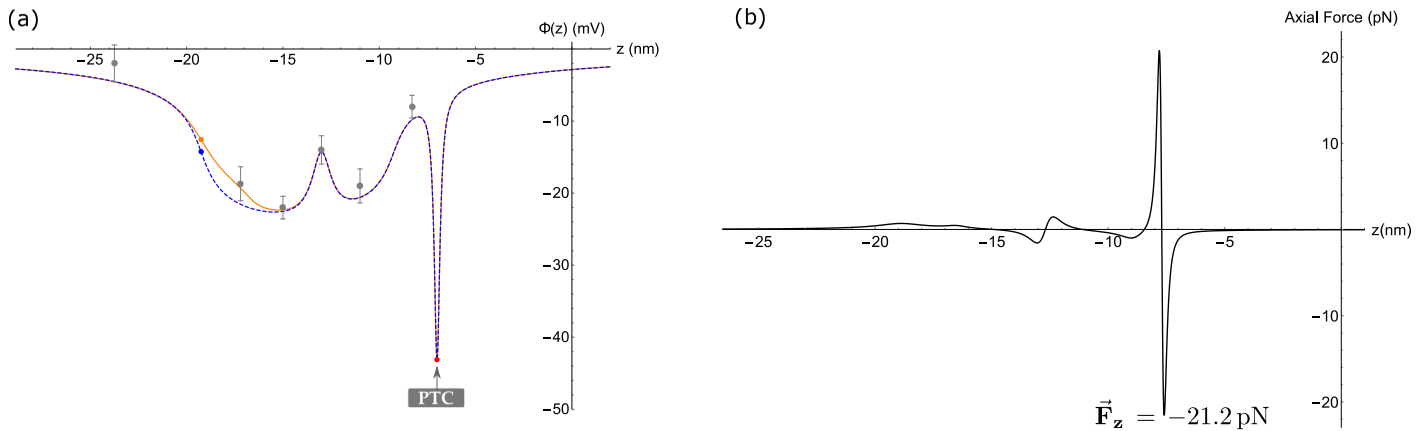


Figure 6.14: Electrostatic potential and axial force profiles around the ribosome PTC and the exit tunnel. **(a)** Electrostatic potential profile $\Phi(z)$ as modeled in [Joiret et al. 2023a; Joiret et al. 2022b]. z is the coordinate along the tunnel center line: $z = -7$ nm is the position of the PTC and $z < -7$ nm is the starting region of the exit tunnel towards negative z. The PTC is ~ 7 nm away from the mRNA decoding center ($z = 0$). The dashed blue line is the potential profile assuming a cylindrical shape at the distal end of the tunnel. Orange line is the potential profile assuming a cone frustum (enlarged tunnel exit). Error bars: experimental data point measurements by Deutsch & coworkers [Lu et al. 2007] of the potential inside rabbit reticulocytes' ribosome exit tunnel. **(b)** Axial force profile along the $z-$axis. The axial force $\vec{\mathbf{F}}_{\mathbf{z}} = q_e \cdot \vec{\mathbf{E}}_{\mathbf{z}}$, where $\vec{\mathbf{E}}_{\mathbf{z}} = -\frac{\partial \Phi(z)}{\partial z}$, the electric field along the $z-$ axis, is the negative of the first derivative of the electrostatic potential $\Phi(z)$ profile (shown in **a**). The axial force on a unit positive charge $|q_e|$ at the P-site is $\vec{\mathbf{F}}_{\mathbf{z}} = -21.2$ pN as calculated in reference [Joiret et al. 2023a].

6.10.2.2 Elongation extended case with oligopeptides of variable lengths and puromycin.

Building on the above explained framework, here, we study the impact of the electrostatic interaction between the nascent chain and the ribosome exit tunnel on the peptidyl transfer kinetics at the catalytic center. To do that, longer oligopeptides are required as P-site peptidyl-tRNA substrates. The global reaction scheme for the elongation extended case is:

$$\underbrace{\text{fMet-oligoP-X-tRNA}}_{P-site} + \underbrace{\text{(Pmn)}}_{A-site} \rightarrow \underbrace{\text{empty tRNA}}_{P-site} + \underbrace{\text{fMet-oligoP-X-Pmn}}_{A-site} \tag{6.40}$$

The number of amino acid residues separating the PTC from the ribosome exit tunnel entry point was estimated to be 5 residues from the P-site or 6 residues from the A-site [Deutsch 2014]. In what follows, three different oligopeptide lengths are used to probe the effect of the electrostatic interaction inside the ribosome exit tunnel. The oligopeptide lengths are chosen such that they extend inside the beginning of the ribosome exit tunnel and almost up to the tunnel vestibule exit end:

- a 10-mer oligopeptide extended length, with charged residues distributed specifically at position 6, 7, 8 from the C-terminal end of the peptidyl-tRNA at the P-site;
- a 22-mer oligopeptide extended length, with charged residues distributed specifically at position 18, 19, 20;
- a 40-mer oligopeptide extended length, with charged residues distributed specifically at position 35, 36, 37, 38, 39 from the C-terminal end of the peptidyl-tRNA at the the P-site.

Except for these locally positioned charged residues, the other amino acid residues are neutral in these oligopeptides.

6.10.2.3 Output variable uncertainty of the model estimated by differential calculus and uncertainty propagation from the input variables

There are a number of assumptions in the electrostatic model and several sources of uncertainties. These were extensively described and discussed in references [Joiret et al. 2023a; Joiret et al. 2022b]. The theoretical uncertainty in the peptide bond formation theoretical rate can be estimated from the differential calculus applied to the electrostatic potential model. The propagation of the numerical errors is detailed in the appendix D.

The experimental uncertainties in the peptide bond formation rate k are around $\Delta k = \pm 1.5\,\mathrm{sec}^{-1}$ [Wohlgemuth et al. 2008]. Given that $\Delta\tau_{1/2} \sim \frac{\Delta k}{k^2}$, it can be estimated that the experimental uncertainties in the median time course measurements are around $1 - 3$ ms.

6.11 CSBJ02 results

6.11.1 Axial force profiles and forces transmitted to the carboxy-terminal end of the tRNA at the P-site

These synthetic peptide oligomers sequences were designed such that they are experiencing qualitatively very different electrostatic potential profiles, resulting in different axial forces. In the first case (10-mer), the tunnel electrostatic potential profile in the $z = [-8.9,\ -9.4]$ nm spatial range will mainly contribute to the axial force transmitted through the backbone of peptidyl-tRNA. In the second case (22-mer), the $z = [-11.95,\ -12.45]$ nm spatial range will mainly contribute to the axial force. In the last case (40-mer), the $z = [-16.20,\ -17.20]$ nm spatial range will mainly contribute to the axial force. In the first case, the electrostatic potential is decreasing, whereas in the second and third cases, the potential is increasing toward the tunnel exit direction. The directions of the axial forces being exerted on positively (or negatively) charged test residues are opposite. The qualitative effects on the kinetics of the peptidyl transfer reaction with puromycin at the A-site are expected to be in opposite directions as well. This is the cause of the reversing in the reaction relative time courses.

In the first case, the 10-mer nascent oligopeptide acylated tRNA at the P-site threads through the ribosome tunnel entry (Figure 6.15 a-c). The axial force acting upon the backbone of the peptidyl-tRNA and transmitted to the tRNA at the P-site caused by three positively charged residues within the oligopeptide is calculated using the algorithm that was earlier exposed [Joiret et al. 2022b] and is $\vec{\mathbf{F}}_{\mathbf{z}} = -2.61$ pN. The mechanical work is estimated to be $W = 2.61\,\mathrm{pN} \cdot 0.25\,\mathrm{nm} = 0.65\,\mathrm{pN} \cdot \mathrm{nm}$ (force and displacement are parallel).

In the second case, the 22-mer nascent oligopeptide acylated tRNA at the P-site threads further through the ribosome tunnel entry (Figure 6.16 a-c). Similarly to the previous case, the axial force acting upon the backbone of the peptidyl-tRNA and transmitted to the tRNA at the P-site caused by three positively charged residues within the oligopeptide is $\vec{\mathbf{F}}_{\mathbf{z}} = +3.42$ pN. The mechanical work is estimated to be $W = -3.42\,\mathrm{pN} \cdot 0.25\,\mathrm{nm} = -0.86\,\mathrm{pN} \cdot \mathrm{nm}$ (force and displacement are antiparallel).

In the third case, the 40-mer nascent oligopeptide acylated tRNA at the P-site threads deeper through the ribosome tunnel down to the vestibule exit (Figure 6.17 a-c). The

axial force acting upon the backbone of the peptidyl-tRNA, caused by the five positively charged residues and transmitted to the tRNA at the P-site is $\vec{\mathbf{F}}_{\mathbf{z}} = +2.1\,\text{pN}$. The mechanical work is estimated to be $W = -2.1\,\text{pN} \cdot 0.25\,\text{nm} = -0.525\ \text{pN} \cdot \text{nm}$ (force and displacement are antiparallel).

6.11.2 Peptide bond formation reaction rates with 10, 22 and 40-mer oligopeptide and puromycin.

We predict the Maxwell-Boltzmann factors and the reaction rate constant values tabulated in Table 6.11.1. The predicted time courses of the 10-mer oligopeptidyl-tRNA reaction with Pmn are shown in Figure 6.15 (f). The red and blue lines are the time courses calculated from equation (6.39) for the positively and negatively charged case respectively as compared to the neutral case (dashed line). The positively charged amino acid C-terminal oligopeptidyl-transfer rate is $k(\mathbf{0}) \times \exp(\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}/k_B T)$ = neutral rate $\times \exp(0.65/4.28)$ = neutral rate $\times$ 1.16 (red line). The negatively charged amino acid oligopeptidyl-transfer rate is $k(\mathbf{0}) \times \exp(-0.65/4.28)$ = neutral rate $\times$ 0.86 (blue line). At their specific positions from $z = -8.95$ nm to $z = -9.45$ nm the three positive charged residues pull the oligopeptidyl-tRNA backbone toward the exit tunnel and the peptide bond formation rate is increased as compared to the neutral oligopeptidyl-tRNA. With three negative charged residues, the resulting axial force points in the opposite direction and the peptide bond formation rate is decreased as compared to the neutral case.

Table 6.11.1: Maxwell-Boltzmann factors $\exp(\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}/k_B T)$ modulating the x-mer (x=10, 22, 40) oligopeptidyl transfer rate constants $k(\vec{\mathbf{F}})$, the waiting time $\tau_{1/2}$ to peptide bond formation event with a probability of 0.5, and theoretical uncertainty on $\tau_{1/2}$. C-terminal residue at P site is phenylalanine. Acceptor substrate at A-site is puromycin. $k(\mathbf{0})$: rate constant in the absence of force (neutral case). fMet: formyl-methionine. Circled dot: neutral; circled plus: positively charged; circled minus: negatively charged oligopeptides respectively, with specific amino acid distributions as detailed in the text.

P-site x-mer oligopeptide with charged residues acting at P-site		Mechanical work (pN.nm)	Maxwell Boltzmann factor (-)	Rate constant $k(\vec{\mathbf{F}})$ (s^{-1})	$\tau_{1/2}$ (ms)	Uncertainty $\Delta\tau_{1/2}^{\text{th.}}$ (ms)
10-mer	$fMet \odot -tRNA\,(0)$	0.0	1	$k(\mathbf{0})$	43.8	
	$fMet \oplus -tRNA\,(+3)$	+0.65	1.16	$1.16 \cdot k(\mathbf{0})$	37.6	4.3
	$fMet \ominus -tRNA\,(-3)$	−0.65	0.86	$0.86 \cdot k(\mathbf{0})$	51	5.9
22-mer	$fMet \odot -tRNA\,(0)$	0.0	1	$k(\mathbf{0})$	43.8	
	$fMet \oplus -tRNA\,(+3)$	−0.86	0.81	$0.81 \cdot k(\mathbf{0})$	53.5	5.3
	$fMet \ominus -tRNA\,(-3)$	+0.86	1.22	$1.22 \cdot k(\mathbf{0})$	35.8	3.6
40-mer	$fMet \odot -tRNA\,(0)$	0.0	1	$k(\mathbf{0})$	43.8	
	$fMet \oplus -tRNA\,(+5)$	−0.525	0.88	$0.88 \cdot k(\mathbf{0})$	49.5	7
	$fMet \ominus -tRNA\,(-5)$	+0.525	1.13	$1.13 \cdot k(\mathbf{0})$	38.7	5.4

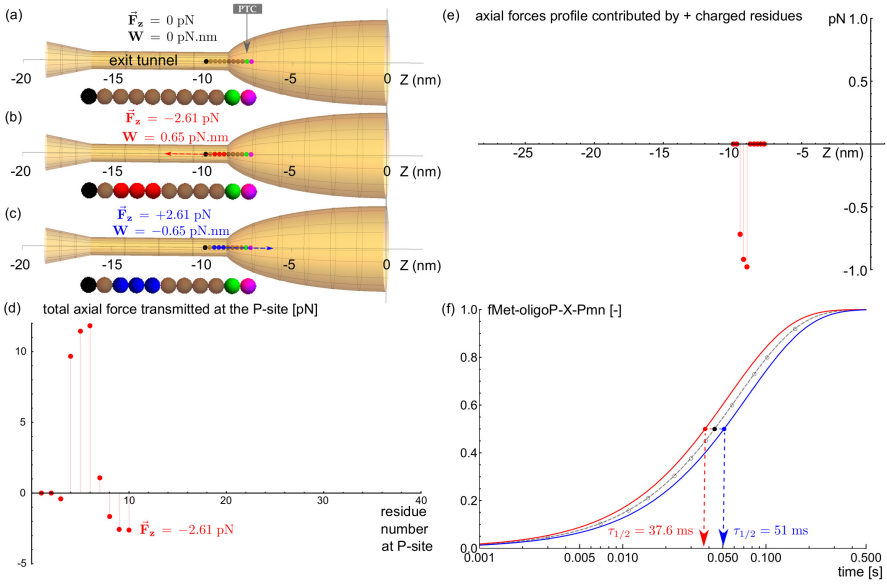


Figure 6.15: Elongation extended case: effect of force on the rate of peptide bond formation when the P-substrate is a 10-mer oligopeptidyl-tRNA (fMet-oligoP-X-tRNA). Electrostatic force acting on X at the P site: **(a)** OligoP=Neutral oligo-alanine (brown sphere). **(b)** OligoP=Positively charged amino acid R at position 6, 7, 8 (red sphere). **(c)** OligoP=Negatively charged amino acid D at position 6, 7, 8 (blue sphere). Methionine (black sphere). X=phenylalanine (green sphere). Puromycin (magenta sphere). **(d)** Total axial force acting on the nascent peptide at the last amino acid residue (counted from N-terminal end) occupying the ribosomal PTC position, as calculated from the algorithm in Joiret *et. al.* Joiret et al. 2022a. **(e)** Axial force profile contributed by positively charged residues when located at indicated $z-$position from the PTC. **(f)** Predicted normalized time courses of the Pmn (20mM) reaction with different oligopeptidyl-tRNAs: calculated theoretical normalized time courses using Maxwell-Boltzmann factors. Neutral oligoP-Phe(◯)($\tau_{1/2}$ = 43.8 ms), dashed line. Positively charged amino acid C-terminal oligopeptidyl-transfer rate ($\tau_{1/2}$ = 37.6 ms), red line. Negatively charged amino acid oligopeptidyl-transfer rate ($\tau_{1/2}$ = 51 ms), blue line.

6.11.3 Estimated theoretical uncertainty for the median time course of the peptide bond formation reaction

The theoretical uncertainties in the output variable are propagated from the electrostatic potential model to the peptide bond formation reaction rate and the theoretical calculation of the median time course is estimated by differential calculus as detailed in the appendix D. Estimates of the uncertainties $\Delta\,\tau_{1/2}^{th.}$ in the median time courses of the peptide bond formation are listed in the last column of table 6.11.1. The upper script th. in $\Delta\,\tau_{1/2}^{th.}$ means it refers to a model-dependent theoretical uncertainty. The predicted time courses of the 22-mer oligopeptidyl-tRNA reaction with Pmn are shown in Figure 6.16 (f). At their specific positions from $z = -11.95$ nm to $z = -12.45$ nm the three positive charged residues push the oligopeptidyl-tRNA backbone toward the P-site and the peptide bond formation rate is decreased as compared to the neutral oligopeptidyl-tRNA. With three negative charged residues, the resulting axial force points in the opposite direction and the peptide bond formation rate is increased as compared to the neutral case. This is qualitatively the opposite situation as the one encountered in the previous case with the 10-mer oligopeptide.

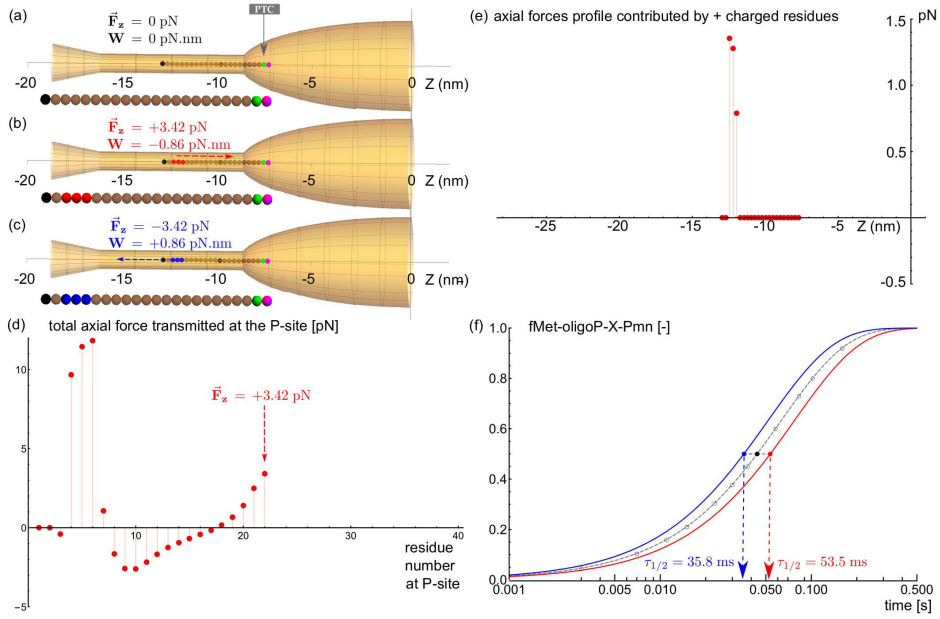


Figure 6.16: Elongation extended case: effect of force on the rate of peptide bond formation when the P-substrate is a 22-mer oligopeptidyl-tRNA (fMet-oligoP-X-tRNA). Electrostatic force acting on X at the P site: **(a)** OligoP=Neutral oligo-alanine (brown sphere). **(b)** OligoP=Positively charged amino acid R at position 18, 19, 20 (red sphere). **(c)** OligoP=Negatively charged amino acid D at position 18, 19, 20 (blue sphere). Methionine (black sphere). X=phenylalanine (green sphere). Puromycin (magenta sphere). **(d)** Total axial force acting on the nascent peptide at the last amino acid residue (counted from N-terminal end) occupying the ribosomal PTC position, as calculated from the algorithm in Joiret *et. al.* Joiret et al. 2022a. **(e)** Axial force profile contributed by positively charged residues when located at indicated $z-$position from the PTC. **(f)** Predicted normalized time courses of the Pmn (20mM) reaction with different oligopeptidyl-tRNAs: calculated theoretical normalized time courses using Maxwell-Boltzmann factors from equations (6.34) and (6.39). Neutral oligoP-Phe(◯)($\tau_{1/2}$ = 43.8 ms), dashed line. Positively charged amino acid C-terminal oligopeptidyl-transfer rate ($\tau_{1/2}$ = 53.5 ms), red line. Negatively charged amino acid oligopeptidyl-transfer rate ($\tau_{1/2}$ = 35.8 ms), blue line.

The predicted time courses of the 40-mer oligopeptidyl-tRNA reaction with Pmn are shown in Figure 6.17 (f). At their specific positions from $z = -16.2$ nm to $z = -17.20$ nm the five positive charged residues push the oligopeptidyl-tRNA backbone toward the P-site and the peptide bond formation rate is decreased as compared to the neutral oligopeptidyl-tRNA. With the five negative charged residues, the resulting axial force points in the opposite direction toward the exit tunnel and the peptide bond formation rate is increased. This is also qualitatively the opposite situation as the one encountered in the first case with the 10-mer oligopeptide. The interesting new predicted hypothetical result is that it should be possible to observe a reversing in the relative time courses of the peptide bond reaction between the two donor and acceptor amino acids at the P- and A-site with different charged amino acid distributions upstream in the probed oligopeptides. This requires the use of oligopeptidyl-tRNAs of different lengths at the P-site with amino acid sequences for which the nature (+ or -) and the space distribution of charges have been carefully positioned by design.

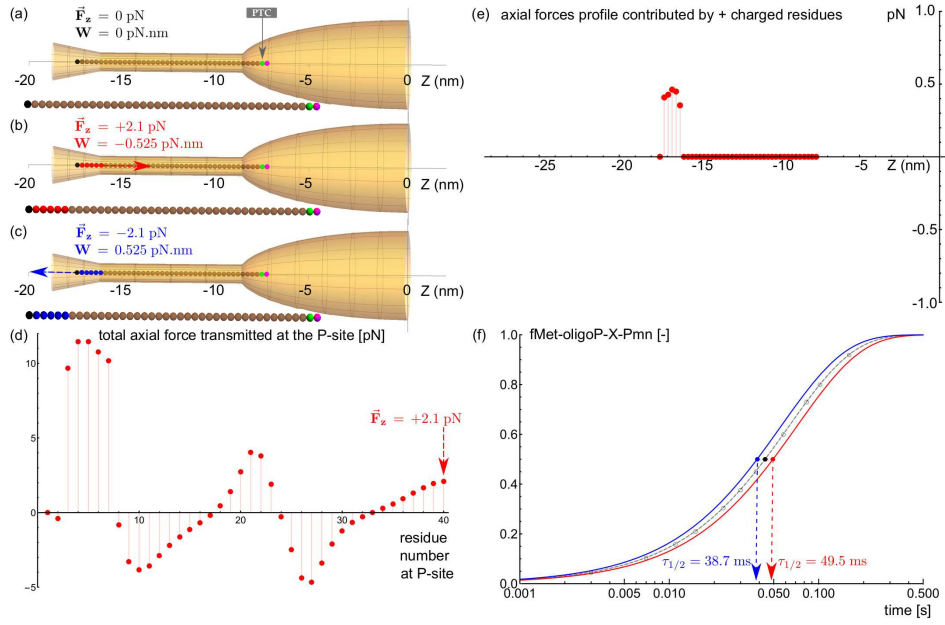


Figure 6.17: Elongation extended case: effect of force on the rate of peptide bond formation when the P-substrate is a 40-mer oligopeptidyl-tRNA (fMet-oligoP-X-tRNA). Electrostatic force acting on X at the P site: **(a)** OligoP=Neutral oligo-alanine (brown sphere). **(b)** OligoP=Positively charged amino acid R at position 35, 36, 37, 38, 39 (red sphere). **(c)** OligoP=Negatively charged amino acid D at position 35, 36, 37, 38, 39 (blue sphere). Methionine (black sphere). X=phenylalanine (green sphere). Puromycin (magenta sphere). **(d)** Total axial force acting on the nascent peptide at the last amino acid residue (counted from N-terminal end) occupying the ribosomal PTC position, as calculated from the algorithm in Joiret *et. al.* Joiret et al. 2022a. **(e)** Axial force profile contributed by positively charged residues when located at indicated $z-$position from the PTC. **(f)** Predicted normalized time courses of the Pmn (20mM) reaction with different oligopeptidyl-tRNAs: calculated theoretical normalized time courses using Maxwell-Boltzmann factors from equations (6.34) and (6.39). Neutral oligoP-Phe(◯)($\tau_{1/2}$ = 43.8 ms), dashed line. Positively charged amino acid C-terminal oligopeptidyl-transfer rate ($\tau_{1/2}$ = 49.5 ms), red line. Negatively charged amino acid oligopeptidyl-transfer rate ($\tau_{1/2}$ = 38.7 ms), blue line.

6.12 CSBJ02 discussion

In our theoretical study, we incorporated as an additional argument of the Maxwell-Boltzmann factor in the Eyring equation, a term accounting for the effect of mechanical forces on a catalyzed reaction. The work of external electro-mechanical forces can affect the Gibbs free energy barrier of the transition state. A modified Michaelis-Menten model is applied to the kinetics of the ribozyme catalyzing the peptidyl transfer reaction in the large subunit of the ribosome. Our theoretical derivation is quantitatively consistent with the large difference in the time courses of the peptide bond formation that was experimentally observed in the dipeptide minimal case between neutral, positively and negatively charged amino residues positioned at the carboxyl-terminal end of the dipeptidyl-tRNA at the P-site [Wohlgemuth et al. 2008]. This confirms that the electrostatic interaction is an important contributing force affecting the transition state barrier for the catalytic reaction taking place at the PTC.

We further predicted that the time courses of the peptide bond formation between a given C-terminal amino acid residue of the peptidyl-tRNA at the P-site with a given aminoacyl-tRNA at the A-site is dependent on the distribution of upstream charged amino acid residues within the peptidyl-tRNA embedded in the exit tunnel. Specifically, depending on whether these charged amino acid residues are positive or negative and are located in the upper tunnel, before or after the tunnel constriction site (where the ribosomal proteins uL4 and uL22 protrude), or in the lower tunnel near the tunnel exit point, the peptide bond formation time courses are predicted to be reversed even if the C-terminal donor residue and acceptor substrates are kept the same in the tested oligopeptides.

We proposed an explanation for how the mechanical forces acting on the nascent protein chain backbone quantitatively affect the rate of the peptidyl transfer reaction. A pulling force exerted on the nascent chain backbone, directed from the PTC toward the exit tunnel, facilitates the peptide bond formation. The Eyring equation (6.34), determining the reaction rate constant where the Gibbs free energy transition barrier is explicitly modulated by the mechanical work of these forces, provides a tool to quantitatively predict the effects of the electrostatic forces on the time course of the peptide bond formation. In our approach, the mechanical work results from the dot product of the axial force with the curvilinear displacement. It is worth noticing that the mechanical work can, equivalently, be expressed as the dot product of a local electric dipole moment with the local electric field at the PTC, as was shown by other authors [Fried and Boxer 2017]. The physical value of the dipole moment is dependent on the type and charge of the amino acid being incorporated at the PTC. Whatever its origin, be it the dot product of a force with a displacement, or the dot product of a dipole moment with an electric field, the mechanical work modulates the Gibbs free energy barrier for the transition state. When the nascent chain threads through the tunnel and elongates toward the ribosome tunnel exit, the electrostatic potential profile along the full length of the

tunnel impacts on the kinetics at the PTC. In this latter case, the mechanical work is best represented by the dot product of an axial force acting on the nascent oligopeptide embedded in the tunnel with the elongation displacement at the PTC.

6.12.1 Connections to previous studies

Free energy barriers for a single positively charged amino acid like lysine (or negatively charged like aspartate) have been calculated in the work of Petrone *et al.* [Petrone et al. 2008]. The authors found a significant free energy barrier along the main axis at the exit of the ribosome tunnel. The free energy barrier was measured in units of $k_B T$ for lysine as $\sim 7\, k_B T$. For aspartate the barrier is smaller $\sim 2\, k_B T$. Petrone *et al.* authors also found that the barrier is much lower at the constriction site. Overall, the free energy barrier profile for a single positively charged amino acid lysine they presented is consistent with the electrostatic profile we estimated from similar structural data of *Haloarcula marismortui* in our previous study [Joiret et al. 2022b].

The mechanical work modulating the activation energy of the peptide bond formation that we estimated in our current manuscript are the following:

- in case of a positively (or negatively) charged amino acid at the carboxy-terminal end of the peptidyl tRNA, the mechanical work is $\pm 5.3\text{pN} \cdot \text{nm}$, i.e., $\pm 1.2 \times k_B T$ (37° Celsius).
- in case of the 22-mer oligopeptide specific sequence described in our manuscript, the mechanical work is $\pm 0.86\text{pN} \cdot \text{nm}$, i.e., $\pm 0.2 \times k_B T$ (37° Celsius).

These values for the mechanical work are smaller than the local free energy barrier values calculated by Petrone *et al.* [Petrone et al. 2008]. They are related to two different phenomena with different consequences. The free energy barriers calculated in Petrone *et al.* [ibid.] refer to a local intermolecular interaction of side chain amino acid residues (called individual chemical probes) with the wall or a local wall binding crevice along the exit tunnel, whereas the values of the mechanical work in our manuscript result from the global electrostatic interaction of charged probes in the nascent chain with the whole ribosome tunnel wall harboring negatively charged phosphates moieties. The former requires a Molecular Dynamics (MD) calculation of the so-called Potential of Mean Force (PMF) which includes non-bonded Van der Waals, London attraction, Pauli repulsion (Lennard-Jones potentials), Coulomb potential (without medium screening), non bonded hydrogen interaction and bonding harmonic interactions. The latter only includes electrostatic interaction calculated with the Yukawa-Debye-Hückel model (with attenuation due to medium electrostatic screening of mobile ions and solvent). The consequence of the electrostatic profile along the tunnel axis results in forces applied axially along the tunnel whereas the consequence of the free energy landscape

as computed by MD simulations results in local attraction or repulsion forces orthogonal to the tunnel wall inner surface. Besides, the values of the permittivity parameter used in MD software tools to calculate the PMF are often the values of standard solvent and were not explicitly indicated in the work of Petrone *et al.* [ibid.]. Comparing the two should be made with caution.

Our results are not in contradiction with the results of Charneski *et al.* [Charneski and Hurst 2013]. Two of our probing oligopeptides, the 22-mer and the 40-mer oligopeptides carry 3 and 5 positively charged amino acids at positions 18, 19, 20, or at positions 35, 36, 37, 38, 39 away from the P-site, respectively. For these two oligopeptide probes, we have predicted a slower rate in the peptide bond formation and hence in the elongation cycle (and hence an expected increased ribosome density footprint). It is only for our 10-mer oligoprobe carrying 3 charged amino acids at position 6, 7 and 8 away from the P-site that we predict a higher rate or a speeding up of the peptide bond formation. Even with this 10-mer, not being able to localize where the $x = 0$ reference is with respect to the C-terminal amino acid at the peptidyl-tRNA (P-site) in the Charneski *et al.* paper , we cannot say that we have contradicting results. Unfortunately, Charneski *et al.* [ibid.] did not extend their analysis to negatively charged amino acids. Charneski *et al.* [ibid.] did not address the global shape or the complete profile of the electrostatic potential along the ribosome exit tunnel axis.

6.12.2 Assumptions, limitations and generalization

The electrostatic forces inside the ribosome exit tunnel are not the only contributing forces. Indeed, when the peptide chain has reached a certain length and its N-terminus has exited the tunnel, it is the structure of the chain itself (captured through the hydropathy) that determines its movement through the tunnel [Dao Duc and Song 2018]. It should be added that the entropy driving forces upon protein folding outside the tunnel or tugging forces exerted from chaperone proteins could also contribute. Furthermore, the secondary structures which build already in the inner tunnel, harbor electric dipole moments interacting with the ambient electric field in the tunnel [Liutkute et al. 2020; Mercier and Rodnina 2018]. The ribosome exit tunnel is also known to be a protein-sensitive channel with gate-latch action and where sequence or specific side chain recognition regulates elongation [Petrone et al. 2008]. These effects have not been considered in our theoretical study. We suggest however that any force, whatever its origin, acting on the backbone of the nascent chain will affect the rate constant of the peptide bond formation. If the mechanical work of these forces can be quantitatively estimated, the impact on kinetics can also be quantitatively estimated through the Maxwell-Boltzmann factor that we have incorporated in the Eyring equation and in the Michaelis-Menten equation.

The oligopeptide probes of different lengths (10, 22 and 40-mer) for which we calculated

the theoretical axial forces in this study are fully embedded in the tunnel and cannot make a complete path through the full length of the tunnel. Moreover, these oligopeptides are supposed to be unstructured (harboring no secondary structures). With these features, they are not concerned with the above limitations.

6.12.3 Theoretical uncertainty calculated on model output based on input and experimental uncertainties

The uncertainty in the calculated rates and median time courses resulting from the model is larger than the experimental uncertainty resulting from the reaction rate experimental measurements using a quench flow apparatus as described in reference [Wohlgemuth et al. 2008]. In the three designed pairs of oligopeptide probes with opposite charges (10-mer, 22-mer and 40-mer), the difference between the positively and the negatively charged probes in the median time courses are 13.4 ms, 17.70 ms, 10.80 ms respectively. For two of the three oligopeptide probes, the effect sizes are slightly larger than the sum of the associated theoretical uncertainties that were estimated in the last column of table 6.11.1: $\Delta\tau^{\text{th.}}_{1/2} = 10.20\,\text{ms}\ (= 4.3 + 5.9), 8.9\,\text{ms}\ (= 5.3 + 3.6), 12.4\,\text{ms}\ (= 7 + 5.4)$. The first two effect sizes in the difference between the median time courses for the 12-mer pair and the 22-mer pair of oligopetides, are larger than the theoretical uncertainties resulting from the electrostatic model and the mechano-chemically modulated model of the peptide bond formation rate. Our main prediction is that we should observe a difference in time course measurements which is much larger than the experimental uncertainty $\sim 3\,\text{ms}$. Figures 2, 3 and 4 (f) show that the expected differences are at least larger than 10.8 ms (one order of magnitude larger than the experimental uncertainty). Moreover, an important result of our contribution is that a reversal in the relative time course should be expected when comparing the 10-mer with the 22-mer oligopeptide probes with the given specific sequences of amino acid residues. This relative comparison has an expected effect size which is larger than the theoretical total propagated uncertainty in the model by the uncertainties in the input variables. The crucial salient feature of the model depends on the bell shape (inverse bell profile and presence of an electrostatic bump near the constriction site) of the electrostatic potential. Indeed, as the axial force (electrostatic field) is the opposite of the first derivative of the potential, the sign of the axial force changes along the two sides of the bell profile, so does the sign of the mechanical work.

In our previously published references [Joiret et al. 2023a; Joiret et al. 2022b], we acknowledged the fact that the electrostatic potential exact profile can be species-dependent. This is one of the reasons why we compared ribosomal structural x-ray crystallographic data across 5 different species. Other studies showed that it is indeed the case that the tunnel geometries can be different across the three domains of life [Dao Duc et al. 2019]. However, there is a consensus regarding similar patterns across

the three domains of life, in the shape of the profile along the central axis of the ribosome exit tunnel. Whatever the detailed profiles in the electrostatic potentials, there appears to be a common pattern in the potential profiles. It is expected, however, that the accuracy of the quantitative predictions is sensitive to the exact potential profile and is sensitive to its uncertainty.

6.12.4 Synthetic oligopeptides to be used as electrostatic probes and future perspectives

Our theoretical results applied to oligopeptides of variable lengths, which are long enough to probe the electrostatic environment inside the ribosome exit tunnel, should help the experimental design of real synthetic oligopeptides (translated from their synthetic transcripts by ribosomes *in vitro*) to experimentally test the validity of our predictions. These synthetic oligopeptides of different lengths and with charged residues specifically positioned at appropriate distances of the carboxyl-terminal end of the t-RNA at the P site, when puromycin is introduced as final acceptor substrate at the A-site, could be used as electrostatic probes of the electrostatic potential profile along the ribosome exit tunnel centerline.

In future studies, the Eyring equation and the modified Michaelis-Menten equation that we used can serve as quantitative tools to improve agent based models of protein synthesis by ribosomes such as the inhomogeneous totally asymmetric simple exclusion process. Specifically, the queueing time of the ribosome for the peptide bond formation (step 2) can be made quantitatively dependent on the charged amino acid residues distributed in the upstream nascent polypeptide, with a mobile window range of 30-70 residues, embedded in the ribosome exit tunnel.

The use of mechano-chemical kinetic models will facilitate the interpretation of optical tweezers experiments assessing the forces exerted on protein nascent chain in the ribosome exit tunnel or forces exerted by the ribosome on the mRNA molecule Desai et al. 2019; Kaiser and Tinoco 2014; Liu et al. 2014b; Wen et al. 2008, as well as the interpretation of single mRNA molecule translation dynamics experiments [Morisaki et al. 2016].

We defer to future studies the application of the mechano-chemical kinetics model to improve the interpretation of Ribo-Seq data, ribosome occupancy maps of given transcripts and of the ribosome residence time on a given codon [Dana and Tuller 2014] as a function of the mRNA sequence upstream or downstream.

Chapter 7

Slow peptide bond formation by proline residues

This chapter is a bibliographical mini-review explaining why proline is slowly incorporated into the polypeptide nascent chain during elongation and why multiproline motifs stall the ribosome. The chapter entails five sections. In the first, we explain why proline is both a slow donor and a slow acceptor substrate at the P and A sites of the peptidyl transferase center. In the second section, we show x-ray structural data of the PTC with C-terminal prolyl-tRNAs at the P or A sites for bacteria. In the third section, we introduce how the specialized **EF-P** elongation factor for bacteria and its archea/eukaryotic ortholog **eIF5A** alleviate the stalling. The active forms of these elongation factor require very specific post-translational modifications that are unique in enzyme biochemistry. In the last section, we derived from the literature the peptide bond formation rates, in vivo, for XP or PX motifs at the P and A-site for the four proline codons, for yeasts and for *E.coli*, when the proline elongation factors are absent, depleted or silenced. These rates provide the corresponding numerical fold change parameters for the ribosome dwell times involving proline-tRNAs at P and/or A-site in the **Ribosomer** model.

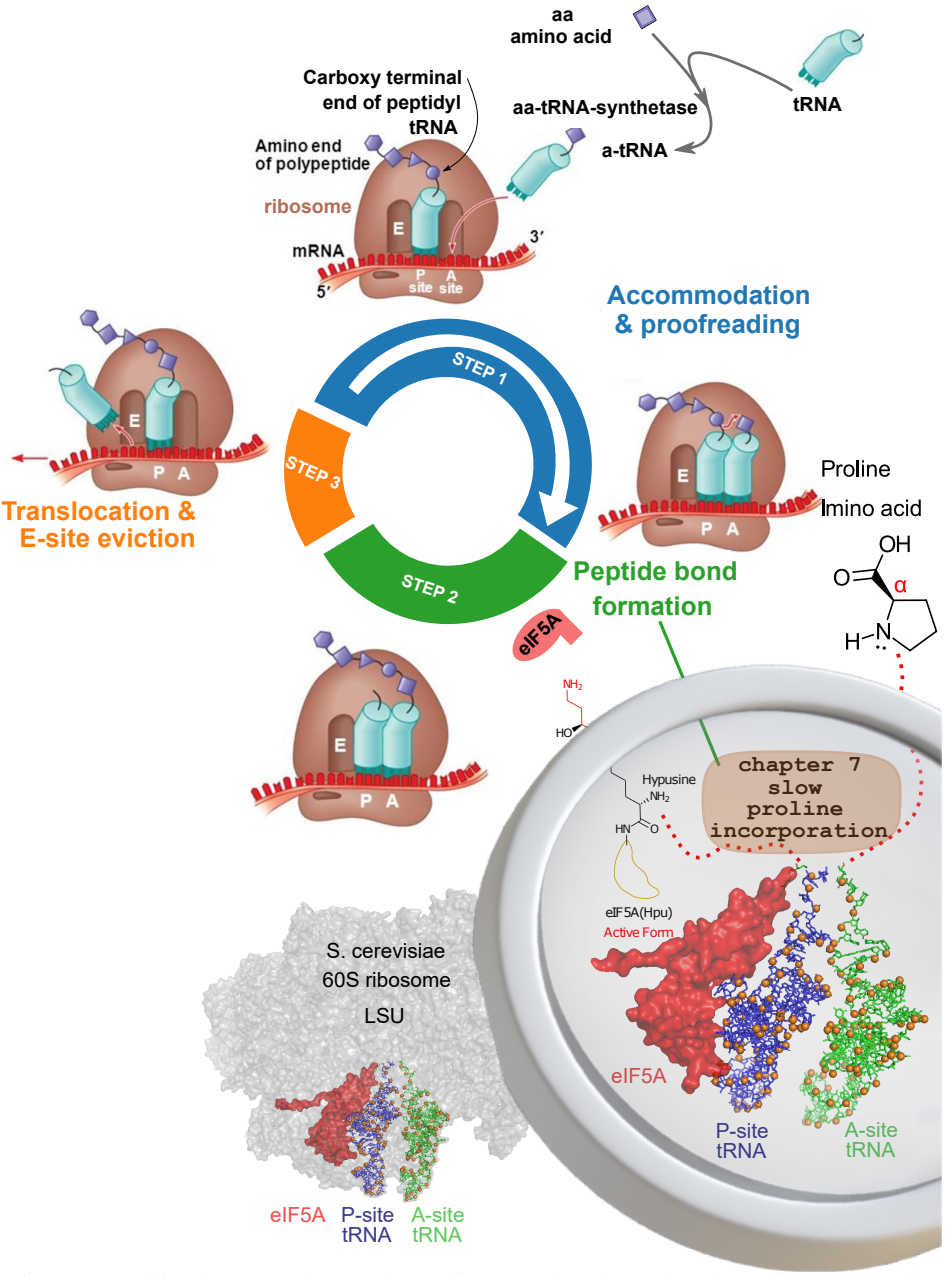


Figure 7.1: This chapter explores why proline can slow down the elongation cycle and potentially cause ribosome stalling unless properly rescued by supplemental factors. Proline acts as a steric hindrance that hampers peptide bond formation. In the agent-based model, the probability distribution of queueing times for STEP 2 is altered at each of the four proline codons, as is the total queueing time for the entire elongation cycle. We calibrated our agent-based model using these changes in hypo-exponential distributions, based on data from the literature. In eucaryotes, depletion of hypusine, inhibition of the polyamine pathway, or silencing of eIF5A affects the incorporation rate of proline.

7.1 Proline is an imino acid

The rate of amino acid polymerization varies for each amino acid, being significantly slower for proline. Translation of stretches of consecutive proline residues lead to ribosome stalling. Ribosome profiling data indicate ribosome accumulation at PPP, PPG as well as PPD and PPE triplets. Proline is a particularly poor substrate for peptide bond formation, both as a donor in the P-site and as an acceptor in the A-site [Peil et al. 2013]. Proline, however, facilitates folding of many proteins by introducing rigid turns into the peptide chain and by setting borders of β-sheets and α-helices [Melnikov et al. 2016]. Also, peptides with consecutive proline residues fold into a characteristic polyproline helix (P_{II}-helix), which constitutes a common protein-protein interaction motif and also endows proteins with unique mechanical properties, e.g. in collagen fibers. Why are proline residues slowly translated amino acids? Proline is not an amino

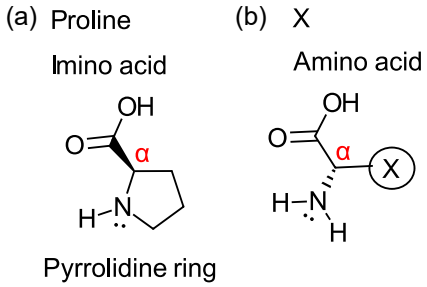


Figure 7.2: Molecular structure of proline (a) as compared to regular amino acid (b). drawing was produced with ChemDoodle 2D.

acid but an imino acid, Figure 7.2. Proline has a rigid cyclic structure where the end of the side chain on the α-carbon (Cα) is covalently attached to the $\alpha-$amine, forming a secondary amine in the pyrrolidine ring: compare (a) and (b) in Figure 7.2. The imino rather than amino group determines proline as a poor A-site acceptor of the peptidyl moiety during peptide bond formation, as well as a poor donor when present in the P-site [Starosta et al. 2014]. It is expected from organic chemistry that the Lewis electron pair on the nitrogen of the imine group (pyrrolidine ring) is less nucleophilic than the homologous electron pair on the nitrogen of the amino group in a regular amino acid, Figure 7.3 [Pavlov et al. 2009]. Consequently, the nucleophilic attack from the imino group to the carboxylic carbon at the C-terminal end of the peptidyl-tRNA at the P-site requires a higher activation energy than for the nucleophilic attack from a regular amino group. The peptide bond formation rate is thus inherently slower when proline acts as the A-site acceptor.

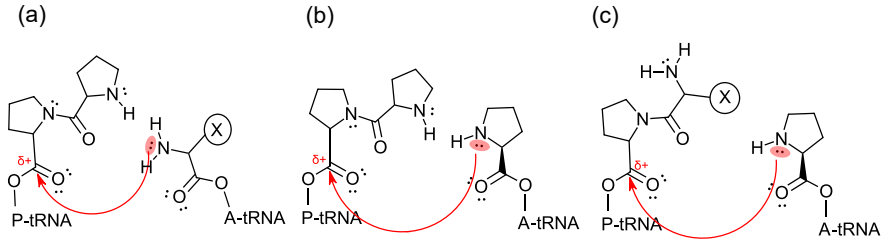


Figure 7.3: Reaction mechanisms of peptide bond formation of regular amino acid as acceptor A-site tRNA with PP donor C-terminal diprolyl-tRNA at P-site (a) as compared to proline imino acid as acceptor A-site tRNA (b), or with XP donor C-terminal X-prolyl-tRNA at P-site (c), drawn with ChemDoodle2D.

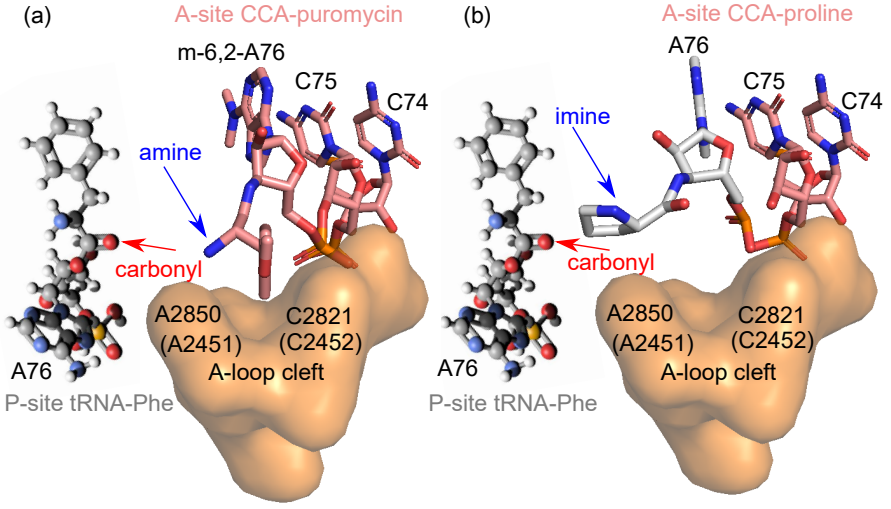


Figure 7.4: Comparison of aminoacyl residues of two A-site substrates: CCA-puromycin (a) and CCA-proline (b). In both panels the A-site is viewed as through the exit tunnel. 23/25S rRNA nucleotides numbering: *S.cerevisiae* and *E.coli* numbering in parentheses. The side chain of a regular amino acid residue is bound to the conserved hydrophobic A-cleft, whereas the side chain of the proline residue creates steric hindrance and interferes with alignment of the reacting groups and proper positioning of the attacking electron pair of the α-imine, adapted from [Melnikov et al. 2016]

Using ribosome substrates CCA-3'end tRNA analogs and solving for the crystal structures after soaking of yeast 80S ribosome, it is possible to infer the position of proline residue in the ribosomal A-site during protein synthesis [Melnikov et al. 2016]. Figure 7.4 shows that proline has an atypical side chain position in the ribosomal A-site. The proline acylated to the 3'end of A76 tRNA-CCA in the A site is badly positioned as compared to a regular amino-acylated to the 3' end of A76 tRNA-CCA in the A-site cleft of the ribosome at the PTC. In the former case, the carbonyl group in the donor peptidyl-tRNA at the P-site is at a larger distance from the nitrogen atom in the acceptor substrate at the A-site. The proline side chain does not occupy the A-side cleft, but instead flips toward the ribosomal P-site. In this conformation, the side chain may prevent proper alignment of the A-site and P-site substrates in the active site of the ribosome.

7.2 Structural data show that diprolyl-peptide is badly positioned at the P-site for peptide bond formation

Expectations from organic chemistry suggest that amino acid/imino acid sterics and basicities affect elongation rates at the peptidyl transfer step [Pavlov et al. 2009]. The reason proline also is a poor P-site donor is determined by the bad position of proline as a substrate at the P-site of the peptidyl transferase catalytic pocket. Structural data with dipeptidyl-tRNA analogs at the P-site allowed to compare diprolyl Pro-Pro-ACC-tRNA ligands with their regular dipeptidyl X-X-ACC-tRNA counterparts [Melnikov et al. 2016]. In the ribosome structure with diprolyl-ACC-tRNA at the P-site, the observed diprolyl structure is bent. The N-terminus of the diprolyl peptide is oriented toward the nascent protein tunnel wall instead of being directed directly into the tunnel. The N-terminus diprolyl moiety lacks conformational flexibility. This bent conformation reflects unique stereochemical constraints for the pyrrolidine ring of a proline residue.

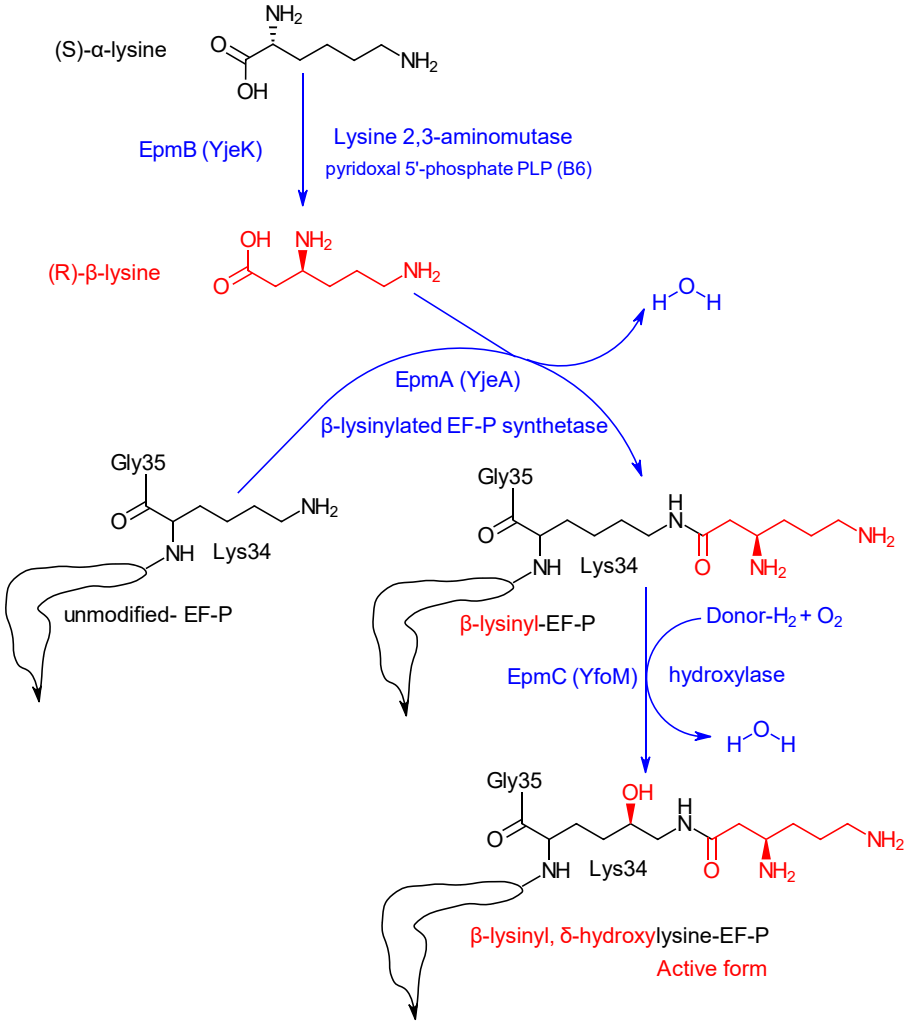


Figure 7.5: Post-translational enzymatic modifications of EF-P elongation factor, adapted from [Lassak et al. 2015b].

7.3 Specific elongation factors are necessary to enhance proline peptide bond formation

The bacterial elongation factor P (EF-P) and its archeal/eukaryotic orthologous elongation factor 5A (a/eIF5A) are universally conserved proteins that specifically accelerate the peptide bond formation when poorly reactive peptidyl-tRNAs, containing C-terminal proline residues, are at the P-site at the catalytic shell cavity of the peptidyl transferase center (PTC) in the large subunit of the ribosome [Doerfel et al. 2012]. In the absence of these elongation factors, the poorly reactive substrates cause ribosome stalling. These elongation factors require very specific and unique post-translational enzymatic modifications to be active. They bind to the stalled ribosome between the peptidyl-tRNA binding (P-site) and the tRNA-exiting site (E-site), and stimulate peptidyl-transferase activity, allowing elongation to resume. In their active form, both **EF-P** and **e/aIF5A** are post-translationally modified at a positively charged residue (lysine for *E.coli* or arginine for *T. thermophilus*), which protrudes toward the peptidyl-transferase center when bound to the ribosome [Lassak et al. 2015b].

7.3.1 Post-translational β-lysinylation of bacterial EF-P elongation factor

Figure 7.5 shows that the post-translational activation of **EF-P** requires at least two enzymes in bacteria: epmb, epma and epmc [ibid.]. In a first step, the lysine 2,3-aminomutase **EpmB** (also known as **YjeK**) converts (S)-α-lysine into (R)-β-lysine [Yanagisawa et al. 2010]. **EpmB** has a [4Fe-4S] cluster and requires pyridoxal-5'-phosphate (PLP) as a cofactor (B6 vitamin). Second, (R)-β-lysine is ligated to the ϵ-amino group of Lys34 by **EpmA** (also known as **YjeA**, **PoxA** and **GenX**) [Pfab et al. 2021]. Phylogenetically, **EpmA** derives from the amino acyl tRNA synthetase (aaRS) **LysRS** that has lost its anticodon binding domain. Accordingly, **EpmA** cannot recognize tRNA but has instead secondarily evolved substrate specificity for **EF-P**. Hence, **EF-P** is a substrate mimicking tRNA. Canonically, tRNA synthetases charge tRNA. However, the lysyl-tRNA synthetase paralog **EpmA** catalyzes the attachment of (R)-β-lysine to the ϵ-amino group of lysine 34 of the translation elongation factor P (**EF-P**) in *Escherichia coli*. Finally, **EpmC** (formerly known as **YfcM**) recognizes **EF-P** only in its modified form and hydroxylates the C5(δ) position of lysine 34 but not the β-lysyl moiety. An electron carrier co-enzyme is required. The hydroxylation stabilize the β-lysinyl moiety and allow proper hydrogen bond interaction with the P-site tRNA. However, hydroxylation is not necessarily required for the β-lysinylated-EF-P to be functional [Lassak et al. 2015b].

EF-P and its β−lysinylated[1] form mimics tRNA in size and shape and binds between the ribosomal P- and E-sites. A loop region in domain I interacts with the CCA-end of the P-site tRNA, presumably restricting mobility of the aminoacyl arm and optimally orienting it for peptide bond formation [Lassak et al. 2015b]. The +144 Da mass spectrometry shift for the modified **EF-P** at Lys34 (K34) of *E. coli* is due to β−lysinylation and subsequent hydroxylation of lysine 34 residue in the unmodified **EF-P**. The hypusinylation of lysine (K50 in human and K51 in yeast) in **eIF5A** causes a mass spectrometry shift of +87 Da.

7.3.2 Post-translational hypusinylation of eukaryotic eIF5A elongation factor

7.3.2.1 Polyamines pathway

The polyamines, putrescine, spermidine, and spermine, exist as polycations under physiological environments and exert numerous effects on nucleic acids and proteins in living cells. Figure 7.6 shows a simplified view of the polyamines pathway. The central molecule in the polyamine pathway is putrescine. Putrescine is metabolized from arginine in two enzymatic steps. In the first step, arginine is hydrolyzed to ornithine and urea by arginase. In the second step, ornithine is decarboxylated by orinithine decarboxylase with the help of pyridoxal-5'-phosphate (**PLP**) as a cofactor (B6 vitamin) to yield putrescine (butane 1,4-diamine) and carbon dioxyde. Putrescine is further converted to spermidine and spermine. These polyamines can interconvert by biosynthetic and catabolytic pathways. Spermidine, N-(3-aminopropyl)butane 1,4-diamine provides the 4-amino-butyl moiety (putrescine) to be ligated on the ϵ-amine group of a specific lysine in the **eIF5A** precursor.

In archea and eukaryotes, a small portion of the polyamine spermidine is metabolized to form an unusual amino acid, hypusine, N^{ϵ}-4-amino-2-hydroxybutyl-lysine, in a single family of cellular proteins: the eukaryotic initiation factor 5A. **eIF5A** was initially thought to function as a translation initiation factor, hence its acronym; but was later shown to be active in elongation and specifically promoting peptide bond formation between consecutive proline residues or facilitating the reactivity of poor substrates like proline in the peptidyl transferase center.

[1] In some bacteria species, like *Thermus thermophilus*, the post-translational modification of **EF-P** is made on arginine (R32 in *T. Themophilus*) instead of lysine (K34 *in E.coli*) and the modification is not a β-lysinylation but a rhamnosylation instead [Lassak et al. 2015a].

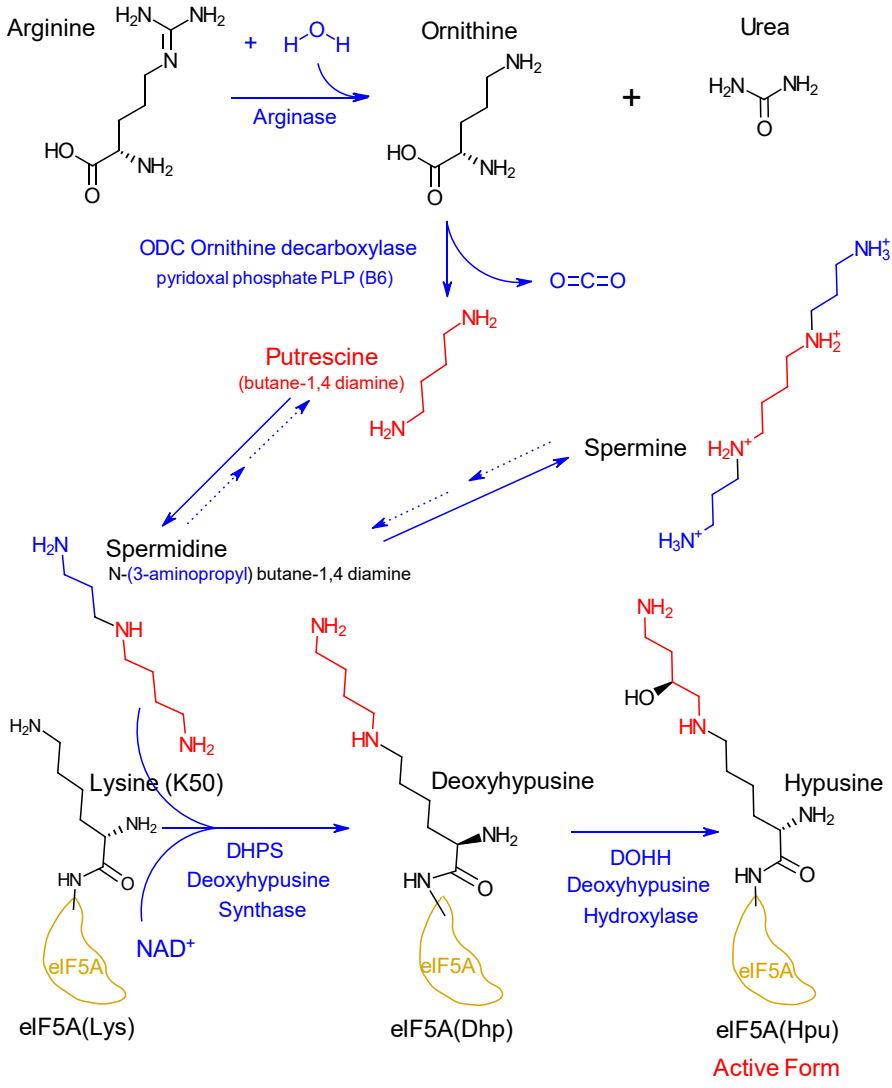


Figure 7.6: Pathways of polyamines, hypusine synthesis, and hypusinylation of eIF5A elongation factor, adapted from [Park et al. 2022; Park and Wolff 2018].

7.3.2.2 Hypusine synthesis and hypusinylation of eIF5A

Hypusine is not formed as a free amino acid. Hypusine synthesis occurs post-translationally in two enzymatic steps [Park and Wolff 2018]. Figure 7.6 shows that deoxyhypusine synthase (DHPS) catalyzes the cleavage of the spermidine and transfer the 4-aminobutyl moiety to the ϵ-amino group of a specific lysine residue of **eIF5A** precursor to form a deoxyhypusine, N^{ϵ}-4-aminobutyl-lysine residue: **eIF5A(Dhp)**. This first step requires NAD^{+} as a cofactor for the oxydation leading to deoxyhypusine, Dhp. In a second step, this intermediate is hydroxylated by deoxyhypusine hydroxylase (DOHH) to form hypusine, Hpu, and yielding the active form of **eIF5A**: **eiF5A(Hpu)**. The hypusine modification occurs exclusively in this single protein, **eIF5A** [ibid.]. It is intriguing that such an elaborate mechanism involving two novel enzymes, DHPS and DOHH, evolved to modify and activate just one protein. This emphasizes the important role **eIF5A** has in stabilizing P-tRNA and facilitating peptide bond formation between P-site poor donor prolyl-tRNA and A-site poor acceptor prolyl-tRNA. It is intresting to note that the frequency of polyproline motifs, like PPP or PPG, potential targets of **eIF5A**, has increased dramatically in proteomes of higher organisms [Mandal et al. 2014]. The human proteome, for example, contains $\sim$ 10,000 motifs with three or more consecutive proline residues, with some proteins having up to 27 consecutive prolines [Morgan and Rubenstein 2013].

7.4 The modified elongation factors rescue the peptide bond formation involving proline residues

Figure 7.7 (a) shows the cryo-electron microscopy cryo-EM crystal structure of the unmodified **EF-P** of *Thermus thermophilus* (PDB id = 4V6A). **EF-P** has 184 amino acid residues. Figure 7.7(b) shows the hypusinylated elongation factor **eIF5A** of *Saccharomyces cerevisiae* (PDB id = 5GAK) at the lysine 51 position. **eIF5A** has 157 amino acid residues and is a bit smaller than **EF-P**. Figure 7.7 (c-d) show the cryo-electron microscopy of the **eIF5A** elongation factor of *Saccharomyces cerevisiae* bound to the 60S large subunit LSU of the ribosome (PDB id **5GAK**) in the presence of P-site tRNA and A-site tRNA [Schmidt et al. 2015]. These structural data revealed how **eIF5A** binds between the exit E-site and the P-site, very close to the P-site tRNA. The hypusinylated side chain of eIF5A-K51 contacts A76 of the CCA-end of the P-tRNA and is capped by 25SrRNA nucleotide A2808 (yeast) (23SrRNA A2439 E.coli). This interaction stabilizes the P-tRNA in the optimal geometry for peptide bond formation with the aminoacylated tRNA at the A-site, but the hypusinylated moiety does not reach the peptidyl transferase center and does not contribute directly to the catalysis. The 4-amino group at the tip of hypusine reaches only so far as to contact the backbone of

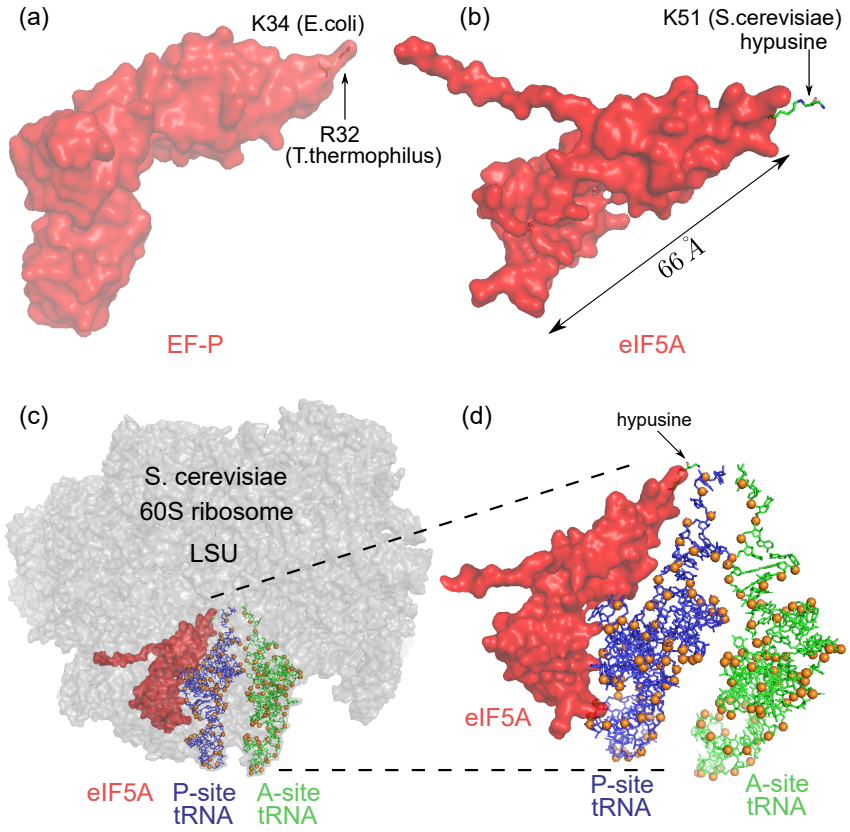


Figure 7.7: Cryo-EM structures of elongation factors EF-P (PDB id 4V6A) of *Thermus thermophilus*, unmodified EF-P (a). Hypusinylated eIF5A (PDF id 5GAK) (b). eIF5A(Hpu) bound to the 60S LSU ribosome of *Saccharomyces cerevisiae* close to the P-site tRNA (c) and showing the position of hypusine in hydrogen bond distance of the non-bridging oxygen of the A76 phosphate group at the tip of the P-site tRNA (d); adapted from [Schmidt et al. 2015].

the CCA-end of P-tRNA and comes into hydrogen bond distance of the non-bridging oxygen `OP1` in the phosphate moiety of adenosine 76 (A76).

The elongation factor `eIF5A` can only be accommodated between the E- and P-sites if the E-site is free.

7.5 Elongation rates for proline tRNAs and their calibration

As reminded in previous chapters (chapter 1 and 4, for the 19 canonical amino acid residues, the elongation rate can be split into three successive steps: accommodation and proofreading (step 1), peptide bond formation (step 2) and translocation (step 3), each step contributing its own rate or queueing time. When an imino acid proline is incorporated, the elongation factor **EF-P** in bacteria or **eIF5A** in archea or eukarya may be recruited. This would require additional queueing times, i.e., a queueing time for the E-site to be free, a queueing time for the elongation factor EF-P or eIF5A to accommodate between E- and P-site and the attached β-lysine or hypusine moiety to interact with the tip of the peptidyl-tRNA at the P-site. The total number of sub-steps in the elongation cycle would be 5 or 6 in this situation.

To keep the structure of our ABM model simple, we maintain the three canonical sub-steps as they are; but the rate k_2 is modulated conditionally on the presence or absence of the elongation factor **EF-P** (bacteria) or **eIF5A** (eukarya). A toggling switch or flag allows to activate the effect of this elongation factor depletion.

A series of in vitro experiments were conducted with a variety of amino acid substrates as carboxy-terminal ends on dipeptidyl-tRNAs at the P-site and with puromycin as final acceptor substrate or full aa-tRNA at the A-site [Wohlgemuth et al. 2008]. The ribosomal cell free extracts were from *E.coli*. In the absence of elongation factor **EF-P**, the peptide bond formation rates were measured and are tabulated in Table 7.5.1 comparing phenyl-tRNA or valyl-tRNA with prolyl-tRNA as donor-substrates.

Table 7.5.1: Experimentally measured dipeptityl transfer rate constants comparing proline with phenylalanine and valine (k_{pep}), apparent affinities of puromycin binding ($K_{\frac{1}{2}}$) and waiting time ($\tau_{1/2}$) to peptide bond formation event with a probability of 0.5; reproduced and adapted from [Wohlgemuth et al. 2008].

P-site dipeptide substrate	k_{pep} (s^{-1})	$K_{\frac{1}{2}}$ (mM)	$\tau_{\frac{1}{2}}$ (ms)	charge
$fMet-Phe-tRNA^{Phe}$	16 ± 1	4 ± 1	43.8	neutral
$fMet-Val-tRNA^{Val}$	16 ± 1	6 ± 1	44	neutral
$fMet-Pro-tRNA^{Pro}$	0.14 ± 0.02	12 ± 4	**5, 102.2**	neutral

The peptide bond formation rate when proline is the P-site donor substrate is 116 fold decreased as compared to phenylalanine or valine.

The prokaryotic ortholog of **eIF5A**, **EF-P**, decreases the activation energy of the peptide bond formation by a value of 2.5 kcal/mol, which is equivalent to 10.5 kJ/mol and 17.3 pN.nm (*in-singulo* molecule) [Doerfel et al. 2015].

At room temperature (296 K), the thermal energy per molecule is given by $k_B \cdot T = 4.1$ pN.nm, where k_B is Boltzmann's constant.

As already introduced in chapter 4, the rate of reaction follows an Arrhenius-like dependence on activation energy (Maxwell-Boltzmann factor), given by

$$e^{-\Delta G^{\ddagger}/(k_B T)} = e^{-17.3/4.1}, \tag{7.1}$$

$$e^{-17.3/4.1} = e^{-4.22},$$

$$e^{-4.22} \approx 0.014,$$

where $\Delta G^{\ddagger}$ is here the decrease in the activation energy (minus sign). This means that without **EF-P**, the reaction occurs at only 1.4% of the enhanced rate.

If the reaction rate in the absence of **EF-P** is k_0, then in the presence of EF-P, the new rate k is given by:

$$k = \frac{k_0}{e^{-4.22}} \tag{7.2}$$

Since $e^{-4.22} = 0.014$, the enhancement factor is $\frac{1}{0.014} \sim 70$.

When the elongation factor is present, this corresponds to a 70-fold increase in the peptide bond formation rate, that aligns well with experimental observations from other studies [Wohlgemuth et al. 2008].

The **EF-P** or **eIF5A** depletion toggling option included in the **Ribosomer** model decreases the peptide bond formation rate, k_2, when proline is the donor amino-acid at the P-site, by a factor calibrated to 90 (median between the two aforementioned published results). A further decrease is optionally added when two proline amino acids occupy both the P and A-sites as donor and acceptor substrates.

7.6 Summary of main findings and insights

This chapter 7 examined the kinetic bottleneck of peptide bond formation when incorporating proline residues, emphasizing the structural and biochemical factors that contribute to this slowdown. Structural data on diprolyl-peptides suggest that proline's unique cyclic structure creates steric hindrance, affecting ribosome dynamics. Specific elongation factors, such as **EF-P** and **eIF5A**, play a crucial role in mitigating this delay by stabilizing transition states and enhancing elongation efficiency. The findings

highlighted the interplay between ribosome mechanics and amino acid properties, with potential implications for translational control in diverse cellular contexts. A key discussion point is whether ribosomes have evolved compensatory mechanisms to minimize proline-induced stalling across different species. The chapter concluded by explaining how proline incorporation modulates the rate of peptide bond formation in the `Ribosomer` ABM model and how this modulation was quantitatively calibrated using literature data.

Chapter 8

mRNA secondary structures

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.
Einstein, A. (1934). On the Method of Theoretical Physics. Philosophy of Science, 1(2), 163–169. Einstein's formulation of **Ockham's razor**.

Everything should be made as simple as possible, ***but not simpler***.
Popular paraphrase summarizing Einstein's view.

During translocation, ribosomes often encounter mechanical barriers due to structures such as hairpins and pseudoknots adopted locally by the mRNA. The mRNA entry channel of the ribosome can only accommodate single-stranded RNA. While ribosomes can intrinsically unwind mRNA secondary structures during translocation, the rate of translation is reduced in front of such barriers. To accurately capture this phenomenon, `Ribosomer` must quantitatively incorporate the effects of mRNA secondary structure on ribosomal movement. This chapter outlines the computational methods assigned to ribosome agents as they interact with transcript agents. Specifically, a secondary structure folding strength is dynamically updated for each transcript in a downstream sequence relative to the current footprint of a ribosome. This strength score is then used to modulate the ribosomal translocation rate accordingly. To achieve this, we implemented a gear-shift algorithm in our ABM, integrating recent biophysical insights obtained from published results of single-molecule experiments combining co-temporal force measurements (optical tweezers) and fluorescence-based techniques (confocal

fluorescence imaging). The computational prediction of secondary structures from an mRNA primary sequence is a well-established problem in Bioinformatics that has been solved by the community. Our solution uses the foundational contributions of the so-called `Vienna RNA package` school.

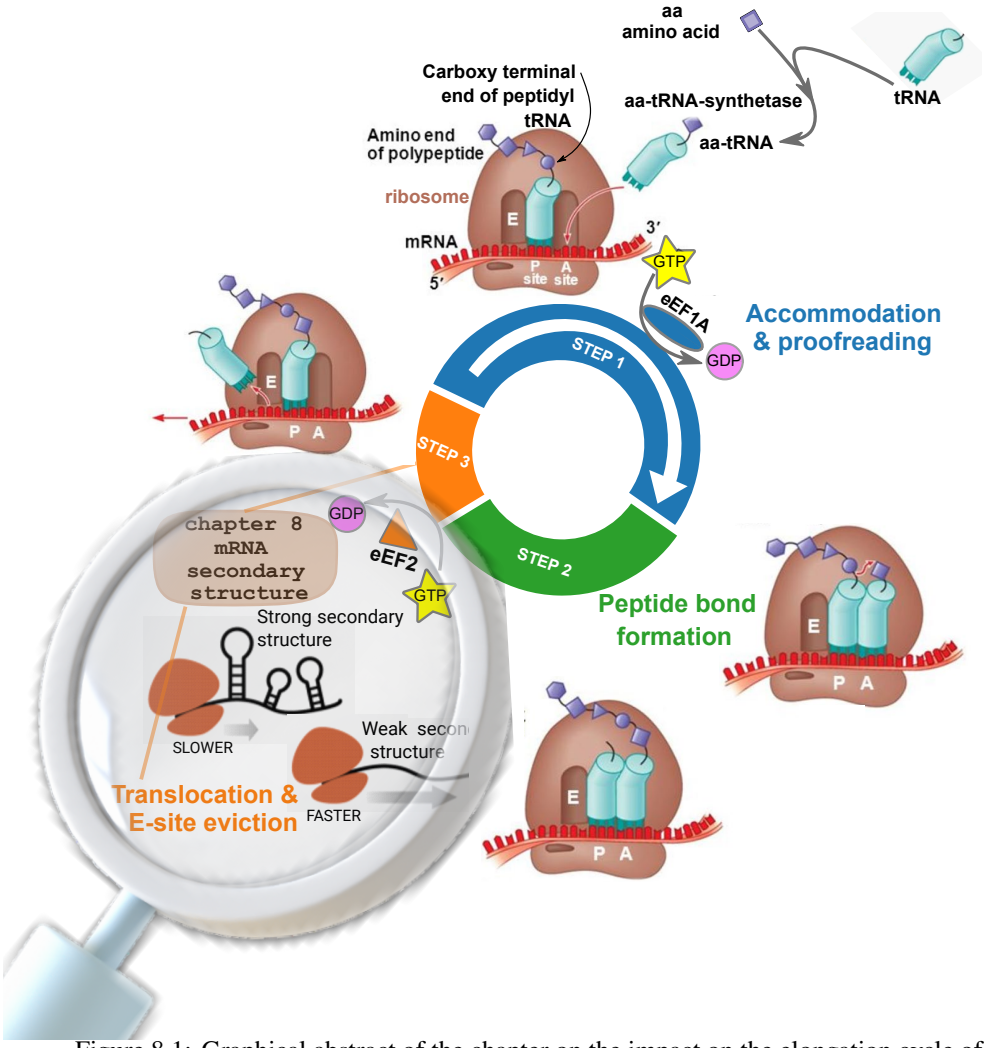


Figure 8.1: Graphical abstract of the chapter on the impact on the elongation cycle of secondary structures downstream of the mRNA. These roadblocks modulate the energy barrier of ribosome translocation, which, in turn, affects the rate of the elongation cycle.

8.1 An introduction to mRNA secondary structures and their effects on the elongation cycle of the ribosome

mRNA secondary structures are locally stable two-dimensional structures, non-covalently interacting elements, within an mRNA molecule (or any RNA molecule), that is otherwise expected to be single-stranded, consisting of contiguous base pairs and loops. Such RNA secondary structures contain both canonical Watson-Crick base pairs and many non-canonical base pairs. Numerous studies have shown that mRNA tends to fold into local secondary structures. These structures are unevenly distributed within mRNA and can act as roadblocks[1] that might influence the protein synthesis rate [Mao et al. 2014]. During elongation, ribosomes moves along mRNA and pause at paired sites until the base pairings are broken. It is known that mRNA secondary structure decreases elongation rate as naively shown on Figure 8.2. In yeasts, local

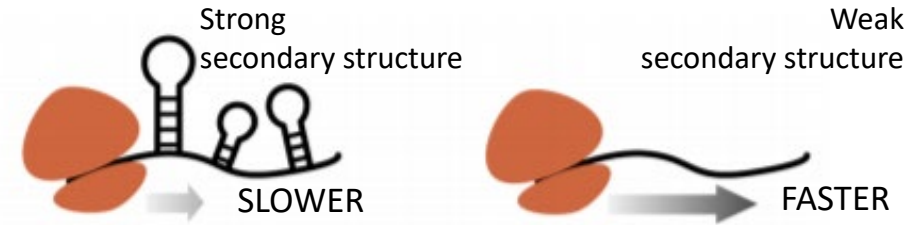


Figure 8.2: Secondary structures. Reproduced from [Gorochowski et al. 2015a].

secondary structures are distributed densely within mRNA. The distance between adjacent structures in vitro is $6 - 7$ nts on average and is significantly shorter than the length of the footprint that a ribosome protects its mRNA [Ingolia et al. 2009]. This suggests that most of the translating ribosomes are located in structural regions and thereby undergo mRNA structural dynamics (folding and unfolding). Two aspects of this mRNA structural dynamics probably play a functional role:

i. mRNA secondary structure blocks ribosomal translocation and further migration on the mRNA being translated. The induced pause of the ribosome might have a co-translational regulation role such as the co-translational folding of proteins or also ribosomal frameshifts.

ii. Translating ribosomes constrain mRNA folding.

[1] We are not talking about target specific post-transcriptional gene silencing with short hairpin RNAs, or about regulatory miRNAs, or RISC (RNA induced silencing complex).

As a ribosome moves along a structural region, mRNA secondary structures disappear, and they may appear after the ribosomes passes through the region. The re-folded structures hold until the next ribosome arrives at this region and re-unwinds them again. This could motivate a distinction between a pre-folded structure of the mRNA before the first ribosome would have engaged in elongation and a folded structure of mRNA between two ribosomes engaged in the translation of the same transcript. The same nucleotide sequence could in principle have different secondary structures according to the difference in the course of events. This approach was followed in Mao et al. 2014. In the approach we will follow in our model, we will only consider re-folded mRNA secondary structures in the sequence between two adjacent ribosomes. The justification of this approach comes from the fact that our dynamical simulations rapidly reach a stationary state, where the average distance between two consecutive ribosomes on the same transcript, depends on the balance between the initiation rate and the elongation rate.

8.2 Predicting secondary structures in single-stranded RNA from their primary sequence

RNAs fold back onto themselves by forming intra-molecular base pairs. The resulting structures are composed of two fundamental building blocks: **paired regions** (mostly A-form helices), and **unpaired loops**. Secondary structure forms an important intermediate level of description of nucleic acids that encapsulates the dominating part of the folding energy, is often well conserved in evolution, and is routinely used as a basis to explain experimental findings. Based on carefully measured thermodynamic parameters, exact **dynamic programming (DP) algorithms** can be used to compute ground states, **base pairing probabilities**, as well as **thermodynamic properties**[Nussinov and Jacobson 1980]. The `Vienna RNA` package has been a widely used compilation of RNA secondary structure related computer programs for nearly three decades. A brief introduction to the existing algorithmic approaches to predict RNA secondary structures is provided here [Lorenz et al. 2016].

8.2.1 Bioinformatics of RNA secondary structures

The approach to treat RNA structures algorithmically is to reduce them to the set of base pairs (which builds the secondary structure), thereby abstracting from the actual spatial arrangement of nucleotides [ibid.].

To determine a valid secondary structure, we require that each nucleotide i interacts with at most one other nucleotide j to form a base pair (i, j). Only the canonical

Watson-Crick base pairs are considered AU, UA, CG, GC as well as the wobble pairs GU and UG. ***Pseudo-knots***, defined as crossing pairs, i.e., (i, j) and (k, l) pairs for which $i < j < k < l$ are excluded. Each base pairs (i, j) in a secondary structure closes a *loop* L, thereby directly enclosing unpaired nucleotides u and possibly further base pairs (p, q). The adverb 'directly' means that there is no other base pair (k, l) with $i < k < l < j$ such that $k < u < l$ or $k < p < q < l$. With these conditions, the number of directly enclosed unpaired nucleotides constitute the *length*, or *size* of L, while the number of directly enclosed base pairs and the enclosing pair determines its degree. Loops of degree 1 are referred to as ***hairpins***. Loops of degree 2 are called ***interior loops*** and loops with degree > 2 are called ***multibranch loops***.

In the following brief overview of the bioinformatics methods, the pseudo-knots, the G-quadruplexes, and ternary structures are neglected to reduce the computational complexity.

8.2.2 Nearest neighbour energy model

Computational predictions of RNA secondary structures are mainly driven by physics based models. The major assumption behind these models is that a good estimate of the overall stability is the Gibbs free energy $E(s)$ of a RNA secondary structure s, calculated as the additive contributions E_L of its individual loops L.

$$E(s) \approx \sum_{L \in s} E_L \tag{8.1}$$

In this approach, the energy contribution of a base pair in a helix depends on the identity of the two adjacent pairs, giving rise to the name *Nearest Neighbour Energy Model*. A large number of studies has been published, determining the free energy parameters from melting experiments for different types of loops with a large variety of sequence compositions [Turner and Mathews 2009]. Modern prediction programs rely on these tabulated databases of free energy melting parameters.

In recent years, several ML/AI inspired methods emerged, which augment or even replace physics-based models through trained parameters. Instead of relying on experimental measurements, these methods require large sets of RNAs with known structure as training data.

8.2.3 Free energy minimization

The simplest type of RNA secondary structure prediction aims at producing a single 'optimal' structure. The **optimality criterion** is the **minimum free energy (MFE)**. Indeed, according to thermodynamics, the MFE structure is not only the most stable,

but also the most probable one in thermodynamic equilibrium. The number of possible secondary structures a particular RNA sequence can adopt grow exponentially with its nucleotide length. It is generally impractical to enumerate all of them in order to assign an optimality score and select the best candidate. However, the problem can be solved efficiently by a bioinformatic technique called dynamic programming (DP), which recursively builds the optimal solution from solutions of smaller sub-problems [Nussinov and Jacobson 1980]. This is easy when pseudo-knots are neglected as each base pair divides the structure into two independent parts, inside and outside of the base pair. The first DP algorithm to compute MFE structure of an RNA was published four decades ago by Zuker and Stiegler [Zuker and Stiegler 1981]. In this algorithm, computation time grows cubic with sequence length. The prediction accuracy (analog of the positive predictive value–PPV), is rather good and around 70% for sequence length smaller than 500 nts.

8.2.4 Thermodynamic ensemble of structures

The probability of a secondary structure s follows the **Boltzmann distribution** from Statistical Physics and the laws of thermodynamics.

$$p(s) = e^{-E(s)/RT} \tag{8.2}$$

where $E(s)$ is the free energy of the structure, $R = Nk_B$ the gas constant and T the temperature of the system is Kelvin scale. The right-hand side of (8.2) of any particular structure is easy to compute, and so is the sum over all possible structures, called the **partition function** Z:

$$Z = \sum_s e^{-E(s)/RT} \tag{8.3}$$

which can be used as a normalization factor for obtaining the equilibrium probability of a secondary structure s:

$$p(s) = \frac{e^{-E(s)/RT}}{Z} \tag{8.4}$$

Equation 8.3 is impractical since it requires summing over all possible structures. But in the nineties, McCaskill realized that a variant of the DP recurvive algorithm for the MFE prediction could solve the problem [McCaskill 1990]. This gave birth to the `RNAfold` program [Ding and Lawrence 2001; Hofacker et al. 1994].

8.2.5 Reliability and prediction performance

How trustworthy or uncertain is a RNA secondary structure prediction? Several reliability measures that are based on base-pairs probabilities and the partition function

can inform the user on the trustworthy of the predictions. There are several reliability measures. We restrict here to two of them:

(i) **Positional entropy:** Positional entropy is a local measure of reliability of the secondary structure prediction for each nucleotide. The positional entropy $S(i)$ of nucleotide i captures whether this particular nucleotide is found mainly in the same configuration, paired or unpaired.

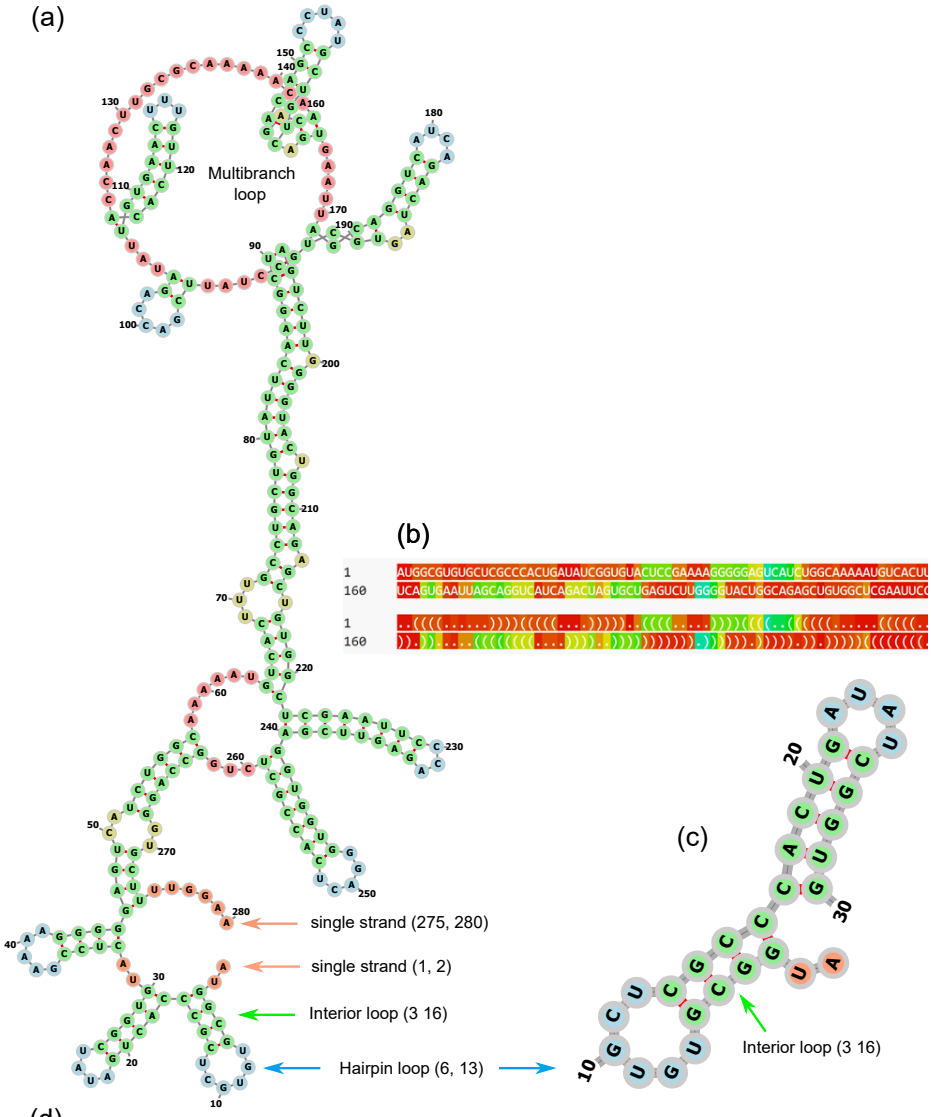
$$S(i) = -\sum_k p_{ik} \log_2 p_{ik} - q_i \log_2 q_i \tag{8.5}$$

where $q_i = 1 - \sum_k p_{ik}$ is the probability that nucleotide i is unpaired. The positional entropy is zero for a nucleotide that is always unpaired or always paired with the same partner. Thus, positions with low entropy are predicted with high confidence.

(ii) **Ensemble centroids:** In probabilistic terms, the minimum free energy (MFE) simply represents the most likely structure in the ensemble. However, other optimality criteria exist and could yield structures more representative of the ensemble. One idea of such a representative is the *centroid* structure s_c. The centroid structure of an ensemble Ω is the structure that minimize the distance to all other structures:

$$s_c = \underbrace{\arg\min}_{s} \langle d(s) \rangle = \sum_{t \in \Omega} p(t) d(s,t) \tag{8.6}$$

s_c becomes trivial when the distance between structures is measured in terms of *base pair distance* $d(s,t)$, i.e., the distance between structure s and structure t which counts the number of pairs present in one, but not both structures. In this case, s_c is equivalent to all base pairs with $p_{ij} > 0.5$.



(d)

>NM_000968.4:66-1349 Homo sapiens ribosomal protein L4 (RPL4), mRNA
ATGGCGTGTGCTCGCCCACTGATATCGGTGTACTCCGAAAAGGGGGAGTCATCTGGCAAAAATGTCACTT
TGCCTGCTGTATTCAAGGCTCCTATTCGACCAGATATTGTGAACTTTGTTCACACCAACTTGCGCAAAAA
CAACAGACAGCCCTATGCTGTCAGTGAATTAGCAGGTCATCAGACTAGTGCTGAGTCTTGGGGTACTGGC
AGAGCTGTGGCTCGAATTCCCAGAGTTCGAGGTGGTGGGACTCACCGCTCTGGCCAGGGTGCTTTTGGAA

Figure 8.3: (a) Minimum free energy mRNA secondary structure of the first 280 nts, as predicted from the Vienna RNA package web server, with (b) minimum free energy predictions and (c) for the first 30 nts of (d) RPL4 first 280 nucleotides [Gruber et al. 2008].

8.2.6 The Vienna RNA package or the Vienna Websuite

What we first need is a model of mRNA secondary structure that predicts the secondary structure from the primary sequence, that can be limited in the nucleotide length to the average distance between two consecutive ribosomes (e.g., ribosome interspacing distance averaged across translatome), that provides a quantitative measure of the local folding strength of the sequence. Efficient computational methods for predicting RNA secondary structures have already been available for a long time [Lorenz et al. 2016].

An application of the `Vienna RNA package` computational methods, in its web-server version (`http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi`), is shown in Figure 8.3 in the case of `RPL4`. `RPL4` is the transcript of the ribosomal protein number 4 in the LSU. Figure 8.3 presents the predicted most likely secondary structure as determined by the minimum free energy (MFE) method using the `RNAFold` webserver [Gruber et al. 2008; Lorenz et al. 2011]. Panel (a) displays the predicted secondary structure of an extended sequence comprising the first 280 nucleotides. Panel (b) provides a visualization of paired and unpaired nucleotides, with a heatmap representation of base-pairing probabilities derived from the thermodynamic ensemble and the partition function (red more likely than blue). Panel (c) depicts the predicted secondary structure when the sequence is restricted to the first 30 nucleotides. Notably, the predicted structures for these first 30 nucleotides are identical in both (a) and (c). Panel (d) displays the coding DNA sequence (CDS) in FASTA format for RPL4 which was used as input data for the `RNAFold` tool used in this analysis.

8.3 More details on the third step of elongation: translocation and the Brownian ratchet mechanism

Translocation occurs concomitantly with the opening of secondary structures (unwinding) of the downstream mRNA being translated, with the GTPase activity that precedes translocation. The detailed sequence of events involved in translocation was reviewed in reference [Liu et al. 2014b] and is summarized here.

After the peptide bond is complete, a deacylated tRNA is at the P-site and the peptidyl-tRNA is at the A-site. In order to start a new elongation cycle, the A-site must be emptied to allow accommodation of the next aminoacyl-tRNA. To achieve this, **the two tRNAs and the mRNA must move relative to the ribosome**. This movement occurs in two steps:

1. In the first step of translocation: the $3'-$ends (the ends that can be acylated with

amino acid or peptide) of the tRNAs in the A and P-sites move relative to the 50S subunit (LSU), into hybrid A/P and P/E states respectively. Formation of these hybrid states can occur spontaneously, reversibly, and **independently of elongation factor G (EF-G)** and is coupled to rotation of the 30S body (body of SSU).

2. In the second step of translocation, irreversible and **EF-G dependent**, **the mRNA is translocated by one codon**, along with movement of the associated anticodon ends of the tRNAs to the classical P and E sites, coupled to an orthogonal rotation of the 30S subunit head domain.

The translocation process also involve other large scale conformational changes in the ribosome, including reverse rotational movements of the 30S (SSU) body and head, and movement of the LSU L1 stalk into the inter-subunit space. Translocation is therefore a highly coordinated and complex process composed of inter- and intra-molecular, force- and torque- generating mechanical movements. During this process, the ribosome must also overcome the mechanical barriers posed by the structured portions of the mRNA being translated.

Translocation works like a Brownian ratchet mechanism. A Brownian ratchet is a biophysical mechanism that drives the unidirectional movement of a molecular-scale object, even though thermal energy induces random motion in both directions. This directionality arises from a coupled chemical reaction that selectively prevents backward movement. In the Brownian ratchet, the system moves back-and-forth spontaneously, driven by thermal energy, along the 'mechanical coordinate', in the Gibbs free energy landscape. However, a chemical transition, that only occurs when the system reaches the forward state, prevents reversal and converts otherwise random motion into directed movement.

The main translocation sub-steps occur as follows. The most likely scenario [Liu et al. 2014b] is that translocation is achieved by two consecutive Brownian ratchet as shown in Fig. 8.4 (a):

(1) *Back and forth relative rotation of the two ribosomal subunits (LSU and SSU).* During the first sub-step, **the ribosomal subunits (LSU and SSU) rotate back and forth** relative to each other along an axis perpendicular to the subunits interface. This process is reversible and activated by **thermal agitation**. It is accompanied by the **repositioning** of the $3'-$acceptor ends of the **tRNAs** initially in the **classical A and P states into the hybrid A/P and P/E states** (and movement of the L1 stalk). The **binding of the** EF • G • GTP **(prokarya) or** eEF2 • G • GTP **(eukarya)** stabilizes the tRNAs in their hybrid states and the counter-clockwise rotation of the SSU subunit relative to the LS subunit. The binding of EF-G (or eEF2) functions as the rectifying

reaction for the first Brownian ratchet.

(2) *Swiveling of the head and passage of the P-site tRNA anti-codon stem loop to the E-site*.
In the second sub-step, the ribosome acts as a **GTPase activator for EF-G (or eEF2) and rapid GTP hydrolysis** catalyzes **conformational changes in the ribosome**: the swiveling of the head opens the way for the passage of the P-site tRNA anti-codon stem-loop to the E-site. The SSU subunit's head rotates forward, moving the tRNAs and the mRNA relative to the SSU body by one codon. Then the SSU detaches from the tRNAs and the mRNA and moves back into the non-rotated position. This functions as a spontaneous and thermally activated transitions between the pre- and post-translocation state. In this second Brownian ratchet, the rectification reaction is the **movement of domain IV of EF-G** (or its **homologous domain in eEF2**) **into the A-site** and the intercalation of two conserved bases of 16S rRNA (or the homologous 18S rRNA) into mRNA. The **orthophosphate released** then induces a **relaxation** of EF-G, destabilizing the contacts between domain III and IV and the ribosome and results in EF-G (or eEF2) **dissociation from the ribosome**.

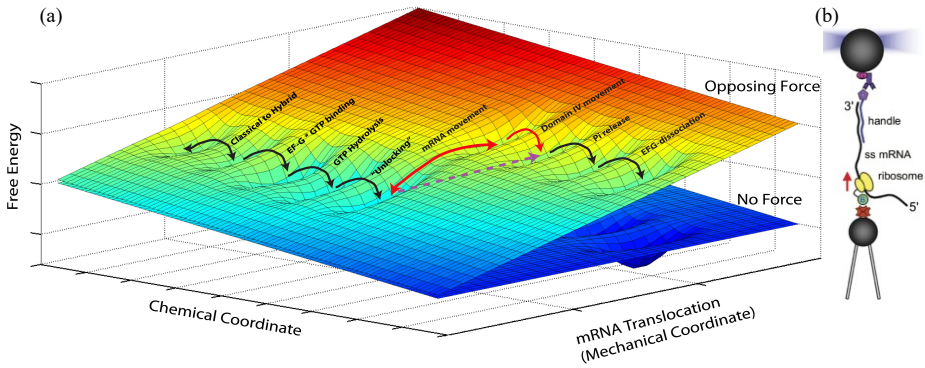


Figure 8.4: (a) Gibbs free energy landscape for translocation: Brownian Ratchet (BR) mechanisms. (b) *In singulo* experimental setting to follow translation on a single mRNA (explanations in the text). Reproduced from [Liu et al. 2014b].

The translocation is a mechano-enzymatic process. Mechano-enzymes in general act by coupling a mechanical task (translocation, force generation) to a downhill chemical reaction, i.e., a reaction that lowers the total Gibbs free energy of the system.

Fig.8.4 (b) shows the geometry of the optical tweezer experimental setting to follow a single-molecule translation by a ribosome. A biotinylated ribosome is loaded onto a

single-stranded mRNA and attached to a streptavidin-coated polystyrene bead fixed to a micropipette. This attachment of the ribosome is made from the back of the ribosome and mechanically constrains the SSU only; the LSU is free in all its movements relative to the SSU. The $3'-$end of the mRNA message is anchored to a second bead through a 1460 bp DNA/RNA hybrid handle. Calibrated forces can be applied to the ribosome by manipulating the second bead with an optical trap, while the translation progress of the ribosome is determined by the change in extension of the tether. During the progress of translation, the length of the mRNA tether decreases and is monitored at nanometer scale resolution.

By the mechanical constraints imposed by the attachment geometry of this *in singulo* experiment using an optical trap (optical tweezer) depicted in Fig. 8.4 (b), the free energy landscape can be reduced to a simplified two-dimensional surface in a three dimensional space. Hence, in Fig. 8.4 (a), one axis represents the mechanical coordinate that describes the movement of the mRNA relatice to the SSU, and the other axis is the chemical coordinate that describes all binding, conformational changes, hydrolysis and dissociation processes.

The most likely path for the reaction occurs along a minimum energy channel on this surface and the events involved in translocation can be described as transitions between minima of this reduced energy surface. The three-dimensional energy surface depicted in Figure 8.4 naturally explains how transition rates are affected when a mechanical force is applied, or when a reaction force is requested to overcome such an applied mechanical barrier. The effect is equivalent to tilting the potential energy surface by rotating the 3D plot around its chemical axis. This will affect the rate and equilibrium constants of reactions along the mechanical coordinate. An external force in the aiding direction (a 'pushing' force) will make translocation more favorable, while a 'pulling' force will make it less favorable. In the experiment shown in Fig. 8.4 (b), the less favorable translocation is when the upper bead is pulled away from the ribosome.

The translocation of mRNA and its two associated tRNA anticodon-stem-loops (ASL) from the A and P sites to the P and E sites of the SSU–the movement along the 'mechanical' axis– must then be coupled to a downhill progress along the 'chemical' axis. There are two ways in which this coupling can occur:

i. PS, the **Power Stroke** mechanism, where the energy released by the chemical transition is directly harnessed to produce the mechanical change. In this case, the system moves diagonally in the energy landscape.

ii. BR, the **Brownian Ratchet** mechanism, where the system moves back-and-forth spontaneously, driven by thermal energy agitation, along the mechanical coordinate, until a chemical transition, that occurs when the system is in the post-translocated state, prevents the back-translocation and rectifies the random

motion into the directed motion. In this case, the system moves in two orthogonal steps in the energy landscape.

While the PS mechanism cannot be fully ruled out, the ribosome likely functions as a BR during translocation as supported by a number of studies [Liu et al. 2014b and references therein].

8.4 Unwinding or melting mRNA secondary structures requires forces

The relative movement of the ribosome with respect to the mRNA occurs during the second step of translocation as was presented above. In section 1.5 of chapter 1, the energy budget of the ribosome available for elongation was detailed. We recall here two of the possible energy sources theoretically available for translocation:

1. *transpeptidation* (ester hydrolysis and peptide bond formation): $\Delta G^0 = -3.7 \pm 1.2\,\text{kcal/mol}$ in thermal units, i.e., $25.7 \pm 8.3\,\text{pN} \cdot \text{nm}$ in mechanical energy units (under assumption of 100% thermodynamical efficiency). This is equivalent to $6.3\,k_BT$.

2. *Hydrolysis of one GTP* by GTP-ase activity of GTP-F (prokarya) or eEF2 (eukarya), the translocation driving fundamental enzymes[2]: $\Delta G^0 = -7.3\,\text{kcal/mol}$ in thermal units, i.e., $50.7\,\text{pN} \cdot \text{nm}$ in mechanical energy units (under assumption of 100% thermodynamical efficiency). This is equivalent to $12.4\,k_BT$.

The maximum mechanical work generated by the ribosome during the translocation step was measured from a set of sophisticated *in singulo* biophysical experiments using optical tweezers and conducted by the group of Ignacio Tinoco, Harry Noller, Carlos Bustamante and their coworkers [ibid.]. The maximal work generated by the ribosome, that is, the product of the force and the step size, is $21.2\,\text{pN.nm} = 5.2\,k_BT$ (or 3.1 kcal/mol).

The energetic source for this mechanical work could be the transpeptidation. In this case the thermodynamical yield would be $\sim 80\%$ of the total energy available from transpeptidation. In principle, it is possible to power translocation from this energy without the need to invoke an energy contribution from the hydrolysis of GTP but such a high thermodynamic efficiency is much higher than observed in most molecular

[2]EF-G and eIF2 are homologous proteins (derived from a common ancestor) that promote translocation after peptide bond formation. Their ribosome-binding domain and GTP-dependent conformational changes are conserved across domains of life.

motors [Bustamante et al. 2004]. Instead, a mechanism in which EF-F (or eEF2) binding and GTP hydrolysis account for the energy of translocation and resetting, with a thermodynamical efficiency of $\sim 42\%$ appears to be more likely.

The translocation events correspond to single codon steps (three nucleotides) along the mRNA as it is convincingly supported by the *in singulo* experiments [Liu et al. 2014b]. The experimental measure of the length of the stepsize was estimated to be 1.42 nm along the mRNA which coincides with the length of three nucleotides or with the distance between the A- and P-site mRNA codons, 1.48 nm, as measured in RX-crystallography solved structures of ribosomes.

The ribosome appears to be able to generate forces as high as 13 ± 2 pN to react against (and unwind) secondary structures.

8.5 Ribosome gear box bifurcation and unwinding forces

The development of a new generation of optical tweezers instruments, endowed with single-molecule fluorescence capability and the co-temporal monitoring of both channels, i.e., the fluorescence signal and the tweezer signal, has opened the possibility of monitoring molecular machine trajectories along two orthogonal reaction coordinates: the mechanical coordinate and the chemical coordinate simultaneously. These instruments make it possible to record and causally relate chemical events (binding of ligands, conformational changes) with the corresponding mechanical events (force, torque generation, and displacement). These instruments are referred to as *fleezers*–fluorescence optical tweezers. Desai *et al.* used such instruments, as shown in Figure 8.5, to investigate the coupling of ribosome translocation to the binding and activity of elongation factor EF-G and to uncover how downstream secondary structures in the mRNA is dealt with by the ribosome [Desai et al. 2019]. This group studied the distribution of dwell times of ribosomes when opposed to the interior loop of an mRNA hairpin. They showed that strong hairpins caused a $\approx 2-$fold decrease in translation rate. The fitting of the observed distribution of the total dwell time required a minimum of two exponentials. This corresponds to a **mixture** of two distributions, as opposed to the **convolution** of two distributions. The latter was extensively described in chapter 4. A **convolution** operation means **two consecutive queueing times (addition of two random variables)**, while a **mixture** distribution means a **bifurcation** between two pathways (two mixed random variables describing **two independent alternative parallel processes**).

Translocation is not a simple linear scheme. It bifurcates into two alternative pathways with rates k_a and k_b. The aforementioned study showed that one of this pathway is fast,

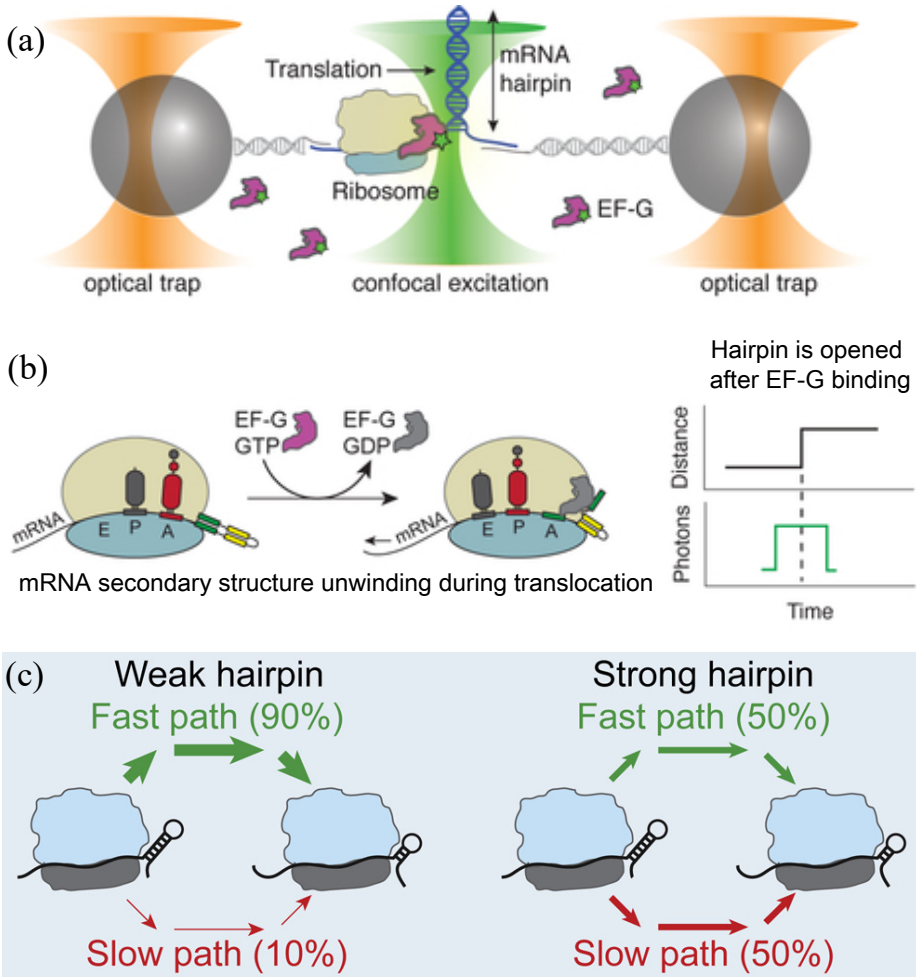


Figure 8.5: Gear box bifurcation pathways (a) Experimental setup of a *fleezer*: co-temporal monitoring of a single molecule optical tweezer and a fluorescence single molecule confocal imaging. (b) Hairpin unwinding opening a variable amount of time after EF-G binding. (c) Ribosome bifurcation in two alternative (fast and slow) gears during translation. Reproduced from [Desai et al. 2019].

the other is slow: $k_a = k_{\text{fast}}$ and $k_b = k_{\text{slow}}$. The values of these fast and slow rates remain relatively constant when the folding strength of the hairpin is altered; however, the fraction of ribosomes instances that enter the slow pathway increases when the hairpin is stronger. The biphasic translation is observed for any individual ribosome,

indicating that the same ribosome can switch between the two pathways. The presence of a strong barrier biases the ribosomes into the slower pathway with minimal alteration of the pathway rates and accounts for a $\sim 2-$fold decrease in the average translation rate observed when a ribosome passes through a strong mRNA interior loop.

The study also showed that the mRNA hairpins are opened a variable amount of time after the EF-G binding on the LSU and the openings are concomitant with translocation. Several kinetics observations lead to the conclusion that there is a force sensitive mechanism by which the ribosome senses the strength of the downstream hairpin and at which point the ribosome bifurcates into the fast or slow pathways, prior to the EF-G binding and converges after EF-G catalyzed hairpin unwinding. The sequence of events appears to be the following, Figure 8.5) (c) and Figure 8.6:

1. There is an hairpin sensor giving rise to the bifurcation of the transition pathway into a fast (green) and slow (red) branch.
2. An intermediate kinetic step occurs in both pathways. This step is insensitive to force and determines the rates through the two paths.
3. The ribosome opens the hairpin in the presence of EF-G via rates $k^{\text{fast}}_{\text{unwinding}}$ and $k^{\text{slow}}_{\text{unwinding}}$, neither of which is rate limiting in the respective pathways. The two pathways converge after the unwinding event, EF-G is released, and the ribosome is reset for another round of elongation.

Taken together, these observations indicate that the presence of a strong barrier at the RNA junction allosterically switches the ribosome prior to EF-G binding from a fast to a slow overall elongation speed. The intermediate transition step mentioned above presumably is the rate of EF-G binding [Desai et al. 2019]. The ribosome responds to stronger barriers by shifting its operation $\sim 50\%$ of the time into an alternative pathway, i.e. a slower gear. Direct measurements of the one codon-displacement of the mRNA, by ~ 1.4 nm, relative to the ribosome have shown that ribosomes are capable of generating forces that can unwind downstream secondary structures during translocation [Liu et al. 2014b]. Such forces could be exerted during either the forward or reverse rotation of the head of the SSU, which results in the coordinated movement of the mRNA codons and the tRNA anticodons in the SSU subunit. The fleezer experiments suggest that the forward rotation of the head of the SSU, which ensues after EF-G binding, is the likely candidate for the force generation step of the ribosome that results in hairpin opening [Desai et al. 2019]. The hairpin opening step of the ribosome catalyzed by EF-G is only marginally affected in the presence of stronger hairpins. Instead, the ribosome responds to stronger barriers by shifting its operation $\sim 50\%$ of the time into an alternative kinetic pathway, i.e., a slower gear.

Shifting into a slower gear to unwind secondary structures appears to be thermodynamically more favorable as the ribosome could take advantage of the thermal

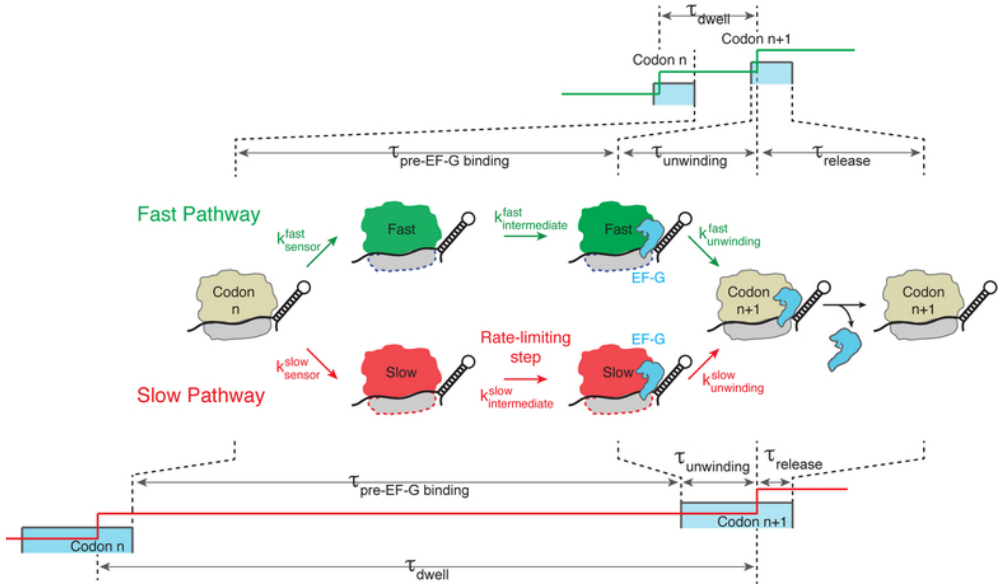


Figure 8.6: Translocation gear shift bifurcation for pathways selection to unwind mRNA secondary structure. The ribosome "senses" the hairpin barrier and irreversibly bifurcates into either fast (green) or slow (red) pathways with rates whose ratio is force sensitive. The ratio is fixed by the force and determines the fraction of translocation events that go through either pathway. Then, in either pathway, the ribosome must undergo an intermediate transition that becomes the rate limiting and determines the overall translocation rate. The unwinding occurs before EF-G is released in a similar fashion in both pathways, suggesting that the bifurcated pathways converged upon unwinding. An illustration of the fleezers trajectory for both pathways is shown above and below the kinetic scheme. Reproduced from [Desai et al. 2019].

fluctuations of the ribosome-mRNA junction, thus minimizing energy dissipation [ibid.]. Several studies have shown that the amount of energy dissipated in a non-equilibrium process can be significantly reduced if the process slows down in those areas of the potential energy surface where there is greater friction [Sivak and Crooks 2016]. This rule has been experimentally confirmed in single molecule studies that used a DNA hairpin [Tafoya et al. 2019]. High thermodynamic efficiency achieved by molecular machines could be explained if these have evolved to switch into a slower gear in regions of their potential energy landscape that are associated with high dissipation, such as interior loops in mRNA hairpins.

8.6 Implementation of the effect of secondary structures and its calibration in the agent-based model

8.6.1 Modulation of the translocation rate (k_3)

As introduced in chapter 4 and already used in chapter 6, we use once again the same recipe. The quantitative link between the rate constant of a (bio)chemical reaction and an (external) mechanical work is established through the Maxwell-Boltzmann (or Arrhenius) factor in the rightmost part of the equation:

$$k(\vec{\mathbf{F}}) = \kappa \cdot \left(\frac{k_B \cdot T}{h}\right) \cdot e^{-\left(\frac{\Delta G^{\ddagger 0}(\mathbf{0})}{N k_B T} - \frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}\right)} = k(\mathbf{0}) \cdot \underbrace{e^{\frac{\int \vec{\mathbf{F}} \cdot d\vec{\mathbf{x}}}{k_B T}}}_{\text{Maxwell-Boltzmann factor}} \tag{8.7}$$

where $k(\vec{\mathbf{F}})$ is the reaction rate constant of the rate limiting step in the presence of an applied force acting on the transition state, $k(\mathbf{0})$ is the reaction rate constant in the absence of applied force. N, k_B, h and κ are Avogadro's number, Boltzmann's constant, Planck's constant and transmission coefficient respectively [Anslyn and Dougherty 2005; Eyring 1935; Laidler and King 1983].

The Maxwell-Boltzmann factor can be larger than 1 and increase the rate, or smaller than 1 and decrease the rate, depending on the algebraic sign of the mechanical work.

Here the rate is always decreased by the mRNA secondary structures roadblocks. Indeed, these roadblock structures act as a friction force for which the displacement of the ribosome is anti-parallel to the applied force. This results in a negative mechanical work. Being negative, the argument of the exponential makes the Maxwell-Boltzmann (or Arrhenius) factor smaller than one.

The movement of ribosomes on mRNA is often interrupted by secondary structures that present mechanical barriers. The energy barrier must be overcome by ribosomes. The transition state model exposed in chapter 4 is again called for to address how to model the secondary structure unwinding. It is hypothesized that step 3 of the elongation cycle—translocation—is modulated by the reaction of the ribosome to the unwinding forces that the ribosome encounters [Liu et al. 2014b]. The gear box mechanism described in reference [Desai et al. 2019] is context dependent. It depends on the mRNA secondary structure downstream of the ribosome current footprint on the transcript being translated.

The algorithm to compute the modulation of the translocation rate at single codon resolution comes from the following basic biophysical reasoning.

It is recognized that the distance between two ribosomes on the same transcript is about 270 nucleotides (or 90 codons) when averaged across the translatome. Secondary structures re-compose within a single-stranded of mRNA in a time scale of microseconds at least three order of magnitude shorter (much faster rate) than for the elongation time scale.

The translocation rate is adjusted based on the gear-box mechanism that was described in section 8.5.

8.6.2 Pseudocode for mRNA secondary structure effect on translocation rate

The likelihood that a secondary structure builds depends on the primary nucleotide sequence. Two situations occurs. In the first, there is no significant secondary structure predicted, i.e., the nucleotides of the local sequence are most likely unpaired, and the standard translocation is unaffected. In the second, a secondary structure is predicted from the sequence and there is a significant probability that the nucleotides in the local sequence are paired within an *interior loop* as the ones indicated by the arrows in Figure 8.3(a) and (c). The probability that a nucleotide in the sequence is paired and is part of a secondary structure is determined via the MFE algorithm from the `Vienna RNA package`.

The base-pairing strength in the local folding of the mRNA structure is quantified as follows.

i. **The next codon is predicted to be unpaired** (i.e., single-stranded; k_3 remains unchanged): If the three nucleotides in the next codon downstream of the ribosome's current footprint[3] are **predicted to be unpaired**, and if the **positional entropy** of each of nucleotide is below a threshold (< 0.3, arbitrarily set), then the codon is classified as **single-stranded**, and the translocation rate remains unaffected. In this case, k_3 retains its default value (or **neutral gear**) corresponding to the codon currently in the P-site.

ii. **The next codon is predicted to be partially or fully paired within an interior loop** (k_3 is modulated): If at least one nucleotide in the next codon is paired, or its positional entropy exceeds the threshold, the **codon is classified as part of an interior loop** within a secondary structure. In this scenario, the ribosome must unwind the paired nucleotides in the next translocation step. The strength of the interior loop is estimated based on its **predicted length**.

[3]The front end of the ribosome is located in the $3'-$direction of the mRNA, at the position of the codon currently at the A site plus a positive offset of two codons (six nucleotides downstream of the A-site).

- If the interior loop length is **strictly greater than two codons** (> 2), it is classified as **strong**.
- If the interior loop length is **less than or equal to two codons** (≤ 2), it is classified as **weak**.

The strength of the secondary structure does not alter the translocation rates; instead, it modulates the **proportion of ribosomes that bifurcate into slow or fast gear** modes, as supported in reference [Desai et al. 2019] and as depicted in Fig. 8.5 (c).

Desai *et al.* quantified, with fleezer experiments, that the maximal forces that the ribosome can exert to unwind secondary structures is ≈ 13 pN [ibid.]. The mechanical displacement associated to translocation is ≈ 1.4 nm [Liu et al. 2014b]. Hence, the mechanical work is $13 \times 1.4 = 18.2$ pN.nm.

As already introduced in chapter 4, and recalled in Equation 8.7 above, the rate of reaction follows an Arrhenius-like dependence on the mechanical work (Maxwell-Boltzmann factor), given by

$$e^{-\Delta G^{\ddagger}/(k_B T)} = e^{-18.2/4.1}, \tag{8.8}$$

$$e^{-18.2/4.1} = e^{-4.44},$$

$$e^{-4.44} \approx 0.012,$$

where $\Delta G^{\ddagger}$, here, is the mechanical work to overcome the energy barrier, and $k_B \cdot T$ is the thermal energy per molecule, given by $k_B \cdot T = 4.1$ pN.nm, at room temperature (296 K). k_B is Boltzmann's constant. The resulting slow-gear modulation factor is $e^{-4.44} = 0.012$ when the ribosome develops this maximum force of 13 ± 2 pN to open the interior loop. This implies that, due to the roadblock caused by the secondary structure, the translocation reaction proceeds at only 1.2% of the neutral rate for ribosomes that shift into the slow-gear state.

The fast-gear modulation factor is when the mechanical energy is of the order of the thermal agitation $\frac{3}{2}\mathrm{k_B T}$ or $\sim e^{-3/2} = 0.22$, corresponding to a mechanical work of ≈ 6.15 pN.nm or a force of ≈ 4.4 pN. In fast-gear mode, the Maxwell-Boltzmann factor is:

$$e^{-\Delta G^{\ddagger}/(k_B T)} = e^{-6.15/4.1}, \tag{8.9}$$

$$e^{-6.15/4.1} = e^{-1.50},$$

$$e^{-1.50} \approx 0.22,$$

The resulting fast-gear modulation factor is $e^{-1.50} = 0.22$ when the ribosome develops forces approximately three times weaker than in slow-gear to open weaker secondary structures.

- **strong secondary structure (long interior loop of at least > 2 codons)**: Increased ribosome proportion in slow gear. The proportion of ribosomes following the fast and slow translocation pathways is evenly distributed, with 50% in fast gear and 50% in slow gear. When a ribosome encounters a **strong** secondary structure, its translocation mode—either fast or slow—is randomly assigned based on a **binomial distribution** (a single Bernoulli trial) with probabilities $p = 0.50$ (fast gear) and $q = 1-p = 0.50$ (slow gear). Depending on the outcome of this binary sampling, the translocation rate (k_3) for the current codon in the P-site is adjusted by a factor of 0.012 (slow gear) or by 0.22 (fast gear) for that specific ribosome instance. Consequently, the global modulation factor applied to the translocation rate (k_3) for a given codon in the P-site, averaged over a large number of ribosomes encountering the same sequence context, is given by $0.50 \cdot e^{-4.44} + 0.5 \cdot e^{-1.50} = 0.12$. **The mean queueing time for translocation of the ribosome at this position, denoted as τ_3, increases by a factor of** 8.3. This results in an approximate ~ 1.7 fold increase in the mean elongation cycle duration, which remains context-dependent on the mRNA secondary structure.
- **weak secondary structure (<= 2 successive paired codons)**: In this scenario, 90% of ribosomes are assigned to fast gear translocation, while 10% adopt the slow gear mode. When a ribosome encounters a **weak** secondary structure, its translocation mode–fast or slow– is randomly assigned based on a **binomial distribution** (a single Bernoulli trial) with probability $p = 0.90$ (fast gear) and $q = 1 - p = 0.10$ (slow gear). Based on this assignment, the translocation rate (k_3) of the current codon in the P-site is adjusted by a factor of 0.012 (slow gear) or by 0.22 (fast gear) for that specific ribosome instance. Consequently, the mean modulation factor applied to the translocation rate (k_3) for a given codon in the P-site, averaged over a large number of ribosomes encountering the same sequence context is $0.10 \cdot e^{-4.44} + 0.90 \cdot e^{-1.50} = 0.20$. **The mean queueing time for ribosomal translocation at this position, denoted as τ_3, increases by a factor of** 5. This corresponds to an approximate ~ 1.4 fold increase in the mean elongation cycle duration which remains context dependent on the local secondary structure.

To summarize, the gearbox mechanism is activated at different levels depending on the predicted interior loop likelihood and the length of base-pairing interactions:

- No secondary structure: The ribosome translocates in a neutral gear without modulation.
- Weak secondary structure: The ribosome is randomly assigned to fast or slow gear, with an odds ratio of $9 : 1$. The probability of adopting the slow gear is 10%.

- Strong secondary structure: The odds ratio shifts to 1 : 1, meaning the ribosome has now a 50% probability of translocating in slow gear.

The modulation factor adjusting the translocation rate k_3 is 1 in neutral gear, $e^{-1.5} = 0.22$ in 'fast' gear and $e^{-4.44} = 0.012$ in slow gear.

8.6.3 Calibration

The calibration of the two-level gearbox mechanism follows the approach established in the literature [Liu et al. 2014b]. The mobile window used for secondary structure prediction spans 90 codons (270 nucleotides). This fixed length is supported by the average spacing between adjacent ribosomes across the translatome, considering mean initiation and elongation rates under normal physiological conditions during exponential growth [Morisaki et al. 2016; Tomuro et al. 2024].

Alternatively, the algorithm can be readily adapted to allow a variable spacing between adjacent ribosomes on the same transcript. The sequence length used for secondary structure prediction could be dynamically updated in real time for each ribosome engaged in elongation during computational simulations.

8.7 Summary of main findings and insights

Chapter 8 explored the impact of mRNA secondary structures on ribosome translocation, highlighting how structural barriers like hairpins and pseudoknots, within the coding sequence, influence translation kinetics. It discussed the ribosome's intrinsic helicase activity and the forces required to unwind stable RNA structures, which can cause translational pausing. A gear-shift model was introduced to describe how ribosomes adjust their elongation speed based on the downstream folding strength of the mRNA. This computational framework integrates experimental insights from single-molecule force spectroscopy, linking biophysical properties to translational efficiency. A key discussion point is the extent to which mRNA structure-mediated regulation contributes to differential gene expression and adaptation across species. The chapter raises broader questions on how ribosome stalling at secondary structures influences co-translational protein folding and cellular stress responses. A crucial insight, which would be worth investigating further, is how the impact of secondary structures varies significantly depending on their location—whether in the 5′−UTR leader sequence or within the coding sequence. Structures in the 5′−UTR can strongly influence translation initiation, while those in the coding sequence primarily affect elongation dynamics and ribosome stalling. In this chapter we limited our study in the latter only. The next chapter will partially cover differential translation initiation effects.

Chapter 9

Limited or abundant ribosomal resources and non-uniform initiation rates

A key feature of our agent-based model (ABM) of mRNA translation by ribosomes is that both the ribosome pool and the general initiation rates are dependent on the species, cell type, and specific conditions. These parameters critically influence experimentally observable biological outcomes. Unlike most computational models of protein synthesis, our ABM not only allows the ribosome pool and general initiation rates to be tunable parameters but also enables the parametrization of individual initiation rates, capturing variable ribosome recruitment preferences across different transcripts. The model takes as input files the coding DNA sequences (CDS) of the transcripts, with their copy numbers, and specific fold-change values of individual initiation rates relative to the general initiation rate. By incorporating these parameters, our ABM provides a framework to explore fundamental biological questions: What are the molecular foundations and significance of the ribosome pool and non-uniform initiation rates in a cell? How are these parameters regulated, and through which pathways? This chapter discusses how known ribosome biogenesis and ribophagy pathways are indirectly integrated into our computational model of protein synthesis, offering insights into their regulatory mechanisms.

This chapter brings us back to the starting point in the overarching figure of the ribosome cycle—translation initiation, as shown in Fig. 9.1.

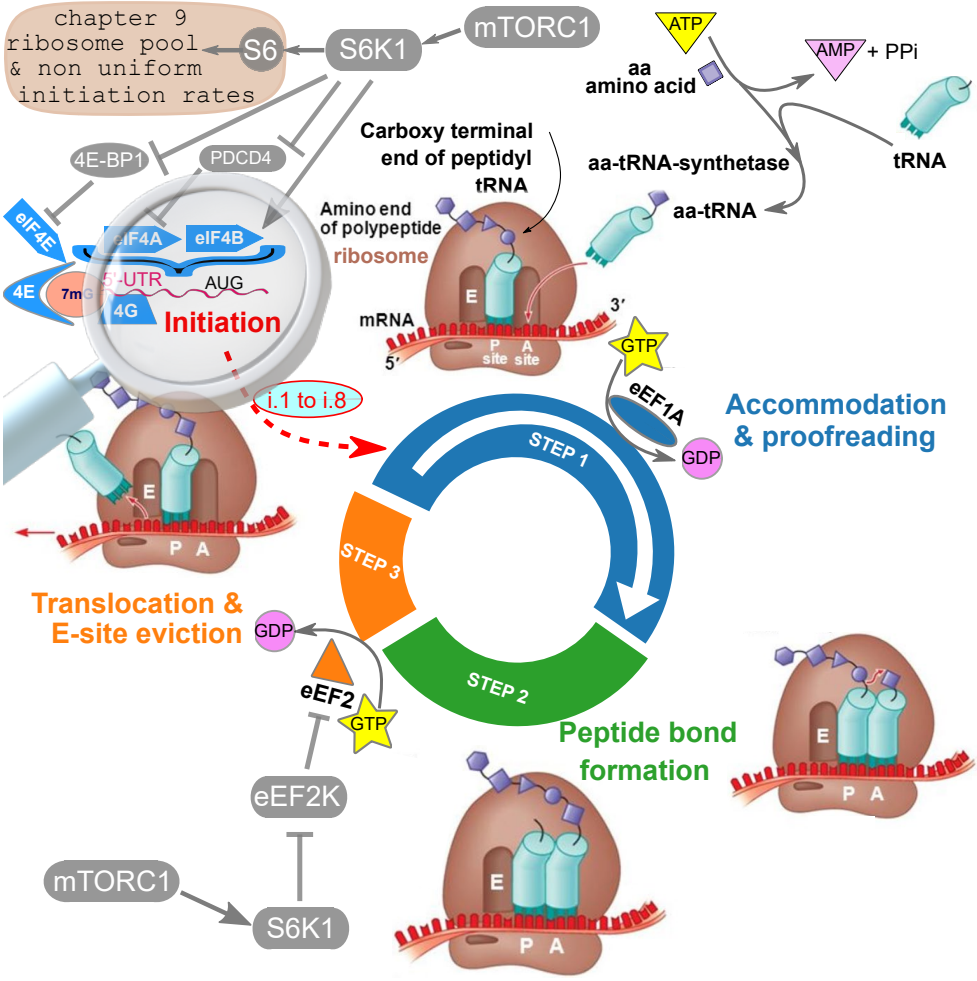


Figure 9.1: Graphical abstract of the chapter focussing on initiation. Before a ribosome can start an elongation round, initiation must occur. `i.1` Met-tRNA binding and formation of ternary complex; `i.2` eIF4A binding to eIF4G; `i.3` eIF4E binding; `i.4` binding to the m^7G-capped mRNA; `i.5` eiF4A helicase and scanning; `i.6` AUG codon recognition; `i.7` eIF5B interacts with 60S subunit; `i.8` eIF6 interacts with 60S subunit recruiting to 48S PIC. Non-uniform initiation rates of the ribosome pool strongly affect the protein landscape. eEF2/GDP dissociation after translocation and tRNA E-site eviction. Ribosome recycling is not represented. Modulators of signaling cascades: `mTORC1`, `S6K1`, `S6`, `PDCD4`, `4E-BP1`, `eEF2K`.

Here, we introduce the final mechanistic factors and parameters that play a critical role in mRNA translation by ribosomes. The control layer described here highlights a fundamental aspect of protein synthesis: the **competition among transcripts** for access to the ribosome pool. How do transcripts compete for ribosomes, and how do **limited or abundant ribosomal resources** shape the proteomic profile?

Initiation is the stage at which ribosomes are recruited to messenger RNAs. The initiation step commits the ribosome to translate the mRNA. Comprehensive and recent summaries of the knowledge about initiation factors implicated in the ribosome recruiting activity as well as reviews of the different mechanisms by which this process is conducted can be found in references [Dmitriev et al. 2020; Pelletier and Sonenberg 2019].

Chapter 1 summarized the steps involved in the canonical cap-dependent (or eIF4E-dependent) route of initiation.

There are three major factors controlling translation initiation [Pelletier and Sonenberg 2019], which are summarized below.

- **(mTOR) Phosphoinositide 3-kinase (PI3K)/mammalian (or mechanistic) target of rapamycin (mTOR) pathway**
  PI3K/mTOR integrates extra and intra cellular cues to effect translation by regulating the association of **eIF4E** with its inhibitory partners, the 4E-binding proteins (**4E-BPs**).
- **(MAPK) Mitogen Activated Protein Kinase signaling cascade** Activation of the MAPK pathway leads to phosphorylation of **eIF4E**, thereby stimulating translation of selected mRNAs.
- **(ISR)** *Integrated stress response*
  The integrated stress response engenders **eIF2**α subunit phosphorylation, **reduces ternary complex (TC) levels**, and causes a depression of general translation, while paradoxically stimulating the translation of selective mRNAs, e.g., general control non-depressible 4 (GCN4) in yeast, activating transcription factor 4 (ATF4) and ATF5 in vertebrates.

The eIF4 class of initiation factors is a key class at the crossroad of initiation mechanisms to recruit ribosome on transcripts. The principal function of eIF4F is to act as a molecular broker in recruiting ribosomes to mRNAs.
eIF4F consists of three subunits:

1. eIF4E, the cap binding protein
2. eIF4A, the DEAD box, RNA-dependent ATPase and RNA helicase

3. eIF4G, the molecular platform with multiple docking sites

RNA chaperones proteins, eIF4B and eIF4H, stimulate eIF4A and eIF4F activities by promoting the coupling of ATP hydrolysis to RNA unwinding (secondary structure melting).

The cap structure assembly of eIF4E family is multifaceted. Different cap-binding assemblies may impart disparate functional outcomes on gene expression:

- stimulating ribosome recruitment and enhancing initiation rates of selected transcripts
- selectively inhibiting of mRNA translation and miRNA-mediated suppression of translation

eIF4E protein levels are generally rate limiting for initiation (level estimated to be 1 eIF4E molecule for 10 ribosomes). Expression levels of eIF4E must be tightly controlled, as an even modest increase in levels (2.5-fold) is sufficient to drive tumorigenesis and drug resistance. eIF4E is amplified in several cancers and can trigger resistance to PI3K/TOR target therapies. Transcription of the eIF4E gene is stimulated by the c-MYC (MYC) proto-oncogene product and MYC translation is eIF4E responsive, establishing an oncogenic self-sustained feedforward loop.

9.1 Ribosome biogenesis and ribophagy: a brief introduction

At the center of the protein synthetic supply chain is the ribosome. In this section, we provide a summary of the highly complex pathway of ribosome biogenesis [Woolford and Baserga 2013]. Ribosomes share a common structure across the three domains of life (eubacteria, archea and eukarya). Chemically, they are composed of four non-coding RNAs (rRNA, ribosomal RNA) and about 80 distinct ribosomal proteins (RPs). The ability of a cell to increase the rate of protein synthesis upon physiological demand is largely mediated at the level of ribosome biogenesis. Ribosome biogenesis is accomplished by a complex integrated molecular signaling network that is far from being completely resolved. Ribosome biogenesis is a huge anabolic investment, a highly coordinated multi-step process involving more than 200 molecular components [Gentilella et al. 2015; Vanden Broeck and Klinge 2023; Vanden Broeck and Klinge 2024]. It is spatially and temporally organized. In eukaryotes, the spatial organization appears to be still more complex. There are quality control checkpoints at different stages of the assembly of the ribosomal subunits. All three RNA polymerases, I, II

and III contribute to the production of nascent ribosomes, by transcribing the different structural and catalytic molecular components required for the ribosome assembly:

- The precursor of the 47S rRNA is transcribed by RNA polymerase I in the ***nucleolus*** and then processed by a specific set of enzymes (produced in the ***cytoplasm*** and transported back to the nucleolus) and small non-coding RNAs into mature 18S, 28S and 5.8S.
- 5S rRNA is transcribed by RNA polymerase III in the ***nucleus***, is first exported to the cytoplasm, processed, reimported in the nucleus, then to the nucleolus before being assembled into a nascent ribosomal subunit. We recall here that tRNA genes are also transcribed by RNA polymerase III. tRNAs may also shuttle in and out of the nucleus to be processed from their precursor to their mature molecules.
- RNA polymerase II is in charge of transcribing all mRNAs. Specifically, RNA polymerase II transcribes the subset of mRNAs that encode for all ribosomal proteins (RPs) as well as the enzymes and small non-coding RNAs required for the processing of both the 40S and 60S ribosomal proteins.

9.2 mTOR and MAPK: central nodes in the signaling cascade of ribosome biogenesis and protein synthesis

Signaling systems controlling ribosomal protein synthesis fulfill an important task: they adjust the rate of translation and thus the state of vitality and the growth rate of cells to the food and energy supply [Marks et al. 2017]. These systems make use of protein phosphorylation, which is catalyzed by kinases serving as sensors of energy and nutriments. Disturbances of this systems rank among major causes of metabolic diseases and proliferation diseases (cancers).

The key signaling mechanism controlling protein synthesis in response to the state of nutrition is played by the serine/threonine-specific protein kinase `mTOR`, an enzyme conserved from yeast to mammals in all eukaryotes. `mTOR` stands for mammalian (or mechanistic) Target Of Rapamycin. Rapamycin is an immunosupressive antibiotic from *Streptomyces hygroscopicus* that in cells interacts with the immunophilin FKBP12.

9.2.1 Cellular effects of mTOR

The protein kinase mTOR is a key regulator of cell growth. The enzyme belongs to the phosphatidyl-inositol-3 (PI3) - kinase like protein kinases. It is a pure kinase and does not exhibit lipid kinase activity. `mTOR` connects PI3-kinase signaling with the regulation of protein synthesis [Marks et al. 2017].

mTOR stimulates mRNA translation. Major substrates of the kinase include:

- ribosomal S6 kinase type p70. S6 kinase (S6K) promotes mRNA translation and cell survival and is activated by `mTOR`.
- 4E-BP1 is a translational repressor that becomes inactivated
- eEF2 kinase becomes inactivated.

9.2.2 Ribosome biogenesis

mTOR also promotes the formation of ribosomes [ibid.]. An important condition of cell growth and proliferation is to adjust the number of ribosomes to the actual physiological requirements. In rapidly growing and proliferating cells, for instance, up to 8,000 new ribosomes are produced per minute, whereas in resting or stressed cells, ribosome biogenesis almost comes to a standstill [ibid.].

Ribosomes synthesis requires the translation of a special family of mRNA encoding all ribosomal proteins and several translation factors. The peculiarity of these mRNA species is that they carry a **5'-TOP (Tract Of Pyrimidines), an oligopyrimidine sequence** adjacent to the 5'-terminal m^7GpppX-cap, that we described in chapter 1. When cells are deprived of energy and amino acids, as well as in the absence of mitogenic and growth-promoting signals (or when treated with inhibitory drugs), polysomes (mRNA occupied by more than one actively translating ribosomes) with **5'-TOP-mRNA** collapse. As a result, formation of ribosomes is blocked and overall protein synthesis comes to a halt. The mTOR signaling pathway is critically involved in the control of 5'TOP-mRNA translation. Although the precise mechanism is still unknown [ibid.], Figure 1.9 in chapter 1 sketched a likely signaling cascade.

In cells, `mTOR` is integrated in two multiprotein complexes, called mTORC1 and mTORC2. mTORC1 is much more sensitive to rapamycin then mTORC2. The canonical functions of mTOR, that is, control of protein synthesis and cell growth are attributed to mTORC1, while mTORC2 controls the dynamics of cytoskeleton, cell shape and cell movements.

9.2.3 mTORC1 signaling pathway is connected to translation

mTORC1 promotes protein synthesis by phosphorylating the eukaryotic initiation factor 4E (eIF4E)-binding protein 1 (4E-BP1) and the p70 ribosomal S6 kinase (S6K1) [Laplante and Sabatini 2009; Marks et al. 2017]. The phosphorylation of 4E-BP1 prevents its binding to eIF4E, enabling eIF4E to promote cap-dependent translation [Richter and Sonenberg 2005a].

The stimulation of S6K1 activity by mTORC1 also leads to downstream pathways that are regulated through the activity of many proteins, such as S6K1, aly/REF-like target (SKAR), programmed cell death 4 (PDCD4), eukaryotic elongation factor 2 kinase (eEF2K) and ribosomal protein S6 [Ma and Blenis 2009]:

- increased transcription (mRNA biogenesis) through S6K1 and SKAR activation
- increased cap-dependent initiation through phosphorylation of 4E-BP1 enabling activation of eIF4E, through double activation of eIF4B and eIF4A or through inhibition of PDCD4 enabling eIF4A
- increased elongation through S6K1 activation, inhibition of eEF2K, desinhibition of eEF2
- increased translation of ribosomal proteins through S6K1 and S6 activation.

A simplified schematic representation of these signaling cascades are shown on Figure 9.1.

The activation of mTORC1 has also been shown to promote ribosome biogenesis by stimulating the transcription of ribosomal RNA (rRNA) through a process involving a protein phosphatase 2A (PP2A) and the transcription initiation factor IA (TIF-IA)[Mayer et al. 2004].

9.2.4 Autophagy and ribophagy

Autophagy and autophagosomes ("self-eating") is a cytoprotective mechanism found in all eukaryotes. It enables the cell to get rid of superfluous material such as defective organelles and macromolecular complexes that when accumulating would have a damaging effect. In its most frequent form, called macroautophagy, the material to be removed and recycled gets packed into double-membrane vesicles called autophagosomes which are transported to lysosomes where their cargo is broken up. When nutrient availability is limited, the degradation of organelles and protein complexes through autophagy provides biological material to sustain anabolic process such as protein synthesis and energy production. mTORC1 inhibition increases

autophagy, whereas stimulation of mTORC1 reduces autophagy. mTORC1 controls autophagy through an unknown mechanism that is insensitive to inhibition by rapamycin. mTORC1 controls autophagy through the regulation of a protein complex unc-51-like kinase 1 (ULK1), autophagy-related-gene 13 (ATG13) and focal adhesion kinase family-interacting protein of 200 kDa (FIP200). mTORC1 represses autophagy by phosphorylating and disactivating ULK1 and ATG13 [Laplante and Sabatini 2009]. Fig. 9.2 is an EM photograph showing Ribophagy with autophagic vacuoles containing ribosomes, in ADAT2 knocked-out human lung cancers cells.

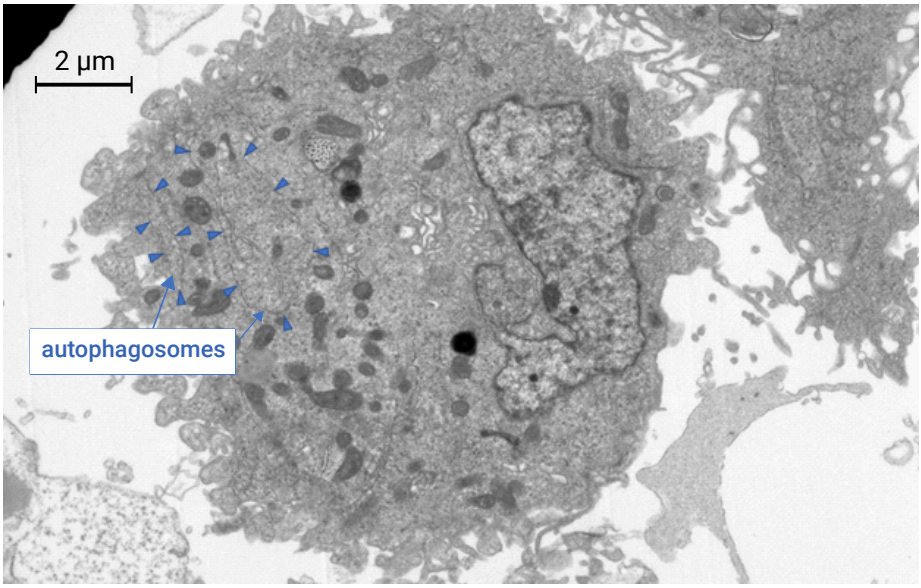


Figure 9.2: Electron micrograph of A549, ADAT2 knocked-out, human lung cancer cells, showing two autophagosomes containing ribosomes. Courtesy of Close, P., Blomme, A., and Thiry, M. Laboratoire de Biologie cellulaire et tissulaire, Liege University.

Ribophagy is a selective autophagy process specifically targeting ribosome turnover, well characterized in yeast and mammals [Kazibwe et al. 2019]. Triggered by nutrient deprivation, ribophagy is the process by which mature ribosomes are delivered to the vacuole or lysosome in an autophagy-dependent fashion, followed by rapid degradation by vacuolar enzymes. In yeast, ribophagy turns over 60S ribosomal proteins.

Fig. 9.3 shows an electron micrograph of autophagosomes containing ribosomes and a mitochondrion, where macroautophagy occurs in a lung cancer cell line knocked-out for ADAT2.

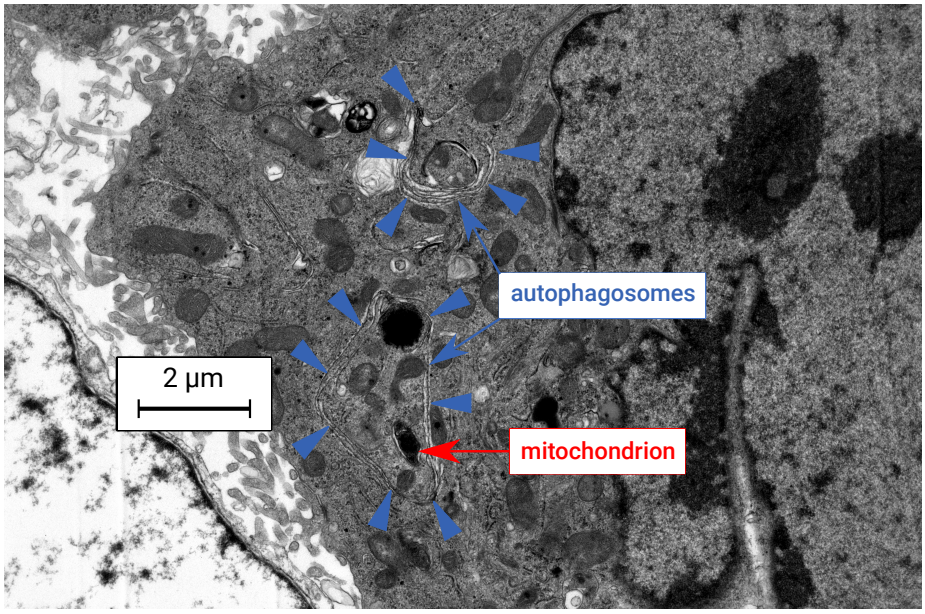


Figure 9.3: Electron micrograph of A549, ADAT2 knocked-out, human lung cancer cells, showing two autophagosomes, one of them containing a mitochondrion. Courtesy of Close, P., Blomme, A., and Thiry, M. Laboratoire de Biologie cellulaire et tissulaire, Liege University.

9.2.5 MAPK signaling and melanoma cancer

The Mitogen Activating Protein Kinase (MAPK) is constitutively activated in 50% of patients with melanoma cancer. It is known that RNA modifications (mRNA, rRNA, tRNA) sustain resistance to targeted therapies, through eIF4F activity [El Hachem et al. 2024]. Activation of the MAPK pathway leads to phosphorylation of **eIF4E**, thereby stimulating translation of selected mRNAs.

9.3 The pool of ribosomes

Ribosomes are the cellular organelles responsible for the translation of the mRNA and polymerization of the amino acid chain during protein synthesis. Ribosomes can lie free in the cytoplasm or can be associated with membranes in the rough endoplasmic reticulum (RER) to produce secreted proteins. Free ribosomes are responsible for the synthesis of cytosol proteins and organelle subunits. They can exist as single free ribosomes or in clusters known as polyribosomes or polysomes. By light microscopy they produce a general low level cytoplasmic basophilia, with more localized and intense basophilia associated with RER as depicted in chapter 1. With the transmission electron micrography, ribosomes appear as small (25-35 nanometer) electron dense particles in the cytoplasm. Polysomes can appear as irregular clumps or organized into rosettes or whorls, Figure 1.2 in chapter 1.

Protein synthesis is an essential function that requires an abundant investment of both energy and resources from cells. Approximately eighty ribosomal proteins, initiation, and elongation factors are necessary for translation to support cell growth and proliferation. Under normal conditions, ribosomes and translation factors are synthesized in coordination with cellular demands. However, under adverse conditions, such as amino acid deprivation or hypoxia, translation is halted, and the energy and resources required to support protein synthesis are redirected to resolve stress-induced cellular damage [Cockman et al. 2020].

9.3.1 Ribosome lifetime

The mean lifetime of a ribosome in a eukaryotic cell during exponential growth is typically on the order of several days. Estimates suggest that in rapidly dividing mammalian cells, ribosomes have a half-life of approximately 5–7 days, though this can vary depending on cell type, metabolic state, and external conditions. In fast-growing cells (e.g., cancer cells, yeast in rich medium), ribosome turnover may be accelerated. In slower-growing or quiescent cells, ribosomes are more stable.

9.4 Non-uniform initiation rates

Deciphering the mechanisms of translational control remains an active area of research. The last two world meetings on translational control, held at Cold Spring Harbor, New York, in September 2022 and 2024, dedicated full sessions—with dozens of talks—to the mechanisms of translation initiation. These discussions have provided new insights into how molecular decisions specifically regulate the initiation of translation for distinct

subsets of transcripts. It has been firmly established that initiation rates vary across different transcripts. In Chapter 1, we summarize the molecular processes involved in translation initiation.

As previously discussed in chapter 1, there are at least two initiation control mechanisms, called the **5m7G Cap-Dependent 5' UTR canonical scanning mechanism** and the **TOP (Terminal Oligopyrimidine Tract) mRNA Translation Initiation mechanism**, respectively. These mechanisms are distinct. The former is a more general mechanism and is less specific in terms of which mRNAs are translated. On the contrary, the latter is more specific and is typically related to the components of the translation machinery, and ribosome biogenesis (ribosomal proteins and elongation factors). The very existence of different 5'UTR sequence upstream the CDS of the ORF of mRNAs imply that different mRNAs are bound to be initiated at different rates. Figure 9.4 illustrates how a selectively increased initiation rate affects translation kinetics, not only by tuning the balance between elongation and initiation rates, but also by selectively enhancing translation efficiency. In particular, the initiation of TOP mRNAs is tightly regulated through secondary structures in their 5' UTRs, which influence ribosome recruitment. Under favorable conditions, increased initiation rates for TOP transcripts facilitate more frequent ribosome loading, leading to a higher ribosome occupancy per transcript. This, in turn, enhances translation efficiency by promoting more rapid and sustained protein synthesis from these mRNAs. For instance, an increase in `eIF4E` amount or activity does not lead to elevated rates of global translation, but instead results in increased translation of subsets of mRNAs. Because some mRNAs require more `eIF4F` owing to strong secondary structure in the `5'-UTR`, over-expression of `eIF4E`, which binds excess `eIF4G` and `eIF4A` (a factor that melts secondary structure upstream of the mRNA open reading frame), can selectively result in their increased translation.

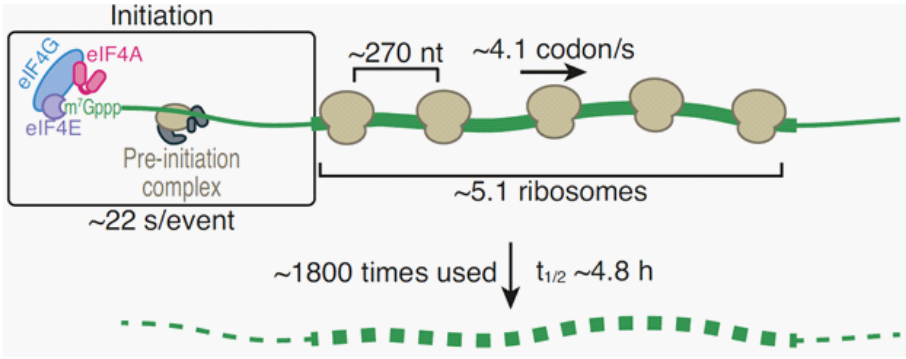


Figure 9.4: Summary of kinetic parameters and average polysome distribution on transcripts. Reproduced from [Tomuro et al. 2024].

Studies have shown that this is clearly the case with ectopically expressed reporter mRNAs containing excessive secondary structure, and also with endogeneous mRNAs encoding such proteins as Myc, fibroblast growth factor (FGF) and vascular endothelial growth factor (VEGF). These and other mRNAs play important roles in controlling cell growth and proliferation, and thus the marked effects of **eIF4E** on transformation could be explained by their elevated rates of translation [Richter and Sonenberg 2005b].

Furthermore, as already presented in chapter 1, mRNAs encoding proteins required for translation are distinguished by a 5' Terminal OligoPyrimidine (5' TOP) motif. The 5' TOP motif begins with a m7G capped C nucleotide followed by a run of approximately 4-15 pyrimidines often followed by a G-rich region. The 5' TOP motif is highly conserved and is found in all 79 human ribosomal proteins as well as non-ribosomal proteins involved in translation including multiple subunits of eIF3, eIF4A, eEF2, and poly(A) binding protein (PABP). This shared TOP motif allows cells to quickly modulate the expression of proteins involved in ribosome production and protein synthesis in response to changes in cellular homeostasis [Cockman et al. 2020].

The relative speed of initiation and elongation determines the rate of protein synthesis and the number and distribution of ribosomes on an mRNA [Arava et al. 2005]. This is true for all transcripts if the ribosome pool is not limited. Slow initiation relative to elongation (and termination) leads to a low density of ribosomes on the mRNA. Faster initiation relative to the previous scenario leads to higher ribosome density. Slower elongation with respect to the first scenario will also lead to a higher ribosome density on the mRNA. It is important to note that an increased ribosome density on a transcript can be the result of either an increased initiation rate relative to the elongation rate or a decrease in the elongation rate with respect to the initiation rate. The key determinant is the **relative speed** between initiation and elongation, rather than their absolute rates.

9.5 Parameter calibration of the ribosome pool and the initiation rates

The recent advances in synthetic biology described in chapter 1 allow high-throughput studies on the determinants of protein production. Sequencing techniques such as ribosomal profiling provide snapshots of the translational machinery in a cell. One way to leverage this new information is to develop computationally tractable models of translation in a cell, to parametrize them from known measurements, and to use them to infer any unknown parameters of global translation dynamics. In combination with ribosomal profiling data, computational models of translation can be used to infer the average initiation rates of all abundant transcripts in a given species such as yeast. Recent studies address the experimental determination of the number of ribosomes per

transcript, their average relative distance on a transcript, the initiation rates and the average number of translation events by ribosomes transcripts [Tomuro et al. 2024].

9.5.1 Calibration of the ratio of the ribosome pool over the number of transcripts

In both cellular and cell-free extract settings, the ribosome pool is limited. An inexhaustible supply of free ribosomes, initiating transcript translation at a constant rate, cannot be assumed. For example, in a single human hepatocyte, the number of ribosomes has been estimated to be around 13 million. With an estimated 100,000 transcript copies, the ratio of ribosomes to transcript copies in a single hepatocyte is approximately 130 : 1. At the other end of the spectrum, in the tip of the axon of a human neuron, the number of ribosomes is much more limited, with roughly one ribosome for every tens of transcript copies. These estimates provide upper and lower bounds for the ribosome-to-transcript ratio, also called the ribosome pool ratio (ratio of the number of ribosomes to the number of transcript copies).

In a pioneering theoretical study on the dynamics of translation in yeast [Shah et al. 2013], along with several experimental studies, the total numbers of ribosomes and mRNAs in an exponential-phase yeast cell were experimentally determined to be 2×10^5 [Warner 1999, Haar 2008] and 6×10^4 [Zenklusen et al. 2008], respectively. This provides a rough estimate for the ribosome-to-transcript copy ratio of approximately 3.3.

In our computational agent-based model, the range of values for the ribosome pool ratio, as detailed in Table 9.5.1, was retrieved from these previously published results [Shah et al. 2013, Warner 1999, Haar 2008, Zenklusen et al. 2008, Tomuro et al. 2024]. This range is considered reasonable and realistic for yeast, fungal species, and even certain mammalian cell types.

Table 9.5.1: Parameter sample space for the ribosome pool ratio: ratio of total number of ribosomes to the total number of transcript copies. The range for the ribosome pool ratio was obtained from references [Warner 1999, Haar 2008, Zenklusen et al. 2008].

Ribosome pool ratio [-]								
$\text{ratio}_{\text{ribo}}$	0.5	1.0	1.5	2.0	2.5	3.0	5.0	10.0

The reference ribosome pool ratio is $\text{ratio}_{\text{ribo}} = 2.5$, i.e., the ratio of the total pool of ribosomes (free + translating) to the total number of transcript copies.

9.5.2 Calibration of the initiation rates

General initiation rates

Implementing a whole-cell model of mRNA translation dynamics at single-codon resolution for a large transcriptome requires accurate values for key parameters, including the total number of available ribosomes in the pool (both free ribosomes and those engaged in translation), the average initiation rate across all transcripts, and the elongation rates of ribosomes at each codon.

Simulations yield quantitative predictions, such as ribosome density on transcripts and the mean distance between consecutive ribosomes, which are now experimentally measurable. When combined with empirical data on protein production rates and mean elongation rates, these simulations allow for the inference of initiation rates.

At least two sources of empirical and experimental data can be used to estimate critical parameters of translation dynamics in living cells, such as the mean initiation rate of transcripts. Here, we compare the range of values inferred for general initiation rates in two literature studies [Shah et al. 2013, Morisaki et al. 2016].

In yeasts, the time between initiation events on any mRNA molecule is between 4 s (5% percentile) and 233 s (95% percentile) with a median time between initiation events of 40 s (50% percentile) [Shah et al. 2013]. A simple calculation allows to determine the dynamics of the pool of free ribosomes. If the time between two initiation events is exponentially distributed, the relations between the initiation rate λ, the mean time t_{mean} and the median time $\tau_{1/2}$ between two initiation events follow form the properties of the exponential probability density function:

$$t_{\text{mean}} = \frac{1}{\lambda} \tag{9.1}$$

$$\tau_{1/2} = ln2 \cdot t_{\text{mean}} \tag{9.2}$$

$$\tau_{1/2} = \frac{\ln 2}{\lambda} \tag{9.3}$$

The units of the initiation rate λ are $[\text{s}^{-1} = 10^{-3}\,\text{ms}]$. Units of initiation rates are inverse of units of time. From an initial condition where all ribosomes are free, the pool of free ribosomes will decay exponentially with a half-life time of $\tau_{1/2} = 40$ s if the median time between two consecutive initiation events is 40 s. Another simple calculation allows to determine, that with a median time between initiation events of 40 s, and with a mean elongation rate of 5.6 amino acid per second (in higher eukarya), the median distance between two consecutive bound ribosomes is around 224 codons (672 nucleotides) under the assumption that the resource pool of ribosomes is unlimited.

Morisaki *et. al* succeeded in achieving real-time in-vivo imaging of mRNA translation with single-molecule resolution. The nascent chain tracking (NCT) technique they developed allowed to determine that initiation occurred stochastically every ≈ 30 s. They observed that polysomes feature a mean distance between two consecutive ribosomes in the range 67 codons (200 nucleotides) to 300 codons (900 nucleotides) [Morisaki et al. 2016].

A straightforward probabilistic and statistical derivation establishes a relationship between the mean initiation rate, the mean elongation rate, and the average number of ribosomes in a polysome as a function of mRNA length, under the assumption of an unlimited ribosome pool.

Assume that ribosomes initiate randomly and uniformly across all transcripts (mRNAs), such that, on average a ribosome elongates n amino acids before another ribosome initiates. To determine the probability of finding M ribosomes on a transcript encoding a protein of length N amino acids (corresponding to N codons in the transcript, excluding the stop codon), we refer to the binomial distribution from probability and statistics theory.

This scenario can be analogized to rolling M sixes in N Bernoulli trials (i.e., N die tosses), where $M < N$. In this analogy, the probability of rolling a six is $p = 1/6$, while the probability of rolling any other outcome is $q = 1 - p = 5/6$. According to Bernoulli trials, the probability of obtaining exactly M successes in N trials is given by:

$$P(M, N, n) = C_N^M \, p^M \, q^{N-M}. \tag{9.4}$$

Following the coin toss analogy, the probability P_{rib} of observing M ribosomes on a transcript encoding N amino acids is given by:

$$P_{rib}(M, N, n) = C_N^M \, (\frac{1}{n})^M \, (\frac{n-1}{n})^{N-M}. \tag{9.5}$$

Here, N choose M (denoted as C_N^M) represents the number of ways to choose M elements from N without considering order in the elements. The probability of initiation within the time required for a ribosome to translate a single amino acid is given by $p = 1/n$, which is equivalent to the ratio of the initiation rate to the elongation rate. Consequently, each time a ribosome translates one amino acid, it is as if a trial (a coin toss) occurs to determine whether another ribosome will initiate. In total, there are N trials (N coin tosses), corresponding to the number of amino acids encoded by the transcript (or equivalently, the number of sense codons in the mRNA sequence) before the ribosome terminates.

From the properties of Bernoulli trials, it is known that the expected number of ribosomes per transcript is given by:

$$< M > = Np \tag{9.6}$$

$$< M > = \frac{N}{n}. \tag{9.7}$$

The longer the transcript (of N codons), the more ribosomes. If the average number of ribosomes engaged in translation (number of translating ribosomes per transcript) is plotted as a function of transcript length (or protein length), N, then the slope will give the ratio of the initiation rate to the elongation rate (assuming an inexhaustible supply of free ribosomes).

The probability P_{poly} to find a polysome, an mRNA with two or more translating ribosomes, is

$$P_{poly} = 1 - P_{rib}(0, N, n) - P_{rib}(1, N, n). \tag{9.8}$$

This provides an estimate of the frequency that polysomes are observed. Finally, the average number of ribosomes per polysome N_{poly} is given by:

$$N_{poly}(N, n) = \frac{1}{P_{poly}(N, n)} \sum_{M=2}^{\infty} M \cdot P_{rib}(M, N, n). \tag{9.9}$$

The reasoning outlined above holds true only under the assumption of an unlimited ribosome pool.

In our computational agent-based model, the range of values for the mean initiation rates in the parameter sample space is derived from previously published results [Shah et al. 2013, Morisaki et al. 2016]. A recent study by Tomuro et al. [Tomuro et al. 2024], published during the preparation of this thesis, provides experimentally measured values that are consistent with this range.

The time between two initiation events on any given transcript (inverse of initiation rate) will be varied in all the computational simulations in the range 4 s to 233 s. More specifically, in the simulation runs, the parameter sample space for the mean initiation rates takes the values in table 9.5.2. It is important to emphasize that, in our agent-based model, the assumption of an unlimited ribosome pool is relaxed. The ratio of the total number of ribosomes in the pool to the number of transcript copies is treated as a tunable parameter in the computational simulations, that must be fixed at the start. We will investigate the impact of a restricted ribosome supply and how the interaction between this limited resource and tRNA modifications—affecting elongation rates—influences protein relative abundances, translation efficiencies, ribosome fragmentation profiles, and ribosome density maps.

Table 9.5.2: Parameter sample space for the mean initiation rates of translation of the most abundant transcripts by ribosomes (*Saccharomyces cerevisiae*). The range of values for the mean initiation rate was adopted from references [Shah et al. 2013, Morisaki et al. 2016].

Initiation rate [$\times 10^6$ ms^{-1}]								
λ	4.3	6	11	16.7	25	50	100	200
Mean time to initiation event by a ribosome [s]								
$\mathbf{t_{mean}}$	232	166	91	60	40	20	10	5
Median time for a free ribosome to engage translation [s]								
$\tau_{1/2}$	161	115	63	41.6	27.7	13.9	6.9	3.5

The reference initiation rate is $\lambda = 16.7\,10^{-6}$ ms^{-1}, i.e., the mean time for a free ribosome to initiate a transcript is 60 s (or $\tau_{1/2} = 41.6$ s, the half-life time of the free ribosomes to engage translation).

Transcript specific fold change of the initiation rates and their calibration

In addition to the transcript copy number associated with each individual transcript's CDS sequence, our ABM model takes as input a file containing the specific fold-change values of individual initiation rates relative to the general initiation rate. This feature enables selective enhancement of ribosome recruitment on preferred transcripts, reflecting the fact that different transcripts have distinct 5' UTR sequences that play a regulatory role.

The precise calibration of these individual initiation rate fold changes remains unknown. As a starting point, users should adopt a default calibration in which all initiation preferences are evenly distributed. It should then be recognized that subsets of transcripts may share the same initiation rate fold change to account for common patterns in their 5' UTR sequences, if applicable. For instance, all ribosomal protein sequences from the Ribosomal Protein Small subunit (RPS) and Ribosomal Protein Large subunit (RPL) families are assumed to be consistently upregulated through the TOP motif, as previously discussed.

An alternative approach to determining individual initiation rates is to empirically tune their values by comparing simulation results with real datasets under control conditions. Additionally, fine-tuning these rates could leverage backpropagation algorithms from machine learning techniques, offering a potential avenue for future research extending the work presented in this thesis.

The key takeaway is that our ABM model provides the flexibility to parameterize individual initiation rates, allowing for the exploration of transcript-specific ribosome

recruitment dynamics.

9.6 Summary of main findings and insights

This chapter 9 explored the impact of **ribosomal resource availability** and **non-uniform initiation rates** on protein synthesis. It discussed how ribosome biogenesis and degradation influence the dynamic balance between translating and free ribosomes. A key insight is how initiation rates vary across transcripts, shaping translation efficiency and proteomic composition. The chapter also described regulatory mechanisms such as the `mTORC1` signaling pathway and ribophagy, linking metabolic states to translation control. Additionally, it presented parameter calibration strategies for quantifying initiation rates and ribosome pool sizes. These findings underscore **the need to rely on computational models to infer biological parameters for which reliable experimental protocols are currently lacking, such as the measurement of global and individual initiation rates across the transcriptome**.

Chapter 10

Model output and flexibility offered by the input data and parameters

In the preceding six chapters, the key causal and influential factors were carefully analyzed, motivated, and justified to determine which input parameters significantly affect the model's behavior. In this chapter, we start by providing a brief summary of the model, its parameters and the ranges of their values as defined in the previous chapter. We then turn our attention to the model's outputs. These outputs correspond to the primary questions that motivated the modeling effort: namely, the predicted protein relative abundances, translation efficiencies, ribosome distribution across transcripts, as well as polysome profiles and ribosome profiling results. Relevant computational procedures and modalities used to generate these outputs throughout the simulations are also explained.

10.1 Summary of the model and its input data and parameters

10.1.1 Ribosomer at a glance

The `Ribosomer` model is a **real-time dynamic framework** that computationally simulates the translation process of multiple mRNA copies by a shared ribosome pool over time. Figures 10.1 and 10.2, replicated from chapter 3, illustrate the main components and influencing factors of the model.

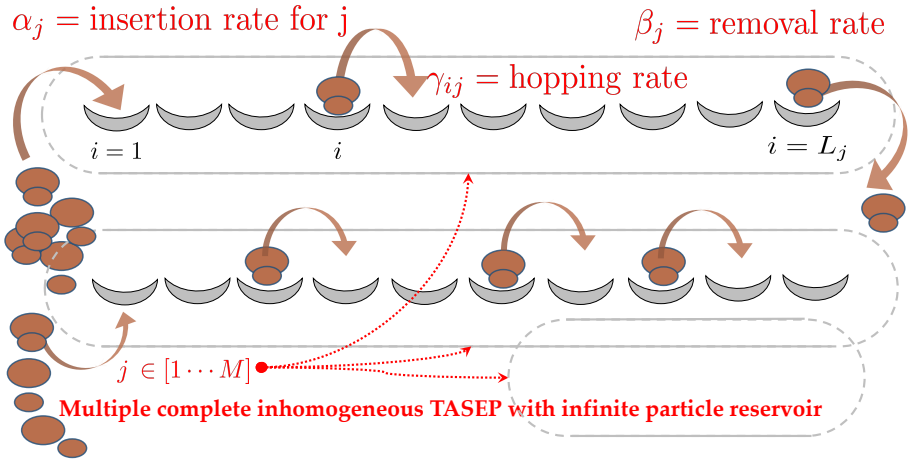


Figure 10.1: Summary of Ribosomer as an extended TASEP model. A ribosome initiates translation with initiation rate (insertion rate) α_j which is transcript dependent, when the first six codon positions after the start codon are not blocked by another ribosome. The ribosome translates the i^{th} codon position with elongation rate γ_{ij} when no downstream ribosome occupies the $(i + 10)^{\text{th}}$ codon position and terminates the translation process with rate β_j. In this figure, the ribosome pool is supposed to be unlimited.

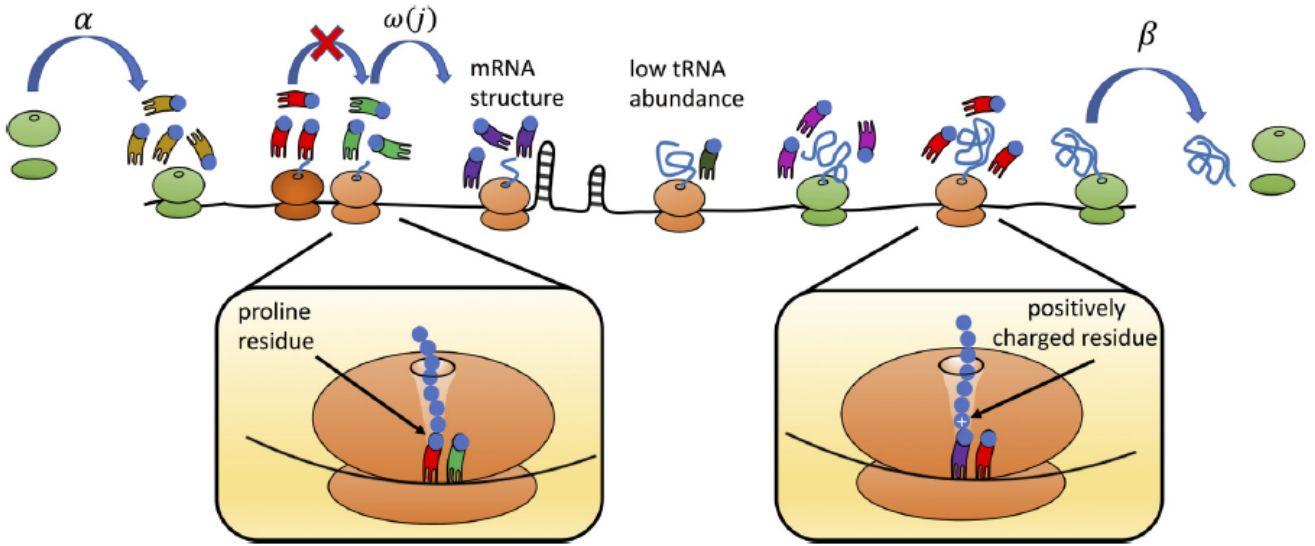


Figure 10.2: Summary of Ribosomer and its contextual factors in the TASEP approach. A ribosome initiates translation with rate α when the first six codon positions after the start codon are not blocked by another ribosome. The ribosome translates the j^{th} codon position with elongation rate ω_j when no downstream ribosome occupies the $(j+10)^{\text{th}}$ codon position and terminates the translation process with rate β. Contextual factors influence the elongation rate: tRNA abundance, tRNA enzymatic modifications, electrostatic interaction with the ribosome exit tunnel, proline residues at the A- or P-sites, mRNA secondary structure roadblocks are accounted for in the model. Ribosomes in green and light-brown are translating fast and slow codons, respectively, whereas the ribosome colored in dark-brown is sterically blocked by a downstream ribosome (traffic congestion). Reproduced from [Sharma et al. 2018].

The `Ribosomer` framework offers several distinctive features, including:

i the ability to constrain the ribosome pool, treating it as a limited resource for which multiple transcripts compete (transcripts are denoted j in Fig.10.1);

ii the option to assign specific initiation rates, α_j, to subsets of transcripts, enabling the simulation of preferential ribosome recruitment and favored initiation (due to shared upstream UTR sequences across subsets of transcripts for instance);

iii the possibility to parametrize the elongation rates of 61 sense codons individually (codons are denoted by their position i on any transcript j in Fig.10.1);

iv the flexibility to activate or deactivate specific contextual factors individually or collectively (represented in Fig.10.2), allowing for the disentanglement of their contributions. In real wet lab experimental settings, these factors are typically intertwined –if not entirely inseparable. However, as a modeling tool, `Ribosomer` allows for targeted, isolated, and computationally controlled analyses of these factors.

10.1.2 List of model input data and parameters

Before running simulations with our model, a key consideration is providing the appropriate input data. These inputs fall into two main categories.

The **first category** consists of experimentally measured data, such as the copy number of transcripts. Specifically, transcript copy numbers can be derived from RNA-Seq read counts within the relevant experimental conditions.

The **second category** includes model-specific parameters, like the global initiation rate. Determining appropriate values for these parameters is a crucial step. While some parameters have been calibrated based on published literature, others lack precise experimental measurements. Certain biologically significant parameters cannot be directly quantified, making computational simulations essential for inferring their plausible values.

To estimate these unknown parameters, **optimization techniques can be employed to minimize the discrepancy between model predictions and experimental data**. However, this is not straightforward: multiple parameters often interact, complicating the calibration process.

A vital first step in this process is conducting a **sensitivity analysis** to assess how each parameter influences the model's output. This will be addressed in the next chapter. The computational experiments presented in the current chapter lay the groundwork for

that analysis. Their purpose is to systematically explore physically meaningful ranges of parameter values.

We recapitulate here the input data and the model parameters that can have an impact on any of the predicted output values of `Ribosomer`:

1. **Relative abundance of transcripts** (distribution of the number of transcript copies across the transcriptome), i.e. RNA-seq profiles.
   By default, all transcripts are considered to have a single copy. The distribution of transcript copies within a sample can be provided via an input file table (`.txt`).

2. **Individual initiation rates of transcripts** (transcript specific preferred initiation).
   By default, the individual initiation rates for each transcript are equal, with no preferential ribosome recruitment on subsets of transcripts. However, it is possible to assign a fold change to the initiation rate of each transcript, along with the previously discussed copy-relative counts. This allows for the assignment of a multinomial distribution of initiation rates within the transcript sample.

3. **Global initiation rate** (rate at which any ribosome starts translation on any transcript). The global initiation rate is the 'insertion' rate in the TASEP 1D-lattice diffusion model, represented by α in Fig. 10.2. Physiologically realistic values for the global initiation rate, determined in Chapter 9, range from 5 s/event to 232 s/event. For yeast cells in the exponential growth phase, the global initiation rate is between approximately 10 and 40 s/event (Chapter 9).

4. **Ribosome pool size** (ratio of the total number of ribosomes to the total number of transcript copies).
   The ribosome pool is a limited resource. The ratio of the total number of ribosomes to the total number of transcript copies, called the ribosome ratio, ranges from 0.5 to 10 (Chapter 9).

5. **Different codon usage or significant bias in the synonymous codons usage** of the nucleotide sequences across the transcriptome.
   The CDS sequences of the transcripts to be computationally simulated are provided in the input data file in FASTA format. Users can provide synthetic sequences of their own design, including custom codon usage. By default, sequences from public repositories can be used. Users can create the FASTA input data file by selecting the gene ID tags for which they wish to perform a protein synthesis simulation using `Ribosomer`. For illustration, the sections below and the next chapter use a set of 27 gene ID transcripts, with sequences available from NCBI or Ensembl repositories. Some of these gene ID tags were chosen for their enrichment in certain codons, such as those in the kinesin family (KIF4A, KIF14, KIF23), ribosomal proteins (RPL4, RPL22), or proline-rich genes (COL1A1).

6. **tRNAs abundance or elongation cycle substeps kinetics parameters**.
   The list of the three rate parameters for the 61 hypo-exponential distributions associated with the 61 sense codons is provided in an input data file, calibrated from a meta-analysis of literature data as detailed in Chapter 4. The kinetics data are stored as a dictionary of key-value pairs, where the keys represent the 61 codons and the values are lists containing three parameters: accommodation, peptide bond formation, and translocation rates.

7. **Amino acid specific sequences** encrypted in the transcriptome (proline enrichment or charged amino acid enrichment or charged amino acid distributions).
   Some transcripts may encode proteins whose sequences are enriched in specific amino acids. For instance, the collagen protein (COL1A1) is particularly enriched in proline. Users can design and engineer their own sequences or use those available in public repositories.

8. **Electrostatic interaction in the ribosome exit tunnel** (model factor).
   This factor is depicted in Fig. 10.2. The electrostatic profile in the ribosome exit tunnel is provided as an input data file, calibrated by default as detailed in Chapter 6. This influential factor can be toggled on or off before a simulation.

9. **Proline slow incorporation** (model factor).
   This factor is represented in Fig.10.2. A slow down caused by ELP or eIF5A depletion can be toggled on or off before a simulation.

10. **mRNA secondary structures** (model factor)
    This factor is represented in Fig.10.2. The mRNA secondary structures factor can be toggled on or off before a simulation. At the thesis publication date, in the available versions of `Ribosomer`, this factor is toggled off.

11. **U34 enzymatic tRNA modifications** on the sensitive codons of K, Q, E (optional experimental factor).
    6 sensitive codons (3 amino acids (E, K, Q) may have their elongation rate kinetics changed upon U34 tRNA hypomodification. This factor can be toggled on or off before a simulation run.

12. **ADAT enzymatic modifications** on the sensitive codons of T, A, P, S, L, I, V, R (optional experimental factor).
    37 sensitive codons (8 amino acids TAPSLIVR) may have their elongation rate kinetics changed upon I34 tRNA ADAT targeted hypomodification. This factor can be toggled on or off before a simulation run.

Some of these parameters are defined as **intrinsic model parameters**, while others depend on the **input data**. Data-dependent parameters include, for example, the **specific**

codon sequence of a given transcript provided in the input data file (`FASTA` format), the **number of copies of a given transcript ID tag** (as determined from RNA-Seq data), or the **fold change in initiation rate of a given transcript** relative to the global initiation rate. In contrast, intrinsic model parameters include the **ribosome pool size** (ribosome ratio) and the **global initiation rate**, which is equal to the inverse of the mean time to an initiation event by a free ribosome.

10.1.3 Steady state achievement and sampling rate used to build aggregated metrics

`Ribosomer` is a real-time dynamical model that simulates the translation of multiple mRNA copies by a fixed pool of ribosomes over time. The total runtime of a simulation must be long enough to achieve a steady state. Pilot simulations have shown that steady state is typically reached after 3-5 minutes. This duration corresponds to 3 to 5 times the average protein translation time. The average length of proteins in the human proteome is approximately less than 400-450 amino acids [The UniProt Consortium 2020]. At an average elongation rate of 5.6 amino acids/s (eukaryotes), it takes about 90 seconds for a single ribosome to translate a 500-codon (1500-nucleotide) transcript. Therefore, a steady state for elongation is expected to be reached after 3-5 minutes.

Furthermore, with a global initiation rate of 20-60 s/event, equivalent to a half-life of $\tau_{1/2} \approx 13.9 - 41.6$ s (see chapter 9), a pool of initially free ribosomes will have reached its steady-state proportion of free and translating ribosomes after 3 minutes. This time corresponds to 4-13 times the half-life of the initially *free* ribosome pool.

In all simulations conducted, the runtime was set to just under two hours (108 minutes). This duration is sufficient not only to allow protein synthesis dynamics to reach a steady state but, more importantly, to collect and aggregate biologically relevant metrics.

During each 108-minute simulation, 36 time checkpoints (one every 3 minutes) are used to record the temporal evolution of several output variables, which are detailed in the following sections. These variables are stored both in memory and in output files.

For specific output variables –particularly those required for polysome and ribosome profiling– the sampling interval is every 10 seconds. In other words, a virtual snapshot of the position of all ribosomes on all transcripts is recorded every 10 seconds. By the end of a single 108-minute simulation (6,840 seconds), 648 snapshots of the translatome are generated. This is equivalent to producing a Ribo-seq analysis for a bulk sample of 648 cells, all under the same physiological conditions as defined by the initial parameter settings. Statistical aggregation of these 648 snapshots is then performed to virtually reconstruct global and per-transcript polysome fragmentation profiles, as well as ribosome density maps.

The snapshots taken during the first 3–5 minutes may not fully represent steady-state conditions; however, they account for only 2.8–4.6% of the sample space for these aggregated output variables.

It is important to emphasize that, in laboratory settings, when cells are sampled for bulk RNA-seq or bulk Ribo-seq analysis, the degree of physiological synchronization within the bulk population is unknown. Some cells may have reached a steady state, while others may not have completed ribosome biogenesis to its full extent. Some may have only just initiated translation, meaning that the full extent of protein synthesis has not yet been achieved. Thus, steady-state sampling conditions in the laboratory cannot be assumed to apply uniformly across the bulk cell population within a sample.

In the computational production of aggregated metrics over the 648 time points discussed above, we are confident that at least 95% of these time points correspond to steady-state conditions within a controlled physiological state.

All endpoint metrics, including the relative distribution of free and translating ribosomes, the proportion of free transcripts, protein relative abundance, and translation efficiency, were calculated only after steady-state conditions had been well established and maintained for an extended period.

10.2 Predicted output variables of the model

The following sections below provide examples of key output variables that `Ribosomer` can predict. These outputs were generated using selected settings that are detailed below and were all based on the same input file of 27 transcripts (27 gene ID tags), for which FASTA sequences were retrieved from public genomic repositories [Ensembl 1999 and National Center for Biotechnology Information 1992]. It is important to emphasize that `Ribosomer` offers a high degree of flexibility, allowing for customization of additional outputs that may be valuable to researchers and end-users.

10.2.1 Distribution of free and translating ribosomes

One of the outputs of `Ribosomer` is the distribution of ribosomes between the free pool and those engaged in translation. At the start of a simulation, all ribosomes are free. After a duration approximately equal to the mean initiation time multiplied by $\ln 2$, half of the initial ribosome pool becomes engaged in translation. After roughly five to ten times the mean initiation time, the system reaches a steady state, and the frequency distribution of free and translating ribosomes stabilizes.

Figure 10.3 illustrates the steady-state distribution of the ribosome pool for various combinations of initiation rates and ribosome pool sizes (expressed as the ribosome ratio, i.e., the total number of ribosomes per transcript copy number). In this example, the simulation input includes 58 transcript copies corresponding to a set of 27 gene ID tags.

A related model output is the average distance between ribosomes on a transcript, calculated across the transcriptome. For instance, when the initiation rate is set to 40 s/event and the ribosome ratio is 3, the average distance between ribosomes is approximately 230 codons (690 nucleotides). In contrast, for a ribosome ratio of 5 and an initiation rate of 25 s/event, the average inter-ribosome distance decreases to 140 codons (420 nucleotides).

It is worth mentioning that, while our model is capable of inferring metrics such as this ribosomal interdistance, recent experimental techniques now allow direct measurement of ribosome spacing on individual transcripts [Morisaki et al. 2016]. This advancement offers an additional opportunity for model validation.

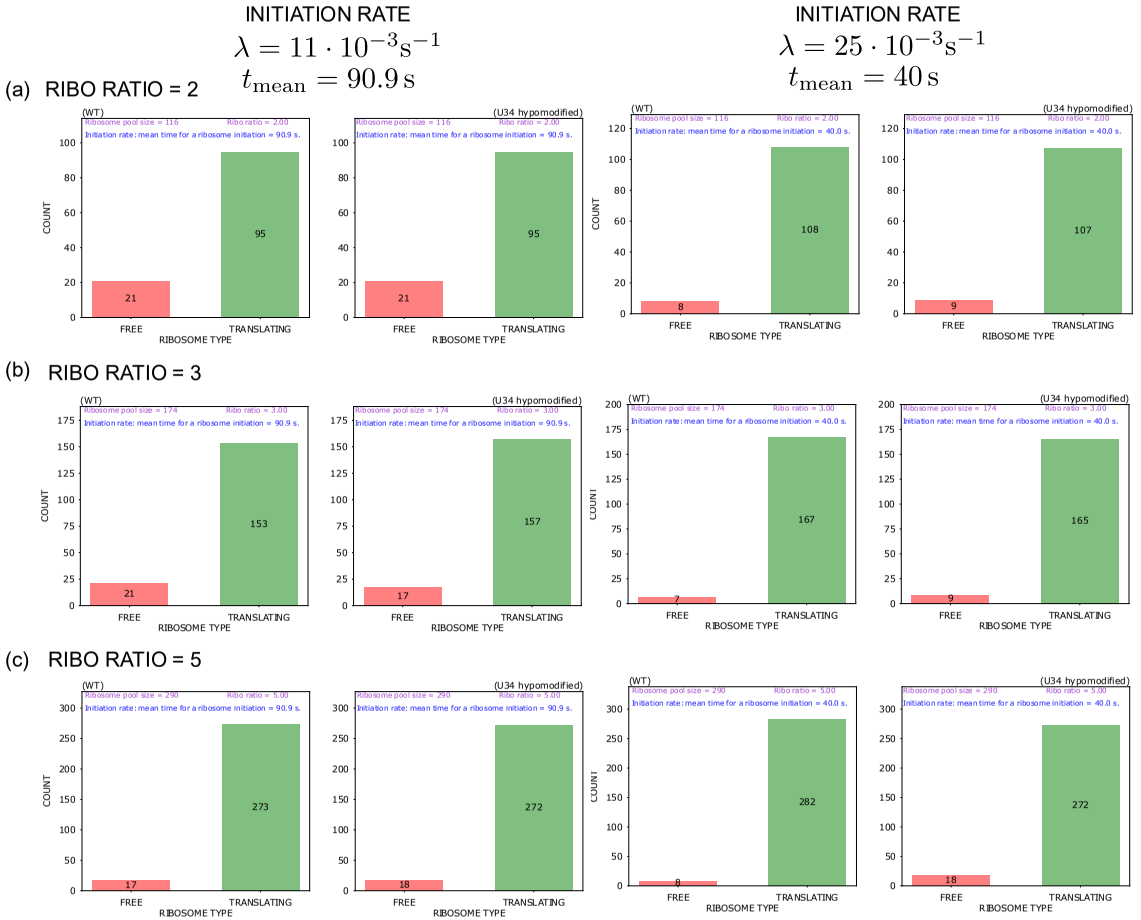


Figure 10.3: Distribution of free and translating ribosomes for two initiation rates (columns) and three ribosome pool sizes (rows). Each panel compares control and treatment conditions. The $y-$axis scale differs across rows and/or columns. The total number of translating ribosomes increases more rapidly with ribosome pool size than with initiation rate.

Another relevant output of `Ribosomer` is the proportion of transcript copies that are not engaged in translation (i.e., those free of ribosomes). A high proportion of transcripts not being translated reflects a poor efficiency of the translation machinery. More on this in the next chapter. A related output is the polysome fragmentation profile that will be described below.

10.2.2 Protein relative abundance

At the end of each simulation run, the protein yield resulting from the translation of all transcripts –provided in the input file and translated by the specified pool of ribosomes– is determined by counting, for each individual copy of a given gene ID tag, the number of ribosomes that successfully completed termination. The total protein abundance associated with each gene ID tag is then obtained by summing the yields across all corresponding transcript copies. This predicted protein abundance can be visualized as a bar plot to highlight the relative abundance within a sample and to facilitate comparisons between samples. An example of such a comparison is shown in Figure 10.4, which contrasts the predicted protein abundances between a wild-type (control) and a U34-tRNA hypomodification (case) setting.

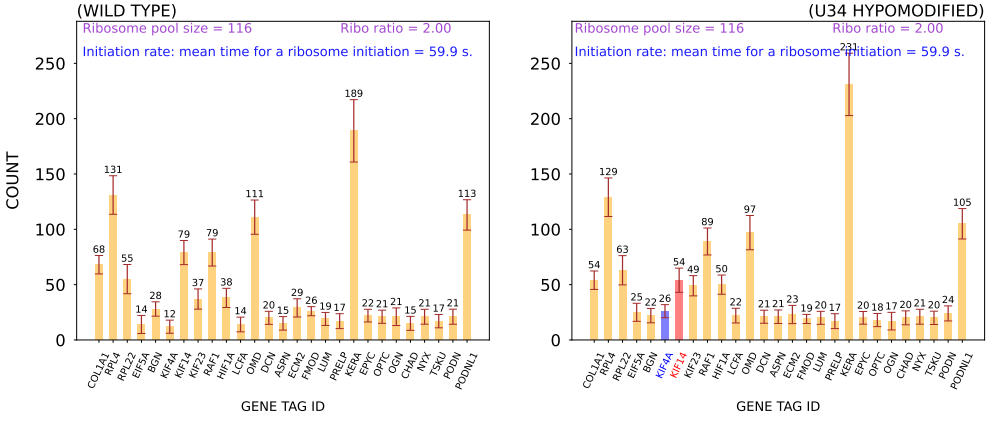


Figure 10.4: Protein relative abundance of 27 gene ID tags at endpoint. Left panel: control. Right panel: U34 tRNA modification. 30 replicates of simulation runs with the same set of parameters in both panels; error bars: standard deviation. Difference in counts larger than $3\,\sigma$ are marked in red (smaller) or blue (larger) if treatment differs from control.

10.2.3 Translation efficiency

Translation efficiencies (TE) are calculated by dividing the protein abundances by the number of transcript copies, normalized per hour and per kilobase of transcript length. An example of this predicted output is shown in Figure 10.5, based on the same simulation settings used to generate the previously discussed protein relative abundances, namely: the same global initiation rate of 60 s/event, ribosome ratio of 2, input file with the same gene ID tags, same transcript copy numbers and individual initiation rate fold changes. 30 replicates were conducted as before but, only one instance is shown here. This is what a typical lab single sample result would look like.

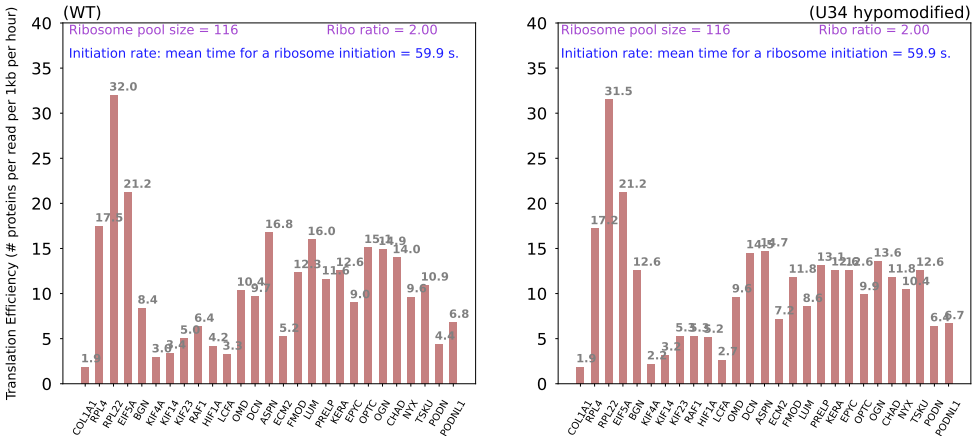


Figure 10.5: Protein translation efficiencies of 27 gene ID tags. Left panel: control. Right panel: U34 tRNA modification. Identical settings as for the relative protein abundances shown before: same global initiation rate of 60 s/event, ribosome ratio of 2, input file with the same gene ID tags, same transcript copy numbers and individual initiation rate fold changes. A single instance of the 30 replicates is shown here.

10.2.4 Polysome fragmentation profile (global)

Contrary to the single-endpoint results for protein relative abundance presented in Section 10.2.2, where 30 replicates were conducted to ensure statistical robustness, the statistical weight for polysome fragmentation profiles is derived from the aggregation of 648 snapshots. At the end of a simulation run, the statistical analysis of the 648 virtual snaphots is conducted by counting the number of transcripts that were free of ribosome, or were footprint with n ribosomes, n ranging from 0 to 20. The polysome

profile predicted output[1] is the bar plot (histogram) of the transcript frequency footprint ordered by polysome size (no ribosome, monosome, disome, n-polysome). Figure 10.6 shows that the polysome multiplicity increases with ribosome pool size and initiation rates. For small ribosome pool size or when the global initiation rate is low, a large proportion of transcripts are free of ribosomes ($n = 0$ in the $x-$axis of Fig. 10.6).

10.2.5 Polysome fragmentation profile (per transcript)

The same procedure as for the previous subsection is applied to produce the polysome profile predicted output for any individual transcript of choice. The 648 snaphots can be filtered by gene tag ID. Figure 10.7 shows the predicted polysome fragmentation profiles of the ribosomal RPL4 transcript taken as example for three different parameter settings of the ribosome pool.

10.2.6 Ribosome density maps per transcript

With a simulation run time of 108 minutes and a sampling frequency of 0.1s^{-1} (i.e., one snapshot every 10 seconds), 648 snapshots are collected, capturing the exact codon positions of all ribosomes in the pool across all transcript copies. The total number of snapshots, multiplied by the number of ribosomes engaged in translation in each snapshot, and then by the ribosome footprint length ($30 - 31$ nucleotides), divided by the total length of the simulated transcriptome, allows to estimate a **virtual equivalent of sequencing depth (or coverage)** in Ribo-Seq profiling[2]. **After normalizing for the length of the transcriptome (or translatome), this corresponds to a minimum fold coverage of 10**. For transcripts averaging 1,500 nucleotides (500 codons) in length, the fold coverage ranges from 10 to 50, depending on the ribosome ratio –ranging from 1 to 5– and the general initiation rate (above 60 s/event), both of which were parametrized at the start of the simulation. The virtual Ribo-Seq coverage is estimated at the end of each simulation and provided along with the ribosome density maps, Figure 10.8.

[1]The polysome profile is built from the statistical aggregation of 648 snaphots collected during 108 minutes in a single simulation run.

[2]Experimental sequencing depths in ribosome profiling correspond to a range of $10 - 50$ in fold coverage, meaning that each nucleotide in the transcriptome was covered by 10-50 ribosome footprints on average.

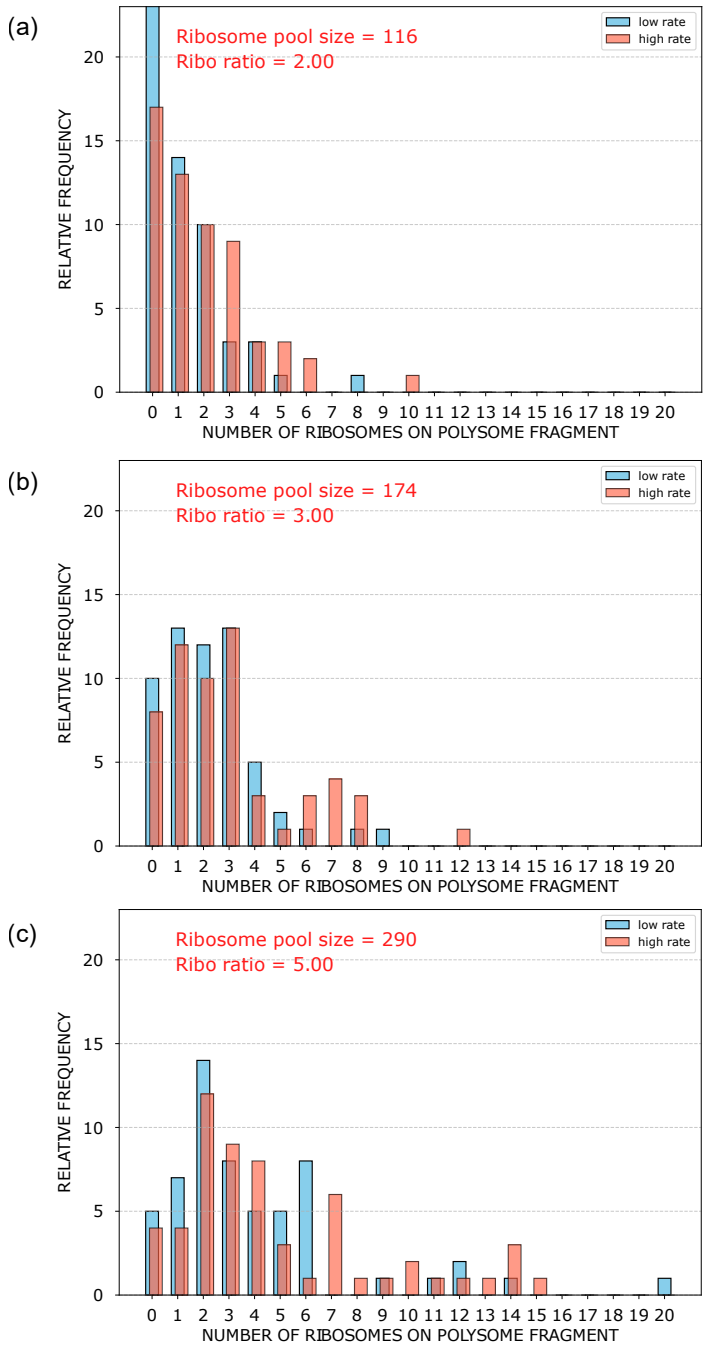


Figure 10.6: Polysome profiles predicted output from six simulation runs conducted with the Python version of **`Ribosomer`**. The relative frequency of transcripts fragments occupied by n ribosomes in the predicted polysome fragmentation profiles are shown for three ribosome pool ratios: (a) 2; (b) 3 and (c) 5. The blue and red colors correspond to a global slow and fast initiation rate of $\lambda = 4.3\,10^{-3}\text{s}^{-1}$ (232 s/event) and $\lambda = 25\,10^{-3}$ s^{-1} (40 s/event), respectively.

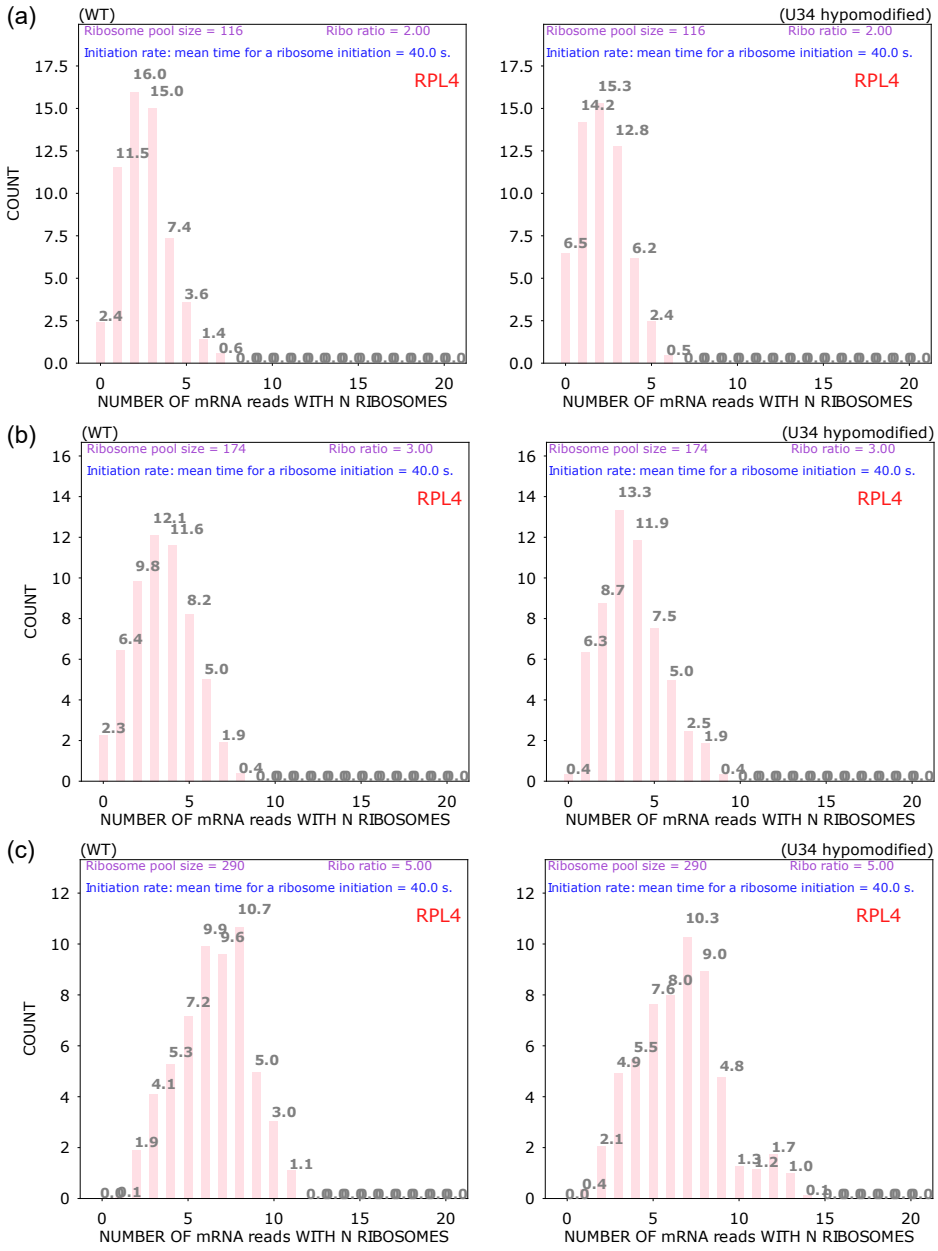
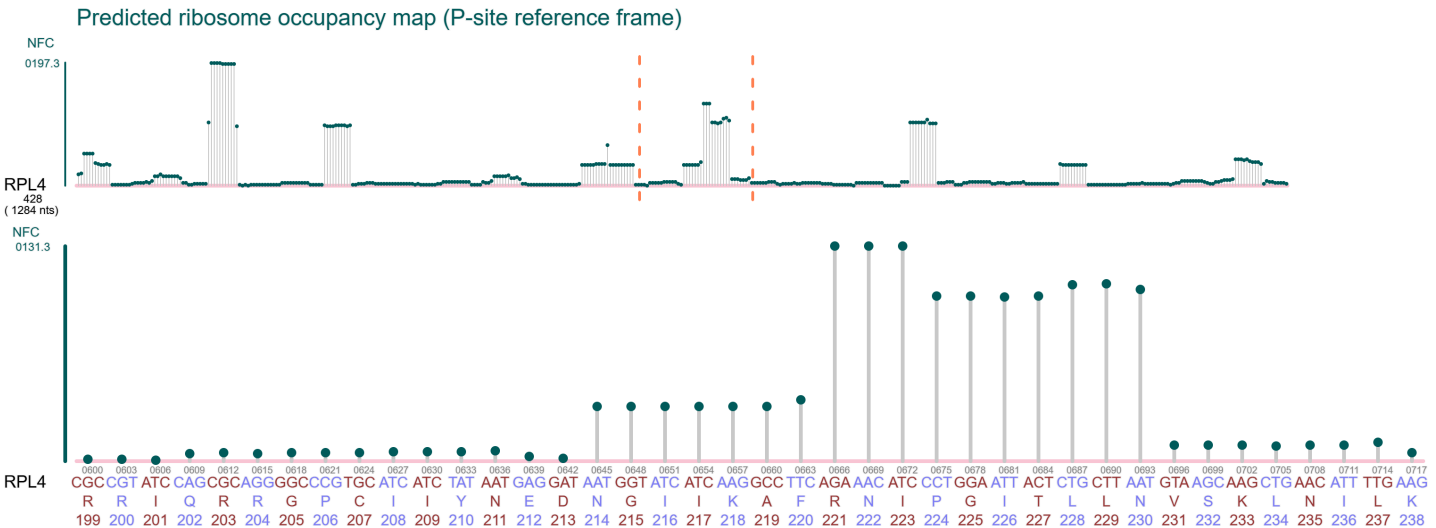


Figure 10.7: Per-transcript polysome profiles predicted output from six simulation runs conducted with the Python version of **`Ribosomer`**. The frequency of the **`RPL4`** transcripts fragments occupied by n ribosomes in the predicted polysome fragmentation profiles are shown for three ribosome pool ratios: (a) 2; (b) 3 and (c) 5. The left and right panels correspond to wild type and U34 tRNA hypomodification settings, respectively. The initiation rate was parametrized to $\lambda = 25\ 10^{-3}\ s^{-1}$ (40 s/event) in all simulations.



Number of ribosome footprint snapshots = 648
Number of ribosome footprint fragments = 46442
Averaged number of ribosomes on all transcripts per snapshot = 071.6
Exact Ribo-Seq coverage = 13.9

Figure 10.8: Ribosome density map of RPL4 at steady state for a ribosome ratio of 2 and a general initiation rate of 60 s/event. Upper panel: Ribosome occupancy across the full transcript length. Lower panel: Zoomed-in view of a 40-codon window, extracted between the orange dashed lines. y-axis: Normalized footprint count (NFC). The NFC is calculated as the total number of ribosome footprint fragments (RPFs) recorded at a specific codon position (P-site) or within a 30-nucleotide fragment window centered on the P-site codon, divided by the transcript copy number and the length of the transcript to which the RPFs belong. x-axis: Codon position in the transcript sequence. **`Ribosomer`** Javascript version.

At the end of any simulation, a **ribosome density map** or **ribosome occupancy map** can be visualized in the *JavaScript* version of `Ribosomer` for any of the transcript gene ID tag of interest that was in the `FASTA` input file, Figure 10.8.

The $y-$axis of a ribosome density or occupancy map typically represents the normalized footprint count (NFC). The NFC is calculated as the total number of ribosome footprint fragments (RPFs) recorded at a specific codon position (P-site) or within a 30-nucleotide fragment window centered on the P-site codon, divided by the transcript copy number and the length of the transcript to which the RPFs belong. The $x-$axis denotes the codon position within the transcript of interest.

These maps can, in principle, be compared to real ribosome density maps available in public Ribo-Seq repositories and visualized using tools such as `GWIPSviz` [GWIPS-viz 2013].

The `Ribosomer` model can also be used to predict ribosome density maps under different conditions, allowing the effect of specific factors on ribosome profiling to be anticipated.

10.3 Verification and sanity checks

10.3.1 Verification of consistent output across the two language implementations

In computational modeling, it is essential to distinguish between **verification** and **validation**. Verification addresses the question: Are we solving the equations or implementing the algorithms correctly? It ensures that the model faithfully follows the intended design and produces consistent, expected outputs based on specified inputs and parameters. Conversely, validation focuses on whether the model accurately represents the real-world system it aims to simulate, typically through comparison with experimental data or established benchmarks. This section focuses exclusively on verification, evaluating the internal consistency and correctness of the code, independent of its real-world applicability.

To this end, a key milestone is confirming that the implemented code behaves as intended. In our case, the algorithms were independently implemented in two different programming languages–*JavaScript* and *Python*–and executed on distinct platforms. This dual implementation provides a natural means of cross-verification: both versions of the model should produce identical results under the same conditions. This is considered to be a valid method of verification as discussed by Musuamba *et al.* [Musuamba et al. 2021].

Accordingly, we conducted a thorough comparison of the outputs, excluding stochastic elements inherent to the model. Given identical inputs and parameter settings, the *JavaScript* and *Python* implementations yielded matching results. Key metrics, such as relative protein abundances or proportion of free ribosomes, or number of ribosomes in polysomes, or average distance between ribosomes, were consistent across both versions. While this does not entirely preclude the possibility that similar coding errors were introduced in both implementations (as they were performed by the same person), it is worth emphasizing that the two versions were developed independently, on different platforms, and at separate stages during the course of this thesis, thereby mitigating this risk.

10.3.2 Time-step adequacy and real-time compatibility in simulations

Another critical verification step involves ensuring numerical consistency across time scales, often referred to as **time-step adequacy** or **real-time compatibility** in numerical simulations. Specifically, this entails verifying that the simulation's iteration cycle (i.e., the program loop time) is appropriately matched to—or shorter than—the smallest time scale of the modeled processes, thereby maintaining numerical stability and accuracy.

In the context of real-time systems, this concept aligns with **real-time scheduling** or real-time performance validation, where the goal is to ensure that computations are completed within a time frame compatible with the dynamics of the simulated system.

In our model, each program loop iterates over hundreds or thousands of agents, such as ribosomes, transcripts, and codons. For the simulation to accurately capture the system's behavior, the total time required to complete one loop must be smaller than the characteristic timescale of the underlying biological processes. For example, ribosomes in the model are assigned timers to simulate their dwell time on codons, with durations ranging from 10 to 200 milliseconds. To preserve temporal fidelity, the program loop time should ideally be at least one order of magnitude smaller than these dwell times.

We verified that this condition is satisfied: the loop execution time was measured between 0.6 milliseconds and 6 milliseconds, approximately one order of magnitude smaller than the biological process timescale being modeled. While the use of high-level, interpreted languages such as *Python* or *JavaScript*[3] could raise concerns regarding computational speed, modern hardware capabilities and CPU frequencies are sufficient to ensure the necessary performance. Nonetheless, shorter loop times could be achieved using compiled languages like C or C++ if required.

[3] In *JavaScript* and *p5.js*, the `HTML` canvas used to visualize the translation process in real time is limited to a maximum refresh rate of 60 frames per second. This corresponds to each visual snapshot having a duration of approximately 17 milliseconds, which is slightly shorter than most ribosome dwell times on individual codons.

10.3.2.1 Artificial time rescaling to mitigate loop time constraints

In scenarios where the program loop time approaches or exceeds the timescale of the modeled processes, one viable strategy is to implement **artificial time rescaling**[4]. This involves proportionally stretching the timescale of all modeled events—effectively running the system in '*slow motion*'–to ensure that the iteration cycle remains shorter than the simulated dynamics.

Such temporal rescaling allows the computational loop to process all agents without violating the causality or sequence of events. While absolute real-time fidelity is sacrificed, relative timings, steady-state behavior, and qualitative trends remain valid. This approach is well-established in the literature, particularly in agent-based simulations and molecular systems where computational speed is a limiting factor.

By artificially stretching the biological or physical time (slowing it down), one ensures that the simulation loop can keep up without sacrificing accuracy or stability. This approach, however, comes with the caveat that absolute timing results should be interpreted carefully, as the simulation no longer operates in '*real time*', but relative dynamics and steady-state properties can still be faithfully captured. Care must be taken, however, to clearly communicate that the model operates on a rescaled temporal axis and that direct comparisons to absolute biological timescales should be made cautiously.

In our `Ribosomer` implementations, we did not have to resort to time rescaling.

10.4 Summary of main findings and insights

This chapter summarizes the key outputs generated by the ABM, including relative protein abundances (both within and between samples), translation efficiencies, polysome profiles, ribosome density maps, inter-ribosome distances, free ribosome proportions, and transcript loading status. These outputs are designed to reflect experimental conditions relevant to molecular biology. The chapter demonstrates that `Ribosomer` is highly flexible, allowing for customized outputs tailored to researchers' and end-users' needs.

[4] This is also known as *slow-motion scaling*, or *temporal rescaling*.

Chapter 11

Sensitivity and uncertainty analysis

Every time a scientific paper presents a bit of data, it's accompanied by an error bar - a quiet but insistent reminder that no knowledge is complete or perfect. It's a calibration of how much we trust what we think we know.
Carl Sagan

The inherent probabilistic nature means that we can never precisely predict stochastic outcomes. This might be why we opt to use the more pretentious-sounding word "stochastic" instead of just saying "random". At least in common parlance, "random" has connotations of hopelessness. To try to understand something random feels futile; it's just random. But of course, we can understand many things about stochastic-random processes, despite this inherent unpredictability, by studying the probability distributions.
Karen Abbott

If you can quantitatively predict the uncertainty in an experiment's outcome before conducting it, you possess a probabilistic model—a known probability density function from which to sample—that identifies the sources of stochasticity. In this sense, you have a causal or mechanistic explanation for the inherent randomness of the phenomenon under study. You have achieved the maximal deterministic reduction, effectively minimizing the unpredictable, noise-driven component.
Anonymous

In statistical inference, the null hypothesis significant testing is cursed by multiple testing. The more you are doing comparisons between outcomes, the more false positives.
Carlo Emilio Bonferroni

This chapter explores the parameter space on model predictions in translation dynamics. It begins by defining uncertainty in the context of stochastic processes and statistical inference, emphasizing the importance of quantifying uncertainty to improve model reliability. The analysis distinguishes between sensitivity (how much an outcome variable changes in response to parameter variation) and uncertainty (the inherent variability in simulation outcomes due to stochastic effects).

A threshold for effect size is established using a three-standard-deviation criterion to compare paired simulations, minimizing false positives in hypothesis testing. Sensitivity analysis is conducted via factorial experimental designs and ANOVA-based statistical methods. Key parameters examined include transcript copy number, ribosome pool size, initiation rates, codon usage bias, tRNA modifications and specific initiation rates of privileged subsets of transcripts.

Eight sensitivity observations are derived from computational simulations. They highlight critical factors affecting translation efficiency. The most significant findings emphasize that (1) ribosome pool size and initiation rates are the primary determinants of protein synthesis efficiency, (2) codon usage bias can be overridden by competition for ribosomes, and (3) transcript length has minimal influence on ribosome occupancy patterns. The chapter underscores that translation is robust to local elongation rate fluctuations but highly sensitive to resource allocation at initiation.

11.1 Threshold for effect size in the comparison of paired simulations

In the following analysis, we consider two outcomes in paired experiments (wild type versus a factor or treatment) to be different if their difference exceeds three standard deviations. All other parameters remain unchanged in the paired computational simulations. Standard deviations were quantified by performing independent replicates of a computational simulation experiment, ensuring that all parameters remained constant across replicates. The number of replicates was set to 30 to produce all output files in the fully-crossed factorial experimental design conducted computationally on the HPC cluster. When relevant for statistical power, all 30 replicates could be used. However, some of the subsequent statistical analysis that were conducted in various sensitivity analysis did not use all replicates available but were restricted to smaller subsets of replicates, e.g., 5 replicates already provided sufficient sample size to achieve statistical power at the requested level of type I error (false positive risk fixed below 5% or 1% as documented below). Using 30 replicates was initially meant to assess the variance in the protein abundance output variables for individual gene ID tags. As will be seen below, 5 replicates in the analysis of variance provide enough statistical power

for null hypothesis testing of fixed effects or for null hypothesis testing of interactions between factors.

11.1.1 Design of computational experiments

In chapter 9, we defined the ranges for two key parameters that are currently not directly measured in the lab:

1. The **ribosome ratio** –the number of ribosomes in the pool divided by the total number of transcript copies.
2. The **general initiation rate**.

For each parameter, we selected eight representative values within the defined intervals. To thoroughly explore the parameter space, we combined these values exhaustively, forming an 8×8 **grid**. Each cell in this grid corresponds to a unique pair of parameter values, and each is replicated multiple times to account for the stochastic variability inherent in the agent-based model (ABM).

This approach follows the **traditional experimental design known as a fully crossed factorial design with balanced randomized replicates** [Dagnelie 1981, Montgomery 2013a,b]. In our example, five replicates per parameter pair result in a total of **320 computational experiments**.

Given that each experiment requires approximately 1 hour and 48 minutes of computing time, conducting all experiments sequentially would be impractical. Thus, high-performance computing (HPC) resources are essential. In Appendix E, we provide a detailed BASH script designed to distribute these jobs efficiently across the nodes and CPUs of the `CéCI` HPC cluster.

The BASH script example in Appendix E is easily adaptable to other scenarios. For instance, the program can be configured to toggle various biological factors, such as the presence or absence of U34 tRNA modification, ADAT A34 tRNA modification, proline-induced translation slowdown, mRNA secondary structure, or ribosomal exit tunnel electrostatic interactions–either individually or in combination. Additionally, alternative input files containing different sets of individual initiation rates per transcript can be used. The potential parameter space is vast, offering **flexibility for diverse experimental setups**.

11.2 Sensitivity and uncertainty analysis

The **sensitivity analysis** usually is a systematic study (alternatively a Monte Carlo study) of the **effect of a change in the parameters space on the output variables** (predicted outcome values) of the model. A parameter p is said to be more sensitive than a parameter q on a predicted outcome variable y, if the same relative change in the parameter's value results in a larger relative change in the outcome variable y. Mathematically if $|\frac{\delta y(p,q)}{y(p,q)\frac{\delta p}{p}}| > |\frac{\delta y(p,q)}{y(p,q)\frac{\delta q}{q}}|$, then y is more sensitive on p than on q. As we are dealing with first order partial derivatives estimations and as y might not be a monotonous function of p and q, the sensitivity analysis is usually conducted around a reference set of values for p and q. The sensitivity analysis itself is dependent on the parameters interval boundaries where it is conducted. Under these circumstances, the **sensitivity analysis is said to be local** (not global).

Statistical methods can also provide more comprehensive sensitivity analyses. In multiple linear regression or **analysis of variance (ANOVA)**, the sensitivity analysis evaluates the **contribution of each factor (independent variable) to the variability of the dependent variable**, aiming to identify which factors most significantly influence the outcome. Different approaches to sensitivity analysis align with traditional experimental designs, such as **fully crossed factorial designs with balanced randomized replicates**, or designs with restricted randomization, like the **Latin square designs** (an extension of the paired Student's t-test for more than two objects) [Montgomery 2013b; Montgomery 2013a; Dagnelie 1981; Dagnelie 1986; Schepdael et al. 2014].

Uncertainty is understood in the sense described by Carl Sagan in his citation at the beginning of the chapter. We use the terms **uncertainty**, **variance**, and its square root, i.e., the **standard deviation**, interchangeably as measures of the **dispersion** of any outcome value obtained by replicating an experiment involving stochastic processes, where the final outcome results from the accumulation of a large number of random events.

We briefly recall the various definitions of heterogeneity or dispersion parameters commonly used in Probability and Statistics, which underpin the null hypothesis significance testing (NHST) method—also known as the frequentist approach—developed by Fisher, Neyman, and Pearson, the renowned founders of statistical inference. Given the available data (sample size n), the primary descriptive statistical parameters of interest

are the sample mean and the sample variance:

$$\overline{x} = \frac{1}{n}\sum_{i=0}^{n} x_i \tag{11.1}$$

$$s^2 = \frac{1}{n}\sum_{i}^{n}(x_i - \overline{x})^2. \tag{11.2}$$

From the sample variance, the sample standard deviation is obtained by taking the square root of the sample variance:

$$s = \sqrt{s^2} \tag{11.3}$$

$$s = \sqrt{\frac{\sum_i^n (x_i - \overline{x})^2}{n}}. \tag{11.4}$$

In statistical inference, we are interested in the population from which the sample was taken, assuming the sample was drawn randomly from a normal (or asymptotically normal) distribution. The sample mean $\overline{x}$ is an unbiased estimate of the population mean μ and we have $\mu = \overline{x}$.

The sample variance s^2, however, is a biased estimate of the population variance σ^2. The population variance unbiased estimate is obtained by the Bessel correction. The population standard deviation follows.

$$\sigma^2 = \frac{n}{n-1}s^2 \tag{11.5}$$

$$= \frac{1}{n-1}\sum_{i}^{n}(x_i - \overline{x})^2. \tag{11.6}$$

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \overline{x})^2}{n-1}}. \tag{11.7}$$

$$\text{s.d.} = \sigma. \tag{11.8}$$

Note that the population standard deviation σ is always a bit larger than the sample standard deviation s.

When a large number of independent random contributions influence the observed final outcome, the **central limit theorem of Laplace-Lyapunov** states that the **sampling distribution of the mean** (e.g., from 30 replicates of a simulation or experimental trial) **asymptotically follows a normal (Gaussian) distribution**. The standard deviation

of the sampling distribution of the mean is called the **standard error (of the mean)**, `s.e.m.`, and is not the same as the standard deviation, `s.d`. The standard error of the mean $\sigma_{\overline{X}}$ decreases when the sample size n increases. It can easily be shown that the standard error of the mean is the ratio of the population standard deviation divided by the square root of the sample size:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{11.9}$$

$$\text{s.e.m.} = \frac{\text{s.d.}}{\sqrt{n}}. \tag{11.10}$$

In the comparison of individual outcome variables in paired experiments of interest, **we consider effect sizes to be highly significant only if the differences exceed three standard deviations. More precisely, we use a threshold of** 3.10 **standard deviations** $(3.10\,\sigma)$ **to account for Bonferroni multiple testing adjustments**, as explained in the next subsection. Since we do not compare means of replicated outcome variables, **we use the larger standard deviation rather than the standard error**. Specifically, the population standard deviation is used instead of the sample standard deviation.

In the computational simulation runs, sensitivity and uncertainty analyses are conducted in parallel. This approach enables us to assess whether **uncertainties can obscure (or eclipse) the effect size of a parameter of interest** on an observed outcome.

11.2.1 The curse of multiple testing

We generally focus on a single comparison of a single pair of outcomes. Although, implicitly, multiple comparisons have truly been conducted. **Testing multiple hypotheses can result in an inflation of the type I error rate (false positive or rejecting the null given that it is true)**. An adjustment for multiple testing should be applied to reduce the false positive risk. Usually, in single null hypothesis significant testing, the false positive risk is set to $\alpha = 5\%$, corresponding to differences larger than two standard deviations $(1.96\,\sigma)$ for an effect size to be detected as significant. Our naive multiple testing adjustment would be to use three standard deviations to keep the false positive risk as reasonably low as possible. How reasonable is that? Recall that, in practice, if α is the type I error for one test, if m independent tests are conducted, the probability that no error is made is $(1-\alpha)^m$. Hence, the probability that at least one hypothesis will be rejected wrongly is $1-(1-\alpha)^m$. In our settings for the simulation runs in the sensitivity and uncertainty analysis, we conduct 27 tests (one for each transcript ID tag) when we compare a wild type simulation with a treatment factor simulation and are interested in the outcome of the 27 relative protein abundance. Then, the probability to reject wrongly at least one test out of the 27 is 0.75 or 0.24, for $\alpha = 0.05$ or 0.01 respectively. We see that, when the number of tests (supposed

here to be mutually independent) is 27, even with $\alpha = 0.01$, or an effect size difference threshold set to three standard deviations, it is still very likely that we will reject wrongly a number of tests and will declare a number (~ 6 out of the 27) of false positives.

In the **Bonferroni multiple testing adjustment**, α is lowered so as to minimize $1 - (1 - \alpha)^m$. For instance, with 27 bilateral tests, α must be set to 0.0019 so that the **multiple type I error** be less than 0.05 (still approximately one false positive to be expected upon conducting 27 paired comparisons), and α must be set to 0.0003722 so that the multiple type I error be less than 0.01 (rather unlikely to have a false positive upon conducting 27 paired comparisons). Setting $\alpha = 0.0019$ requires a difference larger than 3.10 standard deviations. Setting $\alpha = 0.0003722$ would require a difference larger than 3.56 standard deviations.

We should also recall that **minimizing type I error increases type II error** (β). The sensitivity or power ($= 1 - \beta$) to reject correctly a null that is really not true decreases if type I error α is set to lower values. **So, being too conservative in type I error will reduce the power to detect an effect at all, given that there is truly an effect**.

We cannot exclude that adopting a size effect that we consider significant, at 3.10 standard deviation ($3.10\,\sigma$), could actually lead to a false positive. However, the criterion we adopted for the joint sensitivity and uncertainty analysis is to use a threshold of 3.10 standard deviations ($3.10\,\sigma$) to detect an effect size. This choice limits the type I error associated with 27 multiple tests to 0.05.

11.2.2 List of model parameters

All the parameters that can have an impact on any of the predicted output values of the model of protein synthesis by ribosomes are recalled from the previous chapter and enumerated below:

- relative abundance of transcripts (distribution of the number of transcript copies across the transcriptome), i.e. RNA-seq profiles
- global initiation rate (rate at which any ribosome starts translation on any transcript)
- individual initiation rates of transcripts (transcript specific preferred initiation)
- ribosome pool size (ratio of ribosomes total number over transcript copies number)
- different codon usage or significant bias in the synonymous codons usage of the nucleotide sequence across the transcriptome

- tRNAs relative abundance or tRNAs pool sizes
- amino acid specific sequences encrypted in the transcriptome (proline enrichment or charged amino acid enrichment or charged amino acid distributions)
- electrostatic interaction in the ribosome exit tunnel (model factor)
- proline slow incorporation (model factor)
- mRNA secondary structure (model factor)
- U34 enzymatic tRNA modifications on the sensitive codons of K, Q, E (optional experimental factor)
- ADAT enzymatic modifications on the sensitive codons of T, A, P, S, L, I, V, R (optional experimental factor)

Some of these parameters are defined as **intrinsic model parameters**, while others depend on the **input data**. Data-dependent parameters include, for example, the **specific codon sequence of a given transcript** provided in the input data file (`FASTA` format), the **number of copies of a given transcript ID tag** (as determined from RNA-Seq data), or the **fold change in initiation rate of a given transcript** relative to the global initiation rate. In contrast, intrinsic model parameters include the **ribosome pool size** (ribosome ratio) and the **global initiation rate**, which is equal to the inverse of the mean time to an initiation event by a free ribosome.

11.3 Results overview of the sensitive parameters on the model predicted outcomes

We enumerate below eight general observations derived from the sensitivity and uncertainty analyses of our agent-based model. These observations provide insights into translational control factors and their functional roles in protein synthesis. Some parameters exert a direct and explicit impact, such as the ribosome pool size, while others influence translation indirectly, like codon usage in a transcript's nucleotide sequence. The effects of these indirect factors may be overridden or amplified, depending on the values assigned to other parameters. Some outcomes might be sensitive to certain parameters, while other outcomes may not, highlighting the complexity of translational regulation.

11.3.1 Transcript copy number

Sensitivity observation 1. The protein relative abundance is sensitive to the relative transcript copy number provided in the input data along with the transcripts sequences **conditionally on the ribosome pool size and initiation rates**.

This observation underlies the use of transcriptomics to profile gene expression from RNA-Seq data. Our agent-based model (ABM) replicates this feature but also demonstrates that it does not provide a complete picture. The influence of relative transcript copy numbers on protein abundance depends on both the ribosome pool size and initiation rates. Only when the ribosome ratio and global initiation rates are sufficiently high does the relative transcript copy number exert a significant impact on protein abundance.

Figure 11.1 supports this sensitivity analysis result. The relative abundance of proteins belonging to the kinesin family—`KIF4A`, `KIF14`, and `KIF23` (blue-filled ellipses), corresponds closely to the relative abundance of their respective transcripts in the input data file, as shown in Fig. 11.1 (a) and (b). This is generally true for all gene ID tags in the figure. Seemingly contradicting this result, the relative abundance of the ribosomal protein `RPL4` (red-filled ellipses) is significantly higher than that of, for example, collagen `COL1A` (dashed arrows), despite their transcript ratios suggesting the opposite relationship. However, this is only an apparent contradiction, as the `RPL4` **specific initiation rate was set to be twice as high as that of all other transcripts in the input data file**, for which the default fold change of 1 was applied. Note, in Fig. 11.1 (c), that the translation efficiency captures, for `RPL4`, the effect of the individual initiation rate fold change. This apparent contradiction can indeed be restored back to the general sensitivity observation 1 by re-running simulations with the specific initiation rate fold change of `RPL4` reset to 1, i.e., the same default fold change as for all other gene ID tags (result not shown here). The origin of the variability in the translation efficiency (TE) within a given sample but across the gene ID tags will be explored below.

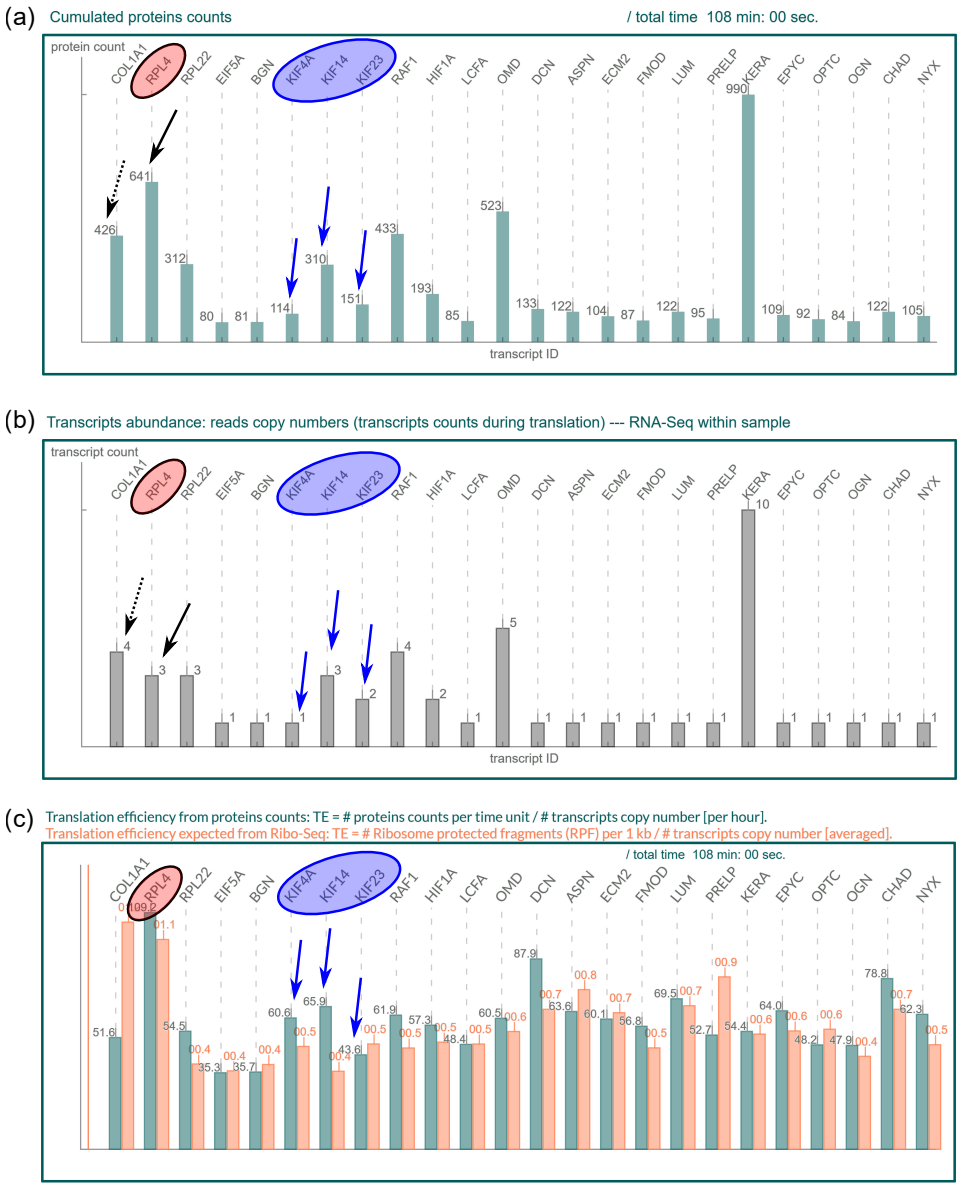


Figure 11.1: The relative protein abundance (a) is sensitive to the relative transcript copy number (b) provided in the input data, conditional on the individual initiation rates. The blue arrows indicate a direct sensitivity of relative protein abundance to transcript copy numbers. However, the black solid and dashed arrows demonstrate that this effect can be overridden by a two-fold increase in the relative initiation rate of `RPL4` (arbitrarily specified in the input data but not shown in the figure). Relative translation efficiencies (c) are also considered. Further explanations are provided in the main text.

11.3.2 Ribosome pool size and general initiation rate

These two parameters are studied jointly as they are key contributors to translation efficiency. All outcomes variables of the agent-based model are highly sensitive to these parameters. A very first naive approach to assess the efficiency in protein translation is to check the balance between supply and demand. Here, the supply is the number of ribosomes that can be provided to engage in the translation of transcripts. The demand is the number of transcripts that need to load ribosomes to be translated. Equivalently, the number of free ribosomes (not engaged in translation) should be a small proportion of the total number of ribosomes (ribosome pool size). A large number of free ribosomes reflects a poor usage of costly resources for the cells. Similarly, the number of transcripts without a single ribosome (unloaded transcripts) should remain a small proportion of the total number of mature mRNAs in the cell (transcriptome size). The mature mRNAs also were costly to produce and are exposed to premature decay if they are not translated by ribosomes. **How sensitive is the efficiency of translation on the ribosome pool and on the general initiation rates?**

How many ribosomes in the ribosome pool are without (or free of) transcripts? How many transcripts in the transcriptome are without (or free of) ribosomes?

Input settings of simulations runs: the same 27 gene ID tags with their CDS sequences are used with the set of transcript copy numbers graphically tabulated in Fig.11.1 (b). All individual initiation rates of the 27 gene ID are defaulted to a fold change of 1, except `RPL4` which was set to 2 in the input data file.

A grid of sixty-four simulation cells (with 5 replicates per cell), with the same input settings as recalled above, in a **fully crossed 2D-factorial design** were conducted to address the question of the proportion of free ribosomes and the proportion of transcripts not engaged in translation. The first factor is the ratio of ribosomes to the number of transcripts copies (factor A) and the second factor is the general initiation rate (factor B). The two factors were split each into the eight levels respectively indicated in chapter 9 and enumerated in the rows and columns in Fig. 11.2. The two heatmaps in Fig. 11.2 shows the results of the sensitivity analysis of these two joined factors on the two outcomes of interest. The structure of each heatmap allows to conduct a more global sensitivity analysis at the scale of the parameters sample range of this 2D-grid. A close look on the heatmaps shows that the two factors interact: the effect of the initiation rate depends on the level of the ribosome tool and vice-versa. We conducted a statistical test to prove that the two factors indeed interact and to quantitatively assess the strength of the interaction. Five computational runs were replicated in each cell of the fully crossed factorial design above ($64 \times 5 = 320$ computational runs of two hours each were submitted on the `CECI` high computing performance (HPC) cluster (in the so called embarrassingly parallel mode). The **fixed effects statistical model (2 way-ANOVA)** or

multiple linear regression [1] **incorporating a first order interaction term** is

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}, \tag{11.11}$$

where $i \in [1, .., 8]$ refers to the $a = 8$ levels of the factor *ribosome pool ratio* (8 rows, factor A); $j \in [1, ..., 8]$ refers to the $b = 8$ levels of the factor *initiation rates* (8 columns, factor B) and $k \in [1, ..., 5]$ refers to the $n = 5$ random replicates in each object cell. y_{ijk} is the observed response (dependent variable) of interest. In what follows, it represents the proportion of free ribosomes and the proportion of free transcripts, respectively. μ is the overall mean effect, τ_i is the effect of the i^{th} level of the row factor, β_j is the effect of the j^{th} level of column factor, $(\tau\beta)_{ij}$ is the effect of the interaction between τ_i and β_j, and ϵ_{ijk} is the residual random error component (supposed to be $\sim N(0, \sigma^2)$ distributed).

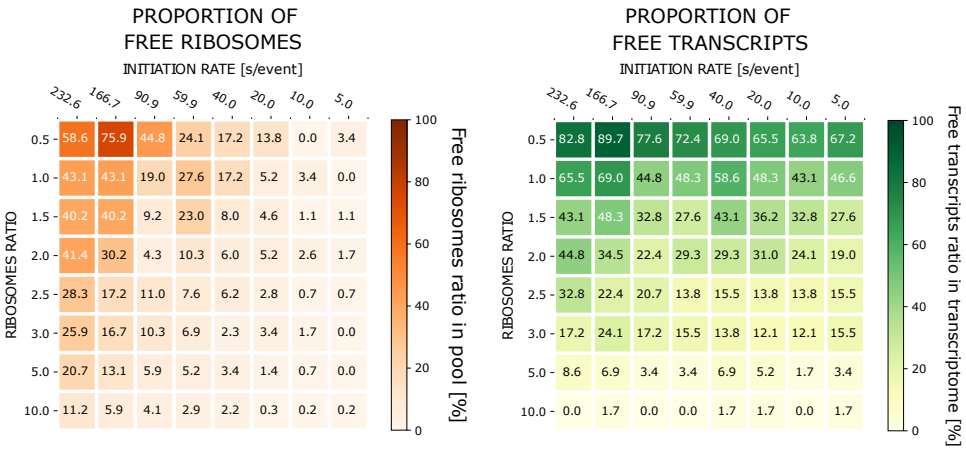


Figure 11.2: Fully crossed factorial design of 64 computational simulations (with 5 replicates in each cell) of the effects of ribosome ratio or ribosome pool (rows) and initiation rate (columns) on translation resources supply (free ribosomes) and demand (free transcripts). Left panel: effect on the proportion of free ribosomes (unused supply) in the pool size. Right panel: effect on the proportion of free transcripts (unmet demand).

[1] When the two factors have quantitative values (numerical values), 2 way-ANOVA and multiple linear regression are completely equivalent and represent the same statistical model.

Specifically we are interested in **testing three hypotheses**:

- testing hypothesis about the equality of row treatment effects (ribosome pool effects),

$$H_0 : \tau_1 = \tau_2 = ... = \tau_8 = 0 \tag{11.12}$$

$$H_1 : \exists i \mid \tau_i \neq 0 \tag{11.13}$$

- testing hypothesis about the equality of column treatment effects (initiation rate effects),

$$H_0 : \beta_1 = \beta_2 = ... = \beta_8 = 0 \tag{11.14}$$

$$H_1 : \exists j \mid \beta_j \neq 0 \tag{11.15}$$

- testing hypothesis whether row and column treatments interact (interaction between the ribosome pool size and the initiation rate),

$$H_0 : (\tau\beta)_{ij} = 0 \forall\, i, j \tag{11.16}$$

$$H_1 : \exists i, j \mid (\tau\beta)_{ij} \neq 0 \tag{11.17}$$

These hypotheses are tested using a 2-way analysis of variance (fixed effects model) with replicates in each combination of the 2 factors (fully-crossed factorial design of factors with replicates: 8 by 8 with 5 replicates, $a = 8$, $b = 8$, $n = 5$). The ANOVA calculation used the Python `statsmodels` library and the `ols` method (ordinary least squares) applied to the linear model equation (11.11). The python script and the complete computer output of the 2-way ANOVA table are given in appendix F both for the outcome response on the proportion of free ribosomes and on the proportion of free transcripts. The appendix also includes the **model adequacy checking** (residual normality analysis, Q-Q plots) for both outcomes. We calculate the percentage of variance explained by each factor to understand its contribution to the overall variability in the response. Finally, we fit a **three-dimensional response surface** with a polynomial of degree 2 and draw the **two-dimensional contour plots** of the outcome response as a function of the two factors, the ribosome pool size (factor A) and the general initiation rate (factor B). The three-dimensional response surface and two-dimensional contour plots are used to determine the domain of optimality where translation efficiency is high, and the domain where translation is impaired due to sub-optimal ribosome resources or initiation rates. We discuss the **biological significance** of these sensitivity results.

Outcome response 1: proportion of free ribosomes in the pool.

A summary of the analysis of variance is given in table 11.3.1.

Table 11.3.1: 2-way ANOVA table and partial sum of squares for proportion of free ribosomes in the ribosome pool.

Source		Sum of Squares	Degrees of Freedom		Mean Square		F Value	P-Value
Model	$SS_T - SS_E$	79,511.48	63		1,262.09		77.61	1.37e-135 (⋆⋆⋆)
Factor A ribosome pool ratio	SS_A	18,309.65	7	MS_A	2,615.66		160.85	6.51e-90 (⋆⋆⋆)
Factor B initiation rate	SS_B	47,716.97	7	MS_B	6,816.71		419.18	2.73e-136 (⋆⋆⋆)
(AB) interaction	SS_{AB}	13,484.86	49	MS_{AB}	275.20		16.92	2.12e-56 (⋆⋆⋆)
Residual	SS_E	4,163.07	256	MS_E	16.26	$\sigma = 4.03$		
Total	SS_T	83,674.55	319					

Effect Sizes (Partial Eta-squared)$[\eta]$:

Source	Sum of Squares	Degrees of Freedom			Eta-Squared, η
Factor A ribosome pool ratio	18,309.65	7		SS_A/SS_T	0.219
Factor B initiation rate	47,716.97	7		SS_B/SS_T	0.570
Factor AB interaction	13,484.86	49		SS_{AB}/SS_T	0.161
Residual	4,163.07	256		SS_E/SS_T	0.050
Model	79,511.48	63	R-squared	$(SS_T - SS_E)/SS_T$	0.95

Methodological note: In the fixed effects model, the Snedecor F statistics are obtained by dividing each of the mean squares by the residual mean square MS_E. If we assume that the model (eq. 11.11) is adequate and that the error term ϵ_{ijk} (the residuals) are normally and independently distributed with constant variance σ^2, then each of the ratios of mean squares MS_A/MS_E, MS_B/MS_E, and MS_{AB}/MS_E is distributed as a Snedecor F with $a-1$, $b-1$, and $(a-1)(b-1)$ numerator degrees of freedom respectively and $ab(n-1)$ denominator degrees of freedom. The critical region is the upper tail of the F distribution. The larger the F statistic, the more likely there is a significant effect, and the less likely the null hypothesis holds true.

The three null hypotheses are all rejected at a very highly significance level.

- **Main effect of ribosome pool ratio**
  Snedecor F statistics $F_{\text{obs.}} = \frac{MS_A}{MS_E} = 160.85$; p-value = $6.51\,10^{-90}$.
  Interpretation and conclusion on null hypothesis: the p-value < 0.001 and the null hypothesis of no effect of factor A is rejected (very highly significant).

- **Main effect of initiation rate**
  Snedecor F statistics $F_{\text{obs.}} = \frac{MS_B}{MS_E} = 419.18$; p-value = $2.73\,10^{-136}$.
  Interpretation and conclusion on null hypothesis: the p-value < 0.001 and the null hypothesis of no effect of factor B is rejected (very highly significant).

- **Interaction of factor A with factor B**
  Snedecor F statistics $F_{\text{obs.}} = \frac{MS_{AB}}{MS_E} = 16.92$; p-value = $2.12\,10^{-56}$.
  Interpretation and conclusion on null hypothesis: the p-value < 0.001 and the null hypothesis of no interaction is rejected (very highly significant).

The local sensitivity of each of the two factors and their interaction is quantified in the coefficient estimates τ_i, β_j and $(\tau\beta)_{ij}$ of the ANOVA table given in appendix F.

The visualization of the sensitivity of each factor on the response can be seen on the Fig. 11.3. The coefficient estimates and the main effect plots show that the ribosome pool ratio (factor A) has the smaller effect on the proportion of free ribosomes and on translation. Factor A (ribosome pool ratio) alone only explains 21.9% of the variability in the proportion of free ribosomes, whereas factor B (initiation rate) explains 57.0% and the interaction of the two factors 16.1% of this variability. Altogether, these factors explain 95.0% of the variability of the proportion of free ribosomes in the computationally simulated translation occurring in cells after 2 hours (end time point in steady-state condition). Note also how the standard deviation is much higher when the levels in either factor A or B decrease. The variability (or uncertainty) in the proportion of free ribosome, as a model output, increases when the ribosome pool ratio decreases, or when the initiation rate decreases as shown in Fig. 11.3.

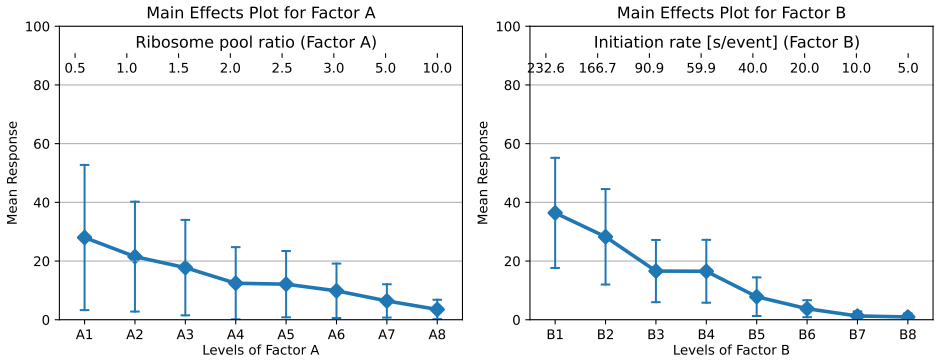


Figure 11.3: Left: Main effect of the ribosome pool ratio on the **proportion of free ribosomes**. Right: Main effect of the initiation rate on the **proportion of free ribosomes**. Error bars: standard deviation ($\pm\sigma$). The ribosome pool ratio has a smaller impact on the proportion of free ribosomes compared to the initiation rate. Notably, for the proportion of free ribosomes to remain below 15% at steady-state, the ribosome pool ratio must exceed 2.0, and the general initiation rate must be faster than $\lambda = 25\,10^{-6}\,\mathrm{ms}^{-1}$, meaning the average time for a free ribosome to initiate translation should be less than 40 seconds.

OUTCOME RESPONSE 2: PROPORTION OF FREE TRANSCRIPTS IN THE TRANSCRIPTOME.

The summary of the analysis of variance is given in table 11.3.2.

Table 11.3.2: 2-way ANOVA table and partial sum of squares for proportion of free transcripts in the transcriptome.

Source		**Sum of Squares**	**Degrees of Freedom**		**Mean Square**		**F Value**	**P-Value**
Model	$SS_T - SS_E$	179,404.42	63		2,847.69		225.1	1.69e-62 (⋆⋆⋆)
Factor A ribosome pool ratio	SS_A	169,981.04	7	MS_A	24,283.01		1919.25	3.41e-217 (⋆⋆⋆)
Factor B initiation rate	SS_B	5,973.01	7	MS_B	853.29		67.44	1.55e-54 (⋆⋆⋆)
(AB) interaction	SS_{AB}	3,450.37	49	MS_{AB}	70.42		5.57	2.22e-20 (⋆⋆⋆)
Residual	SS_E	3,239.0	256	MS_E	12.65	$\sigma = 3.56$		
Total	SS_T	182,643.42	319					

Effect Sizes (Partial Eta-squared)$[\eta]$:

Source	**Sum of Squares**	**Degrees of Freedom**			**Eta-Squared**, η
Factor A ribosome pool ratio	169,981.04	7		SS_A/SS_T	0.931
Factor B initiation rate	5,973.01	7		SS_B/SS_T	0.033
Factor AB interaction	3,450.37	49		SS_{AB}/SS_T	0.019
Residual	3,239.0	256		SS_E/SS_T	0.018
Model	179,404.42	63	R-squared	$(SS_T - SS_E)/SS_T$	0.982

The three null hypotheses are all rejected at a very highly significance level.

- **Main effect of ribosome pool ratio**
  Snedecor F statistics $F_{\text{obs.}} = \frac{MS_A}{MS_E} = 1919.25$; p-value = $3.41\,10^{-217}$.
  Interpretation and conclusion on null hypothesis: the p-value < 0.001 and the null hypothesis of no effect of factor A is rejected (very highly significant).

- **Main effect of initiation rate**
  Snedecor F statistics $F_{\text{obs.}} = \frac{MS_B}{MS_E} = 67.44$; p-value = $1.5\,10^{-54}$.
  Interpretation and conclusion on null hypothesis: the p-value < 0.001 and the null hypothesis of no effect of factor B is rejected (very highly significant).

- **Interaction of factor A with factor B**
  Snedecor F statistics $F_{\text{obs.}} = \frac{MS_{AB}}{MS_E} = 5.57$; p-value = $2.22\,10^{-20}$.
  Interpretation and conclusion on null hypothesis: the p-value < 0.001 and the null hypothesis of no interaction is rejected (very highly significant).

The local sensitivity of each of the two factors and their interaction is quantified in the coefficient estimates τ_i, β_j and $(\tau\beta)_{ij}$ of the ANOVA table given in appendix F. The visualization of the sensitivity of each factor on the response is presented on Fig. 11.4. The coefficient estimates and the main effect plots show that the ribosome pool ratio (factor A) has the largest effect on the proportion of free transcripts and on translation. Factor A (ribosome pool ratio) alone explains 93.1% of the variability in

the proportion of free transcripts, whereas factor B (initiation rate) only marginally explains 3.3% and the interaction of the two factors 1.9% of this variability. Altogether, these factors explains 98.2% of the variability of the proportion of free transcripts in the computationally simulated translation occurring in cells after 2 hours (end time point in steady-state condition). Note that, the larger variances observed for the proportion of free transcripts, across all levels of factor B (initiation rate), in the right panel in Fig. 11.4, confirm these conclusions. It is indeed factor A (ribosome pool ratio) that explains most of the variability in the outcome variable and not factor B (initiation rate).

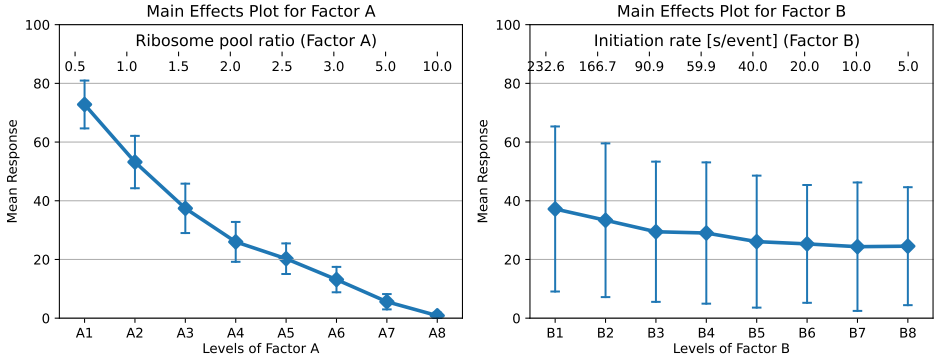


Figure 11.4: Left: Main effect of the ribosome pool ratio on the **proportion of free transcripts**. Right: Main effect of the initiation rate on the **proportion of free transcripts**. The ribosome pool ratio has a greater impact on the proportion of free transcripts than the initiation rate. Notably, once the initiation rate exceeds a certain threshold, further increases have no additional effect on the response. Error bars are standard deviation ($\pm\sigma$).

Fitting Response surfaces As the two factors are quantitative, we can fit a response surface for predicting the outcome variable at various combinations of the design factors. For both outcomes, building further from the previous ANOVA results, a second order polynomial of the two factors was fitted to the data. The quadratic regression model with 6 coefficients was fitted to each of the outcome data, complying to the hierarchy principle [Montgomery 2013a]

$$z = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy. \tag{11.18}$$

The second-order polynomial is used to construct the response surface and the corresponding contour plots. The fitted polynomials and the plots were generated with the python code in appendix F. For both outcome variables, we highlighted two contours where the response represents 15% (in black) and 20% (in red), indicating the threshold proportions of free ribosomes or free transcripts, respectively, separating efficient translation from non-optimal translation, Fig. 11.5. According to previously published

literature, the proportion of free ribosomes in steady-state conditions for healthy yeasts in exponential growth is 15% [Shah et al. 2013]. The significant interactions identified earlier are reflected in the curvature of the response surfaces. These surfaces are commonly utilized for optimization in engineering applications. In our context, we can infer how cells may regulate translation by modulating ribosome biogenesis, ribophagy, or the repression and stimulation of initiation factors that drive initiation.

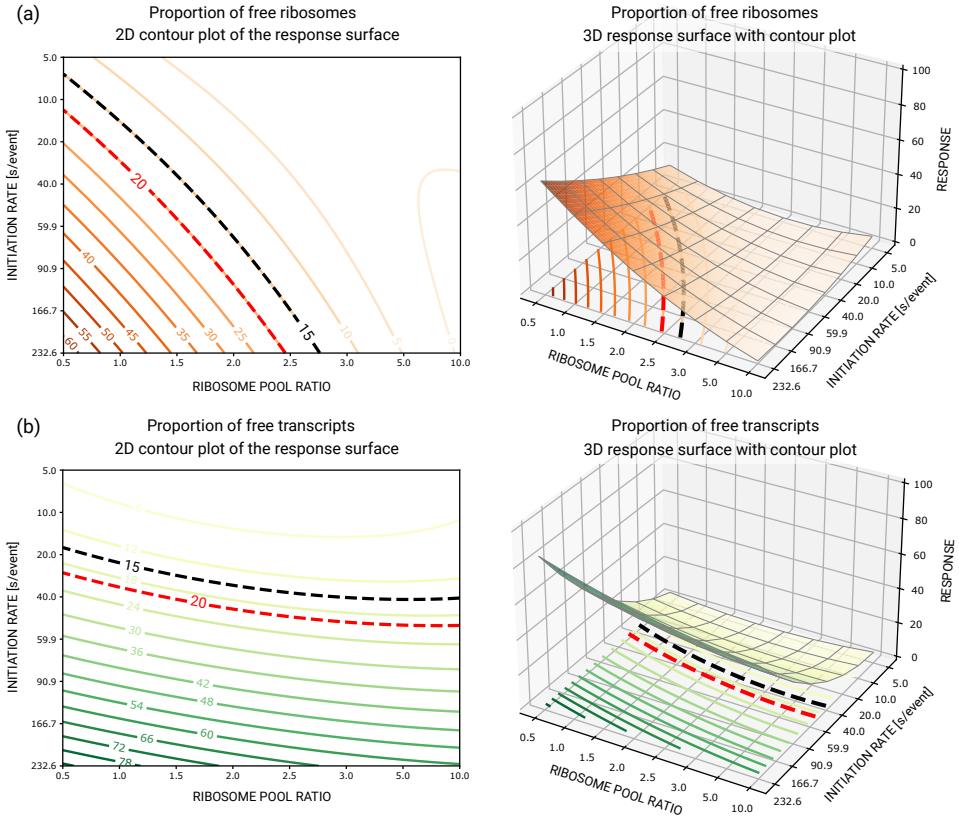


Figure 11.5: (a) Contour plot and response surface of the proportion of free ribosomes in the pool as a function of two factors. (b) Contour plot and response surface of the proportion of free transcripts in the transcriptome. Notably, at a fixed initiation rate, increasing the ribosome pool has only a marginal effect on the proportion of free transcripts. A comparison of panels (a) and (b) reveals that while a ribosome pool and initiation rate combination of 2 and 40 [s/event] keeps the proportion of free ribosomes below 15%, it fails to reduce the proportion of free transcripts below 20%. The optimal adjustment depends on the specific optimization objective pursued by the cells.

The results of this global sensitivity analysis are summarized below.

Sensitivity observation 2.

The effect sizes of each factor are assessed with Partial Eta Squared (η^2). Contribution to the variance: calculate the percentage of variance explained by each factor to understand its contribution to the overall variability in the response. Visualizing Main Effects: plotting the main effects to visualize the impact of each factor level on the dependent variable.

(1) *The translation process is highly sensitive to the ribosome pool size*.
The proportion of free transcripts reflects a poor efficiency in translation. The proportion of free transcripts is considerably reduced if the ribosome pool size increases. It appears that the ribosome pool ratio should at least be higher than 3 in order for more than 85% of the transcripts to be loaded with ribosomes: Fig.11.4.

(2) *The sensitivity of the initiation rate depends on the level of the ribosome ratio (ribosome pool size).*
A high proportion of free ribosomes means inefficient use of translation resources. The proportion of free ribosomes is almost unsensitive to the initiation rate unless the ribosome pool size is lower than $2.5 - 3$: Fig.11.5.

(3) *Fitting a response curve and a contour plot of the outcome: the response surface*.
Contour plots of the response surface for the proportion of ribosomes without a transcript and for the response surface for the proportion of transcripts without a ribosome show the interaction between ribosome pool size and initiation rates: Fig.11.5.

(4) Once the initiation rate exceeds a certain threshold, further increases have no additional effect on the response. This indicates that the initiation rate is the limiting factor in translation as long as it remains below $\lambda = 25 \cdot 10^{-6}\,\mathrm{ms}^{-1}$, corresponding to an average initiation time longer than 40 s for a ribosome (yeast). Once the initiation rate exceeds this threshold, elongation may become the limiting factor in translation under certain conditions. This is true only in the domain where the ribosome pool ratio is larger than $2.5 - 3$ (see sensitivity observation 2.2).

Biological functionality significance: these results provide cues on how ribosome biogenesis and/or ribosome autophagy are used in cells to modulate the proteome. With the initiation rate being the limiting factor, translation

outcomes remain robust to variations in the elongation cycle, which are local and randomly distributed within transcripts (see also sensitivity observation 7).

11.3.3 Sensitivity of ribosome pool size and general initiation rate on the ribosome fragmentation profile

How sensitive is the polysome fragmentation profile to the ribosome pool and on the general initiation rates?

Ribosome fragmentation profiles can be experimentally assessed through polysome profiling via ultracentrifugation in a sucrose gradient cushion, as described in Section 1.6.3 (Chapter 1). Our model enables a distinction, through synthetic simulations, between (i) the number of ribosome-protected fragments (RPFs) aggregated across the transcriptome (polysome profiling) and (ii) the polysome profile of individual transcripts.

Here, we analyze (i) the sensitivity of the aggregated polysome profile and (ii) the sensitivity of the polysome profile per individual transcript in response to changes in the ribosome pool size and general initiation rate. To investigate these issues, we conducted 480 simulation runs of the `Ribosomer` Python version on the HPC. The simulations followed a fully crossed design with 30 replicates, combining 4 ribosome pool ratios and 4 initiation rates ($4 \times 4 \times 30 = 480$ runs), each lasting 108 minutes[2]. We restricted the parameter exploration space to this 4 by 4 combination and selected the chosen range of values for the following reasons. First, as shown in the previous section, a ribosome ratio below 2 was inefficient, as indicated by either an unused ribosome pool or a large number of ribosome-free transcripts (unmet demand). We consider this an unsustainable physiological state for cells. Second, the global initiation rate was limited to four values to ensure clarity in graphical representation, as grouping four different rates per gene ID already results in a visually dense figure, making further subdivisions both impractical and visually uncomfortable. The transcript copy numbers used were the default values from the input data file (`readCounts01.txt`), as shown in Figure 11.1 (b). The relative initiation rate fold changes were set to 1 for all transcripts, except for `RPL4`, which was arbitrarily set to 2, as previously discussed in Figure 11.1. This choice was made to maintain awareness and retain a local reference for potential confounding effects of this parameter, which might otherwise go unnoticed. A dedicated sensitivity analysis for the individual initiation rates (privileged ribosome recruitment on subsets of transcripts) will be conducted in section 11.3.5.

[2] An equivalent computational time of 864 hours (36 days) on a single CPU was required. However, by distributing the job in an *embarrassingly parallel* manner across 30 nodes on the CéCI HPC cluster, the total execution time was reduced to less than 30 hours.

Figures 11.6 and 11.7, on the left and right paired sides below, show the **number of ribosomes per transcript** along with the **transcriptome-wide polysome fractionation profile**, displayed in the upper-right insets. The four figures are arranged in ascending order of the **ribosome pool ratio parameter**, specifically 2, 2.5, 3 and 5, respectively, from left to right and bottom to top (2: bottom left, 2.5: top left, 3: bottom right and 5: top right).

In each figure, **four initiation rates** have been parametrized, i.e., **slow to fast** from $\lambda = 4.3, 11.0, 16.7, 25.0 \times 10^{-3}\text{s}^{-1}$ in **blue, green, orange and red**, respectively. The **box and whisker plots** illustrate the dispersion[3] of results across individual gene IDs for each initiation rate.

How sensitive is the **aggregated polysome fragmentation profile** to the ribosome pool size and on the initiation rate?
How sensitive are the **polysome fragmentation profile per transcript** to the ribosome pool size and on the initiation rate?

We start by estimating the sensitivity of the average number of ribosomes in the polysome aggregated across all transcripts, as represented in the 4 insets in the upper right corner of figures 11.6 and 11.7. The four histograms represents the frequency distribution of the number of transcripts having no ribosome, one ribosome, two, ..., up to 20 ribosomes. In each inset, these histograms are faceted by two colors, one for the fastest initiation rate (red) and one for the slowest initiation rates (blue). The mean, median and standard deviation of the number of ribosomes in the transcripts are tabulated inside the insets and contrasted for the two global extreme initiation rates: fast (red) and slow (blue). The four insets correspond to the four ribosome ratios (2, 2.5, 3 and 5). For clarity the main outcome variables are recapitulated in Table 11.3.3.

Note that the medians are, in most cases, smaller than the corresponding means, indicating assymetry with a positive skewness –associated to distributions skewed to the right tail as seen in the four insets in figures. The right tail of the distribution is longer or extends further than the left tail. This indicates that the majority of data points are concentrated on the left side, with a few higher values pulling the mean to the right. Examples of such positively skewed distributions are exponential distributions or Poisson distributions. The median is a relevant and convenient statistics to describe the outcome of interest here, i.e. the number of ribosomes in the polysome fragments across all transcripts. The median corresponds to the value for which 50% of the observations are smaller and 50% larger.

[3]The box edges correspond to Q1 (first quartile) and Q3 (third quartile), with the box length representing the interquartile range, IQR. The whiskers extend to the largest (or smallest) data point within 1.5 times the IQR. Outliers, beyond this range, are shown as individual star. The + symbol marks the mean, while the – red notch indicates the median.

Table 11.3.3: Number of ribosomes in polysome fragments across all transcript copies.

Number of ribosomes in polysome fragment				
		Median q_{50}		
Initiation	Ribosome ratio			
rate	2	2.5	3	5
Fast ($\lambda = 25 \cdot 10^{-3}\,\mathrm{s}^{-1}$)	1.0	2.0	2.0	3.5
Slow ($\lambda = 4.3 \cdot 10^{-3}\,\mathrm{s}^{-1}$)	1.0	1.0	2.0	3.0
		Mean μ		
Initiation	Ribosome ratio			
rate	2	2.5	3	5
Fast ($\lambda = 25 \cdot 10^{-3}\,\mathrm{s}^{-1}$)	1.88	2.34	2.93	4.83
Slow ($\lambda = 4.3 \cdot 10^{-3}\,\mathrm{s}^{-1}$)	1.79	1.79	2.22	3.97

For case (i), using the definition of the (local) sensitivity as defined in the beginning of section 11.2, we calculate the ratio of the relative variation of the outcome variable, i.e., the median q_{50}, divided by the relative variation of the parameter, the ribosome ratio, associated to the variation of the outcome.

For case (ii), we divide by the relative variation of the initiation rate.

The explicit calculation of sensitivities are as follows:

i **sensitivity of ribosome ratio on number of ribosomes in polysome fragments**. For the fast rate, the relative increase in the parameter when moving from the first column to the last is:

$$\frac{\delta p}{p} = \frac{5-2}{2} \tag{11.19}$$

$$= 1.5. \tag{11.20}$$

Using the median q_{50} in table 11.3.3, the relative increase in the outcome q_{50} is:

$$\frac{\delta q}{q} = \frac{3.5-1.0}{1.0} \tag{11.21}$$

$$= 2.5. \tag{11.22}$$

Hence, the sensitivity to the ribosome ratio is:

$$\frac{\frac{\delta q}{q}}{\frac{\delta p}{p}} = \frac{2.5}{1.5} \tag{11.23}$$

$$= 1.67. \tag{11.24}$$

This means that the sensitivity is 167%: on average, if the ribosome ratio is multiplied by 2, the median number of ribosomes in polysome fragment across fragments will be multiplied by $2 \times 1.67 = 3.34$.
A similar calculation, still using the median as the outcome, but for the slow rate, gives a sensitivity of 133%.

ii **sensitivity of global initiation rate on number of ribosomes in polysome fragments**.
For the low pool of ribosomes (ribosome ratio of 2), the relative increase in the parameter when moving from slow rate to fast rate (rows) is:

$$\frac{\delta p}{p} = \frac{25 - 4.3}{4.3} \tag{11.25}$$

$$= 4.81. \tag{11.26}$$

Using the median q_{50} in table 11.3.3, the relative increase in the outcome q_{50} is:

$$\frac{\delta q}{q} = \frac{1.0 - 1.0}{1.0} \tag{11.27}$$

$$= 0. \tag{11.28}$$

Hence, the sensitivity to the initiation rate is:

$$\frac{\frac{\delta q}{q}}{\frac{\delta p}{p}} = \frac{0}{4.81} \tag{11.29}$$

$$= 0. \tag{11.30}$$

This means that the (local) sensitivity is 0%: an increase in initiation rate as no effect on the median number of ribosome when the ribosome pool size ratio is 2.
A similar calculation, still using the median as the outcome, but for a more abundant ribosome pool (ribosome ratio of 5), gives a sensitivity of 3.5%. The median number of ribosomes barely increases when the initiation rate is increased.

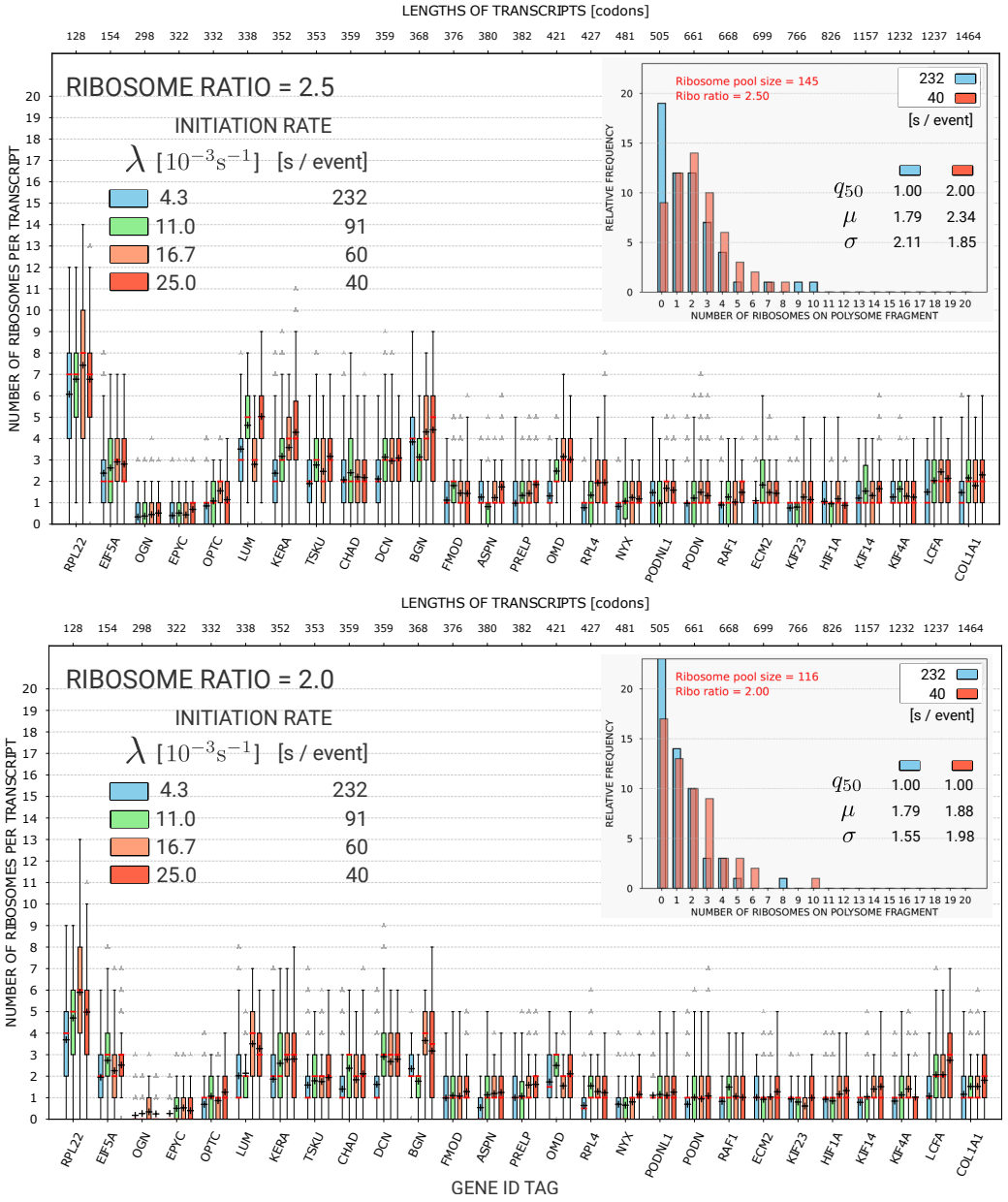


Figure 11.6: The polysome fragmentation profile per transcript and aggregated across all transcripts (upper right insets) as a function of general initiation rates and ribosome pool size (**restricted pool**). Further explanations are provided in the main text.

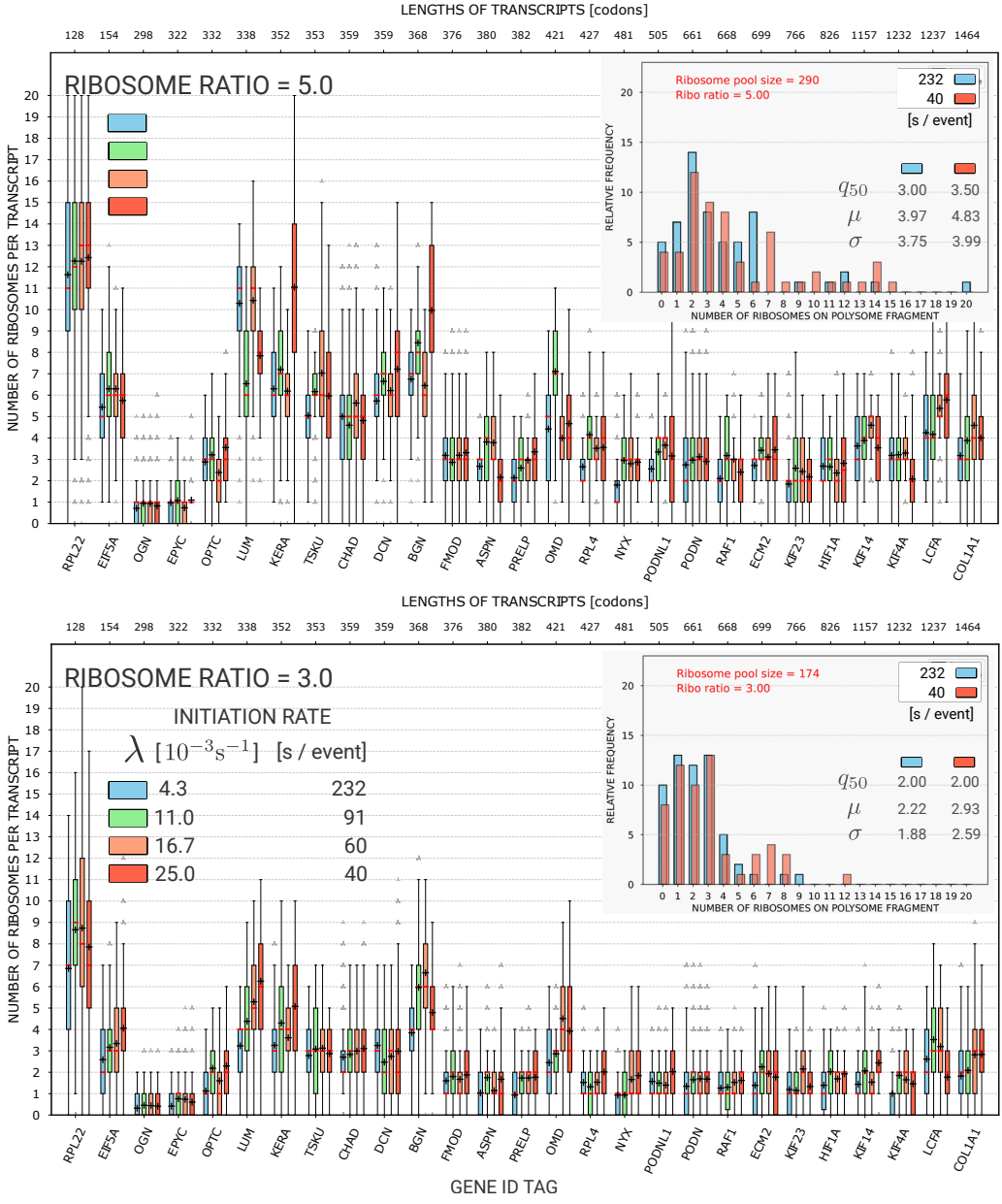


Figure 11.7: The polysome fragmentation profile per transcript and aggregated across all transcripts (upper right insets) as a function of general initiation rates and ribosome pool size (**well-supplied pool**). Further explanations are provided in the main text.

The results of this local sensitivity analysis are summarized below.

Sensitivity observation 3. The number of ribosomes in polysomes is moderately sensitive to the initiate rate but highly sensitive to the ribosome pool size.

The number of ribosomes in polysomes has a relative sensitivity[a] of 1% to 7%, to the general initiation rate **conditionally on the ribosome pool size**, whereas it is highly sensitive –relative sensitivity of 100% to 167%– to the ribosome pool size, unless the initiation rate is extremely low, i.e., below 200 s/event.

These observations collectively underlie the use of polysome fragmentation profiling to assess the global initiation rate or ribosome pool size, i.e., the **upregulation of the translational machinery**.

[a] This is a local relative sensitivity as defined previously.

The biological significance of an increased polysome fraction on transcripts should be interpreted primarily as a sign of enhanced ribosome biogenesis or reduced ribophagy, rather than solely as an increase in initiation rate. An increase in polysome fraction across the transcriptome is expected to correlate with an **enrichment of gene ontology terms related to protein synthesis pathways**.
This sensitivity observation further confirms that **a key determinant of translational control is the ribosome pool size—a seemingly obvious fact that is often overlooked in practice**.

The Sensitivity observation 3 addressed how the polysome profile aggregated across all transcripts was sensitive to the ribosome pool size and the initiation rate. Now we turn to the same question but per transcript.

How sensitive are the **polysome fragmentation profile per transcript** to the ribosome pool size and on the initiation rate?

By examining individual transcripts and visually comparing the four main panels of Figure 11.6 and 11.7, we observe that interquartile positions shift upward along the y-axis as the ribosome pool ratio increases from a restricted to a well-supplied state. This trend is particularly evident if the ribosomal gene ID `RPL22` (the first gene ID tag on the left and the shortest in sequence length) is taken as an example. In the box-and-whisker plots, the box edges represent the first (Q1) and third (Q3) quartiles, with the box length corresponding to the interquartile range (IQR). The whiskers extend to the largest or smallest data points within 1.5 times the IQR, while outliers beyond this range are shown as individual stars. The + symbol marks the mean, while the − red notch indicates the median.

Specifically, for `RPL22` as an example, with a fast initiation rate of $\lambda = 25 \cdot 10^{-3}\,\mathrm{s}^{-1}$ or 40 s/event (red color), the interquartile range Q1-Q3 shifts from 3-6 to 5-8, to 5-10 and finally to 11-15 as the ribosome ratio increases from 2 to 2.5, then to 3 and 5. Similar overall trends are observed for all gene ID tags, despite the considerable dispersions in the distribution of ribosome counts within polysome fragments. The sensitivity analysis (observation 3), made earlier regarding polysome profiles aggregated across all transcripts also holds at the individual transcript level.

At the individual transcript level, increasing the global initiation rate does not have a significant impact, as it is masked by the high variability in outcomes. Comparing the interquartile range (Q1–Q3) within each gene at a fixed ribosome pool ratio, as the global initiation rate increases from blue (slowest) to green, then to orange, and finally to red (fastest), reveals that this parameter has minimal influence on polysome multiplicity. However, some gene ID tags, such as `RPL22`, `EIF5`, `LUM`, and `COL1A`, exhibit a slight positive correlation. The large dispersions in the data obscure the mechanistically expected trend discussed in Chapter 9.

The results of this local sensitivity analysis are summarized below.

Sensitivity observation 4. Polysome fragmentation profiles of individual transcripts shift with ribosome pool size, at fixed global initiation rate, but not to the transcript length.

(1) Sensitivity observation 3 previously addressed how the aggregated polysome profile responded to ribosome pool size and initiation rates. Examining individual transcripts now shows a clear trend: as the ribosome pool shifts from a restricted to a well-supplied state, interquartile positions in polysome fragmentation profiles move upward along the $y-$axis. This effect is particularly visible for most individual transcripts, where increased ribosome availability leads to a higher median and broader interquartile range in ribosome occupancy.

(2) Limited effect of global initiation rate: at the individual transcript level, increasing the global initiation rate has minimal impact on polysome multiplicity due to high variability in ribosome distribution. Only a few gene IDs (`RPL22`, `EIF5`, `LUM`, and `COL1A`) show a slight positive correlation.

(3) High dispersion masks trends: large variability in polysome profiles,as shown in Fig. 11.6 and Fig. 11.7, both within and across transcripts, blurs mechanistically expected trends, complicating straightforward interpretations of initiation and elongation rate effects.

Restricting the ribosome pool available for translation may help cells prevent excessive ribosome loading on transcripts and reduce the risk of ribosome congestion (traffic jams).

In Chapter 9, we established that the number of ribosomes engaged in translation per transcript, when plotted against transcript length, should exhibit a slope proportional to the ratio of the initiation rate to the elongation rate. However, this relationship holds only under the assumption of an unlimited ribosome pool. Figures 11.6 and 11.7 display transcripts ordered by length, ranging from 128 to 1464 codons. Notably, longer transcripts do not consistently contain more ribosomes in their polysome fragments. This suggests that, at least within this transcript length range (which remains relatively small), a ribosome pool size corresponding to a ribosome pool ratio between 2 and 5 does not meet the criteria for an unlimited ribosome pool condition.

Universal scaling as discussed in the literature in the context of protein synthesis is the observation that protein synthesis times scale linearly with coding-sequence length [Sharma et al. 2018]. Across transcripts in an organism, the average codon translation rate exhibits minimal variation, implying a nearly constant elongation rate per transcript. This allows the synthesis time of a protein to be estimated simply by multiplying the transcriptome-wide average elongation rate by the coding sequence length. This is also related to the notion that the dwell time of a ribosome on any given codon is the same within a transcript, on average, across codons of the same type. This is also true for the average dwell time of the ribosome on any given codon across different transcripts throughout the transcriptome. The causes of this universal scaling are multiple:
(i) Despite local fluctuations, codon translation rates are near-randomly distributed across transcripts, meaning slow-translating codons are offset by fast-translating ones. This statistical balance leads to a consistent average translation rate per transcript;
(ii) The law of large numbers–e.g., large number of codons in a transcript–explains the mathematical origin of scaling. The average codon translation rate per transcript follows the law of large numbers, converging toward the transcriptome-wide mean as gene length increases;
(iii) Key molecular determinants of translation speed, such as codon usage, tRNA availability, proline content, and charged residues, scale proportionally with transcript length. This suggests that randomness in these factors drives the overall scaling behavior.

This sensitivity analysis observation is summarized below.

Sensitivity observation 5. Absence of transcript length effect in the ribosome fragment profiles

(1) Deviation from expected ribosome-transcript length relationship: the expected linear relationship between ribosome count and transcript length, predicted under unlimited ribosome availability, is not observed in figures 11.6 and 11.7. This suggests that within the tested range of ribosome pool ratios (2–5), the system does not operate under conditions of an unlimited ribosome supply.

(2) Combined with this operating mode of restricted pool of ribosomes, the universal scaling of protein synthesis time scale and the fact that elongation rates are, on average, smaller than the initiation rates, tend to unravel the crucial importance of competition between transcripts to recruit ribosomes.

This universal scaling confers robustness to mRNA translation and protein elongation, ensuring stability despite biological complexity. The independence of ribosome number per transcript from transcript length, combined with universal scaling, faster elongation rates relative to initiation, and a constrained ribosome pool, leads to the conclusion that transcripts must compete for ribosome recruitment at initiation. This competition plays a dominant role in shaping both polysome distributions and ribosome profiling landscapes.

A last sensitivity observation can be directly drawn based on figures 11.6 and 11.7 and is summarized hereafter.

Sensitivity observation 6. Polysome fragmentation profiles exhibit overdispersion.

The uncertainties in polysome fragmentation profiles as measured by σ^2–**the variance of the number of ribosomes on the transcripts** aggregated across the transcriptome (polysome fragments profiling)–**is at least as large as their mean**, μ–**a characteristic feature of overdispersion**. Ribosome profiling (Ribo-Seq) studies often model ribosome occupancy on transcripts using Poisson-based statistical frameworks. Empirical data exhibit overdispersion, where the observed variance exceeds that predicted by a standard Poisson distribution. To account for this, researchers have employed models such as the Poisson-log-normal hierarchical model, which introduces random effects to accommodate the extra variability observed in ribosome profiling data. The four insets in Figures 11.6 and 11.7 show that the ratios of the variance over the mean, σ^2/μ in all simulations are in the range $\sigma^2/\mu \in [1.3 - 3.5]$. Furthermore, **the variance in the ribosomes fragment profiles is sensitive to the ribosome pool size**: the larger the ribosome pool size, the larger the variance.

Research evaluating the reproducibility of ribosome profiling data across various organisms and protocols found significant variability in ribosomal densities at nucleotide resolution, suggesting that the noise level is higher than previously thought [Diament and Tuller 2016]. While ribosome counts on transcripts might initially be modeled using a Poisson distribution, empirical evidence indicates that these counts are overdispersed, necessitating the use of more sophisticated statistical models to accurately capture the observed variability in ribosome profiling data. The inset in the lower panel of Figure 11.7 even suggests a bimodal distribution, which may indicate ribosome congestion (traffic jam).

Our TASEP-based agent-based model successfully reproduces the pronounced overdispersion observed in ribosome profiling data.

11.3.4 Codon usage bias and U34 tRNA modifications

To investigate the effect of codon usage and the U34 tRNA modifications, we analysed 1440 simulation runs of the Python version of `Ribosomer` on the HPC. The simulation runs were executed in a fully-crossed design with 30 replicates, i.e., 2 conditions (control–wild type, and case–U34 hypo-modification) $\times$ 4 ribosome pool ratios $\times$ 8

initiation rates × 30 replicates = 1440 runs, each lasting 108 minutes. The other input data files were the same as before, i.e same transcripts copy numbers, and same default relative fold-change for the individual initiation rates as before.

Figure 11.8 (a-d) compares the codon usage bias in the U34 sensitive codons between three different transcripts: `KIF4A`, `KIF14`, `KIF23` and the human transcriptome. The percentage of each of the three amino acids K, Q, and E are indicated in the (a-d) panel of the figure, with the percentage of all the U34 sensitive codons in the transcript, as well as the relative unbalance of the A-ending codons in the transcript. The A-ending codons for each pair of codons of the tree amino acid (K, Q, E) have a red color bar in the figure 11.8 (a-d) of the codon usage frequency. The G-ending codons are shown in grey in the figure.

The case setting –U34 tRNA hypo-modification corresponds to `ELP3-URM1` enzyme silencing. The lack of U34 modification or U34 hypomodification (mcm5s2) or lack of s2 (thiolation) slow down the first sub-step in the elongation cycle, for codons ending with A in lysine, glutamate and glutamine amino acids (K, E, Q). It affects all rate kinetics of substep 1 for 6 codons (3×2) for the 3 amino acids (K, E, Q). The G ending codons of these amino acids are read slightly faster in the case of yeast species [Nedialkova and Leidel 2015; Ranjan and Rodnina 2017].

We recall, from chapter 5, the factors in the dictionary that we calibrated from the literature results of Nedialkova and Leidel [Nedialkova and Leidel 2015]: `uridine34SilencedDict` = '**AAA**': 0.33, 'AAG': 1.10, '**CAA**': 0.55, 'CAG': 1.0, '**GAA**': 0.62, 'GAG': 1.18. The values mapped by the codon type keys are the multiplying factors affecting the rate of the accommodation sub-step in the elongation cycle of these codons. The A-ending codons in bold typeface in the above dictionary (keys:values) are the codons for which the accommodation rate is significantly reduced as was explained in chapter 5, section 5.3. Hence the dwell times of ribosomes on these codons are expected to be significantly longer upon U34 tRNA modifications.

Recall from this dictionary and chapter 5 that the G-ending codons of the three amino acids (K, Q, E) are accommodated a bit faster or have unchanged kinetics upon U34 hypo-modification.

Upon U34 tRNA hypo-modification, it is expected that the AA ending codons of the three amino acids lysine, glutamine, glutamate (K,Q,E) will accommodate their cognate tRNA at a slower rate (as was detailed in chapter 5), thereby locally slowing down the elongation rates. Proteins that are enriched in these codons are expected to be more affected by this effect than any other transcript. Based on codon composition of these three transcripts, `KIF4A`, `KIF14`, `KIF23`, it can be calculated that the proportions of U34-sensitive codons in these kinesins are 13.5%, 16.5%, and 14.1% respectively, whereas an average human transcript only contains 6.5% of such codons. The total counts of U34-sensitive codons, K_{AAA} + Q_{CAA} + E_{GAA}, are 166, 191, 108 for `KIF4A`,

KIF14 and KIF23 respectively. They are highly enriched in these codons. Hence, these transcripts are expected to be mostly affected as compared to all others, upon U34 hypo-modification when contrasted to the control case.

Figure 11.8 (e) contrasts the relative protein abundance between case (U34 hypo-modification: right panel) and control (wild type: left panel) for the 30 simulation run replicates where the ribosome pool ratio was 3 and the general initiation rate was 60 s/event (or $\lambda = 16.7 \cdot 10^{-3}\mathrm{s}^{-1}$).

No consistent and significant reduction in protein abundance is observed across the kinesin family in the simulation results shown in Fig. 11.8(e). Among the three kinesins, only KIF4A exhibits a significant reduction –a difference exceeding 3.1σ, in protein abundance upon U34 tRNA hypo-modification between the left and right panels in Fig. 11.8(e). This observation is not reproduced in Fig.11.9(c-d) in the next section below for instance, suggesting a possible false positive result. We recall that, due to the curse of multi-testing, even with an effect size difference of 3.1σ, the multiple comparison of 27 gene ID tags may give 0.05% of false positive results, i.e., at least $\approx$ 1 gene ID tag declared significantly different, while it is not. The inconsistency observed in Figure 11.8(e) is that the protein abundance levels of the other two kinesins are not reduced as was expected upon U34 tRNA hypo-modification.

The reason of this negative result is that the slow-down in the local elongation rates encountered at the U34 sensitive codons are not enough to lower the elongation rate, averaged over the full length of the transcript, below or at the level of the initiation rate of the transcript. Furthermore, the sensitive codons are randomly distributed along the length of the transcript (not shown here).

This result demonstrates that the relative abundance of transcripts is NOT sensitive to the codon usage or to U34 tRNA modifications, even for transcripts with very high enrichments in potentially targeted codons.

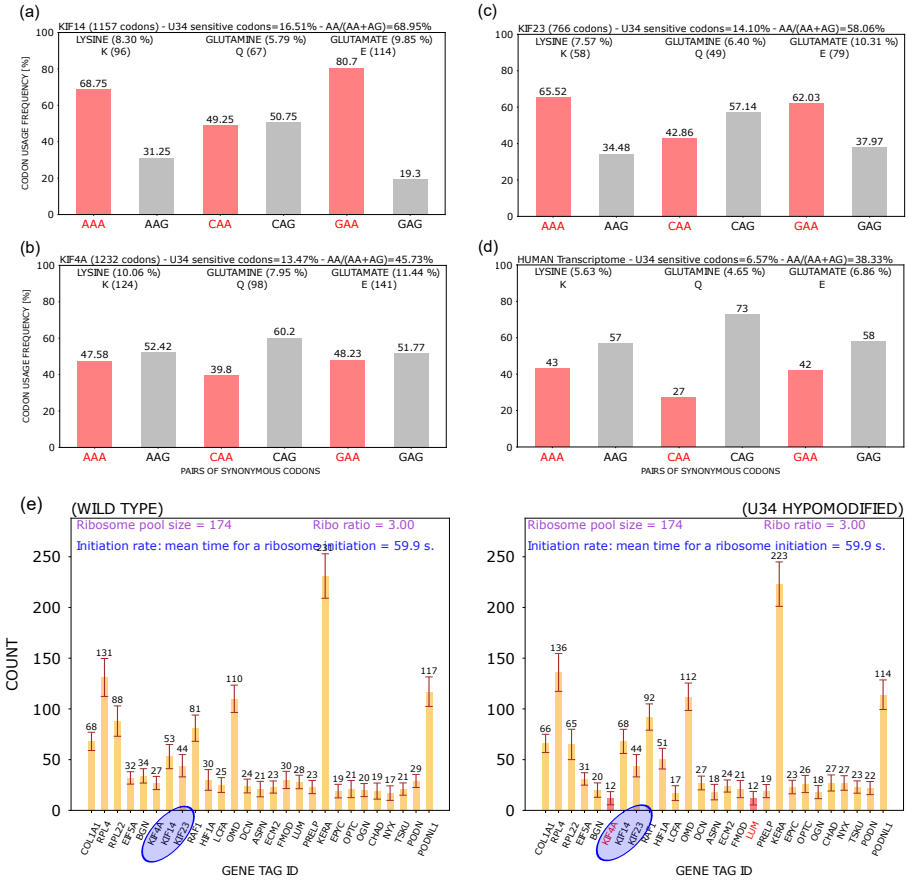


Figure 11.8: Codon usage and U34 tRNA modification sensitivity and uncertainty analysis on protein relative abundance. Twenty-seven geneID tags are compared. The 27 multi-testing includes three proteins belonging to the kinesin KIF family (blue ellipse) highly enriched in lysine, glutamine, glutamate (K, Q, E) and exhibiting highly biased codon usage in favor of AA ending codons: **KIF14** (a), **KIF4A** (b), **KIF23** (c), as compared to the average transcriptome (d). Comparison of relative protein abundance (e) between wild type (control, left panel) and U34 hypomodification (case, right panel). The case gene ID tags labelled in red differ by more than 3.1σ in the protein abundance outcome as compared to control. Further explanations are provided in the main text.

Sensitivity observation 7. Protein relative abundance sensitivity to the codon usage upon tRNA modification is overridden by initiation rates being smaller than elongation rates and by *universal scaling*.

Figure 11.8 (a-c) compares the enrichment of three amino acids and their codon usage biases relative to the average human transcriptome (d). The proportions of U34-sensitive codons in these kinesins are 13.5%, 16.5%, and 14.1%, whereas an average human transcript only contains 6.5% of such codons.

Despite this very high enrichment in codons sensitive to U34 tRNA modifications, no significant effect can be measured on the relative protein abundance outcome upon tRNA hypo-modifcation, leading us to conclude that the protein relative abundance is not sensitive to codon usage or to U34 tRNA modifications.

The local slowdown of elongation, despite more than a twofold increase in the proportion of U34 tRNA modification-sensitive codons, is **overridden by the initiation rate as the limiting factor**.
Furthermore, the inherent variability in protein abundance obscures differences that, in most cases, remain below 3.1σ. In Figure 11.8, it cannot be ruled out that the observed reduction in `KIF4A` protein abundance is a **false positive**–see the discussion above on the curse of multiple testing.
The biological significance of this **robustness in the proteome outcome with respect to local changes in the elongation rates** is due to what Sharma *et al.* called **universal scaling** of protein synthesis times [Sharma et al. 2018].
The term universal scaling refers to the linear relationship between protein synthesis time and coding sequence length. It was demonstrated that, across different organisms (*E. coli, S. cerevisiae, and H. sapiens*), the time required for ribosomes to synthesize a protein is directly proportional to the number of codons in the transcript. This means that if the coding sequence length doubles, the synthesis time also doubles. The study explains this scaling relationship using the law of large numbers: since **codon translation rates are near-randomly distributed across transcripts**, the average translation speed per gene converges to the global transcriptome-wide average as gene length increases. Additionally, the **random distribution of molecular determinants** (such as tRNA concentrations, proline residues, charged amino acid distribution in the exit tunnel, and mRNA secondary structures) **ensures that the scaling relationship remains robust even when specific factors influencing translation speed are altered**. **Despite local fluctuations, codon translation rates are near-randomly distributed across transcripts, meaning slow-translating codons are offset by fast-translating ones.**

11.3.5 Ribosome recruitment on privileged transcripts: fold change in individual initiation rates

To investigate the effect of a privileged ribosome recruitment on a subset of transcripts (KIF family gene ID tags) in conjonction with codon usage and the U34 tRNA modifications, we analysed 960 simulation runs of the Python version of `Ribosomer` on the HPC. The simulation runs were executed in a fully-crossed design with 30 replicates, i.e., 2 conditions (control–wild type, and case–U34 hypo-modification) $\times$ 2 ribosome pool ratios $\times$ 4 initiation rate $\times$ 2 sets of initiation fold change values (default and KIF privileged) $\times$ 30 replicates = 960 runs, each lasting 108 minutes. The other input data files were the same as before, i.e same transcripts copy numbers. The default relative fold-change for the individual initiation rates were the same as before, except for the subset of the KIF family gene ID tags, i.e., `KIF4A`, `KIF14`, `KIF23`, for which the initiation rates specific fold change were set to 4. This scenario simulates a privileged ribosome recruitment process, where initiation on these KIF transcripts occurs four times faster than on all other transcripts.

In this sensitivity analysis, we examine whether variations in individual initiation rate fold changes, relative to the global initiation range, influence the resulting protein abundance profile.

The key aspect of this analysis is that modifications to the input variable were exclusively applied to three gene ID tags corresponding to kinesin family proteins: `KIF4A`, `KIF14`, and `KIF23`. In the `readCounts.txt` input file, the transcript copy numbers for all gene ID tags were maintained identical to those used in previous simulation runs. The transcript copy numbers were already shown in Fig 11.1 (b) and are 1, 3, 2 respectively for `KIF4A`, `KIF14`, `KIF23`.

We changed, for the three `KIF` geneID tags, the fold change affecting the individual initiation rate of these transcripts only (and all their copies). These three transcripts still have their very same codon sequences as in all other simulation runs. The output variables that are examined are all the protein abundance levels. The comparisons in Figure 11.9 should mainly focus between rows (c) and (d), within control (wild type) and within case (U34 hypo-modification). In all panels, the ribosome ratio parameter was set to 3 and the global initiation rate was set to 40 s/event.

Within the wild type, as the fold change in individual initiation rates is increased from 1 to 4, for the subset of the three KIF geneID tags (simulating a priviliged ribosome recruitment during initiation), the outcome in protein abundance of these three geneID tags is increased from 23, 62, 52 to 52, 165, 98, respectively. The corresponding sensitivities to the protein abundance levels are calculated as 31.5%, 41% and 22%, respectively.

Within the U34 hypomodification case, as the fold change in individual initiation rates

is increased from 1 to 4, for the subset of the three KIF geneID tags (simulating a priviliged ribosome recruitment during initiation), the outcome in protein abundance of these three geneID tags is increased from 27, 58, 39 to 50, 162, 102 respectively. The corresponding sensitivities to the protein abundance levels are calculated as 21.3%, 44% and 40%, respectively.

Between wild type and case, the protein abundance levels are not significantly affected. False positive results cannot be ruled out for the two red bars as discussed earlier in relation with multiple testing.

A crucial observation is that gene ID tags with unchanged individual initiation rate fold changes still exhibit significant alterations in protein abundance levels. This is evident for `RPL4`, `OMD`, `KERA`, and `PODNL1` in both wild-type and case conditions. These transcripts experience the strongest competition for ribosome recruitment, as they also have the highest copy numbers, as shown in Fig. 11.1 (b). This finding supports the expected interaction between transcript abundance and ribosome competition.

Figure 11.9 (a-b) compares the codon usage bias in the U34 sensitive codons for transcript: `RPL4` and the human transcriptome. Figure 11.9 (c-d) contrast the relative protein abundance between case (U34 hypomodification: right panel) and control (wild type: left panel) for the 30 simulation run replicates where the ribosome pool ratio was 3 and the general initiation rate was 40 s/event (or $\lambda = 25 \cdot 10^{-3}\text{s}^{-1}$). The comparison of each row of Fig. 11.9 (c) and (d) show that the protein relative abundance profiles is very sensitive to a change in the individual initiation rates. Other ribosome pool ratios and other general initiation rates (not shown here) do not alter the conclusions.

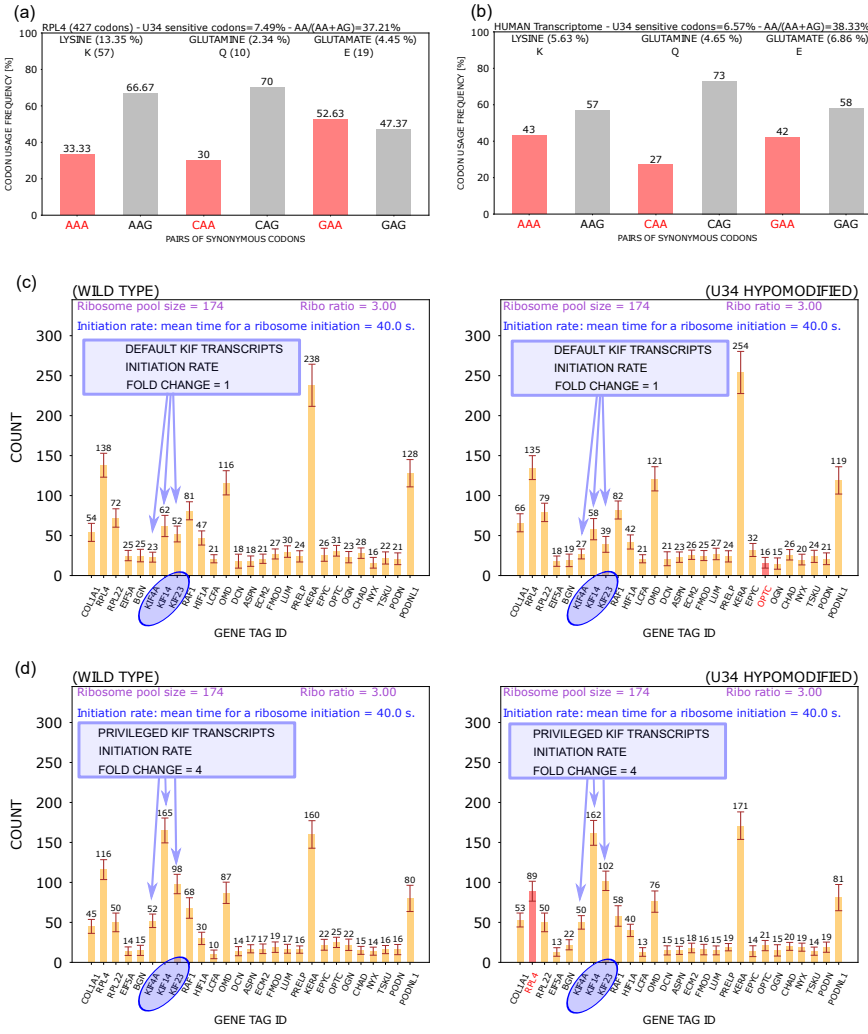


Figure 11.9: Ribosome recruitment on privileged transcripts and uncertainty analysis on protein relative abundance. Twenty-seven geneID tags are compared. The 27 multi-testing set includes three proteins from the kinesin KIF family (blue ellipse)–`KIF4A`, `KIF14` and `KIF23`–which are highly enriched in lysine (K), glutamine (Q), and glutamate (E) and exhibit a strong codon usage bias favoring AA-ending codons. `RPL4` is also highly enriched in lysine but has a lower frequency of AA-ending codons (a) compared to the average transcriptome (b). Panels (c) and (d) compare relative protein abundance between wild-type (control, left panel) and U34 hypomodification (case, right panel). In (c), the KIF family gene ID tags retain the default fold change in initiation rates, whereas in (d), the fold change is set to 4, promoting preferential ribosome recruitment on these transcripts. The case gene ID tags labelled in red differ by more than 3.1σ in the protein abundance outcome as compared to control. Further explanations are provided in the main text.

Sensitivity observation 8. The general profile of protein relative abundance is sensitive to fold changes in the initiation rates of a subset of transcripts. This reflects the strong impact of preferential ribosome recruitment at the initiation stage of these privileged transcripts, which in turn influences the overall proteome landscape.

Even with a ribosome pool corresponding to a ribosome-to-transcript ratio of 3:1–three ribosomes per transcript copy–transcripts still compete for a limited ribosome supply. Comparing rows (c) and (d) in Fig. 11.9 reveals that gene ID tags, for which both transcript copy numbers and initiation rates were kept unchanged, undergo significant variations in protein abundance levels (e.g., `KERA` or `RPL4`). This suggests that **ribosome recruitment competition, driven by higher initiation rates on a subset of privileged transcripts, disproportionately affects the translation of other transcripts**. The effect of codon usage bias appears to be outweighed by this competitive advantage in ribosome allocation. Once again, it cannot be ruled out that the observed reduction in `RPL4` protein levels is a false positive, potentially arising from the multiple testing issue discussed earlier.

Biological significance: Protein synthesis is robust to local variations in elongation rates but is highly sensitive to ribosome pool availability and the competition for ribosome recruitment during the initiation stage.

11.4 Summary

Ribosome recruitment at initiation, governed by ribosome pool size and initiation rates, is the dominant factor shaping protein synthesis efficiency. While elongation dynamics and codon usage play roles, they are secondary to competition for ribosomes. Polysome fragmentation and profiling variability further underscore the dynamic nature of ribosome allocation, particularly under ribosome-limited conditions.

Chapter 12

General discussion

We may regard the present state of the universe as the effect of its past and the cause of its future. An intelligence that could comprehend all the forces by which nature is animated, and all the positions of all the components of the universe at a given moment, would be able to calculate the future and the past with the same certainty.
Pierre Simon Laplace (1814). Philosophical Essay on Probabilities.
Laplace's encapsulation of the concept of **determinism** states that if someone (or something) knew all the forces and states of every object in nature, at a given point in time, they would be able to predict the future and reconstruct the past.

We are drowning in information and starving for knowledge.
Rutherford D. Rogers (1955), Librarian at the New York Public Library.

The situation is dire, my friend. We are drowning in information but starving for knowledge.
John Naisbitt (1982), American author and futurist who echoed and popularized the phrase in his 1982 book 'Megatrends'.

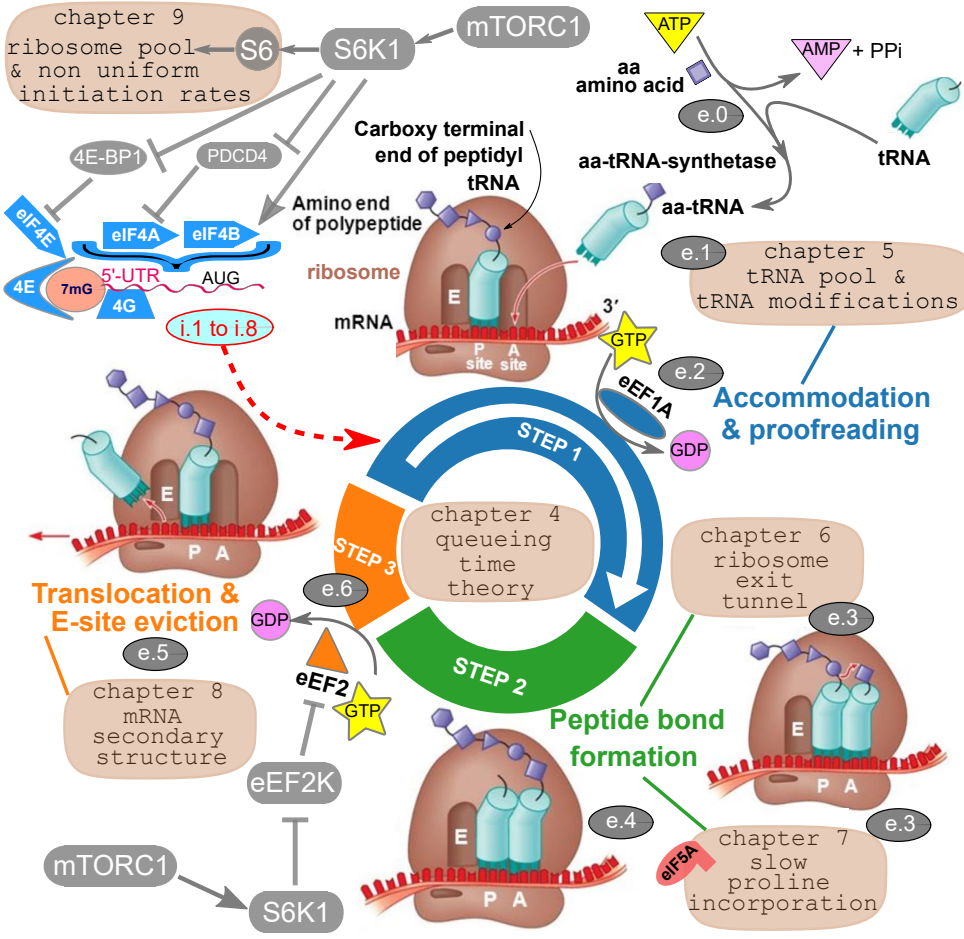


Figure 12.1: Recapitulative overview of the factors that are studied and modeled in this PhD dissertation and how they affect the translation cycle.

12.1 Summary of the work and main contributions

12.1.1 Chapters summary

Chapter 1 –General introduction– provides an overview of ribosome-driven protein synthesis, focusing on the fundamental principles of translation elongation. It explores the roles of codon usage and tRNA adaptation, as well as experimental techniques such as polysome profiling and ribosome footprinting. The chapter outlines key challenges in understanding ribosome kinetics and identifies five contextual factors that contribute to the non-uniform speed of ribosomes along mRNAs. It sets the foundation for computational modeling approaches and highlights the thesis's motivation—bridging molecular biology and quantitative modeling to address gaps in our mechanistic understanding of translation.

In **Chapter 2 –Thesis objectives and methodology–** outlines the objectives of the thesis, focusing on the development of a computational framework to model translation elongation with high biological fidelity. A key methodological approach is the use of an agent-based model (ABM), integrating stochastic processes to capture ribosome dynamics. The research questions are centered on understanding how key factors influence mRNA translation by ribosomes. An agent-based model, `Ribosomer`, is developed to simulate the effects of tRNA modifications, ribosome tunnel interactions, proline incorporation, mRNA structures, and ribosome congestion. Computational simulations are used to explore how these factors shape translation efficiency and protein output. The chapter also introduces the open-source Ribosomer tools designed for further research and visualization.

Chapter 3 –Building the `Ribosomer` framework using a TASEP approach– details the construction of the `Ribosomer` model, based on the Totally Asymmetric Simple Exclusion Process (TASEP). It describes the core computational structure of the ABM, where ribosomes are treated as interacting agents moving along mRNA transcripts. The model is parameterized to capture realistic elongation rates, ribosome queueing behavior, and polysome dynamics. A key contribution is the ability to simulate ribosome competition for limited resources, allowing for the study of translation bottlenecks under different biological scenarios.

Figure 12.1 depicts the relationship between Chapters 4 to 9 and their focus on different aspects of the ribosomal protein elongation cycle during mRNA translation.

Chapter 4 –Mechanochemistry and queueing time statistical theory– introduces the queueing time theory to model the stochastic kinetics of ribosome movement along mRNA. It discusses how ribosome dwell times at codons are governed by probabilistic distributions, informed by mechanochemical constraints. The theoretical framework connects ribosome elongation dynamics with statistical properties of waiting times,

providing insights into how codon identity, tRNA availability, and energy-dependent processes shape translation efficiency.

Chapter 5 –tRNA pool and tRNA modifications– explores the role of tRNA modifications, particularly at the wobble position at the anti-codon loop, in modulating elongation rates. It examines how post-transcriptional modifications influence decoding efficiency and ribosome accommodation times. The chapter also presents an implementation of these effects within the ABM, calibrating the impact of tRNA modification patterns on translation dynamics. The findings suggest that differential tRNA modifications act as a regulatory layer influencing protein synthesis speed and fidelity.

Chapter 6 –Ribosome exit tunnel electrostatic interaction– explores how electrostatic interactions within the ribosome exit tunnel influence elongation kinetics. A computational model of the tunnel's electrostatic potential is developed to predict how charged residues in nascent peptides experience forces that modulate translation rates. Experimental data calibrates the model, providing a mechanistic explanation for sequence-dependent pausing and co-translational folding effects. This factor serves as a contextual element, extending the influence of codons by maintaining a mobile memory window of approximately 50 codons upstream of the nascent peptide elongation site, where past amino acid residues can influence ongoing elongation dynamics.

Chapter 7 –Slow peptide bond formation by proline residues– focuses on the kinetic bottleneck introduced by proline residues in elongation. It explores how the unique structural constraints of proline slow down peptide bond formation and ribosome translocation. The role of elongation factors such as EF-P and eIF5A in alleviating proline-induced stalling is discussed. The findings highlight how specific amino acid sequences can modulate translation speed, with implications for proteome-wide regulation.

Chapter 8 –mRNA secondary structures and ribosome translocation– addresses the impact of mRNA secondary structures on ribosome movement, with an emphasis on hairpins and pseudoknots. The chapter discusses the intrinsic helicase activity of the ribosome and how stable structures can act as roadblocks, influencing elongation rates. A "gear-shift" model is introduced to describe how ribosomes adjust their speed in response to downstream structural constraints. These constraints are contextual elements that extend the influence of codons over a window of more than 90 codons downstream of the decoding site. These findings contribute to understanding how mRNA structure affects translocation dynamics.

Chapter 9 –Ribosome pool and non-uniform initiation rates– explores the effects of ribosome availability and differential initiation rates on translation efficiency. It discusses the balance between free and engaged ribosomes, linking ribosome biogenesis to protein synthesis regulation. Key regulatory pathways, such as `mTORC1` signaling

and ribophagy, are highlighted as mechanisms controlling translation initiation. The chapter also presents strategies for calibrating initiation rates in computational models, emphasizing the role of resource allocation in shaping cellular proteomes.

Chapter 10 –Model output and flexibility offered by the input data and parameters– presents the outputs of the `Ribosomer` model and its flexibility in parameter selection. The model simulates mRNA translation by a shared ribosome pool, allowing for precise control of key parameters such as initiation rates, ribosome pool size, and codon-specific elongation rates. `Ribosomer` also enables the disentanglement of five key influencing factors explored in Chapters 5–9: tRNA modifications, ribosome exit tunnel electrostatics, proline-induced slowdowns, mRNA secondary structure roadblocks, and the interplay between ribosome pool size and initiation rates. Key predicted outputs include protein relative abundances, translation efficiencies, ribosome distribution, polysome profiles, and ribosome density maps. All these outputs are experimentally accessible in the lab, allowing the model to support a more meaningful interpretation of experimental results in similar settings. The chapter also discusses steady-state conditions, sampling rates, and statistical robustness, emphasizing the importance of computational modeling in capturing biological processes that are difficult to measure experimentally. Verification efforts confirm consistency across different programming implementations, and the model's computational efficiency ensures realistic simulation dynamics. Overall, `Ribosomer` offers a powerful and customizable framework for investigating translation kinetics and elucidating biophysical and biochemical mechanisms involved in protein synthesis by ribosomes.

Chapter 11 –Sensitivity and uncertainty analysis– starts by explaining how the computing experiments are designed and how they can be conducted on a high-performance cluster to explore the parameter space. The chapter distinguishes between sensitivity (how much an outcome variable changes in response to parameter variation) and uncertainty (the inherent variability in simulation outcomes due to stochastic effects). The chapter begins by defining uncertainty in the context of stochastic processes and statistical inference, emphasizing the importance of quantifying uncertainty to improve model reliability. It explores through *embarrassingly parallel* simulations, the impact of parameter variability and their interaction on model predictions in translation dynamics. Eight sensitivity basic observations were derived from computational simulations highlighting the critical factors affecting translation efficiency and the proteome output landscape.

12.1.2 Main contributions, key insights and thesis legacy

Throughout this thesis, we aimed to develop a computational tool to address the primary question and several secondary questions. The primary objective was achieved by constructing an integrative agent-based model (ABM) named `Ribosomer`. The

secondary questions focused on evaluating how various contextual factors influence protein synthesis by ribosomes. This evaluation considered their impact not only on the proteome landscape but also on polysome profiles and ribosome profiles at the level of individual transcripts. The ABM enables the computational prediction of ribosome density maps at the scale of individual transcripts.

`Ribosomer`, delivered in two open-source versions as part of **the thesis legacy**, is a computational framework that explicitly incorporates key factors influencing the biological process of protein synthesis. The disentanglement of these intertwined factors was achieved through the modular construction of the agent-based model (ABM).

A key contribution of this thesis is the representation of the elongation cycle as three sequential substeps. Each ribosome, on any codon and any transcript, undergoes an elongation cycle whose duration is governed by a statistical queueing time distribution. The queueing time is sampled from a hypo-exponential distribution, fully determined by the influential factors[1].

The queueing time theory accounts for the inherent stochasticity of the natural protein elongation cycle—the *quantum of reality* we aim to represent and understand[2].

The incorporation of influencing factors into the elongation cycle was motivated and justified based on common mechanochemical arguments, keeping the model conceptually as simple as possible. However, this apparent simplicity at the scale of a single agent gives rise to complex collective behavior when multiple agents interact, producing emergent outcomes that could not be predicted by considering an individual agent alone.

Our results highlight that ribosome pool limitations and the preferential allocation of higher initiation rates to subsets of privileged transcripts are among the most sensitive factors.

Computational simulations further reveal that detecting the effects of codon usage and tRNA modifications at the proteome landscape level is not only challenging but ultimately illusory. This finding is consistent with the long-established rule of *universal scaling* described in [Sharma et al. 2018]. Yet, these effects are predicted to be detectable through ribosome profiling, as long as sequencing depth allows for ribosome density mapping at the level of individual transcripts.

More than ten years ago, researchers already emphasized that accurate computational models of protein translation require an integrative approach [Gritsenko et al. 2015].

[1] Mathematically, the hypo-exponential probability density function is the convolution of three exponentially distributed probability densities, each characterized by a distinct rate parameter. These three rates correspond to the decomposition of the ribosomal elongation cycle into three sub-step mechanisms that biochemists have not yet fully elucidated: (i) accommodation and proofreading, (ii) peptide bond formation, and (iii) translocation.

[2] Understanding in the sense articulated by Richard Feynman: *What I cannot create, I do not understand.*

Our findings align with this conclusion and highlight several key factors:

1. **Translation efficiency requires both initiation and elongation dynamics.**
   Translation efficiency is not fully captured by elongation alone; initiation plays a dominant role. Ribosome profiling data indicate that models must account for both steps to accurately describe translation dynamics.

2. **Codon adaptation to the tRNA pool is insufficient.**
   Codon usage bias alone does not fully explain translation efficiency. tRNA post-transcriptional modifications and variable decoding rates significantly contribute to ribosome elongation dynamics. Our computational simulations show that tRNA modifications at the anticodon loop have more significant effects on Ribo-seq experiment results than on the protein relative abundance results.

3. **Codon context matters beyond individual codon effects.**
   Codon effects extend beyond local interactions, with secondary structures and nascent chain interactions in the ribosome exit tunnel, influencing ribosome progression. These contextual factors must be incorporated into predictive models. Our findings show that codon context add variability and uncertainty in polysome and ribosome profiles.

4. **Ribosome occupancy patterns reveal hidden regulatory features.**
   Ribosome profiling data reveal non-uniform ribosome distributions, indicating that additional regulatory mechanisms, such as initiation rate competition and pausing, shape translation efficiency.

5. **Modeling approaches need to integrate multiple regulatory layers.**
   Simple models based on codon usage or tRNA adaptation fail to fully capture translation dynamics. Integrative models incorporating initiation rates, elongation speed, and context-dependent effects provide better predictive power. TASEP inspired approaches have shown better predictive power as compared to other static models.

6. **Ribosome profiling data interpretation has limitations**
   Sequencing biases and the inability to distinguish between initiation and elongation effects present challenges. Careful normalization and complementary experimental approaches are needed for accurate quantification and to elucidate yet unknown mechanisms. In any case, computational integrative modeling is a crucial tool for interpreting these particularly noisy results. Computational modeling can also extract valuable insights that are currently inaccessible through direct laboratory measurements, such as initiation rates on specific subsets of transcripts and the ribosome pool's size and distribution between free (unassembled mature ribosomal subunits) and actively translating ribosomes.

Our work builds upon these findings. Specifically, we incorporated two context-dependent factors in our quest to address secondary questions related to (i) interactions between the nascent peptide chain and the ribosomal exit tunnel, and (ii) secondary structure effects. Detailed algorithms have been developed for both factors in the thesis. While the first has been fully implemented in our computational model, the second still requires integration into our code, leaving room for future work.

Furthermore, we emphasize the need to include the following finding:

- **The pool of ribosomes matters in the translation efficiency.**
  The ratio of the total number of ribosomes to the total number of transcript copies –the ribosome ratio– is crucial biological information that should be quantified in all laboratory experiments. The fraction of fully assembled ribosomes (80S monosomes or polysomes) engaged in translation, relative to the total number of ribosomes (including free 30S `SSU` and 60S `LSU` subunits in eukaryotes), should be used as a metric to assess translation efficiency. Conversely, translation is considered inefficient if a large proportion of transcript copies are free of ribosomes. Thus, the ratio of transcript copies bound by ribosomes to those that are free should also be a key translation efficiency metric of interest to molecular biologists. The `Ribosomer` integrative ABM can predict these metrics, at least in a stationary state. Conversely, knowing these metrics could help infer relevant biological information –such as the general initiation rate– that is otherwise not measured in the laboratory.

12.2 General reflections on computational biology

12.2.1 The importance of uncertainty quantification

Can we still adhere to Laplace's reductionist view –that if all forces, deterministic causes, and physico-chemical laws were known, we could fully explain the complex data patterns in molecular biology, such as Ribo-seq density maps, ribosome profiles, RNA-seq outputs, and proteomic data in any experimental setting?

Where do we stand between Laplace's determinism and the inherent stochasticity permeating not only fundamental physics but also molecular and structural biology?

History shows that physics did not end with Laplace. Later developments revealed that even systems governed by fully deterministic coupled ordinary differential equations can display deterministic chaos [Lorenz 1963]. Tiny differences in initial conditions or system boundaries can lead to unpredictable, unforeseeable trajectories after some time –a phenomenon famously referred to as the butterfly effect.

Importantly, we don't need to zoom down into atomic-scale physics, where Heisenberg's uncertainty principle renders precise positions and momenta inherently undefined, making individual trajectories nonsensical. Even without reaching this scale, we encounter fundamental uncertainty: **ribosome profiling and Ribo-seq data are intrinsically noisy**.

Laplace also generalized **Bernoulli's law of large numbers**[3] and, in 1810, formulated an early version of the **central limit theorem (CLT)**, later called the **Laplace-Lyapunov theorem**[4]. Interestingly, the stochasticity inherent in ribosomal elongation cycles, when considered over many codons, satisfies CLT assumptions and leads to the **universal scaling behavior of protein synthesis times** [Sharma et al. 2018]. In this sense, both the law of large numbers and the CLT confer **robustness to protein production against local fluctuations in elongation rates**. This explains why variations in codon usage or tRNA modifications –even for transcripts (e.g., KIF family proteins) where up to 15% of codons are sensitive– have little effect on relative protein abundance, though they may significantly alter polysome fragmentation profiles and ribosome density maps.

As Nicolas Ingolia succinctly put it: "*Ribo-seq data are very noisy*"[5]. Noise, variance, and heterogeneity are all manifestations of uncertainty –different faces of the same coin.

There are two principal approaches to quantifying uncertainty.

The inductive method is data-driven, relying on statistical inference. It tells us whether observed differences between control and case are significant but does not necessarily explain them.

Conversely, **the deductive approach is theory-driven**: it builds mechanistic models incorporating known probabilistic structures to generate synthetic data. However, this approach is prone to inductive bias –hypotheses are embedded within the model, and alternative explanations may exist that could reproduce the same data patterns.

Uncertainty is inescapable in both frameworks: the inductive approach estimates it directly from data, while the deductive approach generates it synthetically and compares model predictions to observed outcomes. Crucially, any acceptance criteria for matching synthetic and real data must be grounded in rigorous statistical methods or information theory, such as the Kullback-Leibler divergence criterion.

This dissertation has sought to explore the origins of this "noisy" nature. **The key challenge for the research community is to disentangle inherent stochasticity from causal, deterministic components –ensuring mechanistic explanations reflect the**

[3]The relative frequency of an event converges to its theoretical probability as the number of trials increases (Bernoulli, 1713).

[4]The explicit statement of the Central Limit Theorem is provided in the appendix of this thesis, within the section on statistical queueing time theory.

[5]Nicolas Ingolia (September 2022), personal communication during a poster session at the Cold Spring Harbor Conference on Translational Control, Long Island, NY, USA.

underlying truth and are not confounded by false positives, noise, or spurious correlations as exposed in Chapter 1 [Larsson et al. 2010].

Addressing this requires constructing probabilistic models governing the behavior of biological agents and embedding them, for instance, in agent-based models (ABMs), where mechanistic rules are explicitly separated from stochastic components[6]. In the synthetic model, these elements are distinct, but the predicted outcomes naturally blend them –only the outcomes are compared to experimental data. **This bottom-up approach is not data-driven but hypothesis-driven: starting from a proposed causal explanation and embedding uncertainty within the model**. If the model reproduces the observed data patterns, we can claim to have achieved a reasonable understanding of the phenomenon –following Richard Feynman's dictum: "*What I cannot create, I do not understand*." By constructing an ABM that replicates real data, we lift a corner of the veil concealing biological reality.

Regardless of the approach, **identifying and quantifying uncertainty is essential**. **Ignoring it risks false positives and spurious correlations** [ibid.].

12.2.2 The importance of sharing and accessing data

Genomic and transcriptomic data deposition in public repositories is occurring at a rate of hundreds of terabytes (TB) to petabytes (PB) annually, with proteomics (several petabytes per year) and Ribo-Seq–translatomic (several terabytes per year) following closely behind, though at a slightly slower pace. The data production and deposition rates are expected to continue to accelerate as sequencing and proteomic technologies advance and become more widespread.

The sharing and accessibility of data are crucial for advancing research and fostering collaboration. However, comprehensive and educated access to this data still lags behind the rapid pace of its production. Even more concerning, our ability to analyze the data falls far behind our capacity to generate it. Fortunately, collaborative platforms such as ELIXIR and others are playing a pivotal role in bridging this gap. Initiatives like ELIXIR, in Europe, and the Global Alliance for Genomics and Health (GA4GH) are streamlining data-sharing practices, promoting more accessible and faster deposition of data. The ELIXIR initiative, for example, integrates Europe's life science resources into a unified infrastructure, enabling seamless access to a wide array of biological data [ELIXIR 2025; ELIXIR Belgium 2025]. These platforms are vital for maintaining the momentum of research in an era of exponential data growth.

Key repositories like the Protein Data Bank (PDB) [Protein Data Bank 1999] and the National Center for Biotechnology Information (NCBI) [National Center for

[6] A digital twin is an invention, not a discovery [Viceconti and Emili 2024].

Biotechnology Information 1992] provide invaluable structural and genetic information, respectively. Ensembl, created by the European Bioinformatics Institute (EMBL-EBI) and the Wellcome Sanger Institute, offers comprehensive genome annotations [Ensembl 1999]. GWIPS-viz, developed in Cork University, serves as a specialized genome browser for ribosome profiling data, enhancing our understanding of translation dynamics [GWIPS-viz 2013].

Integrating such tools with resources like the ViennaNGS package [ViennaNGS 2015], the Sequence Read Archive SRA [National Center for Biotechnology Information 2000], and the European Nucleotide Archive ENA [European Nucleotide Archive 2009] streamlines the analysis of high-throughput sequencing data, promoting comprehensive insights across various omics disciplines.

Proteogenomic pipeline tools like the `Proteoformer` facilitate the identification and analysis of protein variants, enriching our understanding of proteomic complexity [BioBix Lab, Ghent University 2015; Crappé et al. 2014; Verbruggen et al. 2019].

In the realm of cancer research, data-sharing platforms are indispensable for advancing our understanding and treatment of the disease. The Cancer Genome Atlas TCGA [The Cancer Genome Atlas 2007], a collaborative effort initiated by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), has been pivotal in this endeavor. By utilizing high-throughput genomic analysis techniques, TCGA has cataloged genetic mutations responsible for various cancers, thereby enhancing diagnostic and therapeutic strategies.

Complementing TCGA, the Human Cell Atlas HCA project aims to create comprehensive reference maps of all human cells, providing insights into cellular functions and disease mechanisms. Recent advancements in HCA have led to detailed maps of cells in the mouth, stomach, intestines, and those involved in bone and joint development, offering valuable data for understanding diseases such as cancer.

The Virtual Physiological Human VPH initiative aims to facilitate the development of integrative models of human physiology, advancing personalized medicine [VPH Institute 2025b]. The EDITH project –Ecosystem for Digital Twins in Healthcare, coordinated by Professor Liesbet Geris, focused on an inclusive ecosystem for digital twins in healthcare across Europe, developing a strategic roadmap for their integration and widespread adoption [VPH Institute 2025a].

Collectively, these platforms and initiatives underscore the critical importance of data sharing and accessibility in driving innovation and collaboration within the life sciences. Moreover, they illustrate that the research community now has access to an unprecedented volume of biological data. However, it is increasingly apparent that the availability of data exceeds its effective utilization. Many datasets and repositories remain underexploited, with their full potential yet to be harnessed by the broader research community. Our ability to analyze data has not kept pace with our capacity to

generate it, highlighting the need for models that can bridge this gap.

12.2.3 The importance and role of *in-silico* modeling as an integrative framework for theory building and mechanisms elucidation

The opening quotations in this final chapter stress the challenge of navigating the modern information age. It emphasizes that while we now have unprecedented access to vast amounts of data and information, true knowledge –the ability to analyze, synthesize, and apply information meaningfully and consistently– remains much harder to attain.

In computational biology, extracting knowledge from the data repositories and datasets enumerated previously still remains a major challenge. **There is a bottleneck transforming data into actionable knowledge**. This is why integrative, computational modeling and frameworks that aim at mechanisms elucidation are important and have a role to play in this huge numerical ecosystem.

We are not talking here of statistical inference, machine learning (ML) or data-driven artificial intelligence (AI) who promise to extract hidden patterns from the data. Without biological interpretability, ML/AI risk being black-box solutions that lack mechanistic insight. We need to make sense of the data as articulated above in the maxim by Richard Feynman.

We are in need of bottom-up, hypothesis-driven approaches and theoretical frameworks for knowledge construction. The `Ribosomer` agent-based model serves as an example of a computational simulation tool that can produce synthetic data to be compared to experimental data deposited in repositories for various purposes:

i. Parameter calibration or optimization;

ii. Validation of model predictions against well-documented datasets;

iii. Cross-validation and performance evaluation across different datasets or environments (training-testing evaluation cycles).

We are in need for cross-disciplinary synthesis. The last two opening quotes of this general discussion chapter underline the importance of integrative approaches that combine data, theory, and computation. Computational biology must move beyond data collection and focus on developing predictive, mechanistic models that generate knowledge, not just information. This doctoral research –combining computational modeling, statistical physics, ribosome biology, and chemistry– exemplifies this approach.

12.3 Perspectives and future work

12.3.1 Perspectives and biotechnology applications

The elucidation of the role and systemic impact of tRNA modifications or relative tRNA abundance in pathological conditions, such as tumors and neurological disorders, is of significant interest in biomedical research and the development of targeted therapies [Torres et al. 2014, Rapino et al. 2017, Rapino et al. 2021, El Hachem et al. 2024].

The computational biology approaches and methods presented in this PhD dissertation are also of interest to the biotechnology, cancer therapy, and pharmaceutical industries, particularly in the domains of anticancer mRNA vaccines and the production of antigenic synthetic proteins. The correct co-translational folding of a protein and the optimization of enzymatic activity depend on the dynamics of mRNA translation. This relationship is well-documented in the literature [Komar et al. 1999, Yu et al. 2015, Buhr et al. 2016, Komar 2019].

There is substantial evidence that changes in synonymous codon usage can influence protein folding, enzymatic activity, and antigenicity by altering elongation rates, which in turn affect epitope configuration. In this context, the agent-based model of protein synthesis developed in this dissertation, alongside species-specific databases of codon usage and tRNA adaptation indices (CAI and tAI), serves as a valuable tool for the rational design of in vitro vaccine production.

In vitro vaccine production systems utilize cell-free extracts containing ribosomes to translate synthetic mRNA candidates and generate antigenic polypeptides. The computational framework developed in this work complements experimental techniques such as RNA-seq, Ribo-seq, and protein mass spectrometry, which are widely employed in vaccine research and development. Notably, *in silico* simulations offer a strategic advantage by facilitating the selection of optimal heterologous expression vectors for large-scale vaccine production.

12.3.2 What comes next for the Ribosomer integrative platform?

To enhance the predictive power of ribosome-driven protein synthesis dynamics, we incorporated two context-dependent factors: (i) interactions between the nascent peptide chain and the ribosomal exit tunnel, and (ii) secondary structure effects. Detailed algorithms for both factors have been developed. While the first has been fully implemented in our computational model, the second still needs to be integrated, offering opportunities for future advancements.

Currently, due to the lack of Ribo-seq data at an appropriate sequencing depth, our ribosome density map predictions have not been compared with independent experimental available results in settings where tRNA modifications have been engineered.

An additional development goal for `Ribosomer` is to incorporate automated optimization techniques for iteratively calibrating initiation rates to align protein abundance predictions with proteomic data, using RNA-seq data from the control sample. The case-control comparison would then serve as a test of `Ribosomer`'s predictive accuracy.

The immediate next step for the `Ribosomer` platform is to apply it to available datasets from human lung cancer cells, where tRNA-modifying enzymes, specifically Adenosine Deaminase Targeting tRNA (ADAT) enzymes, have been knocked out.

Looking further ahead, additional investigations will focus on the large subunit of the ribosome to better characterize the electrostatic environment near the catalytic pocket of the peptidyl transferase center (PTC).

Appendix A

Queueing time statistical theory

A.1 Sum of exponentially distributed random variables with arbitrary parameters, hypo-exponential, gamma and exponentially modified Gaussian density distributions

This appendix reviews important mathematical results of the probability and statistics theory related to the density of sums of independent random variables having each a specific density distribution.

In all probability and statistics introductory courses, the Gaussian distribution plays a key role because of the so-called Central Limit Theorem (CLT). The strong version of the Central Limit Theorem (Laplace-Lyapounov) asserts that if a random variable X is the sum of a large number of independent random variables X_i, each with unknown arbitrary distinct and standardized distributions (re-centered and re-scaled), then $X = 1/n \sum_{i=1}^{n} X_i$ is approximately normally distributed and $X \sim N(x; \mu = 0, \sigma^2 = 1)$ when $n \to \infty$. The CLT theorem does not tell how fast the convergence is to the normal distribution. In fact, for a limited number of independent contributing random variable terms, the normal distribution may not be a good approximation at all, especially if the contributing variables have asymmetric distributions (one-sided tails or skewed to one direction).

Here, we want to focus on cases where the independent random variables taken as the

terms of the sum are present in a limited finite number, e.g. 2, 3 or 5 contributing independent distributions (not a number close to infinity); may have different pairwise means or variances; and may be highly skewed. We are not looking for an asymptotic approximation but for exact or quasi-exact results. More specifically, we re-derive below the hypo-exponential probability density function as being that for the sum of independent exponentials having pairwise distinct arbitrary parameters. We further assess whether highly skewed classical distributions with 2 (or 3) parameters such as the (shifted) Gamma distribution or the exponentially modified Gaussian distribution or the (shifted) hypo-exponential distribution can be suitably fitted to one another. The quality of the fit will be assessed by three criteria, i.e., the Kullback-Leibler (KL) divergence criterion, the Akaike Infomartion Criterion (AIC) and the Bayesian Information Criterion (BIC).

Exponentially distributed random variables are prevalent in the applied fields of probability and in stochastic modeling. One of the most used of such models, the Poisson process, the interarrival times of events are independent exponential random variables as a consequence of its postulates [Levy 2020; Ross (2014)]. Combining such processes in the development of his method of stages (or steps or phases) in queueing models, the Danish statistician Agner Erlang was led to introduce what is now the familiar Erlang distribution in the context of assessing the reliability of communication call centers between 1909 and 1920 [Erlang 1948]. The Erlang distribution is the distribution of an integer sum of independent and identically distributed (*iid*) exponentials with a common rate parameter λ ($\lambda_i = \lambda$, $\forall i$). The Erlang distribution is a particular case of a Gamma distribution for which the shape parameter α is an integer number and the scale parameter $\beta = 1/\lambda$, the inverse of the rate parameter λ, is such that $\beta = 1/\lambda = \sum_{i=1}^{n} 1/\lambda_i$ of the contributing *iid* exponential terms.

A.1.1 Probability density function for the sum of random variables as a convolution product of the probability density functions of the terms in the sum

Let X_f and Y_g be two independent random variables having, as probability distribution functions, respectively $f(x)$ and $g(x)$; then the random variable Z being the sum of X_f and Y_g has a probability density function which is the convolution product of $f(x)$ and $g(x)$:

$$f_Z(x) = (f \star g)(x) = \int_{\mathbb{R}} f(y) \cdot g(x - y) dy \tag{A.1}$$

Note that X_f and Y_g need not to be identically distributed. Formula A.1 can be generalized to more than two random variables and also can be generalized in $\mathbb{R}^n$. Both commutativity and associativity hold for the convolution product operation.

If the family of the distribution for the convolution product belongs to the same family as the contributing distributions in the sum of the random variables, this family is said to be stable. By definition, a distribution is called stable if a linear combination of two random variables drawn from it also has the same distribution (up to location and scale parameters). For instance, a sum of Poisson distributed random variables is still a Poisson random variable. This also holds for the sum of Gaussian random variables, hypo-exponentially distributed random variables (if all the contributing rates are pairwise distinct), chi-squared distributed random variables, Erlang distributed random variables, Gamma distributed random variables. All these families are stable. But the exponentially modified Gaussian family is not stable.

A.1.2 Poisson process, exponential distribution and the memoryless property

Poisson experiments are experiments where the number of outcomes occurring during a given time interval, or in a specified region, are counted. The given time interval may be of any length. The number of outcomes is a random variable called a Poisson random variable and its probability distribution is a Poisson density function. The Poisson distribution is discrete as the outcome count takes discrete integer values. A Poisson experiment is derived from the Poisson process which has the three following properties:

1) The number of outcomes occurring in one time interval (or specified region of space) is independent of the number that occur in any other disjoint time interval (or region). It is said that Poisson process has no memory.

2) The probability that a single outcome will occur during a very short time interval (or in a very small region) is proportional to the length of the time interval and does not depend on the number of outcomes occurring outside this time interval.

3) The probability that more than one outcome will occur in such a short time interval is negligible.

The mean number of outcomes is computed from $\mu = \lambda t$ where t is the specific unit time interval. So, λ is the mean number of outcomes per unit of time and is called the rate (of occurrence of outcomes). From the three above properties, it follows that the Poisson probability density function is given by [Ross (2014)]:

$$f_{X_{POISSON}}(x;\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \tag{A.2}$$

$$f_{X_{POISSON}}(x;\mu) = \frac{e^{-\mu}\mu^x}{x!} \tag{A.3}$$

The salient feature of the Poisson distribution is that it depends on a single parameter μ or the rate λ and that the mean, μ, is equal to the variance, $\sigma^2 = \mu$ (for the Poisson distribution).

The exponential probability density function is related to the Poisson process. Consider a random variable X_{EXP} describing the time required for the first Poisson event to occur in a Poisson process. Using the Poisson distribution, we find that the probability of no events occurring in the span up to time t is given by:

$$P(X_{POISSON} = 0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} \tag{A.4}$$

$$= e^{-\lambda t} \tag{A.5}$$

The probability that X_{EXP} will exceed x is the same as the probability that no Poisson event occurs within the time span x, the latter was just given by $e^{-\lambda x}$. Hence,

$$P(X_{EXP} > x) = e^{-\lambda x}. \tag{A.6}$$

Thus the cumulative distribution function (CDF) for X_{EXP} is given by:

$$P(0 \leq X_{EXP} \leq x) = 1 - P(X_{EXP} > x) \tag{A.7}$$

$$CDF(x) = 1 - e^{-\lambda x}. \tag{A.8}$$

The probability density function for X_{EXP} results from the derivative of its CDF:

$$f_{X_{EXP}}(x) = \lambda e^{-\lambda x}. \tag{A.9}$$

Let X_i be a random variable having the exponential distribution with rate parameter $\lambda_i > 0$. Then its probability density function, $f_{X_i}(t)$, is given by:

$$f_{X_i}(t) = \begin{cases} \lambda_i e^{-\lambda_i t} & t \geq 0 \\ 0 & t < 0. \end{cases} \tag{A.10}$$

The main descriptive statistics parameters of the exponential density are:

$$\mathbb{E}(X_{EXP}) = \beta = 1/\lambda \tag{A.11}$$

$$\text{VAR}(X_{EXP}) = \beta^2 = 1/\lambda^2 \tag{A.12}$$

$$\text{skewness} = 2 \tag{A.13}$$

$$\text{Fisher kurtosis} = 6. \tag{A.14}$$

The median is $\ln 2/\lambda$, which is always smaller than the mean. The exponential distribution is asymmetric with a heavy one-sided right tail (skewed to the right).

The exponential distribution has a memoryless property (lack of memory). This lack of memory means that the conditional probability for X being larger than $t + t_0$ given that X was larger than t_0 is equal to the probability that X was larger than t. So if the waiting time or queueing time 'makes it' to t_0, the probability of waiting an additional t time is the same as the probability of waiting t time in the first place (as if you would have reset the queueing time from the start). There is no 'benefit' from patience or no 'punishment' through wear that may have ensued for lasting the first t_0. Thus the exponential distribution is appropriate when the memoryless property is justified. When patience pays off or ageing effects occur, other distributions than the exponential distribution are more appropriate as will be seen in the next subsections.

A.1.3 Erlang and Gamma distributions and the loss of the memoryless property

Before we proceed to the Gamma distribution and its particular case, the Erlang distribution, the Euler gamma function must be recalled. The Euler's gamma function is defined by:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x} dx, \quad \text{for } \alpha > 0. \tag{A.15}$$

The following properties hold for the gamma function:

(a) $\Gamma(n) = (n-1)!$ for positive integer n.

(b) $\Gamma(1) = 1$.

(c) $\Gamma(1/2) = \sqrt{\pi}$

The Gamma probability density function is defined by:

$$f_{X_{GAMMA}}(x;\alpha,\beta) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x \geq 0 \\ \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{A.16}$$

where the shape parameter is a positive real $\alpha > 0$ and the scale parameter is a positive real $\beta = 1/\lambda > 0$. The main descriptive statistics parameters of the Gamma density are:

$$\mathbb{E}(X_{GAMMA}) = \alpha/\beta \tag{A.17}$$

$$\mathrm{VAR}(X_{GAMMA}) = \alpha/\beta^2 \tag{A.18}$$

$$\text{skewness} = 2/\sqrt{\alpha} \tag{A.19}$$

$$\text{Fisher kurtosis} = 6/\alpha \tag{A.20}$$

The median has no analytically closed form but can be calculated by numerical quadrature.

The maximum likelihood estimators for the shape and rate parameters of the Gamma distribution when a data sample of size n is available are:

$$\hat{\beta} = \frac{n^2}{n\sum_{i=1}^{n} x_i \ln x_i - \sum_{i=1}^{n} \ln x_i \sum_{i=1}^{n} x_i}, \quad \text{biased} \tag{A.21}$$

$$\tilde{\beta} = \frac{n-1}{n}\hat{\beta}, \quad \text{unbiased} \tag{A.22}$$

$$\hat{\alpha} = \frac{\hat{\beta}}{n}\sum_{i=1}^{n} x_i, \quad \text{biased} \tag{A.23}$$

$$\tilde{\alpha} = \hat{\alpha} - \frac{1}{n}\left(3\hat{\alpha} - \frac{2}{3}\frac{\hat{\alpha}}{\hat{\alpha}+1} - \frac{4}{5}\frac{\hat{\alpha}}{(\hat{\alpha}+1)^2}\right), \quad \text{unbiased.} \tag{A.24}$$

Assumption: the data are supposed to have been sampled from a Gamma distribution. Note the differences in the above formula with respect to the classical formula for data that would have been sampled from a Gaussian distribution.

The exponential distribution is a particular case of a Gamma distribution when $\alpha = 1$. The Erlang distribution is a particular case of the Gamma distribution when α is a positive integer. The Erlang distribution results from the sum of α *iid* exponentials having all a common rate $\lambda = 1/\beta$. Erlang distributions are the distributions of the waiting time for a specific number (more than one) of Poisson events to occur with the assumption that all these Poisson events have the same rate λ. Erlang distributions are stable.

The Gamma distribution can be viewed as a generalization of the Erlang when α, the shape parameter is not restricted to be an integer but is extended to real positive values.

The memoryless property does not hold for the Erlang or the Gamma distributions. If a queueing time or an event waiting time is a result of a gradual or slow wear or results from multiple time steps that need to proceed sequentially, then the memoryless property does not hold and Gamma distributions or hypo-exponential distributions can be appropriate stochastic models in these situations.

A.1.4 Hypo-exponential density as a sum of independent exponentials having arbitrary pairwise distinct parameters

The sum of n mutually independent exponential random variables, X_i, with pairwise distinct parameters, $\lambda_i, i = 1, ..., n$, respectively, has the hypo-exponential probability density function, given by [Ross (2014)]:

$$f_{HYPO}(t) = \Big(\prod_{i=1}^{n} \lambda_i\Big) \sum_{j=1}^{n} \frac{e^{-\lambda_j t}}{\prod_{k=1,\neq j}^{n} (\lambda_k - \lambda_j)}, \qquad t \geq 0. \tag{A.25}$$

The condition that the λ_i's be distinct is essential as the formula A.25 is undefined for any instance where $\lambda_i = \lambda_j$ for $i \neq j$. In the case where $n = 3$, we explicitly have:

$$f_{HYPO}(t) = \lambda_1\lambda_2\lambda_3\left(\frac{e^{-\lambda_1 t}}{(\lambda_3 - \lambda_1)(\lambda_2 - \lambda_1)} + \frac{e^{-\lambda_2 t}}{(\lambda_3 - \lambda_2)(\lambda_1 - \lambda_2)} + \frac{e^{-\lambda_3 t}}{(\lambda_2 - \lambda_3)(\lambda_1 - \lambda_3)}\right). \tag{A.26}$$

The main descriptive statistics parameters of the hypo-exponential density are:

$$\mathbb{E}(X_{HYPO}) = \sum_{i=1}^{n} 1/\lambda_i \tag{A.27}$$

$$\text{VAR}(X_{HYPO}) = \sum_{i=1}^{n} 1/\lambda_i^2 \tag{A.28}$$

$$\text{skewness} = 2\left(\sum_{i=1}^{n} 1/\lambda_i^3\right) \Big/ \left(\sum_{i=1}^{n} 1/\lambda_i^2\right)^{3/2} \tag{A.29}$$

$$\text{Fisher kurtosis} = \text{no simple closed form.} \tag{A.30}$$

The median has no analytically closed form but can be calculated by numerical quadrature.

Hypo-exponential distributions are stable if all the contributing exponentials have pairwise distinct rates.

It can easily be shown that the memoryless property does not hold for the distribution of the sum of two independent exponential distribution [Oguntunde et al. 2014].

A supplementary material `.mp4` animation shows the geometric interpretation of the convolution product to calculate the probability distribution function resulting from the sum of three exponentially distributed random variables.

A.1.5 Exponentially modified Gaussian density

The sum of a normally distributed random variable having the two parameters (mean μ and variance σ^2), with an independent exponentially distributed random variable having one parameter (rate λ), is a random variable that has an exponentially modified Gaussian probability distribution (EMG) having 3 parameters μ, σ, λ. The probability density function resulting from this sum of random variables is expressed by:

$$f_{EMG}(x) = \frac{\lambda}{2} e^{\frac{\lambda^2\sigma^2}{2}} e^{-\lambda(x-\mu)} \operatorname{Erfc}\left[\frac{1}{\sqrt{2}}\left(\lambda\sigma - \frac{x-\mu}{\sigma}\right)\right] \tag{A.31}$$

where $\operatorname{Erfc}(x)$ is the complement of the error function:

$$\operatorname{Erfc}(x) = 1 - \operatorname{Erf}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{A.32}$$

$$= \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt. \tag{A.33}$$

The main descriptive statistics parameters of the exponentially modified Gaussian density are:

$$\mathbb{E}(X_{EMG}) = \mu + 1/\lambda \tag{A.34}$$

$$\operatorname{VAR}(X_{EMG}) = \sigma^2 + 1/\lambda^2 \tag{A.35}$$

$$\text{skewness} = \frac{2}{(\sigma\lambda)^3}\left(1 + \frac{1}{\sigma^2\lambda^2}\right)^{-3/2} \tag{A.36}$$

$$\text{Fisher kurtosis} = \frac{3\left(1 + \frac{2}{\sigma^2\lambda^2} + \frac{3}{\sigma^4\lambda^4}\right)}{\left(1 + \frac{1}{\sigma^2\lambda^2}\right)^2} - 3. \tag{A.37}$$

The median has no analytically closed form but can be calculated by numerical quadrature.

Exponentially modified Gaussian random variables are not stable.

The exponentially modified Gaussian distribution is used in chromatography as a theoretical model for the shape of the peak of a given chemical solute during elution on a chromatographic column [Grushka 1972]. It was also applied as a statistical model of the intermitotic time in dividing cells and of the ribosome residence time on specific codons during `mRNA` translation [Dana and Tuller 2014].

A.1.6 A note on the number of the required parameters to determine the density of a distribution, their relation to the mean, variance, skewness and definition domains

A noticeable shortcoming of the exponentially modified Gaussian distribution is that the support of this density is $\mathbb{R}$, i.e. negative or positive real values. The negative values do not make sense in the framework of a queueing time theory as a queueing time is always positive. The exponential distribution, the hypo-exponential distribution and the Gamma distribution have their support on $\mathbb{R}^+$: the probability space is the set of all positive or null real values. This is fully consistent with a queueing time theory. The Gamma distribution requires 2 (3 for shifted gamma) parameters to be unequivocally defined. The hypo-exponential distribution requires 3 parameters (4 if shifted). The exponentially modified Gaussian requires 3 parameters. In all these distributions, the mean and variance are not independent, in contrast to the Gaussian distribution. The mean and the variance are not sufficient to determine these distributions, except for the Gamma distribution (2 parameters). Hence, the measure of asymmetry, i.e., the skewness, of these distributions is the third parameter that can help specifying them completely. In this respect, it is informative to know that the skewness of the exponential distribution is always equal to 2, while the skewness of the Gamma distribution is $2/\sqrt{\alpha}$, and the skewness of the hypo-exponential and the exponentially modified Gaussian are dependent on their other parameters in complex expressions. The skewness of a symmetric distribution is equal to zero. It is also extremely important to be aware that the classical sample mean and variance are not necessarily equal to the maximum likelihood estimators of the true distribution parameters as would be the case for normally (Gaussian) distributed data. Refer for instance to equations (A.21-A.24) for the expressions of the maximum likelihood estimators of the Gamma distribution.

A.1.7 Brute force mutual comparison of distributions and quality of the fit of a (shifted) hypo-exponential or a (shifted) Gamma to the exponentially modified Gaussian density

We will conduct the comparison using two comparison criteria exposed in (1) and (2) below. The comparison using the two methods are carried out after that.

1. Kullback-Leibler (KL) divergence criterion.

   In mathematical statistics and information theory, the Kullback-Leibler divergence, also called relative entropy, noted $KL(P||Q)$ is a statistical distance measuring how one reference probability distribution P is different from a second, probability distribution Q [Kullback and Leibler 1951]. A simple interpretation of the divergence of P from Q is the expected excess surprise from using Q as a model when the actual distribution is P. This distance is not a metric, is asymmetric and the triangle inequality does not hold. In the simple case, a relative entropy of 0 indicates that the two distributions in question convey the exact same information.

   Consider two probability distributions P and Q. Usually, P represents the data, the observations, or a measured probability distribution. Distribution Q represents instead a theory, a model, a description or an approximation of P. The Kullback–Leibler divergence is then interpreted as the average difference of the number of bits required for encoding samples of P using a code optimized for Q rather than one optimized for P. Note that the roles of P and Q can be reversed in some situations where it is easier to compute.

   In the discrete case, the Kullback-Leibler divergence is:

   $$\mathrm{KL}(P||Q) = \sum_{x_i \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)} \tag{A.38}$$

   for P and Q defined on the same probability space $\mathcal{X}$. In other words, it is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P.

   For the continuous case, the KL divergence definition is:

   $$\mathrm{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \ln \left(\frac{p(x)}{q(x)}\right) dx. \tag{A.39}$$

2. Akaike information criterion (AIC) and Bayesian information criterion (BIC).

   The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data [Akaike 1974]. Given a collection of models for the data, AIC estimates the quality of each

model, relative to each of the other models. Thus, AIC provides a means for model selection.

AIC is founded on information theory. When a statistical model is used to represent the process that generated the data, the representation will almost never be exact; so some information will be lost by using the model to represent the process. AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.

Suppose that we have a statistical model of some data. Let k be the number of estimated parameters in the model. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the AIC value of the model is the following:

$$\mathrm{AIC} = 2k - 2\ln(\hat{L}). \tag{A.40}$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.

AIC is founded in information theory. Suppose that the data is generated by some unknown process f. We consider two candidate models to represent f: g_1 and g_2. If we knew f, then we could find the information lost from using g_1 to represent f by calculating the Kullback–Leibler divergence directly, $\mathrm{KL}(f||g_1)$; similarly, the information lost from using g_2 to represent f could be found by calculating $\mathrm{KL}(f||g_2)$. We would then, generally, choose the candidate model that minimized the information loss.

We cannot choose with certainty, because we do not know f. Akaike showed, however, that we can estimate, via AIC, how much more (or less) information is lost by g_1 than by g_2. The estimate, though, is only valid asymptotically; if the number of data points is small, then some correction is often necessary.

The BIC is defined as [Schwarz 1978]

$$\mathrm{BIC} = k\ln n - 2\ln(\hat{L}) \tag{A.41}$$

where

- $\hat{L}$, the maximized value of the likelihood function of the model M, i.e. $\hat{L} = p(x \mid \widehat{\theta}, M)$, where $\widehat{\theta}$ are the parameter values that maximize the likelihood function;
- x, the observed data;
- n, the sample size;

- k, the number of parameters estimated by the model.

How does the exponentially modified Gaussian distribution compare to the hypo-exponential or the Gamma distributions?

We conducted two comparisons:

(a) How similar is the 4-parameter (shifted) hypo-exponential distribution, called Q_1, to the 3-parameter exponentially modified Gaussian, called P?

(b) How similar is the 3-parameter (shifted) Gamma distribution, called Q_2, to the 3-parameter exponentially modified Gaussian, called P?

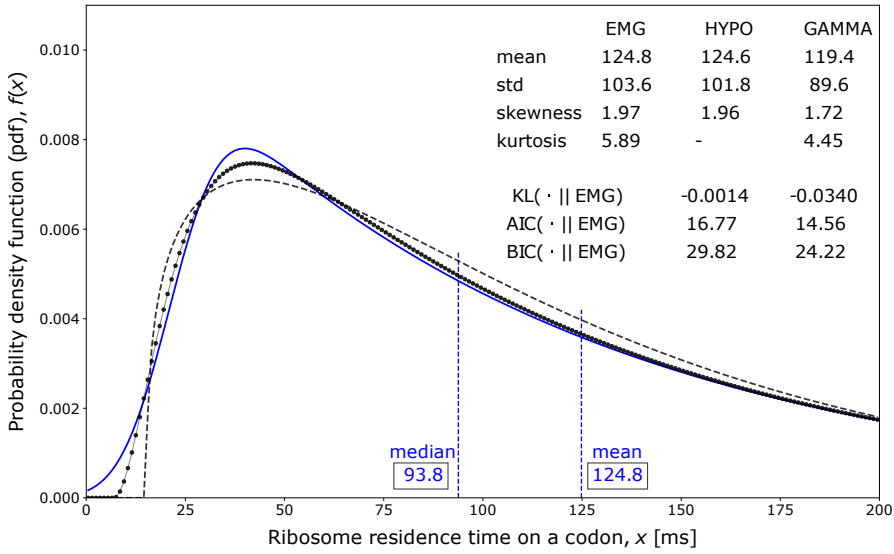


Figure A.1: Best least squared fit of shifted hypo-exponential (black dotted line) or shifted Gamma distribution (dashed line) to the exponentially modified Gaussian distribution (blue line) as empirical model for the ribosome residence time on the GAA codon (in yeast). EMG parameters were taken from the *Saccharomyces cerevisiae* GAA codon reported in Dana and Tuller [Dana and Tuller 2014].

In both cases, we took the same reference EMG density as adapted from the normalized footprint counts of Dana & Tuller to estimate the ribosome residence time on the specific codon **GAA** decoding for Glutamate **Glu-E**, in yeast [Dana and Tuller 2014]. This EMG density, i.e., P, is arbitrary taken as the 'true' data generating model. The Q distributions used in the calculation of the goodness of the fit criteria are the ones for which the parameters were determined either by the maximum likelihood approach or by the non-linear least squared fit obtained with the Levenberg-Marquardt algorithm. The three probability density function plots are represented in Fig.A.1 for comparison. The calculated values of the three information criteria (Kullback-Leibler, AIC and BIC) for the similarity and goodness of the fit are given in the table at the insert of Fig.A.1.

The shifted hypo-exponential distribution essentially conveys the same information as the exponentially modified Gaussian with a Kullback-Leibler divergence $KL(\text{EMG}||\text{shifted} - \text{HYPO}) = -0.0014$. This is slightly better than the Gamma distribution for which $KL(\text{EMG}||\text{shifted} - \text{GAMMA}) = -0.0340$. If the number of parameters of the distributions are taken as a penalty to assess the goodness of the fit, as is the case with AIC and BIC criteria, then the shifted Gamma (3 parameters) ranks better than the shifted hypo-exponential (4 parameters) with respect to the exponentially modified Gaussian distribution (3 parameters). The shifted hypo-exponential extra parameter may cause some overfitting in the goodness of the fit. In any case, any of these three distributions could be used to represent the ribosome residence time data on any specific codon without affecting significantly the conveyed information when moving from one to the other. Other authors have shown that the log-normal distribution can be a good fit as well for the exponentially modified Gaussian distribution [Dana 2014].

A supplementary material **.mp4** animation shows the geometric interpretation of the convolution product to calculate the probability distribution function resulting from the sum of three exponentially distributed random variables. The animation also compares the quality of the fit between the hypo-exponential distribution, the Gamma distribution and the exponentially modified Gaussian distribution.

Appendix B

Supplementary material of the CSBJ01 paper

B.1 Analytical solution for the area of the truncated prolate spheroid as a surface of revolution of a truncated ellipse

This appendix reproduces the first supplementary material of the CSBJ01 paper exposed in chapter 4.

A prolate spheroid is a surface $S \in \mathbb{R}^3$ generated by the revolution of an ellipse about its major axis. A parametric representation of a simple ellipse in the plane Oxy in $\mathbb{R}^3$ having semi-major axis a and semi-minor axis b is

$$\gamma(u) = (a\cos u, b\sin u, 0). \tag{B.1}$$

The parametric representation of the surface of revolution is

$$\phi(u,v) = \big(|\gamma(u) \wedge \vec{e}|\cos(v), |\gamma(u) \wedge \vec{e}|\sin(v), < \gamma(u), \vec{e} > \big). \tag{B.2}$$

for all $(u,v) \in K$, with $K = [u_{lower}, u_{upper}] \times [0, 2\pi]$ and $\vec{e}$ the unit vector about which the revolution takes place. The area of the surface is determined by the formula known in elementary mathematical analysis

$$A_s = 2\pi \int_{u_{lower}}^{u_{upper}} |\gamma(u) \wedge \vec{e}| \sqrt{(D_u|\gamma \wedge \vec{e}|)^2 + (< D_u\gamma, \vec{e} >)^2}\, du. \tag{B.3}$$

Taking the revolution about the semi-major axis, we have $\vec{e} = (1, 0, 0)$ and

$$|\gamma(u) \wedge \vec{e}| = \left| det \begin{pmatrix} a\cos u & b\sin u & 0 \\ 1 & 0 & 0 \\ \vec{e}_x & \vec{e}_y & \vec{e}_z \end{pmatrix} \right| \tag{B.4}$$

$$= |-(b\sin u)\vec{e}_z| = b\sin u. \tag{B.5}$$

$$D_u|\gamma(u) \wedge \vec{e}| = b\cos u \tag{B.6}$$

$$< \gamma(u), \vec{e} >= a\cos u \tag{B.7}$$

$$D_u < \gamma(u), \vec{e} >= -a\sin u \tag{B.8}$$

$$A_s = 2\pi \int_{u_{lower}}^{u_{upper}} b\sin u\sqrt{b^2\cos^2 u + a^2\sin^2 u}\, du \tag{B.9}$$

$$= 2\pi \int_{u_{lower}}^{u_{upper}} b\sin u\sqrt{a^2(1-\cos^2 u) + b^2\cos^2 u}\, du \tag{B.10}$$

Substituting $t = \cos u$, $dt = -\sin u\, du$, we have

$$A_s = 2\pi\, b \int_{\cos u_{upper}}^{\cos u_{lower}} -\sqrt{a^2 - t^2(a^2 - b^2)}\, dt \tag{B.11}$$

$$= 2\pi\, ab \int_{\cos u_{upper}}^{\cos u_{lower}} \sqrt{1 - e^2t^2}\, dt \tag{B.12}$$

where, in the previous line, the eccentricity of the ellipse $e = \sqrt{\frac{a^2-b^2}{a^2}}$ was used. Substituting again $et = \sin w$, $e\,dt = \cos w\,dw$, the last integral turns into

$$A_s = 2\pi \frac{ab}{e} \int_*^* \sqrt{1-\sin^2 w}\cos w\,dw \tag{B.13}$$

$$= 2\pi \frac{ab}{e} \int_*^* \cos^2 w\,dw \tag{B.14}$$

$$= 2\pi \frac{ab}{e} \int_*^* (\frac{1}{2} + \frac{1}{2}\cos 2w)\,dw \tag{B.15}$$

$$= 2\pi \frac{ab}{e} \Big[\frac{w}{2} + \frac{1}{4}\sin 2w\Big]_*^* \tag{B.16}$$

$$= \frac{\pi ab}{e} \Big[w + \frac{1}{2}2\sin w\cos w\Big]_*^* \tag{B.17}$$

$$= \frac{\pi ab}{e} \Big[\arcsin(et) + et\sqrt{1-e^2t^2}\Big]_{t_{lower}=\cos u_{upper}}^{t_{upper}=\cos u_{lower}}. \tag{B.18}$$

The area of the truncated prolate spheroid of semi-major axis $a = 8.9744\,\text{nm}$ and semi-minor axis $b = 2.25\,\text{nm}$ (eccentricity $e = 0.9681$), for t in the range $t_{lower} = \cos u_{upper} = 0.7454$ and $t_{upper} = \cos u_{lower} = 0.9750$, is calculated from this last formula to be $15.75\,\text{nm}^2$. Note that, when $t \in [0,1]$ or $u \in [\frac{\pi}{2}, \pi]$, the surface of the half prolate spheroid is $\frac{\pi ab}{e}\big(\arcsin(e) + e\sqrt{1-e^2}\big) = 102.23\,\text{nm}^2$ with the adopted values of a and b.

Appendix C

Supplementary material of the PRE paper

This appendix reproduces the supplementary material of the PRE paper exposed in chapter 6.

C.1 Hollow straight cylinder model

In a first simplified approach, the ribosome exit tunnel is considered a hollow straight cylinder (Fig. 6.4 (a) left panel). The wall material is not of the conductor type with mobile free charges but is rather a dielectric material harboring fixed partial charges – the fixed phosphate moieties lining the inner wall. As a first reasonable assumption, the fixed charges are supposed to be uniformly distributed on the surface of the inner wall. The size of the hollow cylinder closest to the shape of the ribosome exit tunnel documented in the literature would be $85 - 100$ $(8.5 - 10\,\text{nm})$ in length and $10 - 20$ $(1 - 2\,\text{nm})$ in diameter [Dao Duc et al. 2019; Voss et al. 2006]. The precise length for the ribosomal exit tunnel as measured by cryo-electron microscopy is 9.2 nm on average in prokaryotes and 8.3 nm on average in eukaryotes [Dao Duc et al. 2019]. The *in vivo* lengths are believed to be a bit larger due to thermal dilatation at the higher temperatures prevailing in living organisms as compared to the cryogenic conditions.

For a given uniformly distributed charge density σ on the inner surface wall of the cylinder, the determination of the electrostatic scalar potential $\Phi(\vec{\mathbf{r}})$ and of the electric field $\vec{\mathbf{E}}(\vec{\mathbf{r}})$, at any spatial point close to or far away from the cylindrical surface, are well stated problems in classical electromagnetism [Jackson (1998)]. For the sake of

simplicity, we restrict ourselves here on spatial points located on the axis of the hollow cylinder, lying anywhere inside or outside of the tunnel. In this schematic pictorial description, a new amino acid is incorporated into the nascent protein which gets into the tunnel from one side (conventionally from the right of Fig. 6.4 (left panel). The nascent oligopeptide is then pushed by the multi-tasking ribosomal enzymatic functions inside the tunnel and out of the tunnel at the other side (left side of Fig. 6.4) of the hollow cylinder. The movement is strictly asymmetric as the nascent protein always enters the tunnel from the same side with the amino terminal end of the protein getting in first and the carboxy terminal end of the protein getting in last. Under this idealized model, the hollow cylinder itself is symmetric and has a uniform charge distribution.

The electrical scalar potential $\Phi(\vec{\mathbf{r}})$ at the observed position $\vec{\mathbf{r}}$ is expressed by:

$$\Phi(\vec{\mathbf{r}}) = \frac{1}{4\pi\epsilon} \int \int_S \frac{\sigma(\vec{\mathbf{r}'})\, da}{|\vec{\mathbf{r}} - \vec{\mathbf{r}'}|} \tag{C.1}$$

where $\sigma(\vec{\mathbf{r}'})$ is the surface-charge density (measured in coulombs per square meter) at position $\vec{\mathbf{r}'}$ of the source, da is the two dimensional surface element at $\vec{\mathbf{r}'}$ and ϵ is the permittivity of the dielectric medium (formula 1.23 in Jackson [Jackson (1998)]) with $\epsilon = \epsilon_r \cdot \epsilon_0$, where ϵ_r is the relative permittivity of the medium and ϵ_0 is the permittivity of free space. We can take advantage of the axial symmetry and restrict to the spatial points on the z-axis, i.e. for $\vec{\mathbf{r}} = (0, 0, z)$. The surface integration is conducted on the support of the source charges. The cylinder's thin wall is geometrically generated by the $\gamma(u)$ curve moving axially along the z-axis from $z = -L$ to $z = 0$ as drawn in Fig. 6.4 (a) (left panel) and where L and R are the length and radius of the hollow cylinder respectively:

$$\gamma(u) = (R\cos u,\ R\sin u, L),\ u \in [0,\ 2\pi]. \tag{C.2}$$

The cylinder's surface is written as $S = \phi(K)$ where $K = \{(u, v) \in [0, 2\pi] \times [-1, 0]\}$ and where $\phi : \mathbb{R}^2 \to \mathbb{R}^3 : \phi(u, v) = (R\cos u,\ R\sin u, vL)$. $D_u\phi$ is the first partial derivative of the parametric equation of the surface $\phi(u, v)$ with respect to u. In the general formula (C.1), the surface-charge density $\sigma(\vec{\mathbf{r}'})$ is dependent of the position $\vec{\mathbf{r}'}$ on the support of the source charges. Here, we will take the simple approximation that σ can be considered a constant parameter over a surface of a given shape, e.g. over a cylinder or over a cone. This is the surface charge uniform distribution assumption for a given shape.

The electrostatic scalar potential results from the surface integral calculation:

$$\Phi(z) = \frac{\sigma}{4\pi\epsilon} \int \int_{K=\{(u,v)\in[0,2\pi]\times[-1,0]\}} \frac{1}{\sqrt{(z - vL)^2 + R^2}} \, |D_u\phi \wedge D_v\phi| \, du\, dv \tag{C.3}$$

$$D_u\phi = (-R\sin u, R\cos u, 0) \tag{C.4}$$

$$D_v\phi = (0, 0, L) \tag{C.5}$$

$$|D_u\phi \wedge D_v\phi| = \left| det \begin{pmatrix} \vec{e_x} & \vec{e_y} & \vec{e_z} \\ -R\sin u & R\cos u & 0 \\ 0 & 0 & L \end{pmatrix} \right| \tag{C.6}$$

$$= |(RL\cos u, RL\sin u, 0)| \tag{C.7}$$

$$= \sqrt{R^2\,L^2\,(\cos^2 u + \sin^2 u)} \tag{C.8}$$

$$= R\,L \tag{C.9}$$

$$\Phi(z) = \frac{\sigma\,R\,L}{4\pi\epsilon}\int_0^{2\pi} du \int_{-1}^{0} \frac{dv}{\sqrt{(z-Lv)^2+R^2}} \tag{C.10}$$

$$= \frac{\sigma\,R\,L}{2\epsilon}\int_{-1}^{0} \frac{dv}{\sqrt{(z-Lv)^2+R^2}} \tag{C.11}$$

$$= \frac{\sigma\,R\,L}{2\epsilon}\int_{-1}^{0} \frac{dv}{R\sqrt{\left(\frac{z-Lv}{R}\right)^2+1}} \tag{C.12}$$

The substitution $w = \frac{z-Lv}{R}$ yields $dw = -\frac{L}{R}dv$ and

$$\Phi(z) = -\frac{\sigma\,R\,L}{2\epsilon}\frac{R}{L}\int_*^* \frac{dw}{R\sqrt{w^2+1}} \tag{C.13}$$

$$= -\frac{\sigma\,R}{2\epsilon}\,\text{arg sinh}(w)\Big|_*^* \tag{C.14}$$

$$= -\frac{\sigma\,R}{2\epsilon}\Big[\,\text{arg sinh}\left(\frac{z-Lv}{R}\right)\Big]_{-1}^{0} \tag{C.15}$$

$$= \frac{\sigma\,R}{2\epsilon}\Big[\,\text{arg sinh}\left(\frac{z-Lv}{R}\right)_{v=-1} - \text{arg sinh}\left(\frac{z-Lv}{R}\right)_{v=0}\Big] \tag{C.16}$$

$$= \frac{\sigma\,R}{2\epsilon}\Big[\,\text{arg sinh}\left(\frac{z+L}{R}\right) - \text{arg sinh}\left(\frac{z}{R}\right)\Big] \tag{C.17}$$

As the arg sinh may be expressed as a logarithm (to prove this, recall that if $x = \sinh y$, $y = \arg\sinh x$ and so $\cosh y = \sqrt{1+x^2}$ whence, $\sinh y + \cosh y = e^y$ and we conclude that $y = \log|x + \sqrt{1+x^2}|$), the electrostatic scalar potential finally writes:

$$\Phi(z) = \frac{\sigma R}{2\epsilon} \log \frac{\left|\frac{z+L}{R} + \sqrt{(\frac{z+L}{R})^2 + 1}\right|}{\left|\frac{z}{R} + \sqrt{(\frac{z}{R})^2 + 1}\right|} \tag{C.18}$$

$$= \frac{\sigma R}{2\epsilon} \log \frac{z + L + \sqrt{(z+L)^2 + R^2}}{z + \sqrt{z^2 + R^2}} \tag{C.19}$$

The electric field projected along the cylinder axis can be computed as the opposite of the scalar potential gradient, i.e. by taking the first derivative with respect to z directly from formula (C.17):

$$E_z = -\vec{\nabla}\Phi(z) \cdot \vec{e}_z \tag{C.20}$$

$$= -\frac{\partial \Phi(z)}{\partial z} \tag{C.21}$$

$$= -\frac{\sigma R}{2\epsilon}\left(\frac{1}{\sqrt{R^2 + (z+L)^2}} - \frac{1}{\sqrt{R^2 + z^2}}\right). \tag{C.22}$$

Of course the axial force applied on a test particle is the product of the axial electric field with the charge of the test particle:

$$F_z = q \cdot E_z. \tag{C.23}$$

The plots of electrostatic scalar potential $\Phi(z)$ and of the axial force F_z acting on a unit test charge located on the tunnel axis at any point of coordinate z are displayed in Fig. C.1, with the medium permittivity prevailing inside the ribosome exit tunnel. A negative force means that the test particle is forced to move towards negative z values whereas a positive force means that the test particle is forced to move towards positive z values. In these plots, σ is adjusted so that the potential fits the range of the experimentally measured values given for instance in Lu *et al.* [Lu et al. 2007].

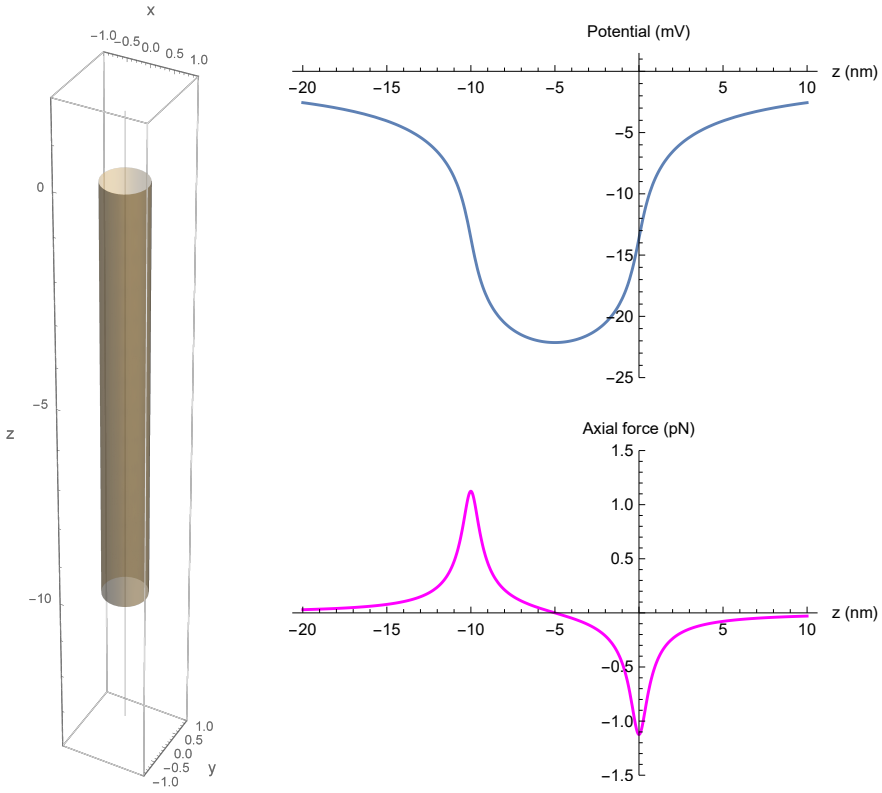


Figure C.1: Electrostatic scalar potential on the axis of the ribosomal exit tunnel (upper panel) and axial force (lower panel) as a function of axial z position for a positively unit charged test amino acid residue on the tunnel axis. Tunnel idealized as a cylinder.

C.2 Normally truncated straight cone model

An alternative approach would depict the tunnel as a hollow cone normally truncated at both ends (Fig. 6.4 (b)). The section radius at the entry point is still equal to R = 0.5 nm but with a section radius twice that value at the tunnel exit point, and equal to R = 1 nm. With the total axial length kept at L = 10 nm, the half opening angle along the axis is $\alpha \sim 0.05$ radian (2.86 arc degrees) and exactly such that $\tan\alpha = \mathrm{R}/\mathrm{L}$ complying with the observation that the diameter at the exit point is around twice the diameter at the entry point of the tunnel. This better reflects the actual geometry of the real ribosomal exit tunnel as reported in the literature [Voss et al. 2006].

To analytically derive the correct equation for the potential and axial electrical field in such a conical tunnel, the procedure is the same as the one previously conducted for the cylinder, but this time with the support of the uniformly distributed charges defined by a cone surface normally truncated at both ends.

The cone's surface is written as $S = \phi(K)$ where $K = \{(u, v) \in [0, 2\pi] \times [-1, 0]\}$ and where $\phi : \mathbb{R}^2 \to \mathbb{R}^3 : \phi(u, v) = \big(R \cdot (1 - \frac{vL\tan\alpha}{R}) \cdot \cos u,\ R \cdot (1 - \frac{vL\tan\alpha}{R}) \cdot \sin u, vL\big)$. The electrostatic scalar potential results from the surface integral calculation:

$$\Phi(z) = \frac{\sigma}{4\pi\epsilon} \int\int_{K=\{(u,v)\in[0,2\pi]\times[-1,0]\}} \frac{1}{\sqrt{(z - vL)^2 + R^2(1 - \frac{vL\tan\alpha}{R})^2}} \, |D_u\phi \wedge D_v\phi| \, du \, dv \tag{C.24}$$

$$D_u\phi = \big(-R(1 - \frac{vL\tan\alpha}{R})\sin u, R(1 - \frac{vL\tan\alpha}{R})\cos u, 0\big) \tag{C.25}$$

$$D_v\phi = (-L\tan\alpha\cos u, -L\tan\alpha\sin u, L) \tag{C.26}$$

$$|D_u\phi \wedge D_v\phi| = \left| \det \begin{pmatrix} \vec{e}_x & \vec{e}_y & \vec{e}_z \\ -R\left(1 - \frac{vL\tan\alpha}{R}\right)\sin u & R\left(1 - \frac{vL\tan\alpha}{R}\right)\cos u & 0 \\ -L\tan\alpha\cos u & -L\tan\alpha\sin u & L \end{pmatrix} \right|$$

$$= \left| \left(RL\left(1 - \frac{vL\tan\alpha}{R}\right)\cos u,\ RL\left(1 - \frac{vL\tan\alpha}{R}\right)\sin u,\ RL\left(1 - \frac{vL\tan\alpha}{R}\right)\tan\alpha \right) \right|$$

$$= \sqrt{R^2L^2\left(1 - \frac{vL\tan\alpha}{R}\right)^2 + R^2L^2\left(1 - \frac{vL\tan\alpha}{R}\right)^2\tan^2\alpha}$$

$$= \sqrt{R^2L^2\left(1 - \frac{vL\tan\alpha}{R}\right)^2(1 + \tan^2\alpha)}$$

$$= R\,L\left(1 - \frac{vL\tan\alpha}{R}\right)\cdot\frac{1}{\cos\alpha} \quad \text{(C.27)}$$

$$\Phi(z) = \frac{\sigma\,R\,L}{4\pi\epsilon\cos\alpha}\int_0^{2\pi} du \int_{-1}^{0} \frac{(1 - \frac{vL\tan\alpha}{R})\,dv}{\sqrt{(z - Lv)^2 + R^2(1 - \frac{vL\tan\alpha}{R})^2}} \quad \text{(C.28)}$$

$$= \frac{\sigma\,R\,L}{2\epsilon\cos\alpha}\int_{-1}^{0} \frac{(1 - \frac{vL\tan\alpha}{R})\,dv}{\sqrt{(z - Lv)^2 + R^2(1 - \frac{vL\tan\alpha}{R})^2}} \quad \text{(C.29)}$$

$$= \frac{\sigma\,L}{2\epsilon\cos\alpha}\int_{-1}^{0} \frac{(R - L\tan\alpha\,v)\,dv}{\sqrt{(z - Lv)^2 + (R - L\tan\alpha\,v)^2}} \quad \text{(C.30)}$$

$$= \frac{\sigma\,L}{2\epsilon\cos\alpha}\int_{0}^{1} \frac{(R + L\tan\alpha\,v)\,dv}{\sqrt{(z + Lv)^2 + (R + L\tan\alpha\,v)^2}} \quad \text{(C.31)}$$

where, in the last line, a dummy integration variable was changed with $\mathrm{v}' = -\mathrm{v} \rightarrow \mathrm{dv}' = -\mathrm{dv}$ and the change of sign was cancelled by the integration limits permutation. The complete derivation is given in the supplemental material following section. To alleviate the notations, the two following substitutions are adopted:

$$f_1(z) = R\cos\alpha - z\sin\alpha \quad \text{(C.32)}$$

$$f_2(z) = R\sin\alpha + z\cos\alpha \quad \text{(C.33)}$$

$f_1(z)$ is always positive for $z \leq 0$ (and even for $z < R/\tan\alpha$, i.e. the virtual z position of the cone summit), which is the domain we are interested in. The z position values are negative in the tunnel and beyond its exit point.

$$\Phi_{\text{cone}}(z) = \frac{\sigma}{2\epsilon}\left\{ f_1(z)\cos\alpha \cdot \log\left[\frac{\left| \frac{L}{\cos\alpha} + f_2(z) + \sqrt{f_1^2(z) + \left(\frac{L}{\cos\alpha} + f_2(z)\right)^2} \right|}{\left| f_2(z) + \sqrt{R^2+z^2} \right|} \right] \right.$$
$$\left. + \sin\alpha \cdot \left[\sqrt{f_1^2(z) + \left(\frac{L}{\cos\alpha} + f_2(z)\right)^2} - \sqrt{R^2+z^2} \right] \right\}$$
$$= \frac{\sigma}{2\epsilon}\left\{ f_1(z)\cos\alpha \cdot \log\left[\frac{\left| \frac{L}{\cos\alpha} + f_2(z) + \sqrt{z^2 + 2L(z+R) + R^2 + \frac{L^2}{\cos^2\alpha}} \right|}{\left| f_2(z) + \sqrt{R^2+z^2} \right|} \right] \right.$$
$$\left. + \sin\alpha \cdot \left[\sqrt{(z+L)^2 + (R + L\tan\alpha)^2} - \sqrt{R^2+z^2} \right] \right\} \quad \text{(C.34)}$$

This last equation (C.34), valid for any conical geometry with entry section of radius R and any cone angle α, replaces equation (C.19) of the cylindrical geometry. Note that the electrostatic potential vanishes at $z = \pm\infty$ as physically expected.

It is also worth noticing that equation (C.34) for the truncated cone restores, as a special case, equation (C.19) for the cylinder when $\alpha = 0$, as expected as well.

The electric field projected along the truncated cone axis can be computed as the opposite of the scalar potential gradient, i.e. by taking the first derivative with respect to z of equation (C.34). The full derivation is provided in the supplemental material

next section and the final result is:

$$
\begin{aligned}
E_{z\,\mathrm{cone}}(z) &= -\vec{\nabla}\Phi_{\mathrm{cone}}(z)\cdot\vec{e}_z \\
&= -\frac{\partial\Phi_{\mathrm{cone}}(z)}{\partial z} \\
&= \frac{\sigma}{2\epsilon}\Bigg\{\sin\alpha\cos\alpha\log\frac{\dfrac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\dfrac{L^2}{\cos^2\alpha}}}{f_2(z)+\sqrt{R^2+z^2}} \\
&\quad + \frac{f_1(z)\cos\alpha\left(\cos\alpha+\dfrac{z}{\sqrt{R^2+z^2}}\right)}{f_2(z)+\sqrt{R^2+z^2}} \\
&\quad - f_1(z)\cos\alpha\frac{\cos\alpha+\dfrac{z+L}{\sqrt{z^2+2L(z+R)+R^2+\dfrac{L^2}{\cos^2\alpha}}}}{\dfrac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\dfrac{L^2}{\cos^2\alpha}}} \\
&\quad - \sin\alpha\left(\frac{z+L}{(z+L)^2+(R+L\tan\alpha)^2}-\frac{z}{\sqrt{R^2+z^2}}\right)\Bigg\}. \qquad \text{(C.35)}
\end{aligned}
$$

Multiplying eq. (C.35) by a positive unit test charge yields the axial forces acting on a positive unit test charge. The plot of the axial forces as a function of the position in the tunnel is displayed in Fig. C.2 (lower panel) for the truncated cone geometry and compared to the cylinder case.

Experimental measurements made on ribosome exit tunnels show that the tunnel exit section radius is around 1 nm, i.e. twice the radius of the innermost part of the tunnel. If the ribosome tunnel were of the cone type, the cone opening angle would be around $\alpha \sim 0.05$ radian (2.86 arc degrees).

The consequence on the electrostatic potential profile is of importance because, with this conical geometry, and if the total charges are kept the same for the two surfaces, the electrostatic potential inside the tunnel will necessarily be algebraically higher than the potential profile in the case of the cylinder as displayed in Fig. C.2 (upper panel) where the analytical equation for the electrostatic potential for the truncated cone was plotted and compared to the cylinder case.

A simple geometrical calculation shows that if the two surfaces support the same total charges $Q_1 = Q_2$, then $\sigma_2 = \mathrm{S}_{\mathrm{cylinder}}/\mathrm{S}_{\mathrm{cone}} \times \sigma_1 = \frac{2}{3} \times \sigma_1$, for a geometry where both tunnels have the same radius at the entry point, the same total lengths L, but where the cone exit section has a radius twice as large as the cylindrical radius. The surface charge density σ_2 on the lateral truncated cone inner surface would be two third of the surface charge density σ_1 prevailing on the lateral inner surface of the cylinder.

Moreover, the potential profile in the conical geometry is skewed to the left as compared to the potential profile for the cylindrical geometry. An asymmetry in the potential profile appears due to the change in radius along the z-axis of the cone. The minimal value of the potential is shifted to the left. The slope of the cylindrical potential profile is steeper than the conical potential at the tunnel exit point, meaning that the electric field intensity will be a bit weaker in that region for the conical geometry as can be seen in Fig. C.2 (lower panel) of the axial forces curves. The axial forces vary more smoothly and are more dispersed in the conical geometry than in the cylindrical geometry.

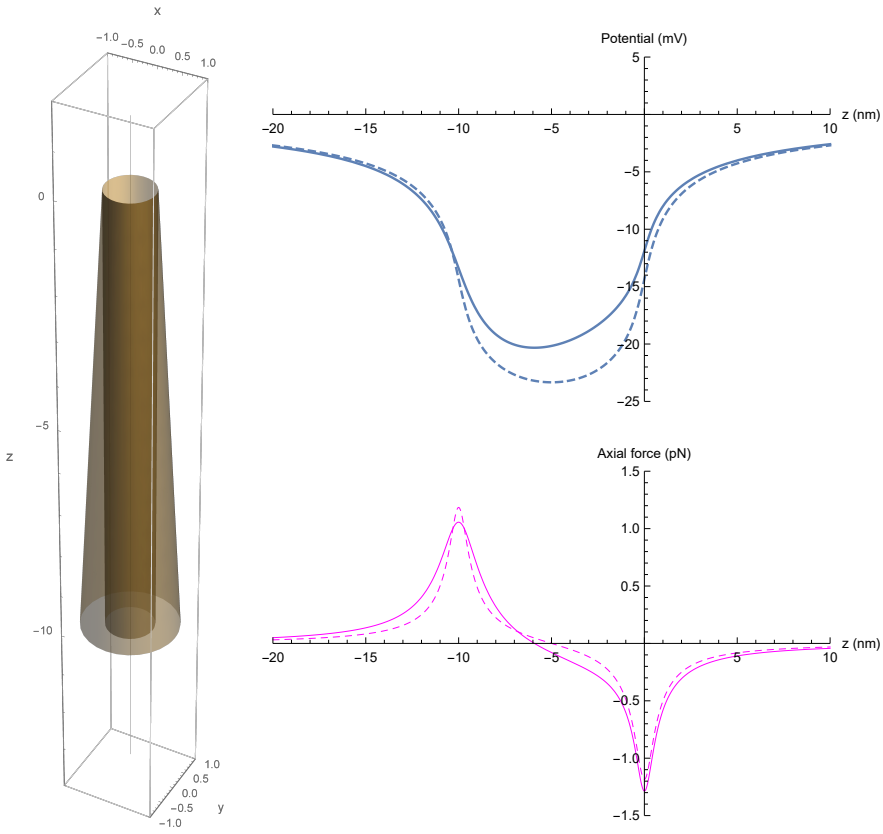


Figure C.2: Electrostatic scalar potential and axial force profiles for a positively unit charged test amino acid residue on the axis of the ribosomal exit tunnel. Comparison of the truncated cone (line) and cylinder (dashed line) geometry with exit section radius of the cone twice as large as the cylinder radius. The lateral surface of the cone is $3/2$ the lateral surface of the cylinder. $\sigma_2 = \sigma_1 \times 2/3$ to keep the same total charges on both surfaces.

C.3 Normally truncated straight cone model complete derivation

C.3.1 Scalar potential

The scalar potential for the normally truncated straight cone model was expressed by equation (C.31) as

$$\Phi_{cone}(z) = \frac{\sigma L}{2\epsilon \cos\alpha} \int_0^1 \frac{(R + L\tan\alpha\, v)\, dv}{\sqrt{(z + Lv)^2 + (R + L\tan\alpha\, v)^2}} \tag{C.36}$$

The expression inside the square root in the denominator of the integrand can be written

$$\begin{aligned}(z + Lv)^2 + \big(R + L\tan\alpha\, v\big)^2 &= z^2 + L^2v^2 + 2zLv + R^2 + L^2\tan^2\alpha\, v^2 + 2RL\tan\alpha\, v \\ &= L^2(1 + \tan^2\alpha)v^2 + 2L(z + R\tan\alpha)v + z^2 + R^2 \\ &= \left[\frac{L}{\cos\alpha}v + \cos\alpha(z + R\tan\alpha)\right]^2 + z^2 + R^2 - \cos^2\alpha(z + R\tan\alpha)^2 \qquad \text{(C.37)}\end{aligned}$$

$$\begin{aligned} &= \big[\,\big]^2 + z^2 + R^2 - \cos^2\alpha\, z^2 - 2zR\sin\alpha\cos\alpha - R^2\sin^2\alpha \\ &= \big[\,\big]^2 + z^2\sin^2\alpha + R^2\cos^2\alpha - 2zR\sin\alpha\cos\alpha \\ &= \big[\,\big]^2 + (z\sin\alpha - R\cos\alpha)^2 \qquad \text{(C.38)}\end{aligned}$$

and so the square root in the above denominator can be rewritten

$$\begin{aligned}\sqrt{(z + Lv)^2 + \big(R + L\tan\alpha\, v\big)^2} &= (z\sin\alpha - R\cos\alpha) \\ &\times \sqrt{\left[\frac{\dfrac{Lv}{\cos\alpha} + (z\cos\alpha + R\sin\alpha)}{z\sin\alpha - R\cos\alpha}\right]^2 + 1} \qquad \text{(C.39)}\end{aligned}$$

To alleviate the notations, we pose as in (C.32) and (C.33)

$$f_1(z) = R\cos\alpha - z\sin\alpha \tag{C.40}$$

$$f_2(z) = R\sin\alpha + z\cos\alpha \tag{C.41}$$

and we pose

$$\begin{aligned} w &= \frac{L\,v/\cos\alpha + (z\cos\alpha + R\sin\alpha)}{R\cos\alpha - z\sin\alpha} \\ &= \frac{L\,v/\cos\alpha + f_2(z)}{f_1(z)} \end{aligned} \tag{C.42}$$

Hence,

$$dw = \frac{L}{\cos\alpha}\,\frac{1}{f_1(z)}dv \tag{C.43}$$

$$dv = \cos\alpha\cdot f_1(z)\frac{1}{L}\cdot dw \tag{C.44}$$

The numerator in the integrand of (C.34) now writes

$$R + L\tan\alpha\, v = R + \big(f_1(z)\,w - f_2(z)\big)\sin\alpha \tag{C.45}$$

and equation (C.31) turns into

$$\begin{aligned} \Phi_{cone}(z) &= \frac{\sigma L}{2\epsilon\cos\alpha}\int_*^* \frac{\frac{1}{L}\Big(R + (f_1(z)w - f_2(z))\sin\alpha\Big)\cos\alpha f_1(z)\,dw}{f_1(z)\sqrt{w^2+1}} \\ &= \frac{\sigma}{2\epsilon}\int_*^* \frac{[R + \Big(f_1(z)w - f_2(z)\Big)\sin\alpha]\,dw}{\sqrt{w^2+1}} \\ &= \underbrace{\frac{\sigma}{2\epsilon}\int_*^* \frac{(R - f_2(z)\sin\alpha)\,dw}{\sqrt{w^2+1}}}_{I} + \underbrace{\frac{\sigma}{2\epsilon}\int_*^* \frac{f_1(z)w\sin\alpha\,dw}{\sqrt{w^2+1}}}_{II} \end{aligned}$$

These two terms easily integrates. The first one (I) is still simplified further through

$$\begin{aligned} R - f_2(z)\sin\alpha &= R - R\sin^2\alpha - z\cos\alpha\sin\alpha \\ &= R\cos^2\alpha - z\cos\alpha\sin\alpha \\ &= \cos\alpha\cdot(R\cos\alpha - z\sin\alpha) \\ &= \cos\alpha\cdot f_1(z) \end{aligned} \tag{C.46}$$

and so,

$$I = \frac{\sigma}{2\epsilon}\int_*^* \frac{\cos\alpha\cdot f_1(z)\,dw}{\sqrt{w^2+1}} \tag{C.47}$$

Substituting $w = \sinh y$, $w^2 + 1 = \cosh^2 y$ and $dw = \cosh y\,dy$, yields

$$\begin{aligned} I &= \frac{\sigma}{2\epsilon}\int_*^* \frac{\cos\alpha\cdot f_1(z)\cosh y\,dy}{\sqrt{\cosh^2 y}} \\ &= \frac{\sigma}{2\epsilon}\int_*^* \cos\alpha\cdot f_1(z)\cdot dy \\ &= \frac{\sigma}{2\epsilon}\,\cos\alpha\cdot f_1(z)\cdot y|_*^* \qquad \text{(C.48)} \\ &= \frac{\sigma}{2\epsilon}\,\cos\alpha\cdot f_1(z)\cdot arg\sinh w|_*^* \qquad \text{(C.49)} \end{aligned}$$

but $\cosh y = \frac{e^y+e^{-y}}{2}$, $\sinh y = \frac{e^y-e^{-y}}{2}$ and $\cosh y + \sinh y = e^y$, so $y = arg\sinh w = \log[\cosh y + \sinh y] = \log(w + \sqrt{1+w^2})$. Hence,

$$\begin{aligned} I &= \frac{\sigma}{2\epsilon}\cos\alpha\, f_1(z)\cdot\log\left[\frac{\frac{Lv}{\cos\alpha}+f_2(z)}{f_1(z)} + \sqrt{1+\frac{(\frac{Lv}{\cos\alpha}+f_2(z))^2}{f_1^2(z)}}\right]_{v=0}^{v=1} \\ &= \frac{\sigma}{2\epsilon}\cos\alpha\, f_1(z)\cdot\log\left[\frac{\frac{L/\cos\alpha+f_2(z)}{f_1(z)} + \sqrt{1+\frac{(L/\cos\alpha+f_2(z))^2}{f_1^2(z)}}}{\frac{f_2(z)}{f_1(z)} + \sqrt{\frac{f_1^2(z)+f_2^2(z)}{f_1^2(z)}}}\right] \\ &= \frac{\sigma}{2\epsilon}\cos\alpha\, f_1(z)\cdot\log\left[\frac{|\frac{L}{\cos\alpha}+f_2(z)+\sqrt{f_1^2(z)+(\frac{L}{\cos\alpha}+f_2(z))^2}|}{|f_2(z)+\sqrt{R^2+z^2}|}\right] \end{aligned}$$

Noticing that

$$f_1^2(z) + f_2^2(z) = R^2\cos^2\alpha - 2Rz\cos\alpha\sin\alpha + z^2\sin^2\alpha$$
$$+ R^2\sin^2\alpha + 2Rz\sin\alpha\cos\alpha + z^2\cos^2\alpha$$
$$= R^2 + z^2 \quad \text{(C.50)}$$

and that

$$f_1^2(z) + \frac{L^2}{\cos^2\alpha} + 2\frac{L}{\cos\alpha}f_2(z) + f_2^2(z) = z^2 + R^2 + \frac{L^2}{\cos^2\alpha}$$
$$+ \frac{2L}{\cos\alpha}\left(R\sin\alpha + z\cos\alpha\right)$$
$$= z^2 + R^2 + \frac{L^2}{\cos^2\alpha} + 2L\left(z + R\tan\alpha\right) \quad \text{(C.51)}$$

results in

$$I = \frac{\sigma}{2\epsilon}\cos\alpha\, f_1(z)\cdot\log\left[\frac{\left|\frac{L}{\cos\alpha} + f_2(z) + \sqrt{z^2 + 2L(z+R) + R^2 + \left(\frac{L}{\cos\alpha}\right)^2}\right|}{\left|f_2(z) + \sqrt{R^2 + z^2}\right|}\right] \quad \text{(C.52)}$$

For the second term (II), substituting $w = \sinh y$, $\cosh^2 y = \sinh^2 + 1 = w^2 + 1$ and $dw = \cosh y\, dy$, we have

$$
\begin{aligned}
II &= \frac{\sigma}{2\epsilon}\int_*^* \frac{f_1(z)\, w\, \sin\alpha\, dw}{\sqrt{w^2+1}} \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon}\int_*^* \frac{f_1(z)\, \sinh y\, \cosh y\, dy}{\sqrt{\cosh^2 y}} \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon}\int_*^* f_1(z)\, \sinh y\, dy \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon} f_1(z)[\cosh y]_*^* \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon} f_1(z)[\sqrt{1+w^2}]_*^* \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon} f_1(z)\left\{\sqrt{1+\left(\frac{\frac{L\,v}{\cos\alpha}+f_2(z)}{f_1(z)}\right)^2}_{v=1} - \sqrt{1+\left(\frac{\frac{L\,v}{\cos\alpha}+f_2(z)}{f_1(z)}\right)^2}_{v=0}\right\} \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon} f_1(z)\left\{\sqrt{1+\left(\frac{\frac{L}{\cos\alpha}+f_2(z)}{f_1(z)}\right)^2} - \sqrt{1+\left(\frac{f_2(z)}{f_1(z)}\right)^2}\right\} \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon}\left\{\sqrt{f_1^2(z)+\left(\frac{L}{\cos\alpha}+f_2(z)\right)^2} - \sqrt{f_1^2(z)+f_2^2(z)}\right\} \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon}\left\{\sqrt{f_1^2(z)+\left(\frac{L}{\cos\alpha}+f_2(z)\right)^2} - \sqrt{R^2+z^2}\right\} \\
&= \frac{\sigma\, \sin\alpha}{2\epsilon}\left\{\sqrt{(z+L)^2+(R+L\tan\alpha)^2} - \sqrt{R^2+z^2}\right\} \qquad \text{(C.53)}
\end{aligned}
$$

Summing the two terms I and II results in equation (C.34).

C.3.2 Electric field

The complete derivation of the electric field projected along the tunnel axis follows from

$$\begin{aligned} E_{z\,cone}(z) &= -\vec{\nabla}\Phi_{cone}(z)\cdot\vec{e}_z \\ &= -\frac{\partial\Phi_{cone}(z)}{\partial z} \\ &= -\frac{\partial I}{\partial z} - -\frac{\partial II}{\partial z} \end{aligned} \tag{C.54}$$

We start with $-\frac{\partial I}{\partial z}$

$$-\frac{\partial I}{\partial z} = \frac{\sigma}{2\epsilon}\Bigg\{ \sin\alpha\cos\alpha\cdot\log\left[\frac{\bullet}{\bullet}\right] + f_1(z)\cos\alpha\cdot\frac{f_2(z)+\sqrt{R^2+z^2}}{\frac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\frac{L^2}{\cos^2\alpha}}}$$

$$\cdot\frac{\partial}{\partial z}\left[\frac{\frac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\left(\frac{L}{\cos\alpha}\right)^2}}{f_2(z)+\sqrt{R^2+z^2}}\right]\Bigg\}$$

$$= \frac{\sigma}{2\epsilon}\Bigg\{ \sin\alpha\cos\alpha\cdot\log\left[\frac{\bullet}{\bullet}\right] + f_1(z)\cos\alpha\cdot\frac{f_2(z)+\sqrt{R^2+z^2}}{\frac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\frac{L^2}{\cos^2\alpha}}}$$

$$\cdot\left[\left(\frac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\left(\frac{L}{\cos\alpha}\right)^2}\right)\left(f_2(z)+\sqrt{R^2+z^2}\right)^{-2}\right.$$

$$\cdot\left(\cos\alpha+\frac{1}{2}(R^2+z^2)^{-1/2}\cdot 2z\right)-\left(f_2(z)+\sqrt{R^2+z^2}\right)^{-1}$$

$$\left.\cdot\left[\cos\alpha+\frac{1}{2}\left(z^2+2L(z+R)+R^2+\frac{L^2}{\cos^2\alpha}\right)^{-1/2}\cdot(2z+2L)\right]\right]\Bigg\}$$

$$= \frac{\sigma}{2\epsilon}\Bigg\{ \sin\alpha\cos\alpha\log\frac{\frac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\frac{L^2}{\cos^2\alpha}}}{f_2(z)+\sqrt{R^2+z^2}}$$

$$+\frac{f_1(z)\cos\alpha\left(\cos\alpha+\frac{z}{\sqrt{R^2+z^2}}\right)}{f_2(z)+\sqrt{R^2+z^2}}$$

$$-f_1(z)\cos\alpha\frac{\cos\alpha+\frac{z+L}{\sqrt{z^2+2L(z+R)+R^2+\frac{L^2}{\cos^2\alpha}}}}{\frac{L}{\cos\alpha}+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+\frac{L^2}{\cos^2\alpha}}}\Bigg\} \quad \text{(C.55)}$$

We go on with $-\frac{\partial II}{\partial z}$

$$
\begin{aligned}
-\frac{\partial II}{\partial z} &= -\frac{\sigma}{2\epsilon}\sin\alpha\cdot\frac{\partial}{\partial z}\left[\sqrt{\bullet}-\sqrt{\bullet}\right] \\
&= -\frac{\sigma}{2\epsilon}\sin\alpha\cdot\left[\frac{1}{2}\left((z+L)^2+(R+L\tan\alpha)^2\right)^{-1/2}\cdot 2(z+L)\right. \\
&\qquad \left.-\frac{1}{2}(R^2+z^2)^{-1/2}\cdot 2z\right] \\
&= -\frac{\sigma}{2\epsilon}\sin\alpha\cdot\left[\frac{z+L}{\sqrt{(z+L)^2+(R+L\tan\alpha)^2}}-\frac{z}{\sqrt{R^2+z^2}}\right] \qquad \text{(C.56)}
\end{aligned}
$$

Summing the two terms yields the final result for $E_{z\,cone}(z)$ as in equation (C.35)

$$
\begin{aligned}
E_{z\,cone}(z) = \frac{\sigma}{2\epsilon}\Bigg\{ &\sin\alpha\cos\alpha\log\frac{L/\cos\alpha+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+L^2/\cos^2\alpha}}{f_2(z)+\sqrt{R^2+z^2}} \\
&+\frac{f_1(z)\cos\alpha\left(\cos\alpha+\frac{z}{\sqrt{R^2+z^2}}\right)}{f_2(z)+\sqrt{R^2+z^2}}- \\
&f_1(z)\cos\alpha\frac{\cos\alpha+\frac{z+L}{\sqrt{z^2+2L(z+R)+R^2+L^2/\cos^2\alpha}}}{L/\cos\alpha+f_2(z)+\sqrt{z^2+2L(z+R)+R^2+L^2/\cos^2\alpha}} \\
&-\sin\alpha\left(\frac{z+L}{(z+L)^2+(R+L\tan\alpha)^2}-\frac{z}{\sqrt{R^2+z^2}}\right)\Bigg\} \qquad \text{(C.57)}
\end{aligned}
$$

C.4 Material and methods for mapping the charged chemical groups on the tunnel inner wall from X-ray solved structures of the ribosomal large subunit

We analyzed the publicly available structure of the large ribosomal subunit of the archeon *Haloarcula marismortui* (PDB code: `4V9F` downloaded from `https://www.rcsb.org/`) obtained from X-ray crystallography at 2.4 Å [Gabdulkhakov et al. 2013]. To find the ribosome exit tunnel and extract the tunnel centerline coordinates, we used a tunnel search algorithm developed by Sehnal *et al* [Sehnal et al. 2013], implemented in MOLE 2.0 and the web-based MOLEonline 2.0 tool publicly available online [Berka et al. 2012; Pravda et al. 2018]. We used PyMOL (The PyMOL Molecular Graphics System, Version 2.3.2) and exported the relevant selected atom positions cartesian coordinates to output files. These files were further processed with mapping algorithms coded in Python to build 2D positional maps of the charged chemical groups on or near the inner surface of the ribosome exit tunnel, as viewed from the tunnel centerline. We used as input starting point to search for the tunnel cavity, the coordinates of one of the 2 non-bridging oxygens bound to phosphorus in the nucleotide G2099 of the 23SrRNA. This point is known to be close to the PTC. Because of the enlargement in the vestibule area and to avoid artifacts in the search algorithm leading to tunnels with a common entry and multiple exits we used as end points of the tunnel cavity, the geometric center of six possible exit points located in the vestibule area, which were the non-bridging oxygen atoms bound to phosphorus of nucleotides C1426, C492, A1442, the `NH2` atom of `R19` in L22, and the `NZ` atoms of `K81` in L24 and `K28` in eL39. In case the search algorithm yielded tunnels with multiple exit points sharing a common entry point, we used the midline obtained at the geometric center of the yielded tunnel centerlines and inspected manually the exit region in PyMOL (The PyMOL Molecular Graphics System, Version 2.3.2). Once a correct tunnel centerline coordinates set was available, we used it as selection reference in PyMOL. We further used Python and Mathematica scripts to calculate different geometrical features of interest, surface charge density estimates and to produce 2D positional maps of the charged moieties as viewed from the tunnel centerline, Fig.6.6. Specifically we translated the crystallographic data model space so that the tunnel entry point would be at the origin and we aligned the direction from the entry point to the exit point along the negative z-axis. We isolated 250 points along the tunnel centerline for which we calculated the curvilinear distance along the tunnel and determined 250 Frenet-Serret frames, i.e the tangent, principal normal and binormal unit vectors forming a right-handed trihedron. We calculated the curvatures and torsions of the three dimensional space curve of the tunnel centerline (not shown here). We used the Frenet-Serret frames to determine 250 normal planes to the centerline three-dimensional

space curve. We algorithmically scanned all the PyMOL selected atoms of interest to calculate the normal distance of the selected atoms to these 250 planes, found the closest intersection points, calculated the radial distance from the centerline and the elevation angle for the virtual line of sight of the selected atoms as viewed from the right-handed trihedron at the centerline points. The elevation angles of the selected atoms were calculated after 2 compound rotations of the Frenet-Serret frames so that the unit tangent was first aligned to the $(0, 0, -1)$ direction, pointing to the negative z-axis; and the principal normal unit vector was then aligned to the $(-1, 0, 0)$ direction, pointing toward the negative x-axis. This procedure is equivalent to a parallel transport of a right-handed trihedron reference frame along the three dimensional space curve of the tunnel centerline when moving from the tunnel entry to the exit points. This parallel transported reference frame is the one used for the angle mapping in Fig.6.6. With this convention, 0 degree points toward the negative x-axis, +90 degrees points toward the positive y-axis, +180 degrees points toward the positive x-axis; whereas -90 degrees points toward the negative y-axis and -180 degrees points toward the positive x-axis. We algorithmically set out the 3D equations of the cylinder and the cone frustum in this reference frame, to calculate the closest distance of the selected atoms to the surface of the model tunnel. The Δ values shown in blue in Fig.6.6 were obtained as the closest (orthonormal) distance of the PyMOL selected atoms either from the model cylinder surface or from the model cone surface, depending on where the atoms are located.

C.5 Comparison of coarse grained electrostatic potential calculated from the observed structural data with the geometrically idealized model

The Coulomb or Yukawa-Debye-Hückel electrostatic potential can be calculated from the X-ray solved exact distribution of the source charges (phosphate moieties and charged amino acids) for which the positional map was shown in Fig.6.6. The dataset includes the exact 3D coordinate positions of a total of 94 charged atoms that are closest to the tunnel centerline.

In the literature, the method to compute the electrostatic potential based on the real observed atom positions belongs to the coarsed-grained modeling methods family. The Yukawa-Debye-Hückel potential is generally used (see for instance Brooks and al 2009) and the exact positions $\vec{r}_i^{\,'}$ of the sources and their charges q_i are summed over all source charges: see equation (C.58) below.

In this formula, two phenomenological parameters are required which are ϵ_r, the relative permittivity of the medium and l_D, the Debye screening length or ξ in our model. The

Coulomb potential is a particular case of the Yukawa potential when the screening length goes to infinity.

In coarse grained modeling, the assumption is made that the two phenomenological parameters are constant (homogeneous) in the media where the potential is computed. The standard or defaulted homogeneous values of these parameters are $\epsilon_r = 78$ (water) and $l_D = 10$. The formula also neglects surface charge polarization effects at dielectric media discontinuities. We showed in the paper that the Debye-Hückel theory is not quite appropriate in the confined environment of the tunnel. In particular, the Debye screening length to be used should be larger considering that, in the vicinity of the tunnel walls, the ions contribute weaker to the screening than they do in the inner core of the ribosome. Much weaker ionic strengths and larger Debye screening lengths should be used in the formula to be able to fit the experimentally observed values of the potential.

The electrostatic screening in the tunnel lumen is due to the permanent electric dipoles of constitutive water molecules or to the induced polarization in dielectric neutral media around the charged sources. In strongly confined environment we should resort to Gouy-Chapman screening lengths when there are much less water molecules or when the permittivity of the medium reaches the minimal value ($\epsilon_{\text{protein}} = 4$).

For these reasons, and in a similar way as what was done with the full analytical idealized solutions of the electrostatic potential, we relaxed the homogeneity assumptions of the two phenomenological parameters and also implemented a piecewise continuous assumption for these parameters in the three tunnel regions of interest: cylinder part, cone part and constriction site.

$$\Phi(z) = \sum_{k \in \text{regions}} \sum_{i \in \text{charged sources}} \frac{q_{i,k}}{4\pi\epsilon_0\,\epsilon_r(k)} \cdot \frac{e^{-\frac{|\vec{r}_i{}' - (0,0,z)|}{\xi_k}}}{|\vec{r}_i{}' - (0,0,z)|} \tag{C.58}$$

In the above Coulomb-Yukawa electrostatic potential formula, different values of $\epsilon_r(k)$ and different ξ_k screening lengths can be used in the different k-indexed regions (or media). The elementary unit charge value of $+|e|$ or $-|e| = -1.602\,10^{-19}$ C is used for each of the charges q_i associated to the positively or negatively charged atoms at their given $\vec{r}_i{}'$ positions. In the model of our manuscript, the sum over the q_i results from a surface averaged ensemble and is computed by an integral of the surface charge density over the surface element of interest. Formula (1) and (25) of our manuscript are the surface integral equivalents of the above discrete sum formula (C.58).

Upon implementing this formula (C.58) in Python and using the exact positions of the Fig. 6.6 mapped 94 charged atoms, we obtained the electrostatic potential along the tunnel centerline shown on Fig. C.3 below.

Table C.5.1: Phenomenological parameter values at standard temperature $T = 298.15\,\mathrm{K}$ in the tunnel regions of interest

Region	k	Medium permittivity	Screening length ξ_k	Ionic strength	ξ_k retained
Tunnel cylinder	1	$\epsilon_k = 8.9$	$\xi_{\mathrm{Debye}} = 2.48\,\mathrm{nm}$	$I = 15\,\mathrm{mM/L}$	$\xi_k = 2.48\,\mathrm{nm}$
Tunnel cone	2	$\epsilon_k = 20.15$	$\xi_{\mathrm{Debye}} = 2.04\,\mathrm{nm}$	$I = 22\,\mathrm{mM/L}$	$\xi_k = 2.04\,\mathrm{nm}$
Constriction site (aa)	3	$\epsilon_k = 4$	$\xi_{\mathrm{Gouy}} = 0.105\,\mathrm{nm}$		$\xi_k = 0.105\,\mathrm{nm}$

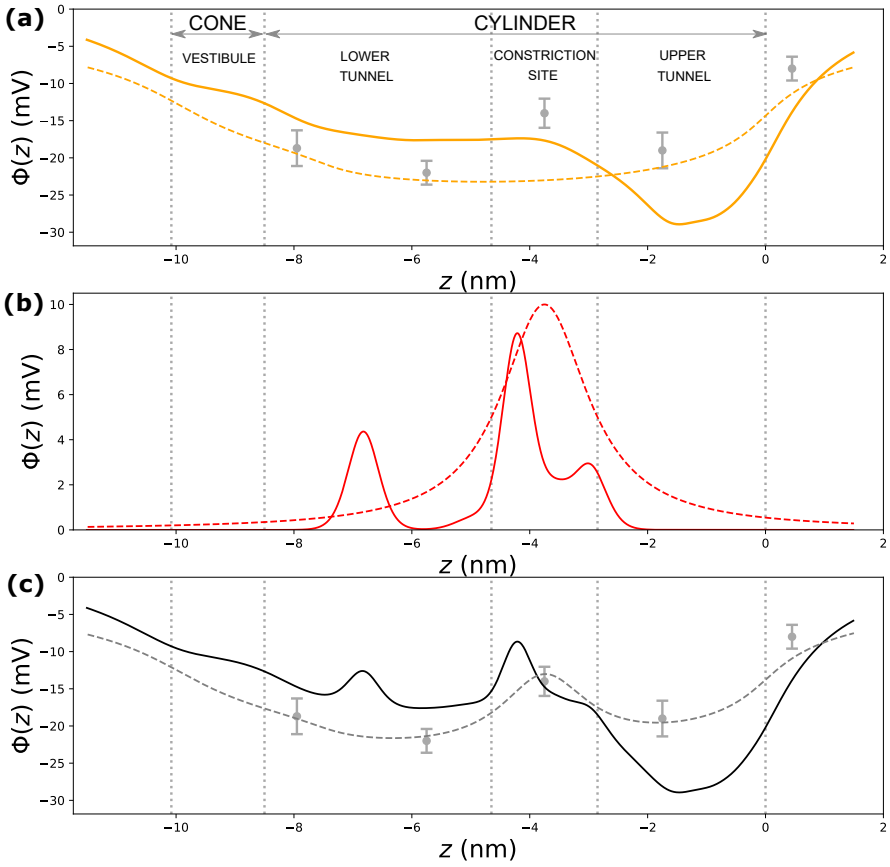


Figure C.3: Electrostatic potential calculated from exact positions of 94 charged atoms close to the tunnel centerline compared to our idealized model. (a) contribution of phosphate moieties only; orange solid line: potential profile calculated from mapped atomic positions; orange dashed line: potential profile calculated from idealized model; gray error bars: measured potential data points in Lu et al. 2007. (b) contribution of the charged amino acid residues around constriction site; red solid line: potential profile calculated from mapped atomic positions; red dashed line: Lorentzian potential profile from idealized model. (c) sum of (a) and (b) for the 94 atoms as mapped on Fig. 6.6 (black solid line); potential profile from idealized model (gray dashed line); measured potential data points in Lu et al. 2007 (gray error bars).

The set of Table C.5.1 parameters was used. We used the same set of dielectric responses as the ones estimated from the idealized model and the data fitting constraints of the observed potential data points. Explicitly, we used $\epsilon_r = 8.3$ in the cylinder, $\epsilon_r = 20.15$ in the cone and $\epsilon_r = 4$ in the constriction site. The screening lengths were adjusted to fit the experimentally observed electrostatic potential data points as well as possible.

It is difficult (or impossible) to get an electrostatic potential profile that fits well the experimentally observed potential data points in Lu et al. 2007 without resorting to different piecewise constant values of the phenomenological parameters. This shows the scope shortcomings or scope limitations of the current models, whether or not based on exact atomic positions, coarse grained or ensemble approaches.

Although an inverse bell shape is observed, the results as plotted on Fig. C.3 (a) to (c) show some discrepancies between the potential profile calculated from the structural data and the data points in ibid. or between the structural data potential and the potential profile as obtained from the idealized model (Fig. 6.4 (a) to (e) in main thesis).

The comparison of the potential profile as obtained from the exact atomic positions (structural data) with the potential profile as calculated from our geometrically idealized model using the piecewise σ/ϵ constant ratio values in the three different regions of concern, calls for the following comments:

1. Except for the upper tunnel, the contribution of the phosphate moieties to the electrostatic potential profile calculated from the structural data (solid orange line) is parallel to the idealized profile where uniformed surface charge distribution is assumed (dashed orange line). The local deeper well of the potential in the upper tunnel is due to the presence of phosphate groups that are closer to the tunnel centerline as compared to other regions: see Fig.6.6(a) and (b) of the paper.

2. The net contribution of the positively charged amino acid residues (minus the negatively charged amino acid residues) calculated from the structural data show that these charged groups cause the peaks increasing the potential locally around the constriction site: see Fig.6.6(b). The profiles of the peaks are very similar to the profiles of Lorentzian peaks as expected. This confirms the Lorentzian function can be used to approximate the contribution of the local enrichment in positively charged amino acid residues at the constriction site. The contribution of the dominant peak in the structural data comes from the close position of R125 (Arg125 of uL22) from the tunnel centerline (R125 is $\sim$ 6 from the centerline). The presence of the 6-7 positively charged residues in general helps explaining the local rise in the potential as initially experimentally measured in [ibid.].

3. The transition from the lower tunnel (cylinder end port) to the vestibule (cone entry port) and further to the exit port of the cone is smooth and the potential calculated from the structural data (solid black line) is parallel to the potential

profile as calculated from our model (dashed gray line) in Fig.6.6(c). The explanation that both potential values diminish in magnitude irrespective of the increase in surface charge density towards the cone end port comes from the fact that 1) most of the phosphate groups tend to get further away from the tunnel centerline; this is also true for the amino acid charged groups: see Fig.6.6.(b) and (d); 2) the dielectric response of the medium goes from an estimated average value of $\epsilon = 8.3$ to $\epsilon = 20.15$ when moving from the lower tunnel to the cone exit port; and 3) the Debye screening length decreases (and the Yukawa exponential factor decreases the potential) due to an expected increase in the ionic strength from the very weak 15 millimole per liter value (lower cylinder estimate) to the still weak 22 millimole per liter value (cone value estimate). The last two facts are both due to the expected increased proportion of water content in the tunnel lumen media and the increased polarizability along the z-axis centerline towards the cone exit port (from a water content of $\sim 6\%$ in the cylinder to $\sim 24\%$ in the cone).

C.6 Specific effect of the tunnel electrostatic interaction on the elongation rate

The electromechanical force due to the tunnel electrostatics acts on the peptide nascent chain and is transmitted inside the ribosomal tunnel up to the peptidyl transfer center (PTC) responsible for the peptide bond formation. The force is transmitted to the PTC through the whole length of the polypeptide chain backbone embedded in the tunnel [Fritch et al. 2018]. At the PTC, the first event that must occur before the peptide bond is built between the peptidyl-tRNA at the P site and the aminoacylated tRNA at the A site is the breaking of the ester covalent bond between the oxygen atom attached on the tRNA 3' end (3' carbon at the CCA terminal ribose) and the carbonyl group of the carboxyl terminal end of the peptide. We presume that a force acting on the peptidyl-tRNA peptide directed from the P site toward the N-terminal end of the peptide would help breaking this ester bond. The chemical reaction rate of this ester bond breaking would be increased in the presence of such a force directed toward the exit tunnel. The ribosome elongation average rate can be quantitatively modulated by applying a Maxwell-Boltzmann factor, i.e $\exp \frac{\int \vec{F}_z \cdot dz}{k_B T}$, resulting from the theoretical treatment of the effect of force on the thermodynamics and kinetics of chemical reactions [Bustamante et al. 2004] or as initially introduced by Bell in a cell to cell adhesion context [Bell 1978; Ribas-Arino and Marx 2012]. This factor correcting the elongation rate specifically accounts for the electrostatic interaction of the nascent chain in the ribosome exit tunnel. This factor is calculated on the basis of all the 50 residues upstream and is updated at each new incorporation. The numerical

value of this factor will be different at each residue incorporation and will always be dependent on the particular amino acid sequence being embedded in the tunnel. For

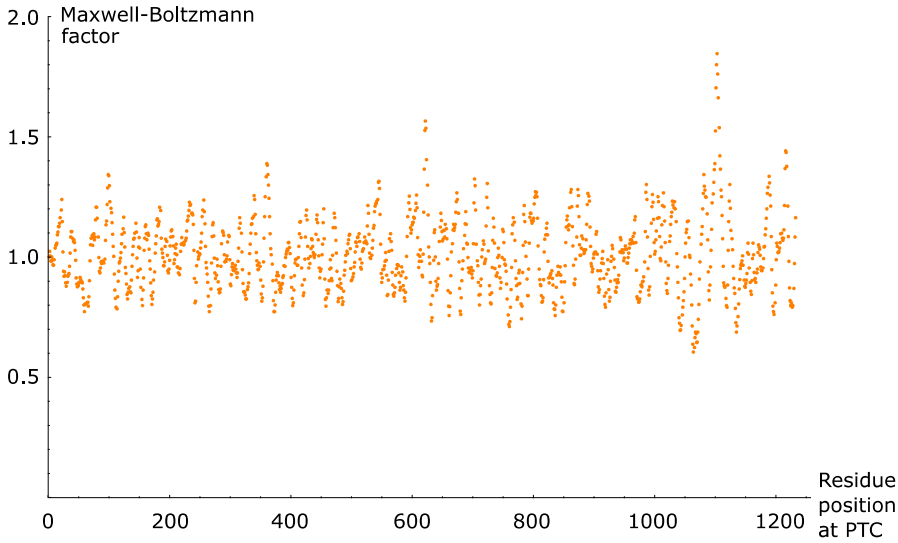


Figure C.4: Maxwell-Boltzmann elongation rate factors weighting for the electrostatic interaction at each incorporation of a new residue at PTC as a function of residue position in the human protein KIF4A. Min = 0.60 and Max = 1.85 occurring at incorporation of residue $E1064$ and $K1103$ respectively. Lower values are associated to smaller elongation rate (slowdown), larger values are associated to higher elongation rates (speeding up).

the arbitrarily chosen protein **KIF4A** (member of the family of human kinesins), all the numerical values for the Maxwell-Boltzmann factors calculated for each residue sequentially incorporated at the PTC are displayed in Fig. C.4. The minimal value is 0.60 and the maximal value is 1.85 for the Maxwell-Boltzmann factor in the particular case of **KIF4A**. 84.5% of the values are in the range [0.80, 1.20]. The mean, 1.01, is very close to 1.0. The minimal value of the Maxwell-Boltzmann factor, 0.60, occurs at incorporation of residue 1064 which is a **E** (negatively charged glutamate), in **KIF4A**, when the axial force on the nascent polypeptide stretch in the tunnel is +8.62 pN. The elongation rate is quantitatively slowed down by a factor 0.60. Equivalently, the time spent by the ribosome on codon 1064 is expected to be larger (average time for this type of codon divided by 0.60) at this position because of the most unfavorable electrostatic interaction occurring at the moment of this residue incorporation. The maximal value of the Maxwell-Boltzmann factor, 1.85, occurs at incorporation of residue 1103 which is a

K (positively charged lysine), in KIF4A, when the axial force on the nascent polypeptide stretch in the tunnel is −10.49 pN. The elongation rate is quantitatively faster by a factor 1.85. Equivalently, the time spent by the ribosome on codon 1103 is expected to be smaller (average time for this type of codon divided by 1.85) at this position because of the most favorable electrostatic interaction occurring at the moment of this residue incorporation. This illustrates how the Maxwell-Boltzmann factors provide a consistent methodological tool to assess quantitatively the contribution to the elongation rate specifically due to the electrostatic interaction occurring in the tunnel, and in a separate way from the other factors affecting the mRNA translation rate.

C.7 Energy sources available for the ribosome

The ribosome is a complex macromolecular machine that requires energy to carry out its multiple tasks. During elongation, a ribosome has to translocate the mRNA each time a codon has been paired to its cognate or semi-cognate tRNA and has to push the nascent protein through the exit tunnel.

The detailed energy balance (energy sources and uptakes) required for elongation has not been fully resolved. Our electrostatic model shows that, in certain situations, a Gibbs free energy fraction in the range 15% − 30% of the total biochemical energy available to the ribosome could be required to move the nascent protein through the exit tunnel.

The energy is found in the biochemical reactions taking place in the ribosome with the help of the associated catalytic sites of enzymes like the elongation factors (eEF in eukaryotes) or ribozymes. The elongation factors (EF and EF-G) are GTPases whose activity is controlled by the ribosome. When an aminoacyl group is hydrolyzed from the loaded tRNA, an ester group is broken and energy is released. For each amino acid incorporation cycle, two GTPs molecules are hydrolysed (one with the help of EF in the ternary complex accommodated at the A-site and one with the help of EF-G required for the mechanical translocation). The peptide bond formation itself requires free energy at each chain elongation by one residue. A very rough estimate of the net change in Gibbs free energy for the net balance between peptide bond formation and ester hydrolysis at pH = 7, 25°C yields $\Delta G^\circ = -3.7 \pm 1.2\,\mathrm{kcal/mol} = -15.5 \pm 5.0\,\mathrm{kJ/mol}$ [Kaiser and Tinoco 2014; Liu et al. 2014a]. This is known as the transpeptidation Gibbs free energy.

Peptide bond formation

the formation of the simplest dipeptide glycylglycine is endergonic and requires 15 kJ/mol (3.6 kcal/mol) per mole of formed peptidic bond:
$\Delta G^\circ = +3.6\,\mathrm{kcal/mol}$ for one residue incorporation (per ribosome cycle).

Hydrolysis of ester bond in aminoacyl-tRNA

the hydrolysis of the ester bond in aa-tRNA is exergonic and releases 30.5kJ/mol(7.3, kcal/mol) per amino acid released from the tRNA:
$\Delta G^\circ = -7.3\,\text{kcal/mol}$ (per ribosome cycle)

Hydrolysis of 2 GTPs

the hydrolysis of 2 GTPs is exergonic and releases 30.5 kJ/mol (7.3 kcal/mol) per mole of GTP. Hence, per residue incorporation cycle (2 GTPs):
$\Delta G^\circ = -14.6\,\text{kcal/mole}$ (per ribosome cycle)

Net Gibbs free energy available to the ribosome per aa residue incorporation

$\Delta G^\circ = -18.3\,\text{kcal/mol}$ (per residue incorporation)

The net result is that one ester bond to the 3'-hydroxyl of a ribose has been broken (locally in the ribosome) and one peptide bond in the nascent protein has been formed, two GTPs have been hydrolyzed, the ribosome has shifted forward the `mRNA` by one codon (translocation distance on `mRNA`, $\Delta x \sim 1.4\,\text{nm}\ (0.9 - 1.8)$, parenthesis indicate 95% confidence limits Liu et al. 2014a) and the nascent peptide has advanced in the ribosome exit tunnel by one residue (nascent peptide chain distance displacement in the tunnel at each translocation, $\Delta z \sim 0.25\,\text{nm}$, which is the estimated distance between two consecutive amino acid $\alpha-$carbons as considered in our model). It is not fully elucidated whether (or how) free energy could be stored in the ribosome and used later to catalyze translocation and possibly assist the progression of the nascent protein through the ribosome exit tunnel when needed. Each step in translation involves intra-subunit or inter-subunit conformational changes [Desai et al. 2019; Kaiser and Tinoco 2014; Liu et al. 2014a]. Such conformational changes could store energy that could be released at a subsequent step, with a thermodynamical yield, providing a conceivable mechanism of harnessing the biochemical energy to use it for mechanical translocation and for moving the nascent peptide through the ribosome exit tunnel when required. The entropy driven spontaneous or chaperones assisted folding of the protein, generating a tugging force outside of the ribosome exit tunnel, might also help the nascent protein to be pulled out of the tunnel [Simpson et al. 2020]. Optical tweezers assays have opened the way to characterizing the ribosome's full mechanochemical cycle [Desai et al. 2019; Liu et al. 2014a]. Recently, such *in vitro* assays provided an estimate for the maximal mechanical energy required per translocation step (near stalling on the `mRNA`), $21.2\,\text{pN}\cdot\text{nm} = 5.2 k_B T$, at 296 K, or $\sim 3.1\,\text{kcal/mol}$ [Desai et al. 2019; Liu et al. 2014a]. As estimated above, the Gibbs free energy available from the transpeptidation step (ester hydrolysis and peptide formation without the help of GTP hydrolysis) is $\Delta G^\circ = -3.7 \pm 1.2\,\text{kcal/mol}$. The mechanical work for translocation would be around 80% of the Gibbs free energy available from the transpeptidation. Such a high thermodynamic efficiency for conversion of chemical energy to mechanical motion is higher than occurs in most molecular motor [Bustamante et al. 2004]. Instead, efficient translocation would require the hydrolysis of at least one GTP with the help of

elongation factor EF-G [Liu et al. 2014a]. EF-G dependent GTP hydrolysis was shown to precede and greatly accelerate translocation [Rodnina et al. 1997]. The mechanical translocation of the ribosome on the mRNA by one codon would take $\frac{3.1\,\text{kcal/mol}}{7.3\,\text{kcal/mol}}$ or 43 % of the Gibbs free energy released by the hydrolysis of one GTP, assisted by elongation factor EF-G. The mechanical energy required to push the nascent peptide chain through the large subunit exit tunnel could be provided by the transpeptidation Gibbs free energy or the hydrolysis of one GTP assisted by elongation factor EF in the ternary complex accommodated in the A site or a combination of both.

Appendix D

Supplementary material of the CSBJ02 paper

D.1 Uncertainty quantification of the theoretical peptide bond formation rate

This appendix reproduces the supplementary material of the CSBJ02 paper exposed in chapter 6.

The classical error estimation of a multivariate function relies on the differential calculus and on the Leibniz derivative chain rule. The differential calculus provides an estimate of the maximum absolute uncertainty on a function and is determined by the absolute uncertainties on the input variables. Let u be a function of multiple independent variables $x, y, z, \cdots, t$.

$$u = f(x, y, z, \cdots, t) \tag{D.1}$$

Suppose that the numerical values of the input variables are known with their uncertainties $\Delta x, \Delta y, \Delta z, \cdots, \Delta t$. The numerical value of u will result with an uncertainty Δu

$$\Delta u = f(x + \Delta x, y + \Delta y, z + \Delta z, \cdots, t + \Delta t) - f(x, y, z, \cdots, t), \tag{D.2}$$

In a first order approximation, if the input variables uncertainties are reasonably small, the total increase in u can be approximated by the total differential of u

$$\Delta u \approx \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z + \cdots + \frac{\partial f}{\partial t}\Delta t. \tag{D.3}$$

Taking the absolute values of the errors (uncertainties) gives the inequality

$$|\Delta u| \leq \left|\frac{\partial f}{\partial x}\right| \cdot |\Delta x| + \left|\frac{\partial f}{\partial y}\right| \cdot |\Delta y| + \left|\frac{\partial f}{\partial z}\right| \cdot |\Delta z| + \cdots + \left|\frac{\partial f}{\partial t}\right| \cdot |\Delta t|. \tag{D.4}$$

It follows that upper limits can be estimated

$$|\Delta^* u| = \left|\frac{\partial f}{\partial x}\right| \cdot |\Delta^* x| + \left|\frac{\partial f}{\partial y}\right| \cdot |\Delta^* y| + \left|\frac{\partial f}{\partial z}\right| \cdot |\Delta^* z| + \cdots + \left|\frac{\partial f}{\partial t}\right| \cdot |\Delta^* t|. \tag{D.5}$$

Three classical results from this chain rule derivation are reminded

1. If $u = x + y + z$, then $|\Delta^* u| = |\Delta^* x| + |\Delta^* y| + |\Delta^* z|$.
2. If $u = x \cdot y$, then $|\Delta^* u| = |x| \cdot |\Delta^* y| + |y| \cdot |\Delta^* x|$.
3. If $u = \frac{x}{y}$, then $|\Delta^* u| = |\frac{1}{y}| \cdot |\Delta^* x| + |\frac{x}{y^2}| \cdot |\Delta^* y|$.

Applying this error calculus to the electrostatic potential profile model for $\Phi(z)$ and $E_z(z)$ that were derived in Joiret et al. 2022a (equations [3], [8], [19] and [20]), we can calculate an estimate of the uncertainty for the theoretical model of the electrostatic interaction in the ribosome exit tunnel. Namely, we can provide uncertainty estimates for the axial electric field E_z, the axial force, the mechanical work for a displacement along the ribosome exit tunnel central path (z displacement). A final estimate is that for the uncertainty on the peptide bond reaction theoretical rate or its inverse, the median time course of the peptide bond formation.

The numerical values adopted for the uncertainties on the input variables resulted from complementary observational constraints. The atomic positions mapped on the tunnel surface, built from high precision x-ray solved structures of real ribosomes, led to bounded ranges in the phenomenological constants σ and ϵ. These ranges are also constrained by the experimental electrostatic potential measurements that were made in the ribosome exit tunnel [Lu and Deutsch 2008; Lu et al. 2007]. Altogether, these constraints jointly set the order of magnitude of the uncertainties on the electrostatic model input variables. A close examination of the electrostatic profile showed in Figure 6.14 suggests that different electrostatic potential profiles could accommodate the experimental confidence intervals for the observed data points. This imposes limits in the lower and upper bounds in the space of values for the input variables of the electrostatic potential model. The numerical values and the uncertainties on the input variables that were used for the output uncertainty calculation are listed in table D.1.1.

Specifically, the experimental uncertainties for the potential show confidence interval values around $\simeq 5$ mV. This corresponds to an uncertainty of $\Delta\Phi(z) = \pm 2.5$ mV. The Yukawa-Debye Hückel equation that models the electrostatic potential $\Phi(z)$, taking

into account the water screening effect in [Joiret et al. 2022b], requires the following

$$\Delta\Phi \approx \left[\frac{R}{2\epsilon_0\epsilon_r}\cdot\Delta\sigma + \frac{\sigma\cdot R}{2}\cdot\frac{\epsilon_0\Delta\epsilon_r}{\epsilon_0^2\epsilon_r^2} + \frac{\sigma}{2\epsilon_0\epsilon_r}\cdot\Delta R\right]\cdot e^{-\frac{\Delta}{\xi}}. \tag{D.6}$$

This shows, for instance, that the uncertainties on Φ, σ, ϵ_r and R are jointly constrained by the model itself, i.e., by the Yukawa-Debye-Hückel laws of electrostatics in dielectric media. The numerical values listed in table D.1.1 comply with these constraints.

Table D.1.1: Numerical values and uncertainties of the input variables of the model of the electrostatic potential. The table rows list the model input variables σ, the surface charge density of the exit tunnel inner wall; ϵ_r, the dielectric response of the medium in the tunnel lumen; R, the tunnel radius; L, the tunnel length; z, the exact position in the tunnel where the potential and electric field are measured; Φ, the electrostatic potential.

Input variable		units [-]	values	error	relative error	absolute error
Unit charge	$q = \|e\|$	C	$1.602\,10^{-19}$			
Surface charge density	σ	C/nm^2	$2.1\|e\|$	$\mathbf{\Delta\sigma}$	20%	$0.4\|e\|$
Vacuum permittivity	$\boldsymbol{\epsilon_0}$	F/m	$8.85\,10^{-12}$			
Relative permittivity (dielectric medium)	$\boldsymbol{\epsilon_r}$	–	78	$\mathbf{\Delta\epsilon_r}$	70%	55
Tunnel radius	$\mathbf{R}$	nm	0.5	$\mathbf{\Delta R}$	10%	0.05
Tunnel length	$\mathbf{L}$	nm	10	$\mathbf{\Delta L}$	5%	0.5
Position	$\mathbf{z}$	nm	−1.2	$\mathbf{\Delta z}$	25%	0.30
	$\mathbf{z}$	nm	−4.2	$\mathbf{\Delta z}$	25%	1.05
	$\mathbf{z}$	nm	−8.7	$\mathbf{\Delta z}$	25%	2.15
Gouy-Chapman screening length	ξ	nm	0.105			
Charge attenuating factor (screening)	$e^{-\frac{\Delta}{\xi}}$	-	0.002			
electrostatic potential	$\mathbf{\Phi}$	mV	−15	$\mathbf{\Delta\Phi}$	17.5%	2.5

z are measured here from the exit tunnel entry port as origin. This corresponds to a shift by 8 nm from the z origin in Figure 6.14.

The sequence of the uncertainty propagation is calculated as follows. The $E_z(z)$ electric field is a function of the minimal set of input variables $z, \sigma, \epsilon, R, L$.

$$E_z = f(z, \sigma, \epsilon, R, L) \tag{D.7}$$

The function f is explicitly derived, for different ribosome exit tunnel geometries, in equations (3), (8), (19), (20) in Joiret et al. 2022a. To calculate an estimate of the uncertainty ΔE_z on the **electric field** E_z, we estimate from equation (3) in Joiret et al. 2022b

$$|\Delta^* E_z| = \left|\frac{\partial f}{\partial z}\right|\cdot|\Delta^* z| + \left|\frac{\partial f}{\partial \sigma}\right|\cdot|\Delta^*\sigma| + \left|\frac{\partial f}{\partial \epsilon_r}\right|\cdot|\Delta^*\epsilon_r| + \left|\frac{\partial f}{\partial R}\right|\cdot|\Delta^* R| + \left|\frac{\partial f}{\partial L}\right|\cdot|\Delta^* L| \tag{D.8}$$

Different geometries lead to more complicated formula but the order of magnitudes for the uncertainties are similar (not shown). The salient profile feature is the inverse bell

shape of the potential. The first derivatives of the terms for the E_z uncertainty are:

$$\left|\frac{\partial f}{\partial z}\right| = \frac{\sigma R}{2\epsilon_0\epsilon_r}\left[- \frac{z+L}{(R^2+(z+L)^2)^{3/2}} + \frac{z}{(R^2+z^2)^{3/2}}\right]$$

$$\left|\frac{\partial f}{\partial \sigma}\right| = \frac{\sigma R}{2\epsilon_0\epsilon_r}\cdot\left(\frac{1}{\sqrt{R^2+(z+L)^2}} - \frac{1}{\sqrt{R^2+z^2}}\right)$$

$$\left|\frac{\partial f}{\partial \epsilon_r}\right| = \frac{\sigma R}{2\epsilon_0^2\epsilon_r^2}\cdot\left(\frac{1}{\sqrt{R^2+(z+L)^2}} - \frac{1}{\sqrt{R^2+z^2}}\right)$$

$$\left|\frac{\partial f}{\partial R}\right| = \frac{\sigma}{2\epsilon_0\epsilon_r}\cdot\left(\frac{1}{\sqrt{R^2+(z+L)^2}}-\frac{1}{\sqrt{R^2+z^2}}\right)+\frac{\sigma R^2}{2\epsilon_0\epsilon_r}\left[\frac{1}{(R^2+z^2)^{3/2}}-\frac{1}{(R^2+(z+L)^2)^{3/2}}\right]$$

$$\left|\frac{\partial f}{\partial L}\right| = \frac{\sigma R}{2\epsilon_0\epsilon_r}\cdot\frac{z+L}{(R^2+(z+L)^2)^{3/2}}$$

where, in all the above formulas, $\sigma = \sigma^* \cdot e^{-\frac{\Delta}{\xi}}$ to take into account the Gouy-Chapman screening length [Joiret et al. 2022b]. Using the numerical values listed in table D.1.1, the estimated maximum absolute uncertainty in E_z is numerically equal to $|\Delta^* E_z| \sim 170\text{MV/cm}$ or $|\Delta^* E_z| \sim 0.3\text{pN}/|e|$ when $z = -1.2\,\text{nm}$.

The uncertainty on the **axial force** is the uncertainty of the electric field multiplied by the test charge. The uncertainty on the **mechanical work** W is obtained in the same way by applying the differential calculus.

$$W = q \cdot E_z \cdot dz \tag{D.9}$$

$$\Delta W = q \cdot dz \cdot \Delta E_z + q \cdot \Delta(dz) \cdot E_z. \tag{D.10}$$

The uncertainties on the **reaction rate** $k(F)$ in the presence of an axial force F and the **median time course** $\tau_{1/2}$ of the reaction are calculated as shown hereafter.

$$k(F) = k(0) \cdot e^{-\frac{W}{k_B T}} \tag{D.11}$$

$$\Delta k(F) = k(0) \cdot e^{-\frac{W}{k_B T}} \cdot \left(\frac{\Delta W}{k_B T}\right) + \Delta k(0) \cdot e^{-\frac{W}{k_B T}} \tag{D.12}$$

$$\tau_{1/2} = \frac{ln\,2}{k(F)} \tag{D.13}$$

$$\Delta\tau_{1/2} = ln\,2 \cdot \frac{\Delta k(F)}{k(F)^2}, \tag{D.14}$$

where $k_0 = 15.82\,s^{-1}$ is the reference median reaction rate for neutral oligopeptide probes and $\Delta k_0 = 1.50\,s^{-1}$ is the experimental uncertainty on experimental rate measurements, corresponding to an experimental uncertainty on the median time course of the peptide bond formation around ~ 3 ms, when puromycin is the A-site acceptor substrate. The theoretical uncertainties for the output variables are calculated from all the equations above and are listed in the right most column of table D.1.2, for three different regions in the ribosome exit tunnel. The calculations were conducted in the three regions of the tunnel covered by the charged amino acid residues of the three designed oligopeptide pairs of probes described in the main manuscript. The values in the left column come from the results in the main manuscript. As far as the uncertainties in the input variables are concerned, they either come from the literature [Joiret et al. 2022b; Lu and Deutsch 2008; Lu et al. 2007] or were estimated by arbitrarily fixing a reasonable relative error (second last column of table D.1.1 in this appendix).

Table D.1.2: Numerical values and calculated uncertainties on the electrostatic field model and on the output variables of the kinetics model.

output variable		units [-]	values	uncertainty	absolute uncertainty
10-mer probe	$z \in [-8.95, -9.45]$ nm		$z_{\text{center}} = -1.2$ nm		
Electric field	E_z	MV/cm		$\lvert\boldsymbol{\Delta}^* \mathbf{E_z}\rvert$	173
Mechanical work	W	pN.nm	0.65	$\lvert\boldsymbol{\Delta}^* \mathbf{W}\rvert$	0.086
Reaction rate	$k(F)$	s^{-1}			
fast		$\oplus$-mer	18.35	$\lvert\boldsymbol{\Delta}^* \mathbf{k}\rvert$	2.11
slow		$\ominus$-mer	13.61	$\lvert\boldsymbol{\Delta}^* \mathbf{k}\rvert$	1.56
Median time course	$\tau_{1/2}^{th.}$	ms			
fast		$\oplus$-mer	37.6	$\lvert\boldsymbol{\Delta}^* \boldsymbol{\tau}_{\mathbf{1/2}}^{\mathbf{th.}}\rvert$	4.3
slow		$\ominus$-mer	51	$\lvert\boldsymbol{\Delta}^* \boldsymbol{\tau}_{\mathbf{1/2}}^{\mathbf{th.}}\rvert$	5.9
22-mer probe	$z \in [-11.95, -12.45]$ nm		$z_{\text{center}} = -4.2$ nm		
Electric field	E_z	MV/cm		$\lvert\boldsymbol{\Delta}^* \mathbf{E_z}\rvert$	39
Mechanical work	W	pN.nm	0.86	$\lvert\boldsymbol{\Delta}^* \mathbf{W}\rvert$	0.017
Reaction rate	$k(F)$	s^{-1}			
slow		$\oplus$-mer	12.81	$\lvert\boldsymbol{\Delta}^* \mathbf{k}\rvert$	1.27
fast		$\ominus$-mer	19.30	$\lvert\boldsymbol{\Delta}^* \mathbf{k}\rvert$	1.91
Median time course	$\tau_{1/2}^{th.}$	ms			
slow		$\oplus$-mer	53.5	$\lvert\boldsymbol{\Delta}^* \boldsymbol{\tau}_{\mathbf{1/2}}^{\mathbf{th.}}\rvert$	5.3
fast		$\ominus$-mer	35.8	$\lvert\boldsymbol{\Delta}^* \boldsymbol{\tau}_{\mathbf{1/2}}^{\mathbf{th.}}\rvert$	3.6
40-mer probe	$z \in [-16.2, -17.2]$ nm		$z_{\text{center}} = -8.7$ nm		
Electric field	E_z	MV/cm		$\lvert\boldsymbol{\Delta}^* \mathbf{E_z}\rvert$	449
mechanical work	W	pN.nm	0.525	$\lvert\boldsymbol{\Delta}^* \mathbf{W}\rvert$	0.20
Reaction rate	$k(F)$	s^{-1}			
slow		$\oplus$-mer	13.92	$\lvert\boldsymbol{\Delta}^* \mathbf{k}\rvert$	1.96
fast		$\ominus$-mer	17.88	$\lvert\boldsymbol{\Delta}^* \mathbf{k}\rvert$	2.51
Median time course	$\tau_{1/2}^{th.}$	ms			
slow		$\oplus$-mer	49.5	$\lvert\boldsymbol{\Delta}^* \boldsymbol{\tau}_{\mathbf{1/2}}^{\mathbf{th.}}\rvert$	7
fast		$\ominus$-mer	38.7	$\lvert\boldsymbol{\Delta}^* \boldsymbol{\tau}_{\mathbf{1/2}}^{\mathbf{th.}}\rvert$	5.4

Appendix E

BASH Script for HPC Job Scheduling and Resource Management

This appendix provides an example of a `UNIX-LINUX` style template for a BASH script used in scheduling and managing jobs on a high-performance computing (HPC) cluster. In the example presented, 64 nodes are requested, with 5 parallel task IDs assigned. Each task is allocated a maximum individual CPU time of 2 hours and 10 minutes. This amounts to a total of $64 \times 5 = 320$ `CPU`s utilized for 2 hours and 10 minutes, equivalent to approximately 693.3 `CPU` hours (28.8 days). Such a computation would take about one month on a single laptop. On the HPC cluster, it will take maximum 2 hours and 10 minutes.

In the BASH script, there are five identical replicates of the input file, each providing read counts for transcript gene ID tags, along with their corresponding individual initiation rate fold changes. These inputs are identical across all five replicates. The number of files in the INPUTDIR directory matches the index range specified in `array=0-4`.

The Python program `myProg02.py` is executed five times (once per replicate) within a double nested loop, using indices i (inner loop) and j (outer loop). The number of files in INPUTDIR corresponds to the array index range. The outer loop index j runs from 0 to 7, corresponding to eight different values of initiation rates listed in `initRATE=(4.3e-6 6e-6 11e-6 16.7e-6 25e-6 50e-6 100e-6 200e-6)`. Similarly, the inner loop index i runs from 0 to 7, corresponding to eight ribosome ratio values specified in `RATIO=(0.5 1.0 1.5 2.0 2.5 3.0 5.0 10.0)`.

The Python program `myProg02.py` requires seven arguments, which are provided as follows:

1. A `.json` file containing the dictionary of elongation cycle kinetics parameters for the 61 sense codons in yeast;
2. The mathematical model describing the electrostatic interactions within the ribosome exit tunnel;
3. The CDS FASTA format sequence of all transcript gene ID tags to be translated;
4. The input files (five replicates) with read count (RNA-Seq) data and individual initiation rate fold changes for each transcript gene ID;
5. The name of the output files, where all requested simulation results will be stored upon job completion. There will be 320 output files;
6. The ribosome ratio value to be used in the current instance of the program;
7. The general initiation rate value to be used in the current instance of the program.

The 320 output files will then be used for further statistical analysis. See for instance chapter 11 and the next appendix.

```
#!/bin/bash
# Submission script for Nic5 (uLiege cluster) with Slurm usage
# Job parameters
#SBATCH --job-name=joiretJOB
#SBATCH --output=joiretJOB.log

# Needed resources
#SBATCH --time=02:10:00 # hh:mm:ss
#
#SBATCH --ntasks=64
#SBATCH --cpus-per-task=1
#SBATCH --mem-per-cpu=100 # megabytes
#SBATCH --array=0-4
#SBATCH --partition=batch
#
#SBATCH --mail-user=marc.joiret@uliege.be
#SBATCH --mail-type=ALL
#
# Operations
module --force purge
#module load LIST_THE_MODULES_YOU_NEED_HERE
#module load releases/2021b
#module load Python/3.9.6-GCCcore-11.2.0
#module load releases/2021b
#module load Scipy-bundle/2021.10-foss-2021b
module load releases/2020b
module load Python/3.8.6-GCCcore-10.2.0
module load Biopython/1.78-foss-2020b
# Paths for relevant directories
WORKDIR=/home/users/m/j/mjoiret
SOFTDIR=/home/users/m/j/mjoiret/myPyPrograms
DATADIR=/home/users/m/j/mjoiret/myData
INPUTDIR=/home/users/m/j/mjoiret/myInput
OUTPUTDIR=/home/users/m/j/mjoiret/myOutput
MYENVDIR=/home/users/m/j/mjoiret/myENV
FILES=($INPUTDIR/*)
RATIO=(0.5 1.0 1.5 2.0 2.5 3.0 5.0 10.0)
initRATE=(4.3e-6 6e-6 11e-6 16.7e-6 25e-6 50e-6 100e-6 200e-6)
# Activate virtual environment
source $MYENVDIR/bin/activate
# Job steps:
# recall that a backslash allows to continue on the next line
echo "Job started at $(date)"
for j in {0..7}
 do
 for i in {0..7}
  do
  srun -N1 -n1 -c1 --exact python ${SOFTDIR}/myProg02.py \
  ${DATADIR}/dataJSONyeast.json ${DATADIR}/tunnelElectrostaticsAF.json \
  ${DATADIR}/CDSfasta01.txt ${FILES[$SLURM_ARRAY_TASK_ID]} \
  ${OUTPUTDIR}/freeRibo${i}init${j}res$SLURM_ARRAY_TASK_ID.txt \
  ${RATIO[$i]} ${initRATE[$j]} &
  done
 done
 wait
echo "Job ended at $(date)"
```

Appendix F

Fully crossed factorial design (fixed effects model)

This Appendix provides the details of the statistical analysis conducted in Chapter 11.

F.1 Proportion of free ribosomes in the pool

F.1.1 Analysis of variance for the two-factor fixed effects model

Python code

```
# A two-way ANOVA in Python
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
from statsmodels.stats.outliers_influence import OLSInfluence
import seaborn as sns

# Define the levels for each factor
factor_A_levels = [f'A{i}' for i in range(1, 9)]
factor_B_levels = [f'B{i}' for i in range(1, 9)]

# Create a DataFrame for the factorial design
data = []

replicate_values_inCell = np.zeros((8, 8), dtype=object)
# we initiated an array 8 by 8 that will contain an object (a list here)
for i in range(8):
     for j in range(8):

 #this is the array of free ribosomes
 # (comment out the next line if you want the free transcripts)
```

```
          replicate_values_inCell[i][j] = np.array(riboFreeProp_arr[i][j][:].tolist())

 # this is the array of free transcripts
 # (comment out the next line if you want the free ribosomes)
          replicate_values_inCell[i][j] = np.array(trFreeProp_arr[i][j][:].tolist())

i_a = -1
for a in factor_A_levels:
     i_a += 1
     i_b = -1
     for b in factor_B_levels:
          # Collect the 5 replicates for each combination of factors
          i_b +=1
          for i in range(5):
               data.append([a, b, replicate_values_inCell[i_a][i_b][i]])

# Convert the data to a DataFrame
df = pd.DataFrame(data, columns=['Factor_A', 'Factor_B', 'Value'])
print(df)

# STEP 1: fit the two-way ANOVA model (fixed model requires ols, ordinary least squares):
model_fixed = ols('Value ~ C(Factor_A) + C(Factor_B) + C(Factor_A):C(Factor_B)',
                  data=df).fit()

# perform the ANOVA:
anova_table = sm.stats.anova_lm(model_fixed, typ=2)

# Display the results:
# Print the ANOVA table and save it to a text file
print(anova_table)
print("ANOVA Table:\n", anova_table)
anova_table.to_csv(fPATHtoFig+'\\'+'anova_table.txt', sep='\t')

# Display all details of the model
print("\nModel Summary:")
print(model_fixed.summary())

# Calculate Total Sum of Squares (SS_total)
SS_total = sum((df['Value'] - np.mean(df['Value']))**2)

# STEP 2: Calculate Partial Eta Squared for each factor
anova_table['Eta_Squared'] = anova_table['sum_sq'] / (anova_table['sum_sq'].sum())
print("\nEffect Sizes (Partial Eta Squared):")
print(anova_table[['sum_sq', 'df', 'Eta_Squared']])
```

F.1.1.1 Proportion of free ribosomes response: 2-way ANOVA table

ANOVA Table:

	sum_sq	df	F	PR(>F)
C(Factor_A)	18309.65	7.0	160.85	6.51e-90
C(Factor_B)	47716.97	7.0	419.18	2.73e-136
C(Factor_A):C(Factor_B)	13484.86	49.0	16.92	2.12e-56
Residual	4163.07	256.0		

Effect Sizes (Partial Eta Squared) % of variance explained:

	sum_sq	df	Eta_Squared
C(Factor_A)	18309.65	7.0	0.219
C(Factor_B)	47716.97	7.0	0.570
C(Factor_A):C(Factor_B)	13484.86	49.0	0.161
Residual	4163.07	256.0	0.050

F.1.2 Diagnostics case statistics and model adequacy checking

Python code

```
# STEP 3: Extract residuals and diagnostic statistics
influence = OLSInfluence(model_fixed)

# Residuals
residuals = model_fixed.resid
print("\nResiduals:")
print(residuals)

# Leverage
leverage = influence.hat_matrix_diag
print("\nLeverage:")
print(leverage)

# Cook's Distance
cooks_d = influence.cooks_distance[0]
print("\nCook's Distance:")
print(cooks_d)

# Outliers (Typically considered if Cook's distance > 4/n
or standardized residuals > 3)
outliers = np.where(np.abs(residuals) > 3 * np.std(residuals))[0]
print("\nOutliers (Index):")
print(outliers)

# Summary statistics for residuals (standardized residuals)
print("\nSummary Statistics for Residuals:")
print("Mean:", np.mean(residuals))
print("Standard Deviation:", np.std(residuals))

# Influence summary (e.g., leverage and Cook's distance threshold)
print("\nInfluence Summary:")
influence_summary = pd.DataFrame({
     'Leverage': leverage,
     "Cook's Distance": cooks_d
})
print(influence_summary.describe())

# Extract and print the fitted values for each factor (predicted values)
print("\nFitted Values:")
print(model_fixed.fittedvalues)

# Extract and display residuals
residuals = model_fixed.resid
print("\nResiduals:")
print(residuals)

# Kolmogorov-Smirnov test to check normality of residuals
from scipy.stats import kstest
ks_stat, p_value = kstest(residuals, 'norm', args=(np.mean(residuals), np.std(residuals)))

# Print the results of the KS test
print("Kolmogorov-Smirnov Test for Normality of Residuals:")
print(f"KS Statistic: {ks_stat}")
print(f"P-value: {p_value}")

# Interpretation
if p_value > 0.005:
     print("The residuals are likely normally distributed (fail to reject H0).")
else:
     print("The residuals are not normally distributed (reject H0).")

# Plotting the residuals
plt.figure(figsize=(10, 5))

# Residual Plot
plt.subplot(1, 2, 1)
sns.residplot(x=model_fixed.fittedvalues, y=residuals, lowess=True,
line_kws={'color': 'red'})
plt.xlabel('Fitted Values')
plt.ylabel('Residuals')
plt.title('Residuals vs Fitted')
```

```
# Q-Q Plot
plt.subplot(1, 2, 2)
sm.qqplot(residuals, line='45', fit=True, ax=plt.gca())
plt.title('Q-Q Plot of Residuals')

plt.tight_layout()
plt.show()

# Residuals distribution and normal fit:
# Fit a normal distribution to the residuals
from scipy.stats import norm
mu, std = norm.fit(residuals)

# Plot histogram of residuals
plt.figure(figsize=(10, 6))
count, bins, ignored = plt.hist(residuals, bins=15, density=True, alpha=0.6,
color='g', edgecolor='black')

# Plot the PDF of the fitted normal distribution
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mu, std)
plt.plot(x, p, 'k', linewidth=2)

# Add labels and title
plt.title('Histogram of Residuals with Fitted Normal Distribution')
plt.xlabel('Residuals')
plt.ylabel('Density')

# Show plot
plt.show()
plt.savefig(fPATHtoFig+'\\'+'residuals_histo.pdf')  # Save residuals histogram
```

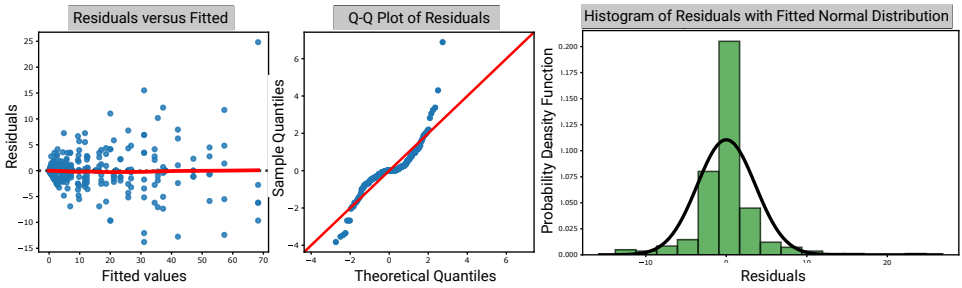


Figure F.1: Left: Plot of residuals versus predicted. Center: Quantile probability plot of residuals (Q-Q plot). Right: Fitted normal probability density to residuals.

The Kolmogorov-Smirnov test on the residuals was rejected at $\alpha = 0.001$, indicating that the residuals are not normally distributed (KS Statistic: 0.1777, p-value: 2.57e-9). Additionally, examining the residuals versus predicted values reveals increasing variance with the response, suggesting a potential violation of the homoscedasticity assumption. A transformation, such as taking the square root, could stabilize the variance. Nonetheless, we proceed with the hypothesis testing and assessment of the main effects without applying this correction.

F.1.3 Conclude on hypothesis tests

F.1.3.1 Main effects

Python code

```
# STEP 5: Plot the main effects plots for Factor A and Factor B

# Custom Labels:
# Specify custom row and column labels
row_labels = [0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 5.0, 10.0] # i
rate_labels = np.array([4.3e-6, 6e-6, 11e-6, 16.7e-6, 25e-6, 50e-6, 100e-6, 200e-6])
column_labels = np.divide(0.001, rate_labels).tolist()
# Format each element to '.1f'
formatted_column_labels = [f'{x:.1f}' for x in column_labels]
# Create the mappings from the numerical factor levels to the custom labels
factor_a_levels = np.linspace(1, 8, 8)  # 8 levels for Factor A
factor_b_levels = np.linspace(1, 8, 8)  # 8 levels for Factor B
row_ticks = np.linspace(min(factor_a_levels), max(factor_a_levels), len(row_labels))
col_ticks = np.linspace(min(factor_b_levels), max(factor_b_levels), len(column_labels))

# Step 5: Plot the main effects plots for Factor A and Factor B
plt.figure(figsize=(10, 4))

# Adjust bottom space to allow space for the second x-axis
plt.subplots_adjust(bottom=0.45)  # Adjust bottom space (increase value if needed)

# Main effects plot for Factor A
ax1 = plt.subplot(1, 2, 1)
sns.pointplot(x='Factor_A', y='Value', data=df, ci='sd', capsize=0.15, ax=ax1,
 markers="D", errwidth=1.5)
ax1.set_title('Main Effects Plot for Factor A')
ax1.set_xlabel('Levels of Factor A', labelpad=2)
ax1.set_ylabel('Mean Response')
ax1.set_ylim(0, 100)

# Add horizontal gridlines for the y-axis (mean response values)
ax1.grid(True, axis='y')
#ax1.grid(True, axis='x')

# Get the position of ax1 and adjust for ax1_bottom
pos1 = ax1.get_position()  # Get the original position of ax1
# Shift the y-position down
new_pos1 = [pos1.x0-0.065, pos1.y0 + 0.15, pos1.width*1.23, pos1.height]

# Add a second x-axis at the bottom
# Create a new set of axes in the same position:
ax1_bottom = ax1.figure.add_axes(new_pos1, frameon=False)
# Set the new x-axis ticks at the bottom
ax1_bottom.xaxis.set_ticks_position('bottom')
# Set the label position for the new axis to the bottom
ax1_bottom.xaxis.set_label_position('bottom')
 # Align limits of the new axis with the original axis
ax1_bottom.set_xlim(ax1.get_xlim())
# Move the second x-axis outward (adjust as needed)
ax1_bottom.spines['bottom'].set_position(('outward', -68))

# Define new labels for the second x-axis for Factor A
# Match the ticks of the new axis to the original axis
ax1_bottom.set_xticks(ax1.get_xticks())
new_labels_A = row_labels
ax1_bottom.set_xticklabels(new_labels_A)
ax1_bottom.set_xlabel('Ribosome pool ratio (Factor A)', fontsize=12, labelpad=-30)
ax1_bottom.set_yticks([])  # Hide y-ticks on the new axis

# Main effects plot for FactorB
ax2 = plt.subplot(1, 2, 2)
sns.pointplot(x='Factor_B', y='Value', data=df, ci='sd', capsize=0.15, ax=ax2,
markers="D", errwidth=1.5)
ax2.set_title('Main Effects Plot for Factor B')
ax2.set_xlabel('Levels of Factor B', labelpad=2)
ax2.set_ylabel('Mean Response')
ax2.set_ylim(0, 100)

# Add horizontal gridlines for the y-axis (mean response values)
ax2.grid(True, axis='y')
#ax2.grid(True, axis='x')
```

```
# Get the position of ax2 and adjust for ax2_bottom
pos2 = ax2.get_position()  # Get the original position of ax2
# Shift the y-position down
new_pos2 = [pos2.x0 + 0.005, pos2.y0 + 0.15, pos2.width*1.23, pos2.height]

# Add a second x-axis at the bottom for FactorB
ax2_bottom = ax2.figure.add_axes(new_pos2, frameon=False)
ax2_bottom.xaxis.set_ticks_position('bottom')
ax2_bottom.xaxis.set_label_position('bottom')
# Align limits of the new axis with the original axis
ax2_bottom.set_xlim(ax2.get_xlim())
ax2_bottom.spines['bottom'].set_position(('outward', -68))

# Define new labels for the second x-axis for FactorB
ax2_bottom.set_xticks(ax2.get_xticks())
new_labels_B = formatted_column_labels
ax2_bottom.set_xticklabels(new_labels_B)
ax2_bottom.set_xlabel('Initiation rate [s/event] (Factor B)', fontsize=12,
labelpad=-30)
ax2_bottom.set_yticks([])  # Hide y-ticks on the new axis

plt.tight_layout()
plt.savefig(fPATHtoFig+'\\'+'mainEffectsPlot2xAxes.pdf')
plt.show()
```

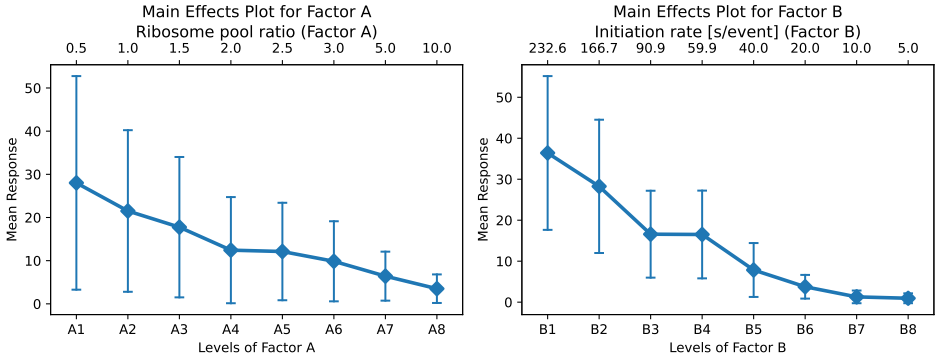


Figure F.2: Left: Main effects of the ribosome pool ratio on the proportion of free ribosomes. Right: Main effects of the initiation rate on the proportion of free ribosomes. Note that in both cases, the standard deviation decreases as the levels of the factors increase, indicating a decrease in the variability of the response with higher factor levels.

F.1.3.2 Interaction

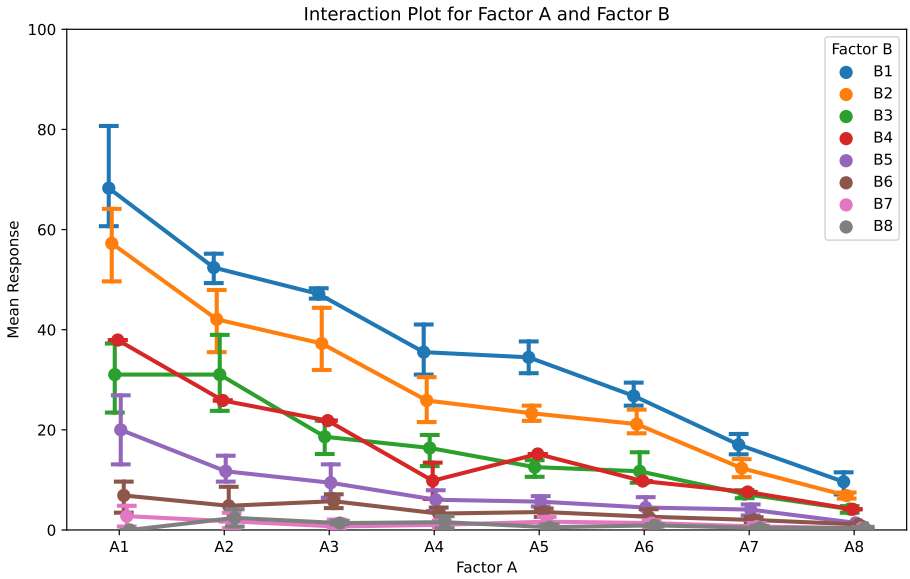


Figure F.3: Interaction plot of ribosome ratio (factor A) and initiation rate (factor B) on the proportion of free ribosomes (response). The strength of the statistical interaction beyond the main effects depends on how the curves converge or diverge relative to each other. The more crossings between the curves, the stronger the interaction.

F.1.4 Model parameters estimation for proportion of free ribosomes response

The local sensitivity coefficients are quantified by the slopes of the response variable as a function of the varied factors. The least squares estimates of the regression coefficients for the fixed effects model as described in eq. (11.11) are tabulated hereafter.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Value   R-squared:                       0.950
Model:                            OLS   Adj. R-squared:                  0.938
Method:                 Least Squares   F-statistic:                     77.61
Date:                                   Prob (F-statistic):          1.37e-135
Time:                                   Log-Likelihood:                -864.57
No. Observations:                 320   AIC:                             1857.
Df Residuals:                     256   BIC:                             2098.
Df Model:                          63
Covariance Type:            nonrobust
=============================================================================================================
                                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                   68.2759      1.803     37.859      0.000      64.724      71.827
C(Factor_A)[T.A2]                          -15.8621      2.550     -6.219      0.000     -20.885     -10.840
C(Factor_A)[T.A3]                          -21.1494      2.550     -8.292      0.000     -26.172     -16.127
C(Factor_A)[T.A4]                          -32.7586      2.550    -12.844      0.000     -37.781     -27.736
C(Factor_A)[T.A5]                          -33.7931      2.550    -13.250      0.000     -38.816     -28.771
C(Factor_A)[T.A6]                          -41.4943      2.550    -16.269      0.000     -46.517     -36.472
C(Factor_A)[T.A7]                          -51.2414      2.550    -20.091      0.000     -56.264     -46.219
C(Factor_A)[T.A8]                          -58.6552      2.550    -22.998      0.000     -63.678     -53.633
C(Factor_B)[T.B2]                          -11.0345      2.550     -4.326      0.000     -16.057      -6.012
C(Factor_B)[T.B3]                          -37.2414      2.550    -14.602      0.000     -42.264     -32.219
C(Factor_B)[T.B4]                          -30.3448      2.550    -11.898      0.000     -35.367     -25.322
C(Factor_B)[T.B5]                          -48.2759      2.550    -18.928      0.000     -53.298     -43.253
C(Factor_B)[T.B6]                          -61.3793      2.550    -24.066      0.000     -66.402     -56.357
C(Factor_B)[T.B7]                          -65.5172      2.550    -25.688      0.000     -70.540     -60.495
C(Factor_B)[T.B8]                          -68.2759      2.550    -26.770      0.000     -73.298     -63.253
C(Factor_A)[T.A2]:C(Factor_B)[T.B2]          0.6897      3.607      0.191      0.849      -6.413       7.793
C(Factor_A)[T.A3]:C(Factor_B)[T.B2]          1.1494      3.607      0.319      0.750      -5.954       8.252
C(Factor_A)[T.A4]:C(Factor_B)[T.B2]          1.3793      3.607      0.382      0.702      -5.724       8.482
C(Factor_A)[T.A5]:C(Factor_B)[T.B2]         -0.1379      3.607     -0.038      0.970      -7.241       6.965
C(Factor_A)[T.A6]:C(Factor_B)[T.B2]          5.4023      3.607      1.498      0.135      -1.701      12.505
C(Factor_A)[T.A7]:C(Factor_B)[T.B2]          6.3448      3.607      1.759      0.080      -0.758      13.448
C(Factor_A)[T.A8]:C(Factor_B)[T.B2]          8.2759      3.607      2.294      0.023       1.173      15.379
C(Factor_A)[T.A2]:C(Factor_B)[T.B3]         15.8621      3.607      4.398      0.000       8.759      22.965
C(Factor_A)[T.A3]:C(Factor_B)[T.B3]          8.7356      3.607      2.422      0.016       1.633      15.839
C(Factor_A)[T.A4]:C(Factor_B)[T.B3]         18.1034      3.607      5.019      0.000      11.001      25.206
C(Factor_A)[T.A5]:C(Factor_B)[T.B3]         15.3103      3.607      4.245      0.000       8.207      22.413
C(Factor_A)[T.A6]:C(Factor_B)[T.B3]         22.1839      3.607      6.150      0.000      15.081      29.287
C(Factor_A)[T.A7]:C(Factor_B)[T.B3]         27.3793      3.607      7.591      0.000      20.276      34.482
C(Factor_A)[T.A8]:C(Factor_B)[T.B3]         31.7931      3.607      8.815      0.000      24.690      38.896
C(Factor_A)[T.A2]:C(Factor_B)[T.B4]          3.7931      3.607      1.052      0.294      -3.310      10.896
C(Factor_A)[T.A3]:C(Factor_B)[T.B4]          5.0575      3.607      1.402      0.162      -2.045      12.160
C(Factor_A)[T.A4]:C(Factor_B)[T.B4]          4.6552      3.607      1.291      0.198      -2.448      11.758
C(Factor_A)[T.A5]:C(Factor_B)[T.B4]         11.0345      3.607      3.059      0.002       3.932      18.137
C(Factor_A)[T.A6]:C(Factor_B)[T.B4]         13.3333      3.607      3.697      0.000       6.230      20.436
C(Factor_A)[T.A7]:C(Factor_B)[T.B4]         20.8966      3.607      5.794      0.000      13.794      27.999
C(Factor_A)[T.A8]:C(Factor_B)[T.B4]         24.8621      3.607      6.893      0.000      17.759      31.965
C(Factor_A)[T.A2]:C(Factor_B)[T.B5]          7.5862      3.607      2.103      0.036       0.483      14.689
C(Factor_A)[T.A3]:C(Factor_B)[T.B5]         10.5747      3.607      2.932      0.004       3.472      17.678
C(Factor_A)[T.A4]:C(Factor_B)[T.B5]         18.7931      3.607      5.210      0.000      11.690      25.896
C(Factor_A)[T.A5]:C(Factor_B)[T.B5]         19.4483      3.607      5.392      0.000      12.345      26.551
C(Factor_A)[T.A6]:C(Factor_B)[T.B5]         25.9770      3.607      7.202      0.000      18.874      33.080
C(Factor_A)[T.A7]:C(Factor_B)[T.B5]         35.3103      3.607      9.790      0.000      28.207      42.413
C(Factor_A)[T.A8]:C(Factor_B)[T.B5]         40.0690      3.607     11.109      0.000      32.966      47.172
C(Factor_A)[T.A2]:C(Factor_B)[T.B6]         13.7931      3.607      3.824      0.000       6.690      20.896
C(Factor_A)[T.A3]:C(Factor_B)[T.B6]         20.0000      3.607      5.545      0.000      12.897      27.103
C(Factor_A)[T.A4]:C(Factor_B)[T.B6]         29.1379      3.607      8.078      0.000      22.035      36.241
C(Factor_A)[T.A5]:C(Factor_B)[T.B6]         30.4828      3.607      8.451      0.000      23.380      37.586
C(Factor_A)[T.A6]:C(Factor_B)[T.B6]         37.2414      3.607     10.325      0.000      30.138      44.344
C(Factor_A)[T.A7]:C(Factor_B)[T.B6]         46.3448      3.607     12.849      0.000      39.242      53.448
C(Factor_A)[T.A8]:C(Factor_B)[T.B6]         52.8621      3.607     14.656      0.000      45.759      59.965
C(Factor_A)[T.A2]:C(Factor_B)[T.B7]         14.8276      3.607      4.111      0.000       7.725      21.931
C(Factor_A)[T.A3]:C(Factor_B)[T.B7]         19.0805      3.607      5.290      0.000      11.978      26.183
C(Factor_A)[T.A4]:C(Factor_B)[T.B7]         31.0345      3.607      8.604      0.000      23.932      38.137
C(Factor_A)[T.A5]:C(Factor_B)[T.B7]         32.6897      3.607      9.063      0.000      25.587      39.793
C(Factor_A)[T.A6]:C(Factor_B)[T.B7]         40.1149      3.607     11.122      0.000      33.012      47.218
C(Factor_A)[T.A7]:C(Factor_B)[T.B7]         49.1724      3.607     13.633      0.000      42.069      56.275
C(Factor_A)[T.A8]:C(Factor_B)[T.B7]         56.2759      3.607     15.602      0.000      49.173      63.379
C(Factor_A)[T.A2]:C(Factor_B)[T.B8]         18.2759      3.607      5.067      0.000      11.173      25.379
C(Factor_A)[T.A3]:C(Factor_B)[T.B8]         22.5287      3.607      6.246      0.000      15.426      29.632
C(Factor_A)[T.A4]:C(Factor_B)[T.B8]         34.3103      3.607      9.512      0.000      27.207      41.413
C(Factor_A)[T.A5]:C(Factor_B)[T.B8]         34.3448      3.607      9.522      0.000      27.242      41.448
C(Factor_A)[T.A6]:C(Factor_B)[T.B8]         42.4138      3.607     11.759      0.000      35.311      49.517
C(Factor_A)[T.A7]:C(Factor_B)[T.B8]         51.6552      3.607     14.321      0.000      44.552      58.758
C(Factor_A)[T.A8]:C(Factor_B)[T.B8]         59.0345      3.607     16.367      0.000      51.932      66.137
==============================================================================
Omnibus:                      105.724   Durbin-Watson:                   2.263
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1326.886
Skew:                           0.968   Prob(JB):                    7.42e-289
Kurtosis:                      12.786   Cond. No.                         79.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

F.2 Proportion of free transcripts in the transcriptome

F.2.1 Analysis of variance for the two-factor fixed effects model

F.2.1.1 Proportion of free transcripts response: 2-way ANOVA table

```
ANOVA Table:
                      Sum Of Squares      DF         F      PR(>F)
C(Factor_A)                169981.04     7.0   1919.25  3.41e-217
C(Factor_B)                  5973.01     7.0     67.44   1.55e-54
C(Factor_A):C(Factor_B)      3450.37    49.0      5.57   2.22e-20
Residual                     3239.00   256.0

Effect Sizes (Partial Eta Squared) % of variance explained:
                      Sum Of Squares      DF      Eta_Squared
C(Factor_A)                169981.04     7.0      0.931
C(Factor_B)                  5973.01     7.0      0.033
C(Factor_A):C(Factor_B)      3450.37    49.0      0.019
Residual                     3239.00   256.0      0.018
```

F.2.2 Diagnostics case statistics and model adequacy checking

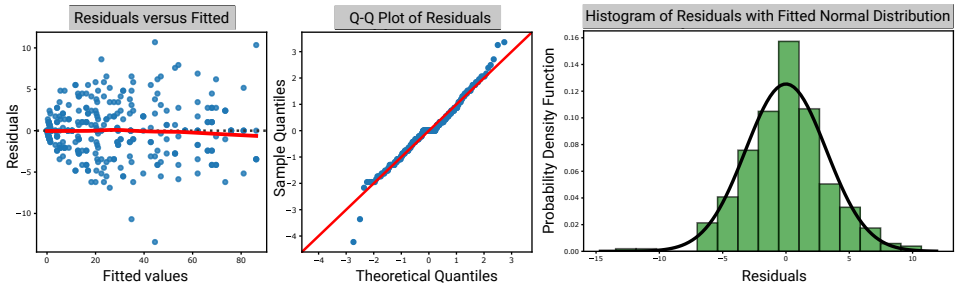


Figure F.4: Left: Plot of residuals versus predicted. Center: Quantile probability plot of residuals (Q-Q plot). Right: Fitted normal probability density to residuals.

The Kolmogorov-Smirnov test conducted on the residuals is not rejected at $\alpha = 0.005$ and we conclude that the residuals are likely normally distributed (KS Statistic: 0.09375; p-value: 0.007).

F.2.3 Conclude on hypothesis tests

F.2.3.1 Main effects

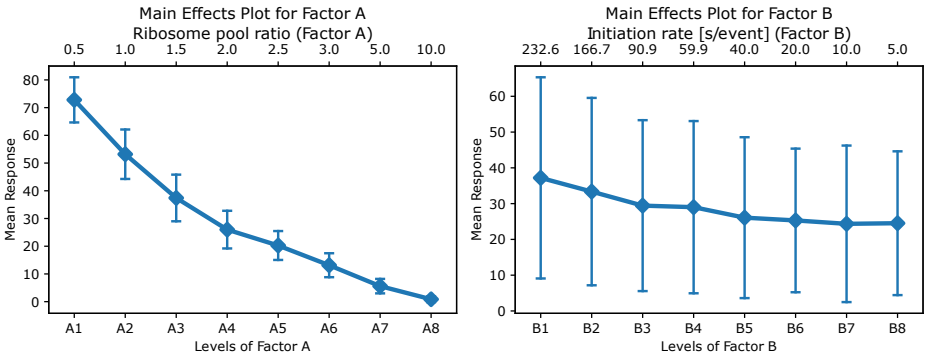


Figure F.5: Left: Main effects of the ribosome pool ratio on the proportion of free transcripts. Right: Main effects of the initiation rate on the proportion of free transcripts.

F.2.3.2 Interaction

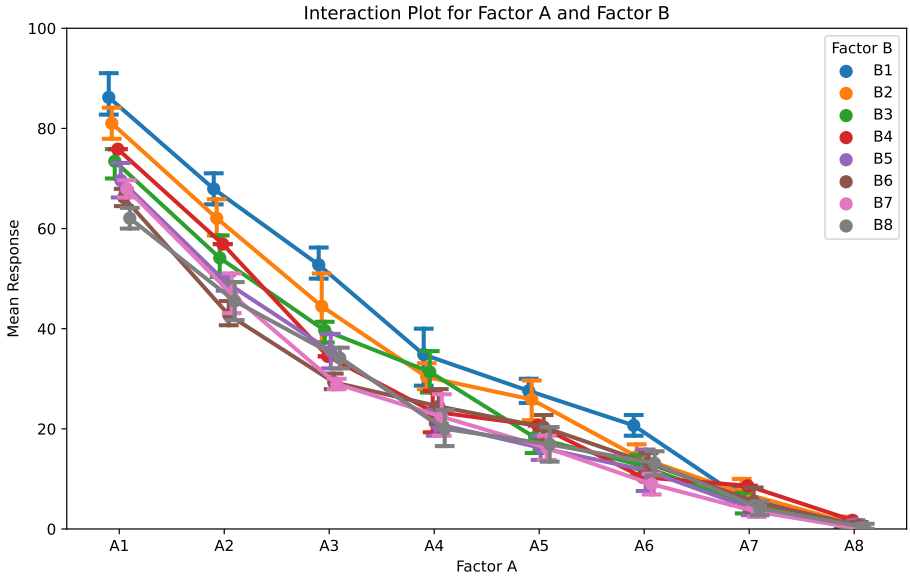


Figure F.6: Interaction plot of ribosome ratio (factor A) and initiation rate (factor B) on the proportion of free transcripts (response). The strength of the statistical interaction beyond the main effects depends on how the curves converge or diverge relative to each other. The more crossings between the curves, the stronger the interaction.

F.2.4 Model parameters estimation for proportion of free transcripts response

The local sensitivity coefficients are quantified by the slopes of the response variable as a function of the varied factors. The least squares estimates of the regression coefficients for the fixed effects model as described in eq. (11.11) are tabulated hereafter.

F.2.5 Fitting Response surfaces

The code to generate a response surface with a second degree polynomial, produce the contour plots, with specified contour lines and response surface is given in the following framed box.

Python code

```
# Fit a response Surface Model
# Use a second-order polynomial regression model to fit the data

# Recall the data:
# Create a grid of Factor_A and Factor_B
factor_a_levels = np.linspace(1, 8, 8)  # 8 levels for Factor A
factor_b_levels = np.linspace(1, 8, 8)  # 8 levels for Factor B

Factor_A, Factor_B = np.meshgrid(factor_a_levels, factor_b_levels)

# Average the response across the replicates (axis=2):
freeRibo_avg = riboFreeProp_arr.mean(axis=2)
freeTr_avg = trFreeProp_arr.mean(axis=2)

data = pd.DataFrame({
    'Factor_A': Factor_A.ravel(),
    'Factor_B': Factor_B.ravel(),
    'Response': freeTr_avg.ravel()
    #'Response': freeRibo_avg.ravel()
})

# Fit a quadratic regression model:
X = data[['Factor_A', 'Factor_B']]
y = data['Response']

# Generate polynomial features (interaction terms, squares, etc.)
poly = PolynomialFeatures(degree=2)
X_poly = poly.fit_transform(X)

# Fit the polynomial regression model
model = LinearRegression().fit(X_poly, y)

# Print the coefficients of the quadratic model
print("Model Coefficients:")
print(model.coef_)
print("Intercept:", model.intercept_)
```

```
# Plot the response surface contour and highlight specific contour levels:

# Generate predictions over a grid for contour plots
x_range = np.linspace(1, 8, 100)
y_range = np.linspace(1, 8, 100)
X_grid, Y_grid = np.meshgrid(x_range, y_range)
Z_grid = model.predict(poly.transform(np.c_[X_grid.ravel(),
Y_grid.ravel()])).reshape(X_grid.shape)

# Create a 2D contour plot
plt.figure(figsize=(8, 6))
contour = plt.contour(X_grid, Y_grid, Z_grid, levels=15, cmap='YlGn')
plt.clabel(contour, inline=True, fontsize=12)
plt.title('2D Contour Plot of Response Surface
[% of free transcripts in the transcriptome]')
plt.xlabel('RIBOSOME POOL RATIO')
plt.ylabel('INITIATION RATE [s/event]')

# --- Custom Labels ---
# Specify custom row and column labels
row_labels = [0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 5.0, 10.0] # i
rate_labels = np.array([4.3e-6, 6e-6, 11e-6, 16.7e-6, 25e-6, 50e-6, 100e-6, 200e-6])
column_labels = np.divide(0.001, rate_labels).tolist()
# Format each element to '.1f'
formatted_column_labels = [f'{x:.1f}' for x in column_labels]
# Create the mappings from the numerical factor levels to the custom labels
row_ticks = np.linspace(min(factor_a_levels), max(factor_a_levels),
len(row_labels))
col_ticks = np.linspace(min(factor_b_levels), max(factor_b_levels),
len(column_labels))

# Set custom ticks for Factor A and Factor B
plt.xticks(row_ticks, row_labels)
plt.yticks(col_ticks, formatted_column_labels)

# Highlight specific contours at 15% and 20%
plt.contour(X_grid, Y_grid, Z_grid, levels=[15], colors='black',
linestyles='dashed', linewidths=2)
plt.contour(X_grid, Y_grid, Z_grid, levels=[20], colors='red',
linestyles='dashed', linewidths=2)

plt.savefig(fPATHtoFig+'\\'+'FreeTrcontour_plot.pdf')
plt.show()

#Generate a 3D plot with a contour plot underneath:

# Create a 3D response surface plot
fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')

# Plot the surface
ax.plot_surface(X_grid, Y_grid, Z_grid, cmap='YlGn', alpha=0.8, edgecolor='None')

# Plot the contour underneath (cmap = 'Oranges' or 'YlGn')
ax.contour(X_grid, Y_grid, Z_grid, levels=15, cmap='YlGn', offset=np.min(Z_grid))

# Highlight specific contours at 15% and 20%
ax.contour(X_grid, Y_grid, Z_grid, levels=[15], colors='black',
linestyles='dashed', linewidths=2.5,
offset=np.min(Z_grid))
ax.contour(X_grid, Y_grid, Z_grid, levels=[20], colors='red',
```

```
linestyles='dashed', linewidths=2.5,
offset=np.min(Z_grid))

# Add the meshgrid (wireframe) on the surface
x_A = np.linspace(1, 8, 8)
x_B = np.linspace(1, 8, 8)
wireA, wireB = np.meshgrid(x_A, x_B)
Z_grid = model.predict(poly.transform(np.c_[wireA.ravel(),
wireB.ravel()])).reshape(wireA.shape)

ax.plot_wireframe(wireA, wireB, Z_grid, color='grey', linewidth=0.5)

ax.set_title('3D Response Surface with Contour Plot')
ax.set_xlabel('RIBOSOME POOL RATIO')
ax.set_ylabel('INITIATION RATE [s/event]')
ax.set_zlabel('RESPONSE')

# --- Custom Labels ---
# Specify custom row and column labels
row_labels = [0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 5.0, 10.0] # i
rate_labels = np.array([4.3e-6, 6e-6, 11e-6, 16.7e-6, 25e-6, 50e-6, 100e-6, 200e-6])
column_labels = np.divide(0.001, rate_labels).tolist()
# Format each element to '.1f'
formatted_column_labels = [f'{x:.1f}' for x in column_labels]
# Create the mappings from the numerical factor levels to the custom labels
row_ticks = np.linspace(min(factor_a_levels), max(factor_a_levels),
len(row_labels))
col_ticks = np.linspace(min(factor_b_levels), max(factor_b_levels),
len(column_labels))

# Set custom ticks for Factor A and Factor B
ax.set_xticks(row_ticks)
ax.set_xticklabels(row_labels)

ax.set_yticks(col_ticks)
ax.set_yticklabels(formatted_column_labels)

# Set the Z-axis limit (min value automatically set to Z.min(), max value to 100)
#ax.set_zlim(Z_grid.min(), 100)
ax.set_zlim(0, 100)

plt.savefig(fPATHtoFig+'\\'+'3d_TrSurface_plot.pdf')
plt.show()

sys.exit()
```

Bibliography

Agarwal, R. and T. Gaddis (Oct. 2014). *Starting Out with Python, 3rd edition*. ISBN: 1292065508.

Akaike, H. (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

An, G., B. Fitzpatrick, S Christley, P. Federico, A. Kanarek, R. Miller Neilan, M Oremland, R. Salinas, R Laubenbacher, and S Lenhart (Nov. 2016). "Optimization and Control of Agent-Based Models in Biology: A Perspective". In: *Bulletin of mathematical biology* 79.

An, G., Q. Mi, J. Dutta-Moscato, and Y. Vodovotz (2009). "Agent-based models in translational systems biology". In: *WIREs Systems Biology and Medicine* 1.2, pp. 159–171.

Anslyn, E. and D. Dougherty (2005). *Energy Surfaces and Kinetic Analysis. In Modern Physical Organic Chemistry*. University Science Books Saucalito CA USA, pp. 355–419.

Arava, Y., F. Boas, and D. Herschlag (Feb. 2005). "Dissecting eukaryotic translation and its control by ribosome density mapping". In: *Nucleic acids research* 33, pp. 2421–32.

Arrasate, M. and S. Finkbeiner (Dec. 2011). "Protein aggregates in Huntington's disease". In: *Experimental neurology* 238, pp. 1–11.

Artieri, C. and H. Fraser (Oct. 2014). "Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation". In: *Genome Research* 24.

Auffinger, P., E. Ennifar, and L. D'Ascenso (Apr. 2021). "Deflating the Mg2+ bubble. Stereochemistry to the rescue!" In: *RNA* 27, pp. 243–252.

Baker, N., D. Sept, S. Joseph, M. Holst, and J. McCammon (Aug. 2001). "Electrostatics of Nanosystems: Application to Microtubules and the Ribosome". In: *Proc Nat Acad Sci USA* 98, pp. 10037–10041.

Barton, G. ((1989)). *Elements of Green's functions and Propagation*. Oxford University Press, pp. 7–38.

Bell, G. (1978). "Models for the specific adhesion of cells to cells". In: *Science* 200.4342, pp. 618–627.

Berg, J. M., J. L. Tymoczko, G. J. G. Jr., and L. Stryer (2019). *Biochemistry*. 10th ed. New York: W.H. Freeman and Company.

Berg, M. D. and C. J. Brandl (2021). "Transfer RNAs: diversity in form and function". In: *RNA Biology* 18.3, pp. 316–339.

Beringer, M. and M. Rodnina (June 2007). "The Ribosomal Peptidyl Transferase". In: *Molecular Cell* 26, pp. 311–21.

Berka, K., O. Hanak, D. Sehnal, P. Banás, V. Navrátilová, D. Jaiswal, C.-M. Ionescu, R. Svobodová Vařeková, J. Koca, and M. Otyepka (May 2012). "MOLEonline 2.0: Interactive web-based analysis of biomacromolecular channels". In: *Nucleic acids research* 40, W222–7.

Bhatt, P. R., A. Scaiola, G. Loughran, M. Leibundgut, A. Kratzel, R. Meurs, R. Dreos, K. M. O'Connor, A. McMillan, J. W. Bode, V. Thiel, D. Gatfield, J. F. Atkins, and N. Ban (2021). "Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome". In: *Science* 372.6548, pp. 1306–1313.

BioBix Lab, Ghent University (2015). *Proteoformer: A Proteogenomic Pipeline*. Retrieved March 7, 2025. URL: `https://biobix.ugent.be/research/downloads/proteoformer/`.

Bonfanti, S., M. C. Lionetti, M. Fumagalli, V. Chirasani, G. Tiana, N. Dokholyan, S. Zapperi, and C. La Porta (Dec. 2019). "Molecular mechanisms of heterogeneous oligomerization of huntingtin proteins". In: *Scientific Reports* 9.

Brooks, B. and et al (2009). "CHARMM: The biomolecular simulation program". In: *J. Comput. Chem.* 30, p. 1545.

Buhr, F., S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M. V. Rodnina, and A. A. Komar (2016). "Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations". In: *Molecular Cell* 61.3, pp. 341–351.

Bustamante, C., Y. R. Chemla, N. R. Forde, and D. Izhaky (2004). "Mechanical Processes in Biochemistry". In: *Annual Review of Biochemistry* 73.1, pp. 705–748.

Campbell, N. A., L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, and J. B. Reece (2020). *Biology: A global approach*. 12th Global edition. Pearson Education Limited.

Cannarozzi, G., N. Schraudolph, M. Faty, P. von Rohr, M. Friberg, A. Roth, P. Gonnet, G. Gonnet, and Y. Barral (Apr. 2010). "A Role for Codon Order in Translation Dynamics". In: *Cell* 141, pp. 355–67.

Charneski, C. A. and L. D. Hurst (Mar. 2013). "Positively Charged Residues Are the Major Determinants of Ribosomal Velocity". In: *PLOS Biology* 11.3, pp. 1–20.

Chen, M. and P. Wolynes (Apr. 2017). "Aggregation landscapes of Huntingtin exon 1 protein fragments and the critical repeat length for the onset of Huntington's disease". In: *Proc Nat Acad Sci USA* 114, p. 201702237.

Cockman, E., P. Anderson, and P. Ivanov (June 2020). "TOP mRNPs: Molecular Mechanisms and Principles of Regulation". In: *Biomolecules* 10, p. 969.

Crappé, J., E. Ndah, A. Koch, S. Steyaert, D. Gawron, S. De Keulenaer, E. De Meester, T. De Meyer, W. Van Criekinge, P. Van Damme, and G. Menschaert (Dec. 2014).

"PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration". In: *Nucleic Acids Research* 43.5, e29–e29.

Cuervo, A., P. Dans, J. Carrascosa, M. Orozco, G. Gomila, and L. Fumagalli (Aug. 2014). "Direct measurement of the dielectric polarization properties of DNA". In: *Proc Nat Acad Sci USA* 111.

Dagnelie, P. (1981). *Principes d'Experimentation*. Presses Agronomiques de Gembloux. ISBN: 2-87016-024-0.

– (1986). "Analyse de la variance à deux critères de classification". In: *Théorie et méthodes statistiques, Vol.2*. Second Edition. Presses Agronomiques de Gembloux. Chap. 15, pp. 153–212. ISBN: 2-87016-010-0.

Dana, A. (Oct. 2014). "Properties and determinants of codon decoding time distributions". In: *BMC Genomics* 15, S13.

Dana, A. and T. Tuller (Nov. 2012). "Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells". In: *PLoS computational biology* 8, e1002755.

– (2014). "The effect of tRNA levels on decoding times of mRNA codons". In: *Nucleic Acids Research* 42, pp. 9171 –9181.

Dao Duc, K., S. Batra, N. Bhattacharya, J. Cate, and Y. Song (Feb. 2019). "Differences in the path to exit the ribosome across the three domains of life". In: *Nucleic Acids Research* 47.

Dao Duc, K. and Y. Song (Jan. 2018). "The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation". In: *PLOS Genetics* 14, e1007166.

Desai, V., F. Frank, A. Lee, M. Righini, L. Lancaster, H. Noller, I. Tinoco, and C. Bustamante (Aug. 2019). "Co-temporal Force and Fluorescence Measurements Reveal a Ribosomal Gear Shift Mechanism of Translation Regulation by Structured mRNAs". In: *Molecular Cell* 75.

Deutsch, C. (2014). "Tunnel Vision: Insights from Biochemical and Biophysical Studies". In: *In: Ito K. (eds) Regulatory Nascent Polypeptides. Springer, Tokyo.*

Dever, T. E. (2002). "Gene-Specific Regulation by General Translation Factors". In: *Cell* 108.4, pp. 545–556.

Diament, A. and T. Tuller (2016). "Estimation of ribosome profiling performance and reproducibility at various levels of resolution". In: *Biol Direct* 11, p. 24.

Ding, Y. and C. Lawrence (Apr. 2001). "Statistical Prediction of Single-Stranded Regions in RNA Secondary Structure and Application to Predicting Effective Antisense Target Sites and Beyond". In: *Nucleic acids research* 29, pp. 1034–46.

Dmitriev, S., K. Lashkevich, K. Akulich, and D. Vladimirov (Nov. 2020). "A Quick Guide to Small-Molecule Inhibitors of Eukaryotic Protein Synthesis". In: *Biochemistry (Moscow)* 85, pp. 1389–1421.

Dobin, A., C. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. Gingeras (Oct. 2012). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics (Oxford, England)* 29.

Doerfel, L., I. Wohlgemuth, C. Kothe, F. Peske, H. Urlaub, and M. Rodnina (Dec. 2012). "EF-P Is Essential for Rapid Synthesis of Proteins Containing Consecutive Proline Residues". In: *Science (New York, N.Y.)* 339. DOI: 10.1126/science.1229017.

Doerfel, L. K., I. Wohlgemuth, V. Kubyshkin, A. L. Starosta, D. N. Wilson, N. Budisa, and M. V. Rodnina (2015). "Entropic Contribution of Elongation Factor P to Proline Positioning at the Catalytic Center of the Ribosome". In: *Journal of the American Chemical Society* 137.40, pp. 12997–13006. DOI: 10.1021/jacs.5b07427.

Doris, S., D. Smith, J. Beamesderfer, B. Raphael, J. Nathanson, and S. Gerbi (Aug. 2015). "Universal and domain-specific sequences in 23S–28S ribosomal RNA identified by computational phylogenetics". In: *RNA (New York, N.Y.)* 21.

El Hachem, N., M. Leclercq, M. Ruiz, R. Vanleyssem, K. Shostak, P.-R. Körner, C. Capron, L. Martin-Morales, P. Roncarati, A. Lavergne, A. Blomme, S. Turchetto, E. Goffin, P. Thandapani, I. Tarassov, L. Nguyen, B. Pirotte, A. Chariot, J.-C. Marine, and P. Close (June 2024). "Valine aminoacyl-tRNA synthetase promotes therapy resistance in melanoma". In: *Nature Cell Biology*.

ELIXIR (2025). *ELIXIR Europe: A Distributed Infrastructure for European Biological Data*. Retrieved March 7, 2025. URL: https://elixir-europe.org.

ELIXIR Belgium (2025). *ELIXIR Belgium: Empowering Life Science Research*. Retrieved March 7, 2025. URL: https://www.elixir-belgium.org.

Ensembl (1999). *Ensembl: A Genome Browser for Vertebrate Genomes*. Retrieved March 7, 2025. URL: https://www.ensembl.org.

Erlang, A. K. (1948). "Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges". In: *in Brockmeyer, E.; Halstrøm, H. L.; Jensen, Arne (eds.), The Life and Works of A.K. Erlang, Transactions of the Danish Academy of Technical Sciences, 2, Akademiet for de Tekniske Videnskaber, pp. 138–155, archived from the original (PDF) on July 19, 2011*.

European Nucleotide Archive (2009). *European Nucleotide Archive: ENA*. Retrieved March 7, 2025. URL: https://www.ebi.ac.uk/ena.

Eyring, H. (1935). "The Activated Complex in Chemical Reactions". In: *Journal of Chemical Physics* 3, pp. 107–115.

Farewell, A. and F. C. Neidhardt (1998). "Effect of Temperature on In Vivo Protein Synthetic Capacity in Escherichia coli". In: *Journal of Bacteriology* 180.17, pp. 4704–4710.

Fluitt, A., E. Pienaar, and H. Viljoen (2007). "Ribosome Kinetics and aa-tRNA Competition Determine Rate and Fidelity of Peptide Synthesis". In: *Comput Biol Chem* 31.5-6, pp. 335–346.

Friberg, M., P. Gonnet, Y. Barral, N. Schraudolph, and G. Gonnet (2006). "Measures of Codon Bias in Yeast, the tRNA Pairing Index and Possible DNA Repair Mechanisms". In: *P. Bucher and B. Moret (eds), Proceedings of the 6th Workshop on Algorithms in Bioinformatics (WABI), vol. 4175 of Lecture Notes in Bioinformatics (Springer Verlag, Berlin)*.

Fried, S., S. Bagchi, and S. Boxer (2014). “Extreme electric fields power catalysis in the active site of ketosteroid isomerase”. In: *Science* 346, pp. 1510–14.

Fried, S. D. and S. G. Boxer (2017). “Electric Fields and Enzyme Catalysis”. In: *Annual Review of Biochemistry* 86.1, pp. 387–415.

Fritch, B., A. Kosolapov, P. Hudson, D. A. Nissley, H. L. Woodcock, C. Deutsch, and E. P. O’Brien (2018). “Origins of the Mechanochemical Coupling of Peptide Bond Formation to Protein Synthesis”. In: *Journal of the American Chemical Society* 140.15, pp. 5077–5087.

Gabdulkhakov, A., S. Nikonov, and M. Garber (June 2013). “Revisiting the *Haloarcula marismortui* 50S ribosomal subunit model”. In: *Acta crystallographica. Section D, Biological crystallography* 69, pp. 997–1004.

Gallez, A., S. Blacher, E. Maquoi, E. Konradowski, M. Joiret, I. Primac, C. Gérard, M. Taziaux, R. Houtman, L. Geris, et al. (2021). “Estetrol combined to progestogen for menopause or contraception indication is neutral on breast cancer”. In: *Cancers* 13.10, p. 2486.

Gandin, V., K. Sikström, T. Alain, M. Morita, S. Mclaughlan, O. Larsson, and I. Topisirovic (May 2014). “Polysome Fractionation and Analysis of Mammalian Translatomes on a Genome-wide Scale”. In: *Journal of visualized experiments : JoVE*.

Gentilella, A., S. Kozma, and G. Thomas (Feb. 2015). “A liaison between mTOR signaling, ribosome biogenesis and cancer”. In: *Biochimica et biophysica acta* 1849.

Gerber, A. and W. Keller (July 2001). “RNA editing by base deamination: more enzymes, more targets, new mysteries”. In: *Trends in biochemical sciences* 26, pp. 376–384. DOI: 10.1016/S0968-0004(01)01827-8.

Gerber, A. and W. Keller (Dec. 1999). “An adenosine deaminase that generates inosine at the wobble position of tRNAs”. In: *Science (New York, N.Y.)* 286, pp. 1146–1149. DOI: 10.1126/science.286.5442.1146.

Goldman, D., C. Kaiser, A. Milin, M. Righini, I. Tinoco, and C. Bustamante (Apr. 2015). “Mechanical force releases nascent chain-mediated ribosome arrest in vitro and in vivo”. In: *Science* 348, pp. 457–60.

Gorochowski, T., Z. Ignatova, R. Bovenberg, and H. Roubos (2015a). “Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate”. In: *Nucleic Acids Research* 43.

Gorochowski, T., Z. Ignatova, R. Bovenberg, and H. Roubos (Mar. 2015b). “Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate”. In: *Nucleic Acids Research* 43.

Greulich, P., L. Ciandrini, R. Allen, and M. Romano (Jan. 2012). “Mixed population of competing totally asymmetric simple exclusion processes with a shared reservoir of particles”. In: *Phys Rev E, Statistical, nonlinear, and soft matter physics* 85, p. 011142.

Gritsenko, A., M. Hulsman, M. Reinders, and D. Ridder (Aug. 2015). "Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data". In: *PLoS computational biology* 11, e1004336.

Gruber, A. R., R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker (Apr. 2008). "The Vienna RNA Websuite". In: *Nucleic Acids Research* 36.suppl2, W70–W74.

Grushka, E. (1972). "Characterization of exponentially modified Gaussian peaks in chromatography". In: *Analytical Chemistry* 44.11, pp. 1733–1738.

GWIPS-viz (2013). *Genome-Wide Information on Protein Synthesis Visualization (GWIPS-viz)*. Retrieved March 7, 2025. URL: `http://gwips.ucc.ie`.

Haar, T. von der (2012). "Mathematical and computational modelling of ribosomal movement and protein synthesis: An overview". In: *Comput. Struct. Biotechnol. J.* 1, e201204002.

Haar, T. von der (2008). "A quantitative estimation of the global translational activity in logarithmically growing yeast cells". In: *BMC Syst Biol* 2.87.

Harris, D. C. ((1988)). *Quantitative Chemical Analysis*. Second edition. W.H. Freeman and Company, pp. 90–92.

Hatters, D. M. (2008). "Protein misfolding inside cells: The case of huntingtin and Huntington's disease". In: *IUBMB Life* 60.11, pp. 724–728.

He, S. L. and R. Green (2013). "Chapter Ten - Polysome Analysis of Mammalian Cells". In: *Laboratory Methods in Enzymology: RNA*. Ed. by J. Lorsch. Vol. 530. Methods in Enzymology. Academic Press, pp. 183–192.

Hofacker, I., W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster (1994). "Fast folding and comparison of RNA secondary structures". In: *Monatshefte für Chemie Chemical Monthly* 125.2, 167 – 188.

Hsiao, C. and L. Williams (Apr. 2009). "A recurrent magnesium-binding motif provides a framework for the ribosomal peptidyl transferase center". In: *Nucleic Acids Research* 37, pp. 3134–42.

Ingolia, N. (Dec. 2010). "Genome-Wide Translational Profiling by Ribosome Footprinting". In: *Methods in enzymology* 470, pp. 119–42.

Ingolia, N., G. Brar, S. Rouskin, A. Mcgeachy, and J. Weissman (July 2012). "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments". In: *Nature protocols* 7, pp. 1534–50.

Ingolia, N., L. Lareau, and J. Weissman (Nov. 2011). "Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes". In: *Cell* 147, pp. 789–802.

Ingolia, N. T., S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman (2009). "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling". In: *Science* 324.5924, pp. 218–223.

Jackson, J. D. ((1998)). *Classical Electrodynamics*. Third edition. Wiley & Sons, p. 32.

Javaux, E. (Aug. 2019). "Challenges in evidencing the earliest traces of life". In: *Nature* 572, pp. 451–460.

Johansson, M., K.-W. Ieong, S. Trobro, P. Strazewski, J. Åqvist, M. Y. Pavlov, and M. Ehrenberg (2011). “pH-sensitivity of the ribosomal peptidyl transfer reaction dependent on the identity of the A-site aminoacyl-tRNA”. In: *Proceedings of the National Academy of Sciences* 108.1, pp. 79–84.

Johansson, R. ((2019)). *Numerical Python*. Second edition. Apress, pp. 267–293.

Joiret, M., F. Kerff, F. Rapino, P. Close, and L. Geris (2022a). “Ribosome Exit Tunnel Electrostatics”. In: *Phys. Rev. E* 105.1, pp. 1–43.

– (2023a). “A simple geometrical model of the electrostatic environment around the catalytic center of the ribosome and its significance for the elongation cycle kinetics”. In: *Computational and Structural Biotechnology Journal* 21, pp. 3768–3795. DOI: 10.1016/j.csbj.2023.07.016.

Joiret, M., J. Mahachie John, E. Gusareva, and K Van Steen (2019). “Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies”. In: *BioData Mining* 12.11.

Joiret, M., F. Kerff, F. Rapino, P. Close, and L. Geris (2022b). “Ribosome exit tunnel electrostatics”. In: *Physical Review E* 105.1, p. 014409. DOI: 10.1103/PhysRevE.105.014409.

– (2024). “Reversing the relative time courses of the peptide bond reaction with oligopeptides of different lengths and charged amino acid distributions in the ribosome exit tunnel”. In: *Computational and Structural Biotechnology Journal* 23, pp. 2453–2464. DOI: 10.1016/j.csbj.2024.05.045.

Joiret, M., M. Leclercq, G. Lambrechts, F. Rapino, P. Close, G. Louppe, and L. Geris (2023b). “Cracking the genetic code with neural networks”. In: *Frontiers in Artificial Intelligence* 6.

Kaiser, C. and I. Tinoco (Jan. 2014). “Probing the Mechanisms of Translation with Force”. In: *Chemical Reviews* 114.

Karlsborn, T., H. Tükenmez, A. K. M. Mahmud, F. Xu, H. Xu, and A. Byström (Jan. 2015). “Elongator, a conserved complex required for wobble uridine modifications in Eukaryotes”. In: *RNA biology* 11.

Kazibwe, Z., A.-Y. Liu, G. Macintosh, and D. Bassham (Dec. 2019). “The Ins and Outs of Autophagic Ribosome Turnover”. In: *Cells* 8, p. 1603.

Kiniry, S., A. Michel, and P. Baranov (Nov. 2019). “Computational methods for ribosome profiling data analysis”. In: *Wiley Interdisciplinary Reviews: RNA* 11.

Klein, D., P. Moore, and T. Steitz (Oct. 2004). “The contribution of metal ions to the structural stability of the large ribosomal subunit”. In: *RNA* 10, pp. 1366–79.

Komar, A. (Nov. 2019). “Synonymous Codon Usage—a Guide for Co-Translational Protein Folding in the Cell”. In: *Molecular Biology* 53, pp. 777–790.

Komar, A. A., T. Lesnik, and C. Reiss (1999). “Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation”. In: *FEBS Letters* 462.3, pp. 387–391.

Kullback, S. and R. A. Leibler (1951). “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86.

Laidler, K. J. and M. C. King (1983). "Development of transition-state theory". In: *The Journal of Physical Chemistry* 87.15, pp. 2657–2664.

Lang, K., M. Erlacher, D. N. Wilson, R. Micura, and N. Polacek (2008). "The Role of 23S Ribosomal RNA Residue A2451 in Peptide Bond Synthesis Revealed by Atomic Mutagenesis". In: *Chemistry and Biology* 15.5, pp. 485–492.

Langmead, B., C. Trapnell, M. Pop, and S. Salzberg (Apr. 2009). "Langmead B, Trapnell C, Pop M, Salzberg SL.. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25". In: *Genome biology* 10, R25.

Laplante, M. and D. Sabatini (Oct. 2009). "MTOR signaling at a glance". In: *Journal of cell science* 122, pp. 3589–94.

Larsson, O., N. Sonenberg, and R. Nadon (Nov. 2010). "Identification of differential translation in genome wide studies". In: *Proceedings of the National Academy of Sciences of the United States of America* 107, pp. 21487–92.

Larsson, O., B. Tian, and N. Sonenberg (Dec. 2012). "Toward a Genome-Wide Landscape of Translational Control". In: *Cold Spring Harbor perspectives in biology* 5.

Lassak, J., E. Keilhauer, M. Fürst, K. Wuichet, J. Godeke, A. Starosta, J.-M. Chen, L. Søgaard-Andersen, J. Rohr, D. Wilson, S. Haussler, M. Mann, and K. Jung (Feb. 2015a). "Arginine-rhamnosylation as new strategy to activate translation elongation factor P". In: *Nat Chem Biol* 11.

Lassak, J., D. Wilson, and K. Jung (Nov. 2015b). "Stall no more at polyproline stretches with the translation elongation factors EF-P and IF-5A". In: *Molecular microbiology* 99, pp. 219–235. DOI: 10.1111/mmi.13233.

Leppek, K., M. Barna, and R. Das (Nov. 2017). "Functional 5'UTR mRNA structures in eukaryotic translation regulation and how to find them". In: *Nature Reviews Molecular Cell Biology* 19.

Levy, E. (Oct. 2020). "On the density for sums of independent exponential, Erlang and gamma variates". In: *arXiv:2006.12428v4*.

Liu, T., A. Kaplan, L. Alexander, S. Yan, J.-D. Wen, L. Lancaster, C. E. Wickersham, K. Fredrick, H. Noller, I. Tinoco, and C. Bustamante (2014a). "Direct measurement of the mechanical work during translocation by the ribosome". In: *eLife* 3.

Liu, T., A. Kaplan, L. Alexander, S. Yan, J.-D. Wen, L. Lancaster, C. E. Wickersham, K. Fredrick, H. Noller, J. Tinoco Ignacio, and C. J. Bustamante (2014b). "Direct measurement of the mechanical work during translocation by the ribosome". In: *eLife* 3. Ed. by X. Zhuang, e03406.

Liutkute, M., M. Maiti, E. Samatova, J. Enderlein, and M. V. Rodnina (2020). "Gradual compaction of the nascent peptide during cotranslational folding on the ribosome". In: *eLife* 9.

LLC, iChemLabs (2021). *ChemDoodle 3D Version 6.2.2*. URL: https://www.chemdoodle.com/3d. Last visited on 2021/03/29.

Lockhart, D. and P. Kim (1993). "Electrostatic screening of charge and dipole interactions with the helix backbone". In: *Science* 260.5105, pp. 198–202.

Lorenz, E. N. (1963). "Deterministic Nonperiodic Flow". In: *Journal of Atmospheric Sciences* 20.2, pp. 130 –141.

Lorenz, R., S. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. Stadler, and I. Hofacker (Nov. 2011). "Vienna RNA package 2.0". In: *Algorithms for molecular biology : AMB* 6, p. 26.

Lorenz, R., M. T. Wolfinger, A. Tanzer, and I. L. Hofacker (2016). "Predicting RNA secondary structures from sequence and probing data". In: *Methods* 103, pp. 86–98.

Lu, J. and C. Deutsch (Oct. 2008). "Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates". In: *Journal of Molecular Biology* 384, pp. 73–86.

Lu, J., W. R. Kobertz, and C. Deutsch (2007). "Mapping the Electrostatic Potential within the Ribosomal Exit Tunnel". In: *Journal of Molecular Biology* 371.5, pp. 1378–1391.

Lucent, D., C. Snow, C. Aitken, and V. Pande (Oct. 2010). "Non-Bulk-Like Solvent Behavior in the Ribosome Exit Tunnel". In: *PLoS Computational Biology* 6, e1000963.

Lyu, X., Q. Yang, L. Li, Y. Dang, Z. Zhou, S. Chen, and Y. Liu (June 2020). "Adaptation of codon usage to tRNA I34 modification controls translation kinetics and proteome landscape". In: *PLOS Genetics* 16, e1008836.

Ma, X. and J. Blenis (May 2009). "Ma XM, Blenis J.. Molecular mechanisms of mTOR-mediated translational control. Nat Rev Mol Cell Biol 10: 307-318". In: *Nature reviews. Molecular cell biology* 10, pp. 307–18.

MacDonald, C. T. and J. H. Gibbs (1969). "Concerning the kinetics of polypeptide synthesis on polyribosomes". In: *Biopolymers* 7.5, pp. 707–725.

MacDonald, C. T., J. H. Gibbs, and A. Pipkin (1968). "Kinetics of biopolymerization on nucleic acid templates". In: *Biopolymers* 6.1, pp. 1–25.

Mandal, A., S. Mandal, and M. H. Park (Nov. 2014). "Genome-Wide Analyses and Functional Classification of Proline Repeat-Rich Proteins: Potential Role of eIF5A in Eukaryotic Evolution". In: *PLOS ONE* 9.11, pp. 1–13.

Mao, Y., H. Liu, Y. Liu, and S. Tao (2014). "Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae". In: *Nucleic Acids Research* 42, pp. 4813 –4822.

Marchi, S. de and S. Page (May 2014). "Agent-Based Models". In: *Annual Review of Political Science* 17, pp. 1–20.

Marks, F., U. Klingmüller, and K. Müller-Decker (2017). *Cellular Signal Processing: An Introduction to the Molecular Mechanisms of Signal Transduction (2nd ed.)* Garland Science, Taylor and Francis Group.

Mayer, C., J. Zhao, X. Yuan, and I. Grummt (Mar. 2004). "mTOR-dependent activation of the transcription factor TIF-IA links rRNA synthesis to nutrient availability". In: *Genes & development* 18, pp. 423–34.

McCaskill, J. S. (1990). "The equilibrium partition function and base pair binding probabilities for RNA secondary structure". In: *Biopolymers* 29.6-7, pp. 1105–1119.

Melnikov, S., J. Mailliot, L. Rigger, S. Neuner, B.-S. Shin, G. Yusupova, T. E. Dever, R. Micura, and M. Yusupov (2016). "Molecular insights into protein synthesis with proline residues". In: *EMBO reports* 17.12, pp. 1776–1784. DOI: https://doi.org/10.15252/embr.201642943.

Mercier, E. and M. Rodnina (May 2018). "Co-Translational Folding Trajectory of the HemK Helical Domain". In: *Biochemistry* 57.

Meyuhas, O. and T. Kahan (2015). "The race to decipher the top secrets of TOP mRNAs". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1849.7, pp. 801–811.

Michel, A. and P. Baranov (Sept. 2013). "Ribosome profiling: A Hi-Def monitor for protein synthesis at the genome-wide scale". In: *Wiley interdisciplinary reviews. RNA* 4.

Milón, P., A. Konevega, C. Gualerzi, and M. Rodnina (July 2008). "Kinetic Checkpoint at a Late Step in Translation Initiation". In: *Molecular cell* 30, pp. 712–20.

Milón, P. and M. Rodnina (Apr. 2012). "Kinetic control of translation initiation in bacteria". In: *Critical reviews in biochemistry and molecular biology* 47, pp. 334–48.

Montgomery, D. C. (2013a). "Introduction to Factorial Designs". In: *Design and analysis of experiments*. Eighth Edition. John Wiley & Sons. Chap. 5, pp. 183–232.

– (2013b). "Randomized Blocks, Latin Squares, and Related Designs". In: *Design and analysis of experiments*. Eighth Edition. John Wiley & Sons. Chap. 4, pp. 139–231.

Morgan, A. and E. Rubenstein (Jan. 2013). "Proline: The Distribution, Frequency, Positioning, and Common Functional Roles of Proline and Polyproline Sequences in the Human Proteome". In: *PloS one* 8, e53785.

Morisaki, T., K. Lyon, K. Deluca, J. DeLuca, B. English, Z. Zhang, L. Lavis, J. Grimm, S. Viswanathan, L. Looger, T. Lionnet, and T. Stasevich (May 2016). "Real-time quantification of single RNA translation dynamics in living cells". In: *Science* 352.

Morris, K. V. and J. S. Mattick, eds. (2012). *RNA Worlds: From Life's Origins to Diversity in Gene Regulation*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Musuamba, F., I. Rusten, R. Lesage, G. Russo, R. Bursi, L. Emili, G. Wangorsch, E. Manolis, K. Karlsson, A. Kulesza, E. Courcelles, J.-P. Boissel, C. Rousseau, E. Voisin, R. Alessandrello, N. Curado, E. Dall'Ara, B. Rodriguez, F. Pappalardo, and L. Geris (July 2021). "Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility". In: *CPT: Pharmacometrics and Systems Pharmacology* 10, pp. 804–825.

National Center for Biotechnology Information (1992). *National Center for Biotechnology Information (NCBI)*. Retrieved March 7, 2025. URL: https://www.ncbi.nlm.nih.gov.

– (2000). *Sequence Read Archive: SRA*. Retrieved March 7, 2025. URL: https://www.ncbi.nlm.nih.gov/sra.

Nedialkova, D. and S. Leidel (June 2015). "Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity". In: *Cell* 161, pp. 1–13.

Nierhaus, K. (Sept. 2014). "Mg2+, K+, and the Ribosome". In: *Journal of bacteriology* 196, pp. 3817–19.

Nieß, A., M. Siemann-Herzberg, and R. Takors (2019). "Protein production in *Escherichia coli* is guided by the trade-off between intracellular substrate availability and energy cost". In: *Microbial Cell Factories* 18(1):8.

Nissen, P., J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz (2000). "The Structural Basis of Ribosome Activity in Peptide Bond Synthesis". In: *Science* 289.5481, pp. 920–930.

Nissley, D. A., Q. V. Vu, F. Trovato, N. Ahmed, Y. Jiang, M. S. Li, and E. P. O'Brien (2020). "Electrostatic Interactions Govern Extreme Nascent Protein Ejection Times from Ribosomes and Can Delay Ribosome Recycling". In: *Journal of the American Chemical Society* 142.13, pp. 6103–6110.

Noller, H. F., V. Hoffarth, and L. Zimniak (1992). "Unusual Resistance of Peptidyl Transferase to Protein Extraction Procedures". In: *Science* 256.5062, pp. 1416–1419.

Novoa, E. M. and L. Pouplana (Aug. 2012). "Speeding with control: Codon usage, tRNAs, and ribosomes". In: *Trends in genetics : TIG* 28, pp. 574–81.

Novoa, E., M. Pavon-Eternod, T. Pan, and L. Ribas de Pouplana (2012). "A Role for tRNA Modifications in Genome Structure and Codon Usage". In: *Cell* 149.1, pp. 202–213.

Nussinov, R and A. B. Jacobson (1980). "Fast algorithm for predicting the secondary structure of single-stranded RNA." In: *Proceedings of the National Academy of Sciences* 77.11, pp. 6309–6313.

Oguntunde, P., O. Odetunmibi, and A. O. Adejumo (Mar. 2014). "ON THE SUM OF EXPONENTIALLY DISTRIBUTED RANDOM VARIABLES: A CONVOLUTION APPROACH". In: *European Journal of Statistics and Probability* 2, pp. 1–8.

Pace, C., G. Grimsley, and J. M. Scholtz (Feb. 2009). "Protein Ionizable Groups: pK Values and Their Contribution to Protein Stability and Solubility". In: *The Journal of biological chemistry* 284, pp. 13285–9.

Pape, T, W Wintermeyer, and M. Rodnina (Jan. 1999). "Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the E. coli ribosome". In: *The EMBO journal* 17, pp. 7490–7.

Park, M.-H., R. Kar, S. Banka, A. Ziegler, and W. Chung (Apr. 2022). "Post-translational formation of hypusine in eIF5A: implications in human neurodevelopment". In: *Amino Acids* 54. DOI: 10.1007/s00726-021-03023-6.

Park, M.-H. and E. Wolff (Sept. 2018). "Hypusine, a polyamine-derived amino acid critical for eukaryotic translation". In: *Journal of Biological Chemistry* 293, jbc.TM118.003341. DOI: 10.1074/jbc.TM118.003341.

Pavlov, M., E. Watts, Z. Tan, V. Cornish, M. Ehrenberg, and A. Forster (Jan. 2009). "Slow peptide bond formation by proline and other N-alkylamino acids in translation". In: *Proceedings of the National Academy of Sciences of the United States of America* 106, pp. 50–4. DOI: 10.1073/pnas.0809211106.

Pavlov, M. Y., G. Ullman, Z. Ignatova, and M. Ehrenberg (Apr. 2021). "Estimation of peptide elongation times from ribosome profiling spectra". In: *Nucleic Acids Research* 49.9, pp. 5124–5142.

Pearson, K. (1896). "On a form of spurious correlation which may arise when indices are used in measurement of organs." In: *Proceedings of the London Royal Society* 60, pp. 489–498.

Peil, L., A. Starosta, J. Lassak, G. Atkinson, K. Virumäe, M. Spitzer, T. Tenson, K. Jung, J. Remme, and D. Wilson (Sept. 2013). "Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P". In: *Proceedings of the National Academy of Sciences* 110. DOI: 10.1073/pnas.1310642110.

Pelletier, J. and N. Sonenberg (June 2019). "The Organizing Principles of Eukaryotic Ribosome Recruitment". In: *Annual Review of Biochemistry* 88, pp. 307–335.

Petrone, P. M., C. Snow, D. Lucent, and V. Pande (2008). "Side-chain recognition and gating in the ribosome exit tunnel". In: *Proc. Natl. Acad. Sci USA* 105, pp. 16549–16554.

Petrov, A., C. Bernier, C. Hsiao, C. Okafor, E. Tannenbaum, J. Stern, E. Gaucher, D. Schneider, N. Hud, S. Harvey, and L. Williams (June 2012). "RNA-Magnesium-Protein Interactions in Large Ribosomal Subunit". In: *J Phys Chem B* 116, pp. 8113–8120.

Pfab, M., P. Kielkowski, R. Krafczyk, W. Volkwein, S. A. Sieber, J. Lassak, and K. Jung (2021). "Synthetic post-translational modifications of elongation factor P using the ligase EpmA". In: *The FEBS Journal* 288.2, pp. 663–677. DOI: https://doi.org/10.1111/febs.15346.

Polikanov, Y., S. Melnikov, D. Soll, and T. Steitz (Mar. 2015). "Structural insights into the role of rRNA modifications in protein synthesis and ribosome assembly". In: *Nature structural & molecular biology* 22.

Polikanov, Y., T. Steitz, and C Innis (Aug. 2014). "A proton wire to couple aminoacyl-tRNA accommodation and peptide-bond formation on the ribosome". In: *Nature structural & molecular biology* 21.

Power, L. (Mar. 2022). "Beginners guide to ribosome profiling". In: *The Biochemist* 44.

Pravda, L., D. Sehnal, D. Toušek, V. Navrátilová, V. Bazgier, K. Berka, R. Svobodová Vařeková, J. Koča, and M. Otyepka (Apr. 2018). "MOLEonline: A web-based tool for analyzing channels, tunnels and pores (2018 update)". In: *Nucleic acids research* 46.

Protein Data Bank (1999). *Protein Data Bank (PDB)*. Retrieved March 7, 2025. URL: https://www.rcsb.org.

Rafels-Ybern, A., C. Stephan-Otto Attolini, and L. Pouplana (Aug. 2015). "Distribution of ADAT-Dependent Codons in the Human Transcriptome". In: *International journal of molecular sciences* 16, pp. 17303–14. DOI: 10.3390/ijms160817303.

Rafels-Ybern, A., A. G. Torres, X. Grau-Bove, I. Ruiz-Trillo, and L. R. de Pouplana (2018). "Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla". In: *RNA Biology*

15.4-5. PMID: 28880718, pp. 500–507. DOI: 10.1080/15476286.2017.1358348. URL: https://doi.org/10.1080/15476286.2017.1358348.

Railsback, S. F., S. L. Lytinen, and S. K. Jackson (2006). "Agent-based Simulation Platforms: Review and Development Recommendations". In: *SIMULATION* 82.9, pp. 609–623.

Ramírez, V., B. Gonzalez, A. López, M. J. Castelló, M. J. Gil, B. Zheng, P. Chen, and P. Vera (Oct. 2015). "Loss of a Conserved tRNA Anticodon Modification Perturbs Plant Immunity". In: *PLOS Genetics* 11.10, pp. 1–27.

Ranjan, N. and M. V. Rodnina (2017). "Thio-Modification of tRNA at the Wobble Position as Regulator of the Kinetics of Decoding and Translocation on the Ribosome". In: *Journal of the American Chemical Society* 139.16, pp. 5857–5864.

Rapino, F, Z Zhou, A. Roncero Sanchez, M Joiret, C Seca, N El Hachem, G Valenti, S Latini, K Shostak, L Geris, and P. t. Li (Apr. 2021). "Wobble tRNA modification and hydrophilic amino acid patterns dictate protein fate." In: *Nat Commun.* 12(1), p. 2170.

Rapino, F., S. Delaunay, Z. Zhou, A. Chariot, and P. Close (Mar. 2017). "tRNA Modification: Is Cancer Having a Wobble?" In: *Trends in Cancer* 3.

Raveh, A., M. Margaliot, E. D. Sontag, and T. Tuller (2016). "A model for competition for ribosomes in the cell". In: *Journal of The Royal Society Interface* 13.116, p. 20151062.

Reis, M., R. Savva, and L. Wernisch (Feb. 2004). "Solving the riddle of codon usage preferences: A test for translational selection". In: *Nucleic acids research* 32, pp. 5036–44.

Requião, R., H. Souza, S. Rossetto, T. Domitrovic, and F. Palhano (Apr. 2016). "Increased ribosome density associated to positively charged residues is evident in ribosome profiling experiments performed in the absence of translation inhibitors". In: *RNA Biology* 13.

Requião, R. D., L. Fernandes, H. J. A. de Souza, S. Rossetto, T. Domitrovic, and F. L. Palhano (May 2017). "Protein charge distribution in proteomes and its impact on translation". In: *PLOS Computational Biology* 13.5, pp. 1–21.

Reynolds, C. W. (1987). "Flocks, herds and schools: A distributed behavioral model". In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '87. New York, NY, USA: Association for Computing Machinery, 25–34.

Rhodes, G. ((2006)). *Crystallography made crystal clear*. Third edition. Academic Press, pp. 183–185.

Riba, A., N. Di Nanni, N. Mittal, E. Arhné, A. Schmidt, and M. Zavolan (2019). "Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates". In: *Proc Nat Acad Sci USA* 116.30, pp. 15023–15032.

Ribas-Arino, J. and D. Marx (2012). "Covalent Mechanochemistry: Theoretical Concepts and Computational Tools with Applications to Molecular Nanomechanics". In: *Chemical Reviews* 112.10, pp. 5412–5487.

Richter, J. and N. Sonenberg (Mar. 2005a). "Regulation of cap-dependent translation by eIF4E inhibitory proteins". In: *Nature* 433, pp. 477–80.
– (Mar. 2005b). "Regulation of cap-dependent translation by eIF4E inhibitory proteins". In: *Nature* 433, pp. 477–80.
Rodnina, M. (May 2016). "The ribosome in action: Tuning of translational efficiency and protein folding". In: *Protein Science* 25.
– (Apr. 2018). "Translation in Prokaryotes". In: *Cold Spring Harbor Perspectives in Biology* 10, a032664.
Rodnina, M., M. Beringer, and W. Wintermeyer (Sept. 2006). "Mechanism of peptide bond formation on the ribosome". In: *Quarterly reviews of biophysics* 39, pp. 203–25.
Rodnina, M., A. Savelsbergh, V. Katunin, and W. Wintermeyer (Feb. 1997). "Hydrolysis of GTP by elongation factor G drives tRNA movement on the ribosome". In: *Nature* 385, pp. 37–41.
Rodnina, M. and W. Wintermeyer (Feb. 2001). "Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanism." In: *Annu Rev Biochem* 70, pp. 415–35.
Roij, R. van (16 July 2009). "Electrostatics in liquids: electrolytes, suspension, and emulsions". In: *Lecture Notes, Institute of Theoretical Physics, Utrecht, The Netherlands (unpublished)*.
Ross, S. ((2014)). *Introduction to Probability Models*. 11th ed. Academic Press San Diego CA USA, pp. 282–361.
Rozov, A., I. Khusainov, and K. t. El Omari (2019). "Importance of potassium ions for ribosome structure and function revealed by long-wavelength X-ray diffraction". In: *Nature Communications* 10.2519.
Sabi, R. and T. Tuller (2014). "Modelling the efficiency of codon-tRNA interactions based on codon usage bias". In: *DNA research* 21(5), 511–526.
Sabi, R., R. Daniel, and T. Tuller (Oct. 2016). "stAIcalc: tRNA Adaptation Index Calculator based on Species-Specific weights". In: *Bioinformatics (Oxford, England)* 33.
Sabi, R. and T. Tuller (Oct. 2015). "A comparative genomics study on the effect of individual amino acids on ribosome stalling". In: *BMC Genomics* 16.
Sansom, M., G. Smith, C. Adcock, and P. Biggin (1997). "The dielectric properties of water within model transbilayer pores". In: *Biophys J* 73.5, pp. 2404–2415.
Schepdael, A. V., A. Carlier, and L. Geris (2014). "Sensitivity Analysis by Design of Experiemnts". In: *Uncertainty in Biology (a computational modeling approach)*. Ed. by L. Geris and D. Gomez-Cabrero. Springer. Chap. 13, pp. 327–366.
Schmeing, T., K. Huang, D. Kitchen, S. Strobel, and T. Steitz (2005). "Structural insights into the roles of water and the 2' hydroxyl of the P site tRNA in the peptidyl transferase reaction". In: *Mol. Cell* 20, pp. 437–448.
Schmidt, C., T. Becker, A. Heuer, K. Braunger, V. Shanmuganathan, M. Pech, O. Berninghausen, D. Wilson, and R. Beckmann (Dec. 2015). "Structure of the

hypusinylated eukaryotic translation factor eIF-5A bound to the ribosome". In: *Nucleic acids research* 44. DOI: 10.1093/nar/gkv1517.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464.

Sehnal, D., R. Svobodová Vařeková, K. Berka, L. Pravda, V. Navrátilová, P. Banáš, C.-M. Ionescu, M. Otyepka, and J. Koča (Aug. 2013). "MOLE 2.0: Advanced approach for analysis of biomacromolecular channels". In: *Journal of cheminformatics* 5, p. 39.

Shah, P., Y. Ding, M. Niemczyk, G. Kudla, and J. Plotkin (June 2013). "Rate-Limiting Steps in Yeast Protein Translation". In: *Cell* 153, pp. 1589–1601.

Sharma, A. K., N. Ahmed, and E. P. O'Brien (2018). "Determinants of translation speed are randomly distributed across transcripts resulting in a universal scaling of protein synthesis times". In: *Phys. Rev. E* 97 (2), pp. 022409–20.

Sharma, P. K., Y. Xiang, M. Kato, and A. Warshel (2005). "What are the roles of substrate-assisted catalysis and proximity effects in peptide bond formation by the ribosome?" In: *Biochemistry* 44(34), 11307–11314.

Sharp, K. and B. Honig (Feb. 1990). "Electrostatic Interactions in Macromolecules: Theory and Applications". In: *Annu rev biophys and biophys chem* 19, pp. 301–32.

Sharp, P. and W.-H. Li (Mar. 1987). "The codon Adaptation Index—A measure of directional synonymous codon usage bias, and its potential applications". In: *Nucleic acids research* 15, pp. 1281–95.

Shaw, L., R. Zia, and K. Lee (2003). "Totally asymmetric exclusion process with extended objects: A model for protein synthesis". In: *Phys Rev E* 68, p. 021910.

Shifman, D. (2018). "Coding challenge 124: Flocking simulation". In: Last accessed on 2024-06-01. URL: https://thecodingtrain.com/challenges/124-flocking-simulation.

Sievers, A., M. Beringer, M. Rodnina, and R. Wolfenden (June 2004). "The ribosome as an entropy trap". In: *Proceedings of the National Academy of Sciences of the United States of America* 101, pp. 7897–901.

Simonovic, M. and T. Steitz (Aug. 2009). "A structural view on the mechanism of the ribosome-catalyzed peptide bond formation". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1789, pp. 612–623.

Simpson, L. J., E. Tzima, and J. S. Reader (2020). "Mechanical Forces and Their Effect on the Ribosome and Protein Translation Machinery". In: *Cells* 9.

Sivak, D. and G. Crooks (Aug. 2016). "Thermodynamic geometry of minimum-dissipation driven barrier crossing". In: *Physical Review E* 94.

Soheilypour, M. and M. Mofrad (June 2018). "Agent-Based Modeling in Molecular Systems Biology". In: *BioEssays* 40.

Sonenberg, N. and A. Hinnebusch (Mar. 2009). "Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets". In: *Cell* 136, pp. 731–45.

Starosta, A. L., J. Lassak, L. Peil, G. C. Atkinson, K. Virumäe, T. Tenson, J. Remme, K. Jung, and D. N. Wilson (Aug. 2014). "Translational stalling at polyproline stretches

is modulated by the sequence context upstream of the stall site". In: *Nucleic Acids Research* 42.16, pp. 10711–10719. DOI: 10.1093/nar/gku768.

Szavits-Nossan, J. and L. Ciandrini (Aug. 2020). "Inferring efficiency of translation initiation and elongation from ribosome profiling". In: *Nucleic Acids Research* 48, pp. 9478–9490.

Tafoya, S., S. Large, S. Liu, C. Bustamante, and D. Sivak (Mar. 2019). "Using a system's equilibrium behavior to reduce its energy dissipation in nonequilibrium processes". In: *Proceedings of the National Academy of Sciences* 116, p. 201817778.

The Cancer Genome Atlas (2007). *The Cancer Genome Atlas (TCGA)*. Retrieved March 7, 2025. URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

The UniProt Consortium (Nov. 2020). "UniProt: the universal protein knowledgebase in 2021". In: *Nucleic Acids Research* 49.D1, pp. D480–D489.

Thommen, M., W. Holtkamp, and M. V. Rodnina (2017). "Co-translational protein folding: progress and methods". In: *Current Opinion in Structural Biology* 42, pp. 83–89.

Tinoco, I. and J.-D. Wen (2009). "Simulation and analysis of single-ribosome translation". In: *Physical Biology* 6.2, p. 025006.

Tirumalai, M., M. Rivas, Q. Tran, and G. Fox (Nov. 2021). "The Peptidyl Transferase Center: a Window to the Past". In: *Microbiology and Molecular Biology Reviews* 85.

Tomuro, K., M. Mito, H. Toh, N. Kawamoto, T. Miyake, S. Chow, M. Doi, Y. Ikeuchi, Y. Shichino, and S. Iwasaki (Aug. 2024). "Calibrated ribosome profiling assesses the dynamics of ribosomal flux on transcripts". In: *Nature Communications* 15.

Torres, A. G., E. Batlle, and L. Ribas de Pouplana (2014). "Role of tRNA modifications in human diseases". In: *Trends in Molecular Medicine* 20.6, pp. 306–314.

Trobro, S. and J. Åqvist (2005). "Mechanism of peptide bond synthesis on the ribosome". In: *Proc. Natl. Acad. Sci. USA* 102, pp. 12395–12400.

– (2006). "Analysis of Predictions for the Catalytic Mechanism of Ribosomal Peptidyl Transfer". In: *Biochemistry* 45.23, p. 7049.

Tuller, T., A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel (Apr. 2010). "An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation". In: *Cell* 141, pp. 344–54.

Tuller, T., I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin, and M. Ziv-Ukelson (Nov. 2011). "Composite Effects of Gene Determinants on the Translation Speed and Density of Ribosomes". In: *Genome Biology* 12, R110.

Turner, D. H. and D. H. Mathews (Oct. 2009). "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure". In: *Nucleic Acids Research* 38.suppl1, pp. D280–D282.

Vanden Broeck, A. and S. Klinge (2023). "Principles of human pre-60S biogenesis". In: *Science* 381.6653, eadh3892.

– (May 2024). "Eukaryotic Ribosome Assembly". In: *Annual review of biochemistry* 93, pp. 189–210.

Velde, M. Van de, M. Ebroin, T. Durré, M. Joiret, L. Gillot, S. Blacher, L. Geris, F. Kridelka, and A. Noel (2021). "Tumor exposed-lymphatic endothelial cells promote primary tumor growth via IL6". In: *Cancer Letters* 497, pp. 154–164.

Verbruggen, S, E Ndah, W Van Criekinge, S Gessulat, B Kuster, M Wilhelm, P Van Damme, and G Menschaert (Aug. 2019). "PROTEOFORMER 2.0: Further Developments in the Ribosome Profiling-assisted Proteogenomic Hunt for New Proteoforms". In: *Mol Cell Proteomics* 18, S126–S140.

Viceconti, M. and L. Emili (2024). "Toward Good Simulation Practice". In: *Synthesis Lectures on Biomedical Engineering*. Ed. by M. Viceconti and L. Emili. Synthesis Collection of Technology. Springer, pp. 1–144. DOI: `https://doi.org/10.1007/978-3-031-48284-7`.

ViennaNGS (2015). *ViennaNGS: A Toolbox for Building Efficient Next-Generation Sequencing Analysis Pipelines*. Retrieved March 7, 2025. URL: `https://pubmed.ncbi.nlm.nih.gov/26236465/`.

Voss, N., M. Gerstein, T. Steitz, and P. Moore (Aug. 2006). "The Geometry of the Ribosomal Polypeptide Exit Tunnel". In: *Journal of Molecular Biology* 360, pp. 893–906.

VPH Institute (2025a). *EDITH: Ecosystem for Digital Twins in Healthcare*. Retrieved March 7, 2025. URL: `https://www.vph-institute.org/news/the-vph-institute-leads-a-european-coordination-and-support-action-to-build-an-ecosystem-for-digital.html`.

– (2025b). *Virtual Physiological Human (VPH) Initiative*. Retrieved March 7, 2025. URL: `https://www.vph-institute.org`.

Wallin, G. and J. Aqvist (Feb. 2010). "The Transition State for Peptide Bond Formation Reveals the Ribosome as a Water Trap". In: *Proceedings of the National Academy of Sciences of the United States of America* 107, pp. 1888–93.

Wang, J., B.-S. Shin, C. Alvarado, J.-R. Kim, J. Bohlen, T. Dever, and J. Puglisi (Nov. 2022). "Rapid 40S scanning and its regulation by mRNA structure during eukaryotic translation initiation". In: *Cell* 185.

Wang, J., C. Karki, Y. Xiao, and L. Li (2020). "Electrostatics of Prokaryotic Ribosome and Its Biological Implication". In: *Biophysical Journal* 118.5, pp. 1205–1212.

Warner, J. R. (1999). "The economics of ribosome biosynthesis in yeast". In: *Trends in Biochemical Sciences* 24.11, pp. 437–440.

Watson, Z. L., F. R. Ward, R. Méheust, O. Ad, A. Schepartz, J. F. Banfield, and J. H. Cate (2020). "Structure of the bacterial ribosome at 2 Å resolution". In: *eLife* 9. Ed. by S. H. Scheres, C. Wolberger, S. H. Scheres, B. Klaholz, and I. S. Fernández, e60482.

Wen, J.-D., L. Lancaster, H. C. Hodges, A. Zeri, S. Yoshimura, H. Noller, C. Bustamante, and I. Tinoco (May 2008). "Following translation by single ribosomes one codon at a time". In: *Nature* 452, pp. 598–603.

Wohlgemuth, I, S Brenner, M Beringer, and M. Rodnina (2008). "Modulation of the rate of peptidyl transfer on the ribosome by the nature of the substrates". In: *Journal of Biological Chemistry* 283.47, pp. 32229–32235.

Wolf, J., A. Gerber, and W. Keller (Aug. 2002). "tadA, an essential tRNA-specific adenosine deaminase from Escherichia coli". In: *The EMBO journal* 21, pp. 3841–51. DOI: 10.1093/emboj/cdf362.

Woolford, J. and S. Baserga (Nov. 2013). "Ribosome Biogenesis in the Yeast Saccharomyces cerevisiae". In: *Genetics* 195, pp. 643–81.

Yanagisawa, T., T. Sumida, R. Ishii, C. Takemoto, and S. Yokoyama (Sept. 2010). "A paralog of lysyl-tRNA synthetase aminoacylates a conserved lysine residue in translation elongation factor P". In: *Nature structural & molecular biology* 17, pp. 1136–43. DOI: 10.1038/nsmb.1889.

Yang, J.-R., X. Chen, and J. Zhang (July 2014). "Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding". In: *PLOS Biology* 12.7, pp. 1–14.

Yu, C.-H., Y. Dang, Z. Zhou, C. Wu, F. Zhao, M. Sachs, and Y. Liu (2015). "Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding". In: *Molecular Cell* 59.5, pp. 744–754.

Zappia, J., M. Joiret, C. Sanchez, C. Lambert, L. Geris, M. Muller, and Y. Henrotin (2020). "From translation to protein degradation as mechanisms for regulating biological functions: a review on the SLRP family in skeletal tissues". In: *Biomolecules* 10.1, p. 80.

Zenklusen, D., D. Larson, and R. Singer (Dec. 2008). "Single RNA Counting Reveals Alternative Modes Of Gene Expression In Yeast". In: *Nature structural and molecular biology* 15, pp. 1263–71.

Zhang, H., Y. Wang, and J. Lu (Apr. 2019). "Function and Evolution of Upstream ORFs in Eukaryotes". In: *Trends in Biochemical Sciences* 44.

Zia, R, J. Dong, and B Schmittmann (Aug. 2011). "Modeling Translation in Protein Synthesis with TASEP: A Tutorial and Recent Developments". In: *J Statist Phys* 144.

Zinshteyn, B. and W. V. Gilbert (Aug. 2013). "Loss of a Conserved tRNA Anticodon Modification Perturbs Cellular Signaling". In: *PLOS Genetics* 9.8, pp. 1–12.

Zuker, M. and P. Stiegler (Jan. 1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information". In: *Nucleic Acids Research* 9.1, pp. 133–148.

Zur, H. and T. Tuller (Sept. 2016). "Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution". In: *Nucleic Acids Research* 44.

Publications and curriculum vitae

F.3 Publications

F.3.1 Original contributions

Three articles are integral parts of the thesis and were published in peer-reviewed journals. Each of these articles are original and multidisciplinary contributions to the field. They constitute two important chapters of the thesis (chapter 4 and chapter 6).

- **Joiret, M.**, Kerff, F., Rapino, F., Close, P. and Geris, L. Ribosome exit tunnel electrostatics. Physical Review E (2022), 105(1), p.014409 [Joiret et al. 2022b].
- **Joiret, M.**, Kerff, F., Rapino, F., Close, P. and Geris, L.. A simple geometrical model of the electrostatic environment around the catalytic center of the ribosome and its significance for the elongation cycle kinetics. Computational and Structural Biotechnology Journal (2023), 21, pp.3768-3795 [Joiret et al. 2023a].
- **Joiret, M.**, Kerff, F., Rapino, F., Close, P. and Geris, L. Reversing the relative time courses of the peptide bond reaction with oligopeptides of different lengths and charged amino acid distributions in the ribosome exit tunnel. Computational and Structural Biotechnology Journal (2024), 23, pp.2453-2464 [Joiret et al. 2024].

F.3.2 Additional publications

During the course of the thesis, collaborations with different research teams, external to the Biomechanics Research Unit, led to the six following publications. In chronological order:

- **Joiret, M**., Mahachie John, J.M., Gusareva, E.S. et al. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. BioData Mining 12, 11 (2019). [Joiret et al. 2019].
- Zappia, J., **Joiret, M.**, Sanchez, C., Lambert, C., Geris, L., Muller, M. and Henrotin, Y. From translation to protein degradation as mechanisms for regulating biological functions: a review on the SLRP family in skeletal tissues. Biomolecules (2020), 10(1), p.80. [Zappia et al. 2020].
- Van de Velde, M., Ebroin, M., Durré, T., **Joiret, M.**, Gillot, L., Blacher, S., Geris, L., Kridelka, F. and Noel, A. Tumor exposed-lymphatic endothelial cells promote primary tumor growth via IL6. Cancer Letters (2021), 497, pp.154-164. [Velde et al. 2021].
- Rapino, F., Zhou, Z., Roncero Sanchez, A.M., **Joiret, M.**, Seca, C., El Hachem, N., Valenti, G., Latini, S., Shostak, K., Geris, L. and Li, P. Wobble tRNA modification and hydrophilic amino acid patterns dictate protein fate. Nature Communications (2021), 12(1), p.2170. [Rapino et al. 2021].
- Gallez, A., Blacher, S., Maquoi, E., Konradowski, E., **Joiret, M.**, Primac, I., Gérard, C., Taziaux, M., Houtman, R., Geris, L. and Lenfant, F. Estetrol combined to progestogen for menopause or contraception indication is neutral on breast cancer. Cancers (2021), 13(10), p.2486. [Gallez et al. 2021].
- **Joiret, M.**, Leclercq, M., Lambrechts, G., Rapino, F., Close, P., Louppe, G. and Geris, L. Cracking the genetic code with neural networks. Frontiers in Artificial Intelligence (2023), 6, p.1128153. [Joiret et al. 2023b].

F.4 Curriculum vitae

Marc Joiret

237 | box 11 rue Walthère Jamar, 4430 Ans, Belgium
Phone: (+32) 0476 73 93 24 | Email: marc.joiret@icloud.com

PhD | Liège University | Biomechanics Research Unit | Giga Molecular & Computational Biology, Belgium
PhD defense date: April 22th, 2025

- 2018–2025: ***PhD Thesis : Protein synthesis by ribosomes - Agent-based modeling of mRNA translation rates incorporating tRNA modification effects***
  Supervisors: Prof. Liesbet Geris (KU Leuven & ULiège) & Prof. Pierre Close (ULiège)
  Researcher & **Computational biologist** on Prof. Liesbet Geris FWO and ERC funded programs.

Work experience 4

- 2007–2018: **Consultant**
  Aquation s.a. Ans, Belgium
  Founded AQUATION s.a. and established an independent consultancy specializing in life science engineering and modeling for health, agro-industry, environment, and water treatment, offering knowledge-based services to public and privately owned companies.
  - Auditing water source treatment in the brewing industry and designing statistical experiments to measure the effect of the sulfate-to-chloride ratio in water on the bitterness of Trappist beers, as evaluated by a panel of rating judges
  - Water treatment engineering projects for general contractors and public tenders in Belgium and abroad (Vietnam, Tunisia, Indonesia, Algeria, Sri Lanka)
  - Water Demineralisation engineering projects for the industry
  - Sea Water Thermal Desalination research projects for the industry
  - Anaerobic water treatment process engineering projects for the industry
  - External reviewer for the European Commission to rank applicants calling for FP7 funding under water or circular economy thematic

Third Master (while being employed)
Hasselt University

- 2017–2018: ***Master thesis : The impact of correlated genetic markers on large-scale DNA-based gene-gene interaction studies.***
  Supervisors: Prof. Kristel Van Steen (KU Leuven & ULiège) & Prof. Ziv Shkedy (UHasselt).
  Genome-wide association simulation study of a pair of interacting functional variants associated to human complex diseases taking into account linkage disequilibrium as confounding.
- 2015–2018: **Master of Biostatistics, specialization Bioinformatics** | Interuniversity Institute for Biostatistics and Bioinformatics, Hasselt, Belgium | *Magna cum Laude:*
  2^{nd} Master: **Great Distinction**
  1^{st} Master: **Distinction**

Work experience 3

• 1999-2007: **R&D Manager and Senior Project Engineer**
Balteau, John Cockerill group Sprimont, Belgium

- Research program coordinator responsible for public fundraising and serving as a senior project engineer
- Design, engineering, construction oversight, commissioning, and start-up of wastewater treatment plants
- Design of activated sludge processes for municipal wastewater treatment, focusing on applied microbiology processes for carbon, nitrogen, and phosphorus removal
- Hydraulic and sanitary engineering for wastewater treatment processes
- Potabilization processes and water conditioning for industrial applications, including softening and demineralization

Work Experience 2

• 1992-1999: **Project Engineer and business development**
Watco, Tractebel, Suez Environment Welkenraedt, Liège, Brussels, Belgium

- Rehabilitation schemes in Landfill Engineering and Biogas to Energy Sector
- Landfill leachates and waste water treatment process design and operation
- Solid waste anaerobic processes optimization
- Ultrafiltration membrane coupling to nitrification & denitrification bioreactors

Second Master (while being employed)
Liège University

- 2004: ***Master thesis : The speed of gravity in General Relativity***
  Supervisors: Prof. Jean Surdej & Yves De Rop & Jean-René Cudell & André Burnel.
  The M. Sc. thesis in Physics focused on Relativistic Gravitation and investigated the gravitomagnetic effects of matter currents on the time delay of electromagnetic signals.
- 2002–2004: **Master of Science in Physics** *specialization in Theoretical Physics*, Université de Liège, ULiège | Liège, Belgium.
  Great Distinction
- 1996–1998: **Bachelor of Science in Physics**, Université de Liège, ULiège | Liège, Belgium

Work experience 1

• 1992: **Analyst Programmer, IT Dept.**
Morgan Guaranty Trust Company of New York Brussels, Belgium

- Database queries programming on the Eurobonds clearing organization (EUROCLEAR) for the Information Management
- IBM 3090 Mainframe application programming (RDBMS, SQL, DB2 and PL1)
- English speaking environment

Military Duty | Second Lieutenant reservist officer (matricule 046028)

- 1990–1991: Completed 5 months of military instruction and officer training in Leopoldsburg (Belgium), followed by 8 months of service in the 3rd Lancers Tank Battalion, in Spich (Germany).

Liaison Officer for the Belgian Armed forces at the 7th US Army training command base in Grafenwoehr, Germany ***1991***

Served for eight months between March and October 1991 in the 3rd Lancers Tank Battalion, stationed in Spich, Germany, as part of the Belgian Armed Forces. During this period, volunteered as a liaison officer to the US Army at Grafenwoehr, the largest American combat maneuver and simulation training center in Europe at the time. The role specifically involved coordinating and facilitating shared use of live-fire ranges during fire training sessions with Leopard tanks as part of **NATO's 1991 military maneuvers**.

First Master — Gembloux Agro Bio Tech

- 1990: ***Bio-engineer Thesis : Modeling and Production of Poly-β-hydroxybutyric Acid (PHB) by Alcaligenes eutrophus in a pilot fermentor*** — **Solvay Research Center**
  Supervisors: Prof. Philippe Thonnart & Raphaëlle Rikir & Eric de Buyl.
  The thesis focused on developing indirect (software) sensors for monitoring microbiological processes. PHB was utilized as a source of enantiomerically pure monomers, which are employed in pharmacology as a cardiotonic molecule when esterified to the arginine amino acid.
- 1987–1990: **ir., Dipl. in Chemistry and Biomolecular Engineering** | Faculté universitaire des Sciences agronomiques, FUSAGx, Gembloux, Belgium | **Great Distinction**
- 1985–1987: **Bachelor of Agriculture and Life Sciences** | Faculté universitaire des Sciences agronomiques, FUSAGx, Gembloux, Belgium | **Distinction**
- July 1985: **Entrance Admission test to Civil Engineering** | Faculté polytechnique de Mons, FPMS, Belgium | **Ranked 4th among 260 candidates**

Computer and software skills

- **Advanced** UNIX/Linux scripts, SLURM jobs for HPC, **Python**, C++, **R**, Mathematica, Matlab, HTML5, CSS, JavaScript, LaTeX, DBMS MS Access, SQL, Excel, Visual Basic, FORTRAN, PASCAL, SAS Statistical Analysis (SAS Certified Base Programmer), OpenBUGS and JAGS (Bayesian methods), PLINK, simuPOP, MB-MDR (GWAS), Ingenuity Pathway Analysis, Netlogo,
  Python libraries (BioPython SeqIO, regular expression, Scikit-Learn, Numpy, Pandas, PyTorch, TensorFlow, Database querying API, Matplotlib).
- **Artificial Intelligence, Deep Learning and data mining** General knowledge and projects experience in ML/AI methods, e.g., neural networks, computer vision, CNN, auto-encoders, GAN.
- **Classical Biostatistics advanced knowledge** of classical statistical inference methods and tools.
- **Bioinformatics and Computational Biology** QC | alignment | mapping tools | genomic repositories | downstream analysis of NGS repositories SRA, ENA, fastqc, trimmomatic, cutadapt, bowtie2, STAR, RNA-Seq (single cell and bulk), RiboProfiling tools, Snakemake data analysis workflows, PyMol for X-Ray crystallography or EM biomolecular structural analysis, Alphafold.

First author publication record in reviewed journals

The scientific seniority is relatively recent, having begun only five years ago, with contributions that span across **interdisciplinary** research in Biophysics, Bioinformatics and Computational biology:

- **Joiret, M.**, Rapino, F., Close, P., Geris, L. Reversing the relative time courses of the peptide bond reaction with oligopeptide of different lengths and charged amino acid distributions in the ribosome exit tunnel, *Comput Struct Biotechnol J.* 2024 May. https://doi.org/10.1016/j.csbj.2024.05.045
- **Joiret, M.**, Rapino, F., Close, P., Geris, L. A simple geometrical model of the electrostatic environment around the catalytic center of the ribosome and its significance for the elongation cycle kinetics, *Comput Struct Biotechnol J.* 2023 Jul 26;21:3768-3795. https://doi.org/10.1016/j.csbj.2023.07.016
- **Joiret, M.**, Leclercq, M., Lambrechts, G., Rapino, F., Close, P., Louppe, G., Geris, L. Cracking the genetic code with neural networks, *Frontiers in Artificial Intelligence*, 6, 2023. https://doi.org/10.3389/frai.2023.1128153
- **Joiret, M.**, Kerff, F., Rapino, F., Close, P., Geris, L. Ribosome Exit Tunnel Electrostatics, *Physical Review E*, 105, 2022. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.105.014409
- **Joiret, M.**, Mahachie John, J.M., Gusareva, E.S, Van Steen, K. Confounding of linkage disequilibrium patterns in large scale DNA based gene-gene interaction studies. *BioData Mining*, 2019

Articles in scientific journals as co-author

The scientific seniority is relatively recent, having begun only five years ago, with contributions that span across **interdisciplinary** research:

- Gallez, A., Blacher, S., Maquoi, E., Konradowski, E., **Joiret, M.**, Primac, I., Gérard, C., Taziaux, M., Houtman, R., Geris, L., Lenfant, F., Marangoni, E., Sounni, N. E., Foidart, J.-M., Noël, A., Pequeux, C. Estetrol combined to progestogen for menopause or contraception indication is neutral on breast cancer. *Cancers*, 2021. doi:10.3390/cancers13102486
- Rapino, F. , ZHOU, Z. , RONCERO SANCHEZ, A. M. , **Joiret, M.**, Seca, C., El Hachem, N., Valenti, G., Latini, S., Shostak, K., Geris, L., Li, P., Huang, G., Mazzucchelli, G., Baiwir, D., Desmet, C., Chariot, A., Georges, M., Close, P. Wobble tRNA modification and hydrophilic amino acid patterns dictate protein fate. *Nature Communications*, 12 (1), 2170, 2021. doi:10.1038/s41467-021-22254-5
- Van de Velde, M., Ebroin, M., Durré, T., **Joiret, M.**, Gillot, L., Blacher, S., Geris, L., Kridelka, F., Noël, A. Tumor exposed-lymphatic endothelial cells promote primary tumor growth via IL6. *Cancer Letters*, 497, 154-164, 2021. doi:10.1016/j.canlet.2020.10.020
- Zappia, J., **Joiret, M.**, Sanchez, C., Lambert, C., Geris, L., Muller, M., Henrotin, Y. From translation to protein degradation as mechanisms for regulating biological functions: A review on the SLRP family in skeletal tissues. *Biomolecules*, 10, 80, 2020. doi:10.3390/biom10010080
- Google Scholar: https://scholar.google.com/ marc joiret
- Orcid ID: https://orcid.org/0000-0001-5381-4196

Mentoring and Teaching Experience

- 2022-2023: Graduate student master thesis mentor in Biomedical Engineering, KUL Leuven university.
- 2021-2024: Teaching Assistant of Prof. Liesbet Geris at ULiege in Biophysics: lecture on the application of PyMol and Alphafold for the study of biomolecules and biological structures (4 hours/year)
- 2020-2021: Teaching Assistant of Prof. Liesbet Geris at ULiege in Biophysics: lecture on optical tweezers and applications to the study of biomolecules and biological structures (4 hours/year)
- 2017-2022: Invited lecturer for Energy course at HEC ULG Liege in the Master in Environmental Management (18 hours/year)
- 2007–2024: Invited lecturer for water treatment and environmental sciences at the Polygone de l'Eau, Verviers & Mons (60 hours/year)

Societies and professional membership or ambassadorship

- Belgian Biophysical Society (BBS)
- American Physical Society (APS)
- **Master Brewers Association of the Americas (MBAA)**
  Actively engaged in fostering connections and collaborations with fellow MBAA members to exchange brewing techniques and recipes, with a particular focus on mastering the art of brewing in the tradition of Belgian Trappist monks. Enthusiastic about contributing to folkloric brewing competitions in the U.S., which aim to replicate the unique characteristics of Trappist-style beers. Currently working on a manuscript and contributing as an author to the MBAA's Technical Quarterly journal.

Languages

- French: Mothertongue
- English: Professional proficiency |CEFR C2
- German: Elementary | A2 | Zertificat Deutsch als Fremdsprache, Goethe-Institut
- Dutch: Elementary | A2

Interests

- Reading | Learning | Writing | Science outreach | Epistemology
  Swimming | Mountain hiking with teammates | Sailing

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF AEROSPACE AND MECHANICS
BIOMECHANICS & COMPUTATIONAL TISSUE ENGINEERING - GIGA INSTITUTE MOLECULAR & COMPUTATIONAL BIOLOGY
Avenue de l'Hôpital, 1 (B34)+5CHU
B-4000 Liège, Belgium
http://www.biomech.ulg.ac.be

LIÈGE université
Sciences Appliquées

universitas leodiensis
UNIVERSITÉ de Liège