OctarellinVII

A new generation of *de novo* designed $(\beta/\alpha)_8$-barrel proteins

Cristina Elisa Martina

Protein Folding and Enzymology Laboratory
Centre for Protein Engineering (CIP)

Molecular Biomimetic and Protein Engineering Laboratory
GIGA-Research

University of Liège, Liège, Belgium

To Cécile, Maxi and André.
Thanks for welcoming me in Belgium
on this super cool project!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 *De novo* Design

De novo protein design is a branch in the field of protein engineering that started at the end of the '70s. Following many successful cases it developed quickly, and nowadays more than 100 papers are published every year on the subject. The number of publications per year on "*de novo* protein design" according to PubMed is reported in Figure 1.1.

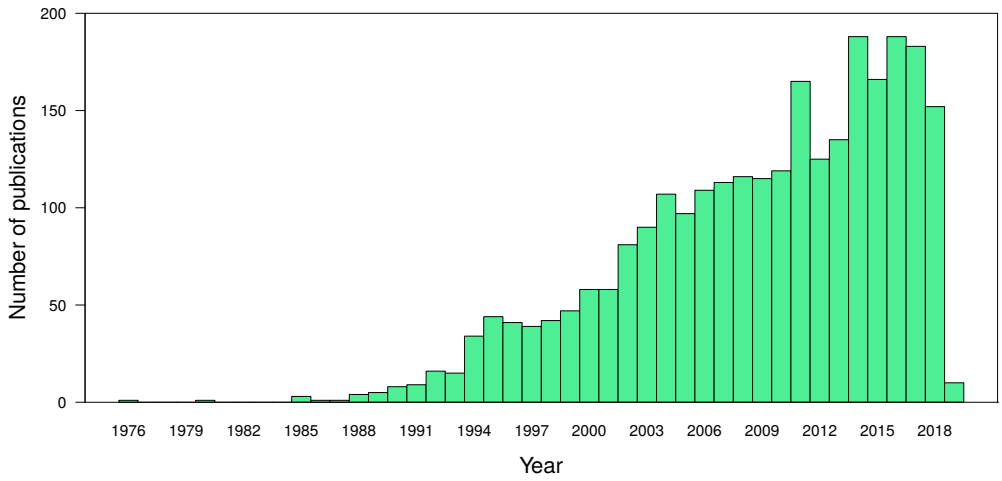


Figure 1.1: *De novo* protein design publications
Number of articles published every year with the key-words "*de novo* protein design". The results were taken from PubMed in January 2019.

Despite its large diffusion, it is quite difficult to describe and explain what exactly protein design is. In general, everything that brings something new (from Latin, *de novo*) in the field of protein science can be classified in this branch: the creation of a new protein topology [7], a new enzymatic activity [8] or an amino acid sequence that is not present in nature [9]. However, the boundaries between protein design and protein engineering

are blurred. For example, a single point mutation in a protein might yield a sequence that is not present in nature, but it will not be defined as *de novo* design. The lack of clear definitions and boundaries is mainly due to its recent development. Moreover the definition of "*de novo*" is time-dependent: designs and methodologies of the '80s may be considered obsolete nowadays thanks to the acquired knowledge and the development of new technologies.

1.1.1 The origins of the *de novo* design

The scientific community generally attributes to Brend Gutte the beginning of the *de novo* design era. In fact, he is considered as the father of the field since he published three publications between the end of the '70s and the beginning of the '80s.

The first article, published in 1975, reports the synthesis by solid phase method of an analog of Ribonuclease S [10]. The wild-type protein is 124 residue long and presents a wide loop on its surface. In order to study the importance of loops in protein folding, Gutte reduced Ribonuclease S to a 70-residue analog, which misses 5 loops not involved in the enzymatic activity. Surprisingly, despite the removal of 54 residues, the first "*de novo*" protein retained 4% of the enzymatic activity and specificity of the wild type enzyme.

The second article was published in 1979 [11]. The group wanted to design an artificial peptide with nucleic acid binding activity. They designed a 34-residue peptide in two steps: first they determined that the minimal structure for DNA-binding should contain a β-strand, a reverse turn, an anti-parallel β-strand, another reverse turn and an α-helix (shown in Figure 1.2). Following the design of the backbone model, they applied the rules for secondary structure prediction in order to find the best amino acid sequence to fit the model.

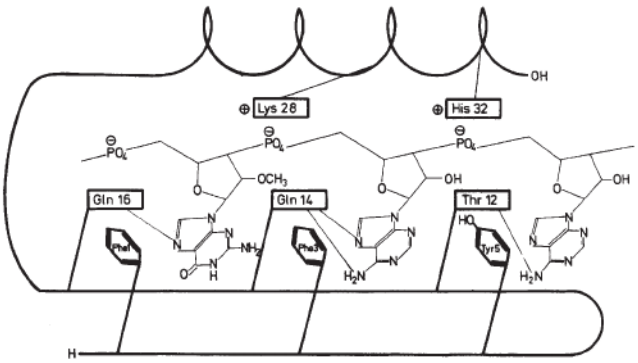


Figure 1.2: Model of an artificial peptide

Model representation of an artificial 34-residues polypeptide interacting with the trinucleotide GAA (from Gutte et al., 1979 [11]).

The peptide was produced by solid phase synthesis and showed binding to single strand DNA. Furthermore, the dimeric peptide also showed ribonuclease activity. Despite the lack of a 3D structure to confirm the design, the results that were obtained at that time are extremely remarkable, especially considering that: 1- computers were not used, 2- the number of available pdb structures was limited (42 according to RCBS-PDB) and 3- the advances in molecular biology, such as recombinant DNA, cloning, sequencing or DNA synthesis were not well-established. Not only they designed and produced artificial sequences, but they were also able to retain the original activity and to introduce a new functionality.

The third major achievement of Gutte was published in 1983 [12]. It describes the design of a 24 residue β-sheet able to bind DTT on both sides of the sheet. A K_d value of 20 μM was determined and crystals were obtained, thus leading to the first artificial crystallized protein. However, the structure of the protein, if any, was never deposited on RSCB-PDB.

1.1.2 Inverse protein folding

In the three *de novo* works of Gutte, we can discriminate among two kinds of designs: in the first example there is a simple deletion of parts of the protein (the loops) that brought to a new sequence; in the remaining two examples there is a rational design of the structure first, and then of the sequence. This last approach was defined by Pabo as "inverted" [13], and later on took the name of "inverse protein folding". In the classic Anfinsen view of protein folding, the amino acid sequence of a protein defines its final 3D structure (Figure 1.3), while in the inverse protein folding concept, a sequence is designed to achieve a given backbone structure [14].

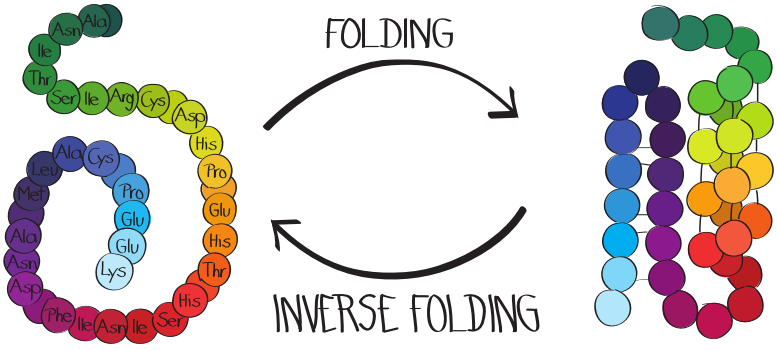


Figure 1.3: Inverse protein folding

Schematic representation of folding vs inverse folding: on the left, an amino acid sequence and on the right its protein structure. Arrows describe the relationships among sequence and structure in the normal folding and in the inverse protein folding approach. The picture was designed for this thesis by Ruth Kellner.

Most examples of *de novo* design that are reported in the literature fall in the inverse protein folding domain. The target backbone may belong to a natural protein or it may be designed from scratch, but the sequence is always optimized according to the structure.

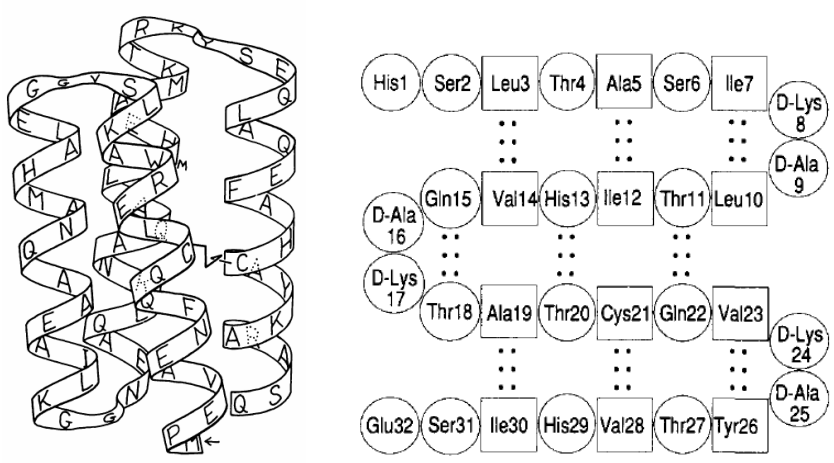
The research work presented in this manuscript, fully refers to the inverse protein folding domain, and the following sections will describe some examples of this class of designs. However, for completeness, we should mention that there are other options for generating artificial proteins, i.e. the directed evolution. This method is widely used to improve properties of natural proteins and enzymes, such as the solubility, the stability and the activity. Ward and co-workers used directed evolution to create an artificial metalloenzyme able to catalyze a reaction that is not present in nature, the olefin metathesis [15]. Another example of alternative *de novo* design is the screening of libraries of artificial sequences for *in vivo* activity [16]. The libraries are inserted in auxotrophs mutants of *E. coli* and only the sequences that recover the lacking functionality allows the survival of the cell. In this way it was possible to select new sequences with enzymatic activity and also to discover alternative metabolic pathways that are not present in nature.

These two examples are defined as combinatorial designs because they imply the experimental screening of a large number of sequence combinations (up to 10^6 for the directed evolution). In comparison to the inverse protein design approach, the combinatorial one has more chances of success thanks to its strong selection method. However, due to its experimental nature (intended as wet-lab experience), it is time consuming and more expensive. For this reason, there are much more examples in the literature of inverse folding designs (often less successful) than combinatorial ones. For example, a PubMed search for "Protein engineering [title]" gave 935 outputs (March 2019). The words "Protein engineering [title] AND design", mainly used for inverse folding designs, found 190 outputs, while "Protein engineering [title] AND libraries", mainly used for combinatorial designs, found only 50 outputs. If both words are restricted to be present in the title of the articles ("Protein engineering [title] AND design [title]" and "Protein engineering [title] AND libraries [title]"), the output numbers are reduced to 44 and 9, respectively, showing that the combinatorial approach is less used compared to the rational one.

Manual design

As described in the previous section, the inverse protein folding is the design of artificial sequences starting from a 3D structure. After Gutte, many other examples were published in the literature: the design of helix-bundles (Figure 1.4a) [17, 18], β-sheet folds (Figure 1.4b) [19], fibrous proteins [20, 21], coiled coil [22] and membrane channels [23]. Attempts to design artificial TIM-barrels have been carried out by the group of Joseph

Martial, at the University of Liège, and they are described in Chapter 1.3, page 19) [1–3].



(a) **(b)**

Figure 1.4: Examples of manual designs

(**a**) The model of the protein Felix, an artificial 4-helix bundle (from Hecht et al. 1990 [18]) and (**b**) the model of the Betabellin, and artificial β-sheet protein (from Yan and Erickson, 1994 [19]).

All these examples are classified as "manual" *de novo* designs, in opposition to the "computational" ones that arrived later on (described in the next section). Until 1997, researchers in the field performed manually all the steps of design, from the alignment of protein sequences, to the creation of backbone models (Figure 1.4) and the optimization of the sequence. The reasons are bound to the technological limitation of that time: despite computers with discrete computational power were available, specialized programs for protein manipulation were missing or had a limited diffusion in the scientific community.

For example, the first algorithms for global sequence alignment of proteins was published in 1970 [24], the first one for local sequence alignment in 1981 [25], and the first one for multiple sequence alignment in 1986 [26]. The first program for homology modelling is dated 1988 [27]; the first one for structural alignment is in 1989 [28], and the first computational method to design proteins *de novo* was published in 1994 [29]. Many of these programs are widely used nowadays, and they are essential to the field of *de novo* design.

Computational design

The jump from "manual" to "computational" designs dates back in 1997. In this year, Dahiyat and Mayo published the first paper describing a revolutionary method for *de novo* design and sequence selection [9]. As mentioned in the previous section, an algorithm for *de novo* design was already published in 1994 [29]. However it was based exclusively on

statistic data extracted from native structures deposited on RCSB-Protein Data Bank. This means that the designs are biased by the database content (i.e. proteins that are soluble enough for the crystallization process only). As a consequence, many folds or proteins were excluded (i.e. membrane proteins). The revolutionary algorithm proposed by Dahiyat and Mayo is based on physical and chemical properties of amino acids, in order to determine with high precision the structure and the stability of any kind of fold. In particular, they introduced the concepts of "scoring functions" and "energy calculation" that are widely used nowadays.

Differences among manual and computational design are not limited to the historical period or the technological advances only. In the manual design there is a limited number of sequence combinations that researchers may take into account. In the computational approach this number is limited only by the computational power. Dahiyat and Mayo calculated the energy score for 1.9 x 10^{27} sequences for the target fold (a zinc binding domain of 28 residues). It took 90 hours of computational calculations. It is not possible to obtain the same results by manual design.

Following the work of Dahiyat and Mayo, software for energy calculations and protein design were developed,that gave rise to the computational era of *de novo* protein design. Some of the most interesting examples are: the re-design of globular proteins [30] and enzymes [31, 32], the design of a new protein fold [7], of an enzyme with new activity [33] or with an unnatural reaction [8], the thermo-stabilization of enzymes [34], the incorporation of unnatural amino acids [35] and the self-assembling of protein nanostructures [36]. The review of Woolfson [37] summarizes most of the successful design obtained so far. In general, small proteins (less than 100 aa) or repeat proteins have an higher rate of success compared to larger ones. Proteins designed by inverse protein folding usually do not have an enzymatic function. Once the structure of the designed protein is experimentally validated, it is possible to add a function, as an enzymatic activity or a binding site. However the success rate in the design is very low. For example, in almost 30 years of design of the TIM-barrel fold, only 1 design out of 43 was successful (see Chapter 1.3, page 19). Regarding the thioredoxin fold (see Section 1.1.2, page 7), only 1 protein out of 48 was experimentally validated. The reasons of this low rate are probably bound to the limitation in our knowledge of proteins. Many aspects in the stability, solubility, folding and function of proteins are not yet fully understood, in particular for protein larger than 100 residues. This led to limitations in software and tools for protein modelling and design. However, these software are constantly improved, with information obtained by both natural and artificial proteins.

Thioredoxin re-design

Among all the available examples on the *de novo* protein design, I will describe in more details the work of Winther and co-workers on the re-design of the thioredoxin protein [31]. The goal of their research was to test the capabilities of a modelling software, Rosetta [38], to recreate the target fold.

The thioredoxin fold was chosen because it is a small and rigid protein (106-112 aa), highly conserved in nature, with 90% of well-defined secondary structures. As for the TIM-barrel fold, the amino acid sequences of the thioredoxin family may be highly different (thus, have low sequence identity). Moreover, the protein was already successfully engineered in previous works. The researchers selected 8 natural structures with a thioredoxin fold from Protein Data Bank (shown in Figure 1.5a), and they performed geometry optimization with RosettaRelax, generating 6 backbones for each natural structure (48 models in total).

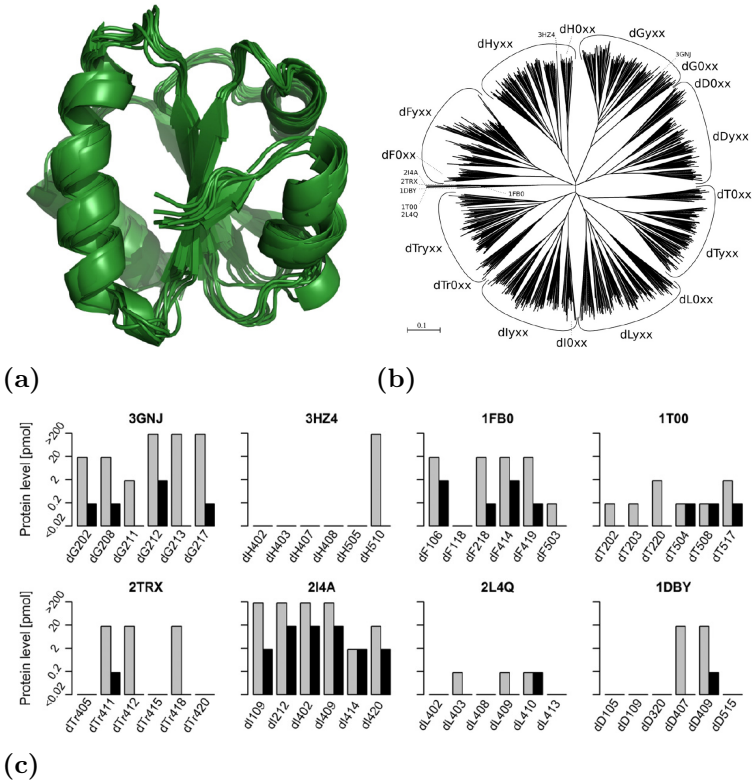


Figure 1.5: Thioredoxin fold, design and expression

(**a**) Structural overlay of 8 natural thioredoxin folds and (**b**) Dendrogram showing the distance relationships after sequence alignment of the 960 designs and (**c**) Analysis of the expression and of the solubility of the 48 designs of the thioredoxin. The gray bars indicate the insoluble fraction (inclusion bodies) while the black bars indicate the soluble fraction of the crude extract. Two pmol of protein corresponds to ∼0.2 mg/mL of protein. The nomenclature used is dNyxx, where d is design, N the first(s) letter in the PDB ID code of the initial templates, y the geometry optimization output and xx the sequence optimization output (from Winther et al., 2016 [31]).

For each of the 48 backbone structures, the sequence optimization generated 20 models, for a total of 960 designs. The sequence relationship of the models after alignment is shown in Figure 1.5b. For each of the 8 initial templates, the group chose six designs for experimental validation according to the lowest Rosetta energy score (so, the ones with higher stability), for a total of 48 models. Their synthetic genes were inserted in a plasmid with an HisTag at the C-terminal part of the protein, and cloned in *E. coli*. Expression trials couppled to western blot analysis indicates that 16 out of the 48 proteins did not expressed at all, 11 were expressed in inclusion bodies only and the remaining ones were partially found in the soluble fraction of the crude extract (Figure 1.5c).

Among the soluble ones, only 2 were successfully purified. The other ones were not stable or soluble enough, with some of them aggregating on-column. Only one of the two purified proteins was able to form crystals, and its structure was solved. The X-ray structure is in good agreement with the computational model, and the average RMSD is of 1.8-2.0$\mathring{A}$ (Figure 1.6).

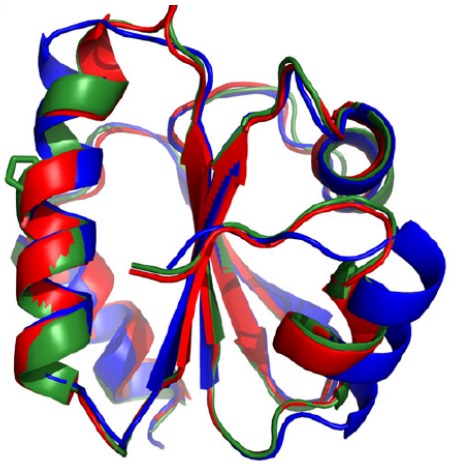


Figure 1.6: Structure resolution of the thioredoxin design

Overlay of thioredoxin structures: the initial template (obtained from Protein Data Bank) is shown in green, in red the designed model and in blue the X-ray structure. Its resolution is 2.4$\mathring{A}$ (from Winther, 2016 [31]).

This work on the redesign of the thioredoxin protein is an excellent example of inverse protein folding. Moreover it is one of the few publications on *de novo* protein design in which both the positive and the negative results are reported. In general, only successful designs are published and this makes more complicated to understand and discuss any unexpected results.

The design of Winther has also many similarities with the work presented in this thesis, for example the use of the Rosetta software. Thanks to these similarities it is possible to

compare and discuss the two designs despite of significant differences in the target folds, thioredoxin on the one side and TIM-barrel on the other.

In the next chapters I will first describe the subject of this thesis project, the TIM-barrel fold, and then I will illustrate results that were obtained in the group of Joseph Martial (the Octarellin project) at the University of Liège.

1.2 Natural TIM-barrels

The TIM-barrel fold can be found in 10% of all known proteins and in 5 out of 6 classes of enzymes, representing the most common fold in nature [39, 40]. This chapter will describe features such as structure, function, folding, stability and evolution of the TIM-barrel fold.

1.2.1 Structure, topology and classification

The TIM-barrel fold owes its name to triosephosphate isomerase (TIM), the first enzyme discovered to have a $(\beta/\alpha)_8$-barrel fold. Its structure was solved in 1975 and it is shown in Figures 1.7a and 1.7b.

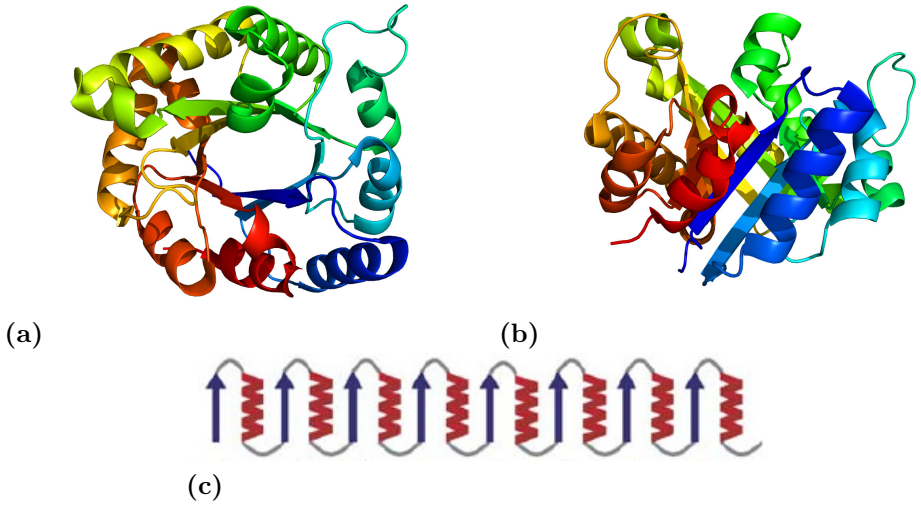


Figure 1.7: Structure of triosephosphate isomerase

(**a**) Top view and (**b**) side view of the triosephosphate isomerase (TIM) structure, solved in 1975 [41] and deposited on RCSB-PDB under the ID "1TIM". (**c**) Topology model of the TIM-barrel fold (from Höcker et al., 2005 [42]).

The TIM-barrel fold is composed of 8 β-strands forming a closed central β-sheet (β-barrel), surrounded by 8 α-helices forming an external barrel (α-barrel). The topology model of TIM-barrels is a 8-fold repetition of the β/α motif (Figure 1.7c). All β-strands are oriented in the same direction, forming a parallel β-sheet. With the exception of the TIM-barrel family, all the other β-barrel proteins contain anti-parallel or mixed β-sheets only [43].

CATH classification

The TIM-barrel family is classified according to the Class, Architecture, Topology and Homology (CATH) classification [44], as 3.20.20: the Alpha Beta Class (3.), the Alpha-

Beta Barrel Architecture (20.) and the TIM-barrel fold (20). Up to date (January 2019), 37 Superfamilies and 15671 Domains are included in the TIM-barrel family.

The β-barrel structure

Many protein folds containing a β-barrel structure, TIM-barrels included, have been deeply studied and classified according to two parameters: the number of strands (n), and the shear number (S) [45–49]. The shear number is defined as "the change of residue numbers on a β-strand when a point moves in the left hydrogen bond direction back to the same β-strand" [45]. Figure 1.8 reports an example of the 3D β-barrel structure of the TIM-barrel fold and its "unrolled" representation [5].

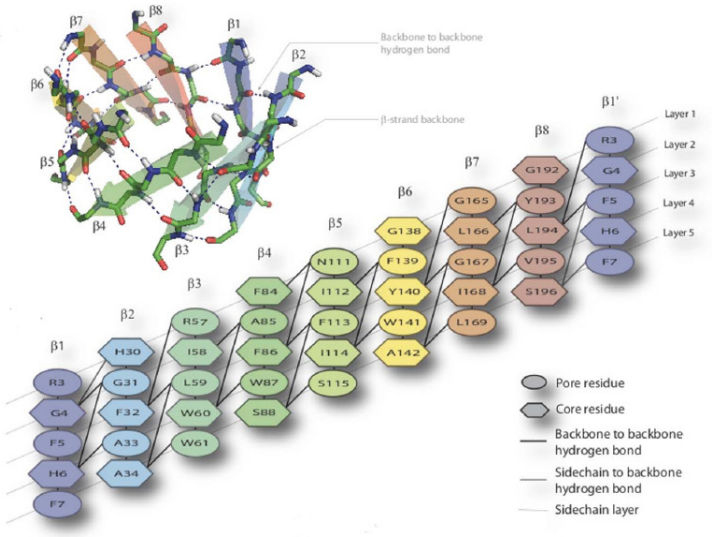


Figure 1.8: Shear number in β-barrels proteins

Characterization of the β-barrel of the TIM-barrel fold according to the hydrogen bond connection and the Shear number calculation (from Figueroa et al., 2013 [5]).

The TIM-barrel family is characterized by n=8 and S=8, since 8 "jumps" among residues are necessary to complete a full turn of the barrel.

The number of strands (n) and the shear number (S) are useful information to characterize β-barrel proteins. When S is positive, the barrel is right-stranded (as in the TIM-barrel fold). Moreover, when n=S, the side-chains of the β-barrel point alternatively towards the center of the barrel (pore residues) and towards the α-helices (core residues). In layer 1 of Figure 1.8, the residues of odd strands (β1, β3, β5 and β7) are pore residues. They points towards the center of the barrel, while the even strands (β2, β4, β6 and β8) are core residues and point outside the barrel. The layer 2 has the opposite combination: odd strands have core residues, even ones pore residues, and so on for the following layers. This difference of odd and even strands set the limits for the symmetry of the TIM-barrels,

which is not a mere repetition of 8 (β/α) motives (8-fold symmetry), but a repetition of 4 ($\beta/\alpha/\beta/\alpha$) motives (4-fold symmetry) [50].

Although the 8 β-strands of the TIM-barrel fold are in general fully connected by H-bonds, we can find examples of proteins that contains "open" β-barrels. One significant example is the methylmalonyl CoA mutase, an enzyme that contains a TIM-barrel fold among other domains. When the substrate is absent, the β-barrel is split apart in an "open" conformation, and the active site is accessible to solvent. After binding, the β-barrel closes around the substrate recovering a fully connected conformation [51].

The α-barrel structure

The α-barrel structure of the TIM barrel fold is less characterized than its β-counterpart. A single α-helix is exposed to the solvent for $1/3^{rd}$ of its surface, while the remaining $2/3^{rd}$ are involved in hydrophobic interaction with the β-barrel and with the neighboring helices. The hydrophobic core of the TIM-barrel fold is not composed by the residues in the pore of the TIM-barrel but by the ones situated at the interface of the β-barrel and the α-barrel (hydrophobic ring) [49]. H-bond interaction among helices are present but, in contrast with the H-bond network of the β-barrels, they are not necessary. Multiple examples of missing helices are reported in the literature: the phosphoinositide-specific phospholiphases C (PI-PLCs) enzymes lack 2 helices in the prokaryotic homologous (α4 and α5) and a single helix (α5) in the eukaryotic one [52]. The structures of cellobiohydrolase II [53] and of E2CD endocellulase [54] are also lacking 2 helices (α7 and α8). These enzymes present irregularities (distortions and missing elements) in the TIM-barrel fold to better adapt to the large substrates [52].

The loop region

TIM-barrels contain 15 loops: 8 to connect β-strands with α-helices (called C-term loops) and 7 for the opposite direction (N-term loops) [43]. C-term loops are usually longer than N-term loops and can include both additional secondary structures and entire extra domains [55]. The reason why C-term loops are longer and contain extra-domains compared to their N-term counterpart is that the active site of TIM-barrel enzymes is always located at the C-term of the barrel. C-term loops not only form the active site but they are also extremely important in the substrate recognition and binding. On the contrary, N-term loops are associated to the stability of the overall structure [56].

1.2.2 Function

TIM-barrels are involved in 5 out of 7 classes of the Enzyme Commission (E.C.) classification for enzymatic activities: 1-oxidoreductase (15% of the TIM-barrel enzymes), 2-transferase (11%), 3-hydrolase (49%), 4-lyase (12%) and 5-isomerase (11%) [57]. The 6-ligase and 7-translocase classes are the only ones in which the fold is not present. However, not all the TIM-barrels have an enzymatic activity: the narbonin, the concavalin B and a chitinase-like protein have no enzymatic function [58–60].

The TIM-barrels sub-families can be bound to a single class of enzymatic reactions (i.e. the racemase family catalyzes only isomerase reactions), while other are spanning among multiple E.C. classes (i.e. the PP-binding family catalyzes oxidoreductase, transferase, lyase and isomerase reactions) [57]. There is also an example of a single protein with 2 activities: the N-(5'-phosphoribosyl)-anthranilate isomerase/indole-3-glycerol-phosphate synthase of *E. coli* catalyze 2 different reactions (isomerase and hydrolase) in the biosynthesis of tryptophan. Indeed the enzyme is composed by two TIM-barrel domains that do not present significant sequence identity [61].

Some of the most efficient enzymes belong to the TIM-barrel family. For example, the 5'-monophosphate decarboxylase has the largest enhancement rate among any other enzymes [62], and the triose phosphate isomerase (the original TIM) works at the diffusion rate limit [63].

At biological level, TIM-barrels are involved in the small molecule metabolism (47%), macromolecule metabolism (25%), energy metabolism (20%), DNA/RNA information pathways (2%) and ion channel transport (1%) [64].

1.2.3 Evolution

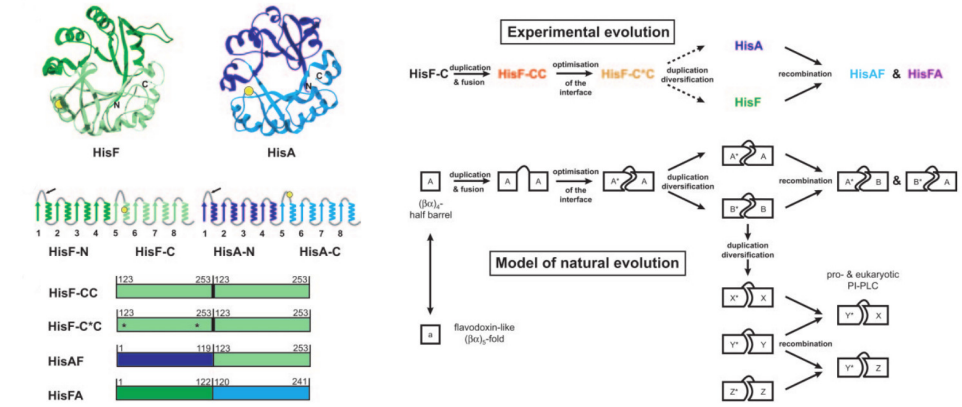
The evolution of the TIM-barrel fold has been extensively discussed during the '90s. The first theory that has been proposed is that the TIM-barrel fold arose from convergent evolution to a stable fold. This theory was mainly supported by the observation that the similarities in sequences, functions and ligands of the known TIM-barrels were too low to support divergent evolution from a common ancestor [65].

On the other side, evidences to support the divergent evolution theory were three: all the TIM-barrels has an enzymatic activity (the narbonin was discovered 5 years later), the active site is always located at the C-term of the protein and similar structural patterns

are present in different structures, including TIM-barrels with different functions [65, 66].

Recent evidences suggested that not only the TIM-barrels are the result of a divergent evolution from a common ancestor, but also that the ancestor was half the size of a modern TIM-barrel and that the corresponding ancestral gene underwent to duplication, fusion and evolutionary divergence [49, 67–70].

In 2004, the Sterner group experimentally mimicked the evolution of the TIM-barrel fold [71]. Starting from the natural proteins HisF and HisA they generate three novel enzymes by duplication of the C-terminal half of HisF (HisF-CC) and by combination of the N-terminal half of one protein with the C-terminal half of the other protein (HisAF and HisFA), shown in Figure 1.9a. Both HisF-CC and HisFA had tendency to aggregate, while HisAF was found stable and monomeric in solution. Improvements in the packing of HisF-CC led to a stable and monomeric protein (HisF-C*C), supporting the theory of divergent evolution by duplication of a half-barrel, followed by diversification and eventually recombination, Figure 1.9b.



(a) (b)

Figure 1.9: Evolution of TIM-barrels from half-barrel domains

(**a**) Structure and topology representations of the natural protein HisF and HisA and their fusion construct; (**b**) proposed evolution model of natural TIM-barrels (from Hocker et al., 2004 [71]).

Höcker et al. [71] also suggested that the $(\beta/\alpha)_4$ ancestor of modern TIM-barrels derived from a flavodoxin-like fold, which is formed by an $(\beta/\alpha)_5$ domain. Evidences of sequence identity among HisF and HisA with the flavodoxin-like proteins support this hypothesis, (Figure 1.9b).

1.2.4 Folding

TIM-barrel proteins all share homologous 3D structure but display very low sequence identity. Is their folding driven by the native topology or by their individual amino acid sequence?

Since the first structure was published [41], many studies have been performed to characterize the folding of TIM-barrel proteins, using different techniques and approaches: circular permutation [72], fragmentation [73, 74], chemical-induced unfolding [75–77], temperature-induced unfolding [78–80] and hydrogen exchange coupled to mass-spectrometry [81]. Difficulties in thermal unfolding studies arose because many TIM-barrel proteins have irreversible denaturation transitions, usually linked to protein aggregation [82, 83]. However, different models of folding pathways have been proposed: from the simple two-state transition [81, 84] to more complex ones that involve intermediates with multiple oligomerization states [85, 86]. An exhaustive review about the proposed models of unfolding was published by Zarate-Perez in 2008 [87].

Intermediates observed in the folding of TIM-barrels are mainly composed by (β/α) modules. The phosphoribosyl anthranilate isomerase (TrpF), from both yeast and *E. coli*, populates a stable intermediate formed by the first six (β/α) subunits when tested *in vitro* (model 6+2) [73]. Surprisingly, when tested *in vivo* it is formed by the first four subunits (model 4+4) [70]. Triose-phosphate isomerases (TIMs) from *E. coli* and rabbit show a 4+4 model, but the homologous enzyme in yeast presents a 3+3+2 folding model [88].

The folding of TIM-barrel proteins can also be mediated by chaperons. In *E. coli* the chaperonin GroEL and its cofactor GroES are involved in the folding pathway of more than 250 proteins (30% to 50% of which contain a TIM-barrel fold) [89, 90]. For example, folding of dihydrodipicolinate synthase (DapA) from *E. coli* goes 30 time faster when associated to the GroEL/ES machinery than with spontaneous folding [91].

1.2.5 Stability

It is difficult to understand which factors are contributing to the stability of TIM-barrel fold, mainly because the protein sequences are highly variable. Two general approaches were adopted: site-directed mutagenesis to produce mutants and comparison of TIM-barrels from different environment (psychrophilic, mesophilic and thermophilic).

In 1998, the group of Wierenga compared the stability of two triose-phosphate isomerases, from the psychrophilic *Vibrio marinus* (vTIM) and from the mesophilic *Escherichia coli* (eTIM) [82]. The first organism has an optimal growth at 15°C, while the latter at 37°C. vTIM and eTIM have 66% of sequence identity, similar k_{cat} values at their temperature optimum (10°C and 25°C respectively) and a temperature of half-denaturation (T_d) of 41°C and 54°C respectively. The group showed how a single mutation of vTIM (A238S) was increasing the stability of the protein of 5°C. However, the catalytic efficiency was affected, showing that enhancing the stability can lead to a lack of efficiency.

Anyway, this is not always the case. The mutation E65Q of the triose-phosphate isomerase from *Leishmania mexicana* (lTIM), increases the stability by 26°C, converting a mesophilic protein to a thermophilic one, without losses in the enzymatic activity [92].

Another remarkable example is the mutation of the first residue (valine) of the xylanase from *Bacillus sp.* [93]. This residue is not involved in the formation of secondary structures. Its mutation in leucine (V1L), increases the overall stability by 5°C, while the mutations V1A and V1G decrease it by 2 and 12°C respectively. The mutations affect in a different way the interactions among the N-term and the C-term of the protein, suggesting that the stability of the TIM-barrel fold can be correlated with the degree of interaction among its extremities. This theory was demonstrated by Ramakumar group that showed how increasing non-covalent interactions in the N- and C-term is enhancing the protein stability, not only in the xylanase (TIM-barrel fold) but also in other protein folds that have connected N- and C-term [94].

A broader study on the adaptation of the TIM-barrel fold to different temperatures was performed in 1999 by Maes group [95]. They compared 10 different TIMs from psychrophilic, mesophilic, thermophilic and hyperthermophilic organisms according to different parameters: **a-** sequence alignment, **b-** volumes, **c-** cavities, **d-** hydrophobicity, **e-** hydrogen bonds, **f-** salt bridges and **g-** charge distribution in the helices. Surprisingly, they did not find evidences for correlation between stability and sequence, volumes, cavities or hydrogen bonding. The psychrophilic and thermophilic TIMs display more salt-bridges compared to mesophilic TIMs. The hyperthermophilic protein enhanced this trend by tetramerization: extra salt-bridges are present at the interface between monomers. Thermophilic TIM-barrels, but not psychrophilic and mesophilic ones, are characterized also by higher hydrophobicity in the protein core. It suggests that folding is mainly driven by hydrophobic collapse at high temperature [95].

Some of the features and characteristic of the natural TIM-barrel family are important in the context of *de novo* design. In the next chapter I will describe the first attempts made in Liège towards the design of artificial TIM-barrel proteins.

1.3 The Octarellin story

At the end of the '90s, the group of Joseph Martial at the University of Liège in Belgium, started a project dedicated to the *de novo* design of proteins with a $(\beta/\alpha)_8$-barrel fold. This is the base of the present thesis project.

Over the years, six artificial TIM-barrels (named Octarellins) were designed from scratch following different methods.

Octarellin I [1, 2], II and III [3] were designed in a "manual" way, with a limited use of computational support. With the advance of computing power and the development of programs dedicated to the calculation of atomic interactions, the research group switched from "manual" to "computational" design and designed Octarellin V [4], Octarellin VI [5] and the present work.

Octarellin V.1 [6] was created through directed evolution of Octarellin V and can be referred to the result of a "combinatorial" method.

1.3.1 Octarellin I

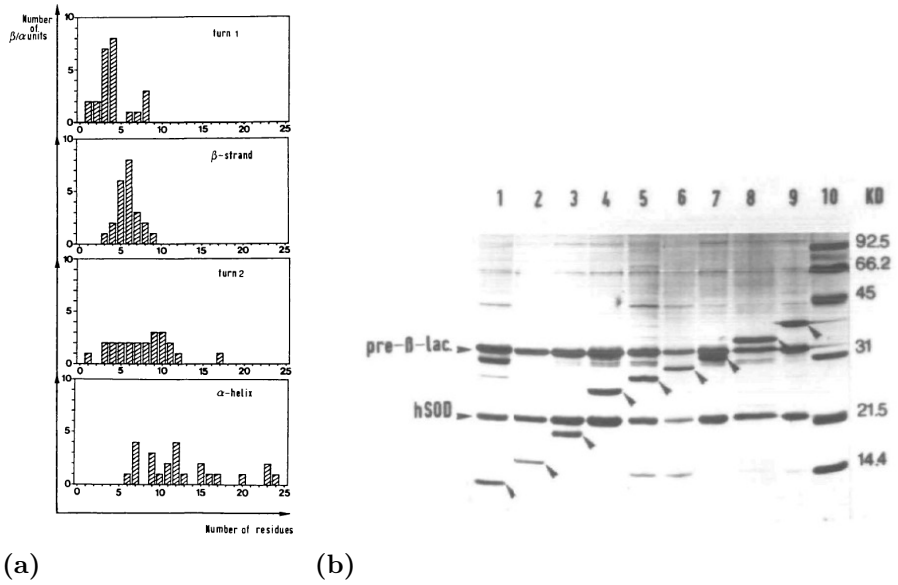
The design of the first artificial TIM-barrel was published in 1990 by Goraj [1] and was followed in 1991 by its biophysical characterization [2]. The group aimed to recreate the TIM-barrel fold by the repetition of 8 structural β/α-units (that gave the name of Octarellin to the proteins). The repetition unit is composed by a first turn, a β-strand, a second turn and an α-helix. In order to find the amino acid sequence for the structural unit, the group analyzed the sequences and the structures of the three natural TIM-barrels known at the time: triose phosphate isomerase (TIM) [41], KDPGaldolase [96] and xylose isomerase [97]. The analysis took into account the length of secondary structure elements (Figure 1.10a), the residue frequency and the β/α-packing.

Analysis of the amino acid composition led to the design of a 30 aa sequence unit that was then repeated 8 times to form the $(\beta/\alpha)_8$-barrel. It is composed by 4 residues for the first turn, 6 for the β-strand, 7 for the second turn and 13 for the α-helix, with the following residues:

DARS - GLVVYL - GKRPDSG - TARELLRHLVAEG

A first plasmid was constructed with the single structural unit, and it was then replicated to obtain 2 to 12 unit repetitions. All the 12 constructs were inserted in *E. coli*, and expressed. The expression profile of proteins with 5 to 12 repetition units is shown in Figure 1.10b. The proteins with 2 to 12 unit repetitions were all produced in inclusion

bodies. Only the 7-, 8- and 9-fold were extracted from the inclusion bodies by denaturation, refolded by dialysis and characterized by far-UV circular dichroism (CD), infrared (FTIR), Raman and UV-absorption spectroscopies. Octarellin I was found to consist of 30% of helix- and 40% of strand-content, and a loosely packed 3D structure.



(a) (b)

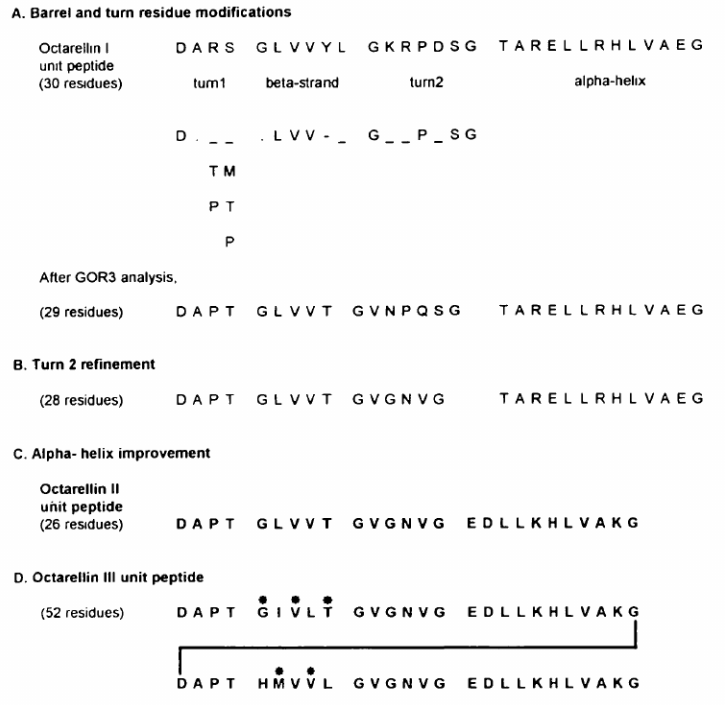
Figure 1.10: Octarellin I, design and production

(**a**) Secondary structure lengths in natural TIM-barrels; (**b**)SDS-PAGE of the artificial polypeptides with 5 to 12 repetition units. Lanes 1-9: inclusion bodies containing the artificial polypeptides (arrows), from 4 to 12 repetition units respectively. Octarellin I is shown in lane 5. Figures are extracted from Goraj et al., 1990 [1].

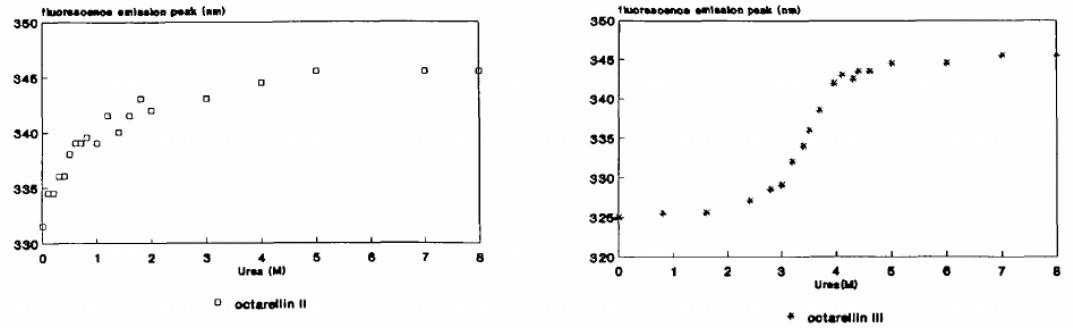
1.3.2 Octarellin II and III

The 4-fold symmetry is now commonly associated to the TIM-barrel fold, but was proposed for the first time in 1989 by Lesk [50]. It was not taken in account by the Martial group until 1995, with the upcoming of the second generation of Octarellins [3]. The design aimed to improve the sequence of Octarellin I, with the help of 5 new natural TIM-barrel structures [98–102] and the use of a 4-fold symmetry instead the 8-fold one. Based on the analysis of natural TIM-barrel structures, sequences of the first turn, the beta strand and the second loop were modified (see Figure 1.11). Octarellin II and III differ by 20 aa due to symmetry: the second kept a 8-fold repetition symmetry, with identical strands, while the third makes a distinction between even and odd strands, introducing the 4-fold symmetry.

Both proteins were produced in inclusion bodies. After refolding, they showed higher solubility compared to Octarellin I. Biophysical characterization by FTIR, CD and fluo-



(a)



(b)

Figure 1.11: Octarellin II and III, design and characterization

(**a**) Evolution of Octarellin I unit peptide (A) to Octarellin II (C) and III (D). The unit sequence was shorten of 4 residues, one from the β-strands, one from the second turn and two from the helix. (**b**) Urea-induced denaturation of Octarellin II and III. Figures are extracted from Houbrechts et al., 1995 [3].

rescence spectroscopies indicated the presence of both helices (30%) and strands (30%) and weekly packed 3D structures.

Urea-induced denaturation showed a significant difference between Octarellin II and III, as shown in Figure 1.11b. Octarellin II is partially unfolded at very low concentration of urea (0.1 M) and fully unfolded at 2-3 M of urea. In contrast, Octarellin III unfolds according to a sigmoidal cooperative transition that starts ∼2.5 M and ends at ∼4.5 M urea. There are 20 residues of difference between Octarellin II and Octarellin III, and all

of them were designed in β-strand regions. It is interesting to notice how these mutations positively affect the stability of the protein upon urea-induced unfolding.

1.3.3 Octarellin V

The Octarellin V [4] was the first artificial $(\beta/\alpha)_8$-barrel protein to be fully designed with computational assistance. The new generation of Octarellins introduced two differences with the previous designs. First, the repetition of a structural β/α-unit was no longer used, and a full polypeptide of 216 residues was designed with no sequence symmetry. Second, the design was divided in 2 phases, i.e. the backbone design and the sequence optimization. This was possible thanks to the creation of specialized programs for the modelling of peptides and for the calculation of atomic interactions [103–107].

The backbone design was divided in 3 steps: assembly of the 8 β-strands (picture A in Figure 1.12), assembly of the 8 α-helices (picture B), and connection of α and β elements with loops (picture C).

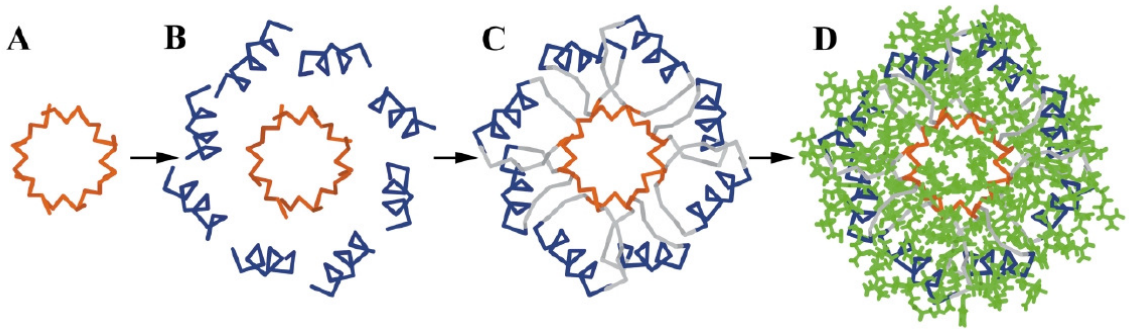


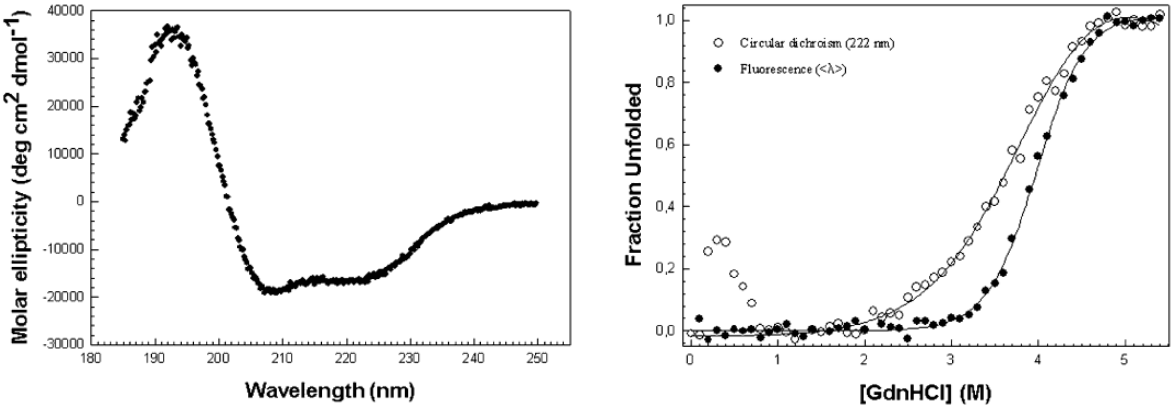
Figure 1.12: Octarellin V, backbone and sequence design
(**A**)Assembly of β-strand elements and (**B**) α-helix elements; (**C**) loop connection formation and (**D**) sequence optimization. Figures are extracted from Offredi et al., 2003 [4].

Once the full scaffold was assembled, the sequence was optimized for the target backbone (picture D) [108], resulting in:

```
MAFLIVEGLSEKELKQAVQIANEQGLRAIAFLKQFARNHEKAERFFELLVREGVEAIIIARGVS
EREIEQAAKLAREKGFEALAFLAEYERRDRQFDDIIEYFERYGFKAVIVATGLDEKELKQAAQK
IEEKGFKALAFLGRIDQENRKIKDIFELLQRQGLRAIIAATGLSERELSWALRAARQYGLDIIF
AYGQFDEQDNQFKHFLELIRRLGAA
```

Octarellin V was produced in inclusion bodies, refolded and characterized by DLS, far- and near-UV CD, thermal and chemical unfolding and ^{1}H-NMR. The far-UV analysis is

shown in Figure 1.13a, and GdmCl-induced unfolding is shown in Figure 1.13b. All the experimental results were encouraging, with the correct secondary structure percentages, a compact 3D structure and a cooperative thermal unfolding with a melting temperature (T_m) equal to 65°C. However, it was not possible to solve the structure of the protein by X-Ray crystallography, probably due to its low solubility.



(a) (b)

Figure 1.13: Octarellin V, characterization

(**a**)Far-UV CD spectrum of Octarellin V and (**b**) its chemical unfolding in GdmCl by far-UV CD and fluorescence. Figures are extracted from Offredi et al., 2003 [4].

1.3.4 Octarellin VI

Octarellin VI was designed in 2013 by Figueroa et al. [5], using the Octarellin V backbone as a starting scaffold. Sequence optimization was performed by layers using the Rosetta software [109]: residues of the hydrophobic ring of the protein (in yellow in Figure 1.14) were allowed to be changed by hydrophobic residues (M,A,F,L,I,V,W,Y,G and H) only, while residues of the central pore (in red) and of the surface (in blue) did not have restrictions for the substitution.

The resulting sequence of Octarellin VI is:

```
MSRGFHFGGPASEEWWERFLRHGEEANYHHGFAAPTGGNYEEARKLAKQVWNNSGGRILWWGQG
GNLQAAHQGARYGREGAGQFAFWSDGTGGLQELYKYFQGVHNFNNHNIFISGNGGDDNTRKKAL
ELIARLNGKGFYWADAGNGYQLWLAWLQHVQQGNGGGLGILGNLGHWQRITEFLEWAKKHQGSG
YLVSNGGGNHQQALAFFEWIRQSS
```

For the first time in the Octarellin history, molecular dynamic simulation (MD) in explicit solvent was performed for 5 nanoseconds with the GROMACS package [110], in order to verify the stability overtime of the Octarellin VI. Analysis of backbone RMSD vs time (in Figure 1.15a), RMS fluctuation per residue (in Figure 1.15b), radius of gyration and secondary structure content were performed.

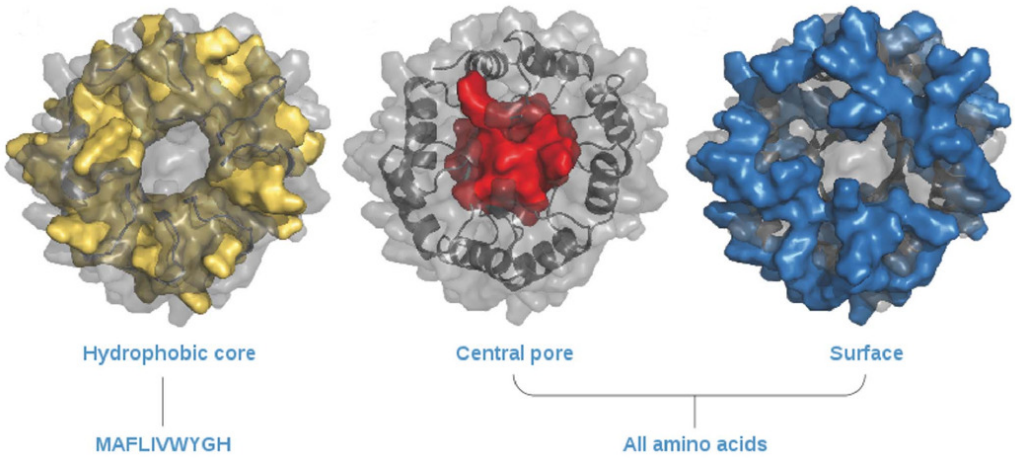
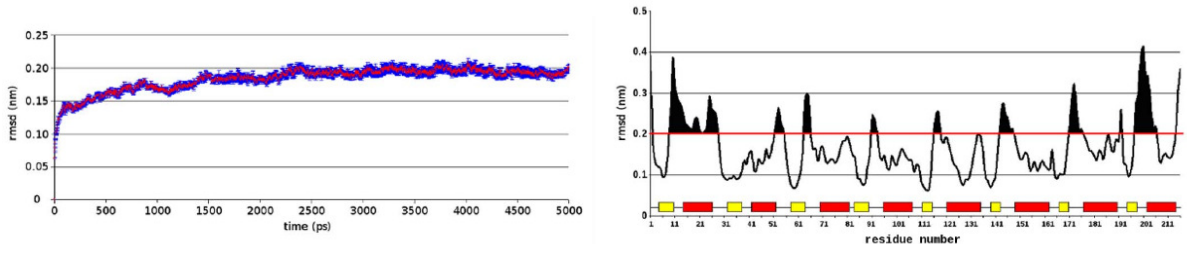


Figure 1.14: Octarellin VI, sequence optimization

Layer definition for the OctarellinVI sequence optimization: hydrophobic core (yellow), central pore (red) and surface (blue). Figures are extracted from Figueroa et al., 2013 [5].

The protein was expressed in *E. coli* and produced in inclusion bodies, as for the previous Octarellins. After refolding, DLS analysis showed the protein to be monomeric, and far-UV CD indicated a structural content of 34% helices, 18% strands, 19% turns and 29% unordered regions. Near-UV CD spectrum suggested the presence of a stable tertiary structure. Thermal unfolding showed protein stability up to 70°C followed by irreversible denaturation. Chemical unfolding by urea showed a non cooperative transition. However, as for the previous Octarellins, the poor solubility of Octarellin VI did not allowed the formation of crystals for structure determination.



(a) **(b)**

Figure 1.15: Octarellin VI, molecular dynamic simulation

(**a**) RMSD vs simulation time and (**b**) RMS fluctuation per residue of the designed model of Octarellin VI. Figures are extracted from Figueroa et al., 2013 [5].

1.3.5 Octarellin V.1

In 2016 Figueroa et al. [6] returned to Octarellin V in order to improve its solubility and, hopefully, solve its structure. They used the Green Fluorescent Protein (GFP)

reporter method developed by Waldo's group in 1999 [111].

In this method, 20 proteins of *Pyrobaculum aerophilum* were selected for their different solubility upon expression: from fully soluble proteins to fully unsoluble ones (i.e. inclusion bodies). The proteins were then expressed in fusion with the GFP in *E. coli*. The ones that were produced in inclusion bodies showed almost no fluorescence emission, while fully soluble proteins were highly fluorescent. The researchers demonstrated how the GFP can be used as solubility reporter, *in vivo*, since its fluorescence is directly correlated with the solubility of the fusion protein, (shown in Figure 1.16).

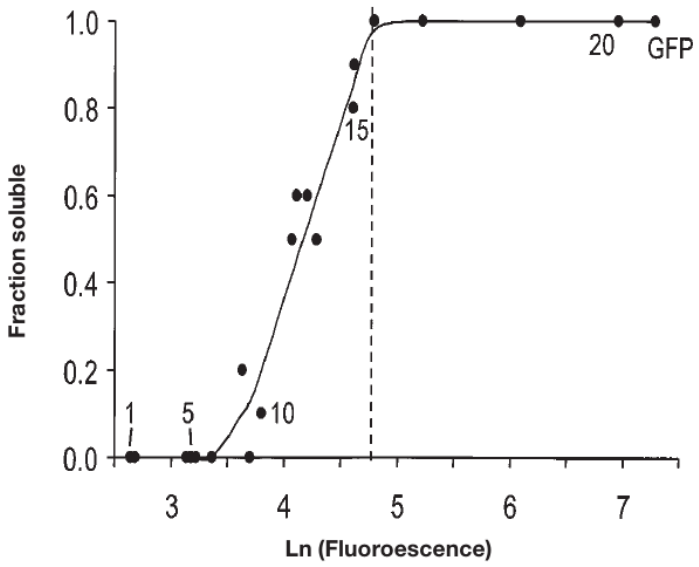


Figure 1.16: Solubility of GFP-fusion products

20 proteins of *Pyrobaculum aerophilum* are over-expressed in fusion with GFP in *E. coli*, and their solubility is directly related to the GFP fluorescence. Details about the proteins are in Waldo et al., 1999 [111].

Figueroa et al. performed directed evolution by using error prone PCR on the gene of Octarellin V, which was then inserted in a plasmid in fusion with the GFP. They were able to easily screen *in vivo* thousands of clones and select the most promising in terms of solubility (Figure 1.17a, extracted from Figueroa, 2016 [6]). Following 8 rounds of directed evolution, the expression of Octarellin V partially shifted from the inclusion bodies to the soluble fraction of the cell extract (Figure 1.17b).

The result of the directed evolution is Octarellin V.1, that contains 16 mutations mainly located at the N- and C-terminal part of the protein (93% of sequence identity). Biophysical characterization (CD, fluorescence and SAXS among the others) showed no significant differences at both secondary and tertiary structure levels between Octarellin

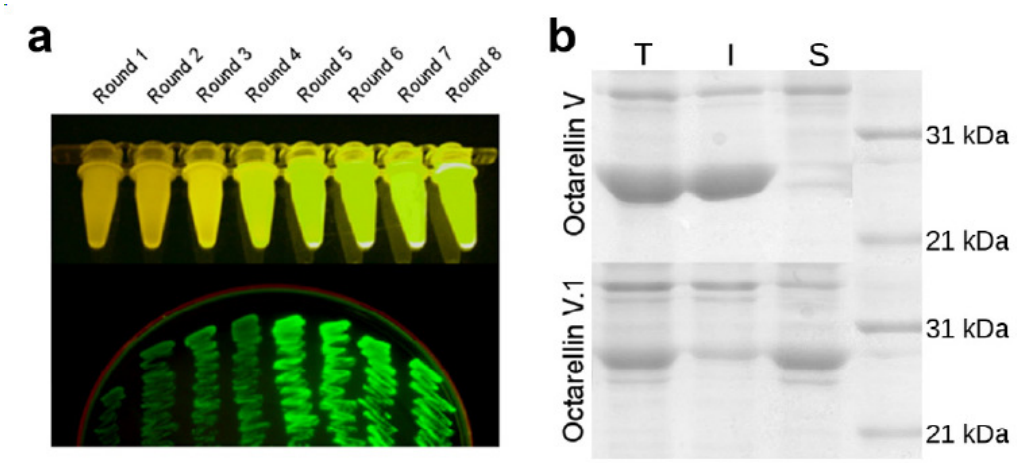


Figure 1.17: Directed evolution of Octarellin V to Octarellin V.1

(**a**) *In vivo* fluorescence of Octarellin V in fusion with GFP reporter through 8 cycles of directed evolution and (**b**) expression of Octarellin V and V.1 in the total fraction (T), insoluble fraction (I) and soluble fraction (S) of the crude extract. Figures are extracted from Figueroa et al., 2016 [6].

V and Octarellin V.1.

Octarellin V.1 was then crystallized in the presence of two crystallization helpers: α-reps (Figure 1.18a) [112, 113] and nanobodies (Figure 1.18b) [114, 115]. α-Reps are based on natural HEAT repeat proteins, and are composed by α-helices only. From a library of 1.7 x 10^9, four different α-reps were selected for their interaction with Octarellin V.1. Nanobodies are single-domain fragments derived from the heavy chain-only antibodies that are naturally produced by *Camelidae*. They are formed mainly by β-sheets and have a molecular mass of 15 kDa. Llama immunization was obtained with 6 injections of 1 mg of purified Octarellin V.1 over a period of 6 weeks. A blood sample of 100 mL was obtained by the immunized llama, and the Peripheral Blood Lymphocytes (PBLs) were isolated. Following total RNA extraction and cDNA synthesis, the genomic sequences encoding the nanobodies were obtained. They were used for phage display selection against the purified OctarellinV.1. Seven nanobodies were found to stable complexes with the protein. Their genomic sequence was then identified.

Both nanobodies and α-reps were produced in *E. coli* BL21(DE3) and purified in two steps: IMAC and size exclusion chromatography.

Crystals were formed with 1 out of the 7 nanobodies and with 1 out of 4 α-reps. The two helpers were found to bind different sites of the protein, and to provide sufficient stabilization for crystallization. In both cases, the obtained structure is not complete, however they are partially complementary and they agree on the position of all the secondary structures. Following modelling, the 3D structure of Octarellin V.1 was obtained.

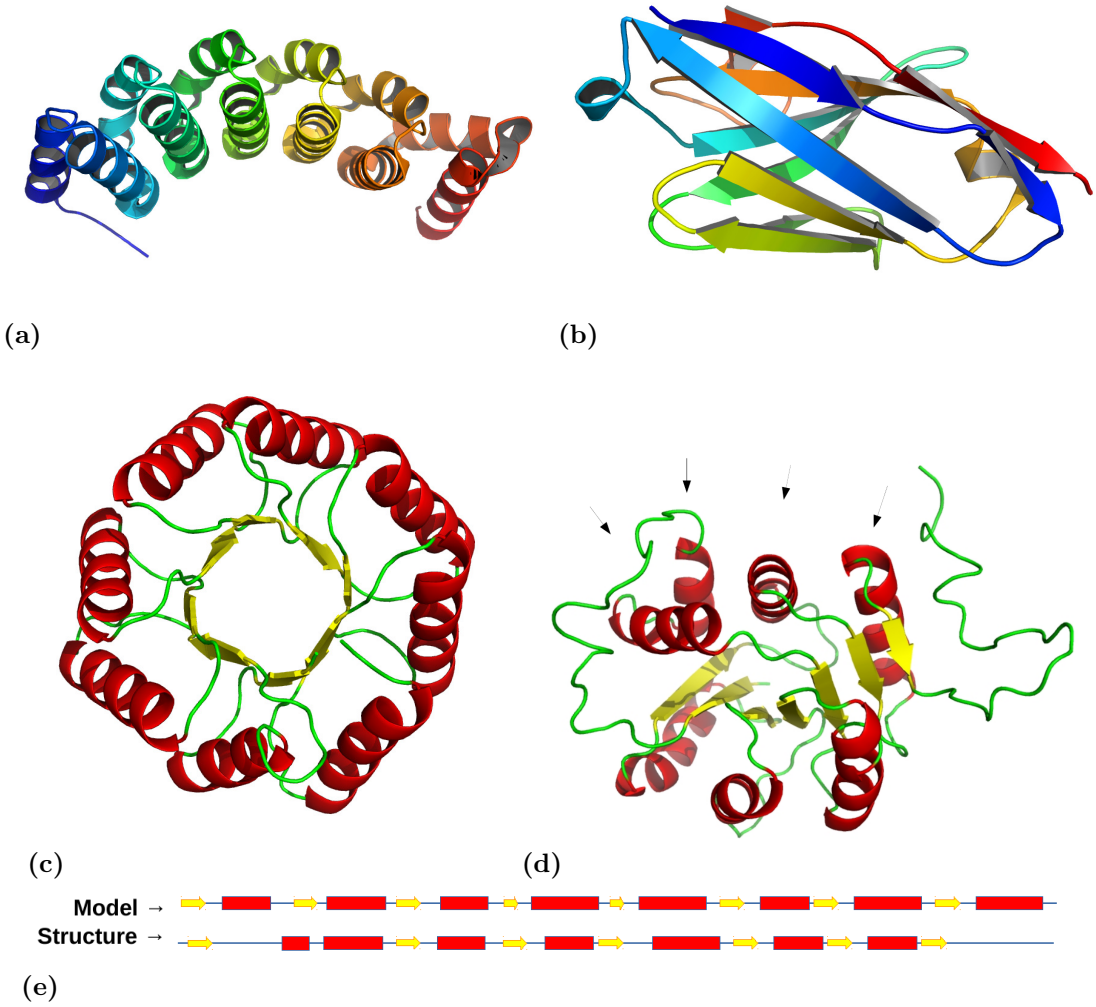


Figure 1.18: Octarellin V.1, model vs structure

(**a**) Crystal structures of the α-rep and (**b**) the nanobody that bind Octarellin V.1. (**c**) Model and (**d**) structure of Octarellin V.1, the black arrows indicate the binding site of the crystallization helpers; (**e**) comparison of their secondary structure topologies: β-strands (yellow arrows) and α-helices (red bars).

Unexpectedly, the experimental fold resulted to be an $\alpha\beta\alpha$-sandwich (Figure 1.18d) instead of the designed TIM-barrel (Figure 1.18c). The nanobody and the α-rep interact with 3 and 4 helices, respectively, on the same side of the $\alpha\beta\alpha$-sandwich (black arrows in Figure 1.18d). Analysis of the secondary structures (Figure 1.18e) showed that the position of 7 out of 8 β-strands and 6 out of 8 α-helices in the primary structure of the protein corresponded to the designed one, but their 3D arrangement resulted to be wrong.

In conclusion, none of the 6 artificial TIM-barrels designed by the Martial's group were fully successful. All of them were produced in inclusion bodies and displayed low solubility. Resolution of the X-ray structure of Octarellin V.1 showed that the design of

secondary structure elements was reasonably good, and that troubles arrive with their arrangement in a 3D structure.

1.4 Other *de novo* TIM-barrels

During the first 2 years of my PhD, two papers were published on the *de novo* design of artificial TIM-barrels [116, 117]. Both describe the design and the characterization of the proteins, but the second only reports structural data at high resolution. Different methodologies were used in the designs and a short description of both works is reported in this chapter.

1.4.1 Symmetrins

In 2015 the group of Rao, Bangalore (India), published a paper [116] on the design of artificial and symmetric TIM-barrel proteins called Symmetrins. The design protocol is different in comparison to the one described in this work for the Octarellins. It is reported in Figure 1.19.

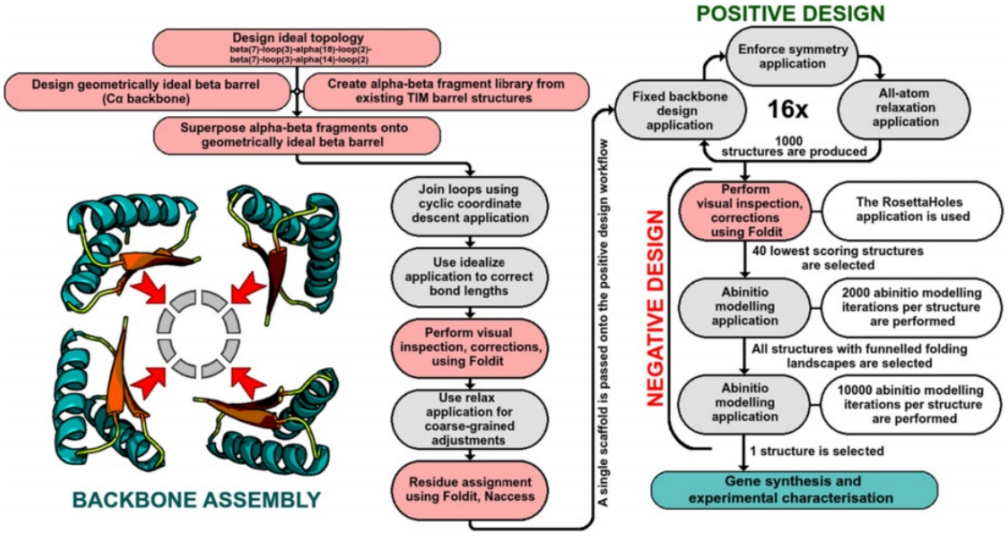


Figure 1.19: Design protocol of Symmetrin proteins

Schematic representation of the design steps for the Symmetrin protein. The picture is taken from Nagarajan et al., 2015 [116].

Backbone Design

Two ideal $\beta\alpha\beta$ motives of different size were selected as a starting point. The two motives were then integrated to form a single $\beta\alpha\beta\alpha$ unit of 56 residues that formed the single unit repeat of the design. The backbone of the repetition unit is assembled with the use of $\beta\alpha\beta$ fragments obtained from natural TIM-barrel structures found in the CATH database [44]. The unit is then repeated 4 times to form a full TIM-barrel, with a total of 224 residues.

Sequence Design

The backbone structure is subjected to 16 cycles of design and energy minimization, performed with the Rosetta software. This protocol is quite standard in the protein design field, and it is used also in the design of the Octarellins. What is differing is the symmetric restriction that imposes the same amino acid mutations in all the 4-repeated unit of the Symmetrin protein. 1000 models were created during the sequence design.

Refinement

40 out of the 1000 models are selected based on their lowest Rosetta energy scores and are subjected to *ab-initio* folding with Rosetta. This technique is not well performing on proteins with more than 100 residues, but it is well-suited for symmetric protein that can be split into smaller sub-units. The 40 models were simulated for *ab-initio* folding with a 95 residue sub-unit. The best one was selected for experimental validation and called Symmetrin-1. Symmetrin-2, -3 and -4 were designed from Symmetry-1 in order to test different kinds of protein stability through mutations in pore residues.

Experimental Validation

Among the 4 Symmetrins, only one is produced in inclusion bodies, while the remaining three are produced in the soluble fraction. They present an alpha-beta spectrum by CD spectroscopy, and two of them are monomeric. Surprisingly, the two monomeric Symmetrins have a T_m of 44°C, while the oligomeric one reaches T_m = 63°C. The 1D-NMR spectra is well resolved for Symmetrin-1 but it indicates that the protein is in a molten globule state. No protein structure was published.

1.4.2 sTIM-11

In 2016, the groups of Baker (Seattle, USA) and Höcker (Tubingen, Germany) published a paper on the design of another symmetric TIM-barrel, called sTIMs [117]. The design of the protein is similar to the Symmetrins one described in the previous chapter. The groups succeed in obtaining the structure of one of their models, sTIM-11, and their design will be used in this work as a positive control during the *in silico* validation (Section 2.3.7, page 61).

Backbone Design

For the backbone design, the groups first focused on a single $\beta\alpha\beta\alpha$ unit. They decided a fixed length for the β-strands (5 residues) and sampled different lengths for N- and

C-term loops (from 2 to 3 residues) and for the α-helices (from 10 to 14). The best combination they obtained is a 46 residue repeat unit, that is repeated 4 times to form the full TIM-barrel model.

Sequence Design

Sequence design is done with 10 cycles of Design and Energy Minimization, the classic protocol of the Rosetta package. As for the work on the Symmetrin, the group used restrictions in order to actively design only a fourth of the whole structure, corresponding to the $\beta\alpha\beta\alpha$ unit. Any mutation in the sub-unit is automatically adjusted in the remaining 3 sub-unit of the TIM-barrel.

Refinement

The best 22 models according to the Rosetta energy score are selected for experimental validation, without any refinement step.

Experimental results

All the 22 proteins are highly expressed in *E. coli* and purified. 5 out of them showed cooperative thermal denaturation and only one was crystallized, sTIM-11. Its structure resolution is 2 Å, with an overall Cα-RMSD of 1.28 Å and deviations mostly in C-terminal loops (Figure 1.20). A disulfide bridge was designed in sTIM-11 model in order to connect the first and the last units of the TIM-barrels. However, the solved structure shows that the two cysteines do not form a disulfide bond.

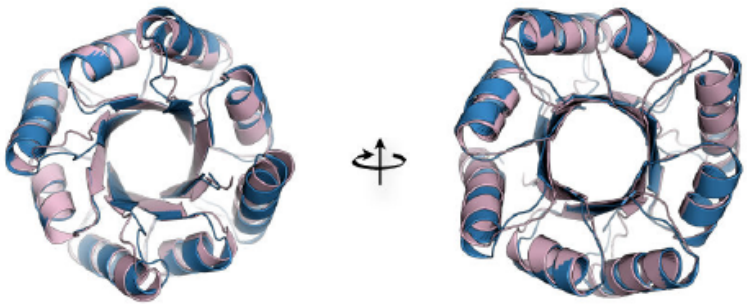


Figure 1.20: Comparison of the model and the structure of sTIM11
X-ray crystal structure (in blue) and the designed model (in pink) of sTIM11, top and bottom views. The picture is taken from Huang et al., 2016 [117].

Chapter 2

Results and Discussion

2.1 Analysis of natural TIM-barrel proteins

This chapter describes the generation and the analysis of a collection of natural TIM-barrel structures. Information collected from natural proteins is relevant in both the design of artificial TIM-barrels (Chapter 2.2, Protein design, page 45) and their validation (Chapter 2.3, *In silico* validation, page 55).

Guidelines for the generation of the collection are two: variability and resolution. As described in Chapter 1.2, page 11, natural TIM-barrels differs widely in function, 3D structure and amino acid sequence, and the collection should take into account all the differences and should not promote one family over the others (variability). At the same time the collection should contain only high resolution structures, that are qualitatively better for analysis, modelling and for a later comparison with artificial models.

Once the collection is set, its analysis is performed in terms of dimension, energy and composition. Amino acid full length distributions of the total protein and of individual strand, helix and loop will define the geometries for the design of artificial TIM-barrels (see Section 2.2.1, Parametric backbone design, page 45). The energy profile and the amino acid composition will be used as guidelines for the selection of the artificial models that will be tested experimentally (see Sections 2.3.1 to 2.3.5, starting at page 55).

2.1.1 Collection of natural TIM-barrels

A simple Google search for "TIM-barrels" lead to the DAtabase of Tim barrel Enzymes (DATE), a ready-to-use collection of natural TIM-barrel enzymes. The web-site was created by S. Kumar Singh and M. Madan Babu at the MRC Laboratory of Molec-

ular Biology in Cambridge (UK) with the aim of providing a quick and comprehensive information about the TIM-barrel of interest. The database includes 85 TIM-barrel enzymes, that are analyzed for composition and residue conformation (Ramachandran plot). Information as sequence, length, oligomeric status, function and metabolic pathways is included in the description. There are however two major disadvantages of the website: only enzymes are included in the list and it was last updated in July 2001. Although it is a good starting point to find natural TIM-barrel structures, it is obviously obsolete.

The 85 enzyme structures were downloaded from the web-site and manually visualized with Pymol [118]. 52 pdb files were selected and used as input on PDBeFOLD [119], a web-service for structural alignment developed in 2003 by E. Krissinel and K. Henrick at the European Bioinformatic Institute in Cambridge (UK). It can perform 3D structure comparison between two structures (pairwise) or more (multiple), and it can look for protein similarity within the whole RCBS-Protein Data Bank archive (RCBS-PDB) [120], which is exactly what we need in order to find more recent structures and improve our collection of natural TIM-barrels. The 52 structures obtained from DATE were loaded on the PDBeFOLD web-site for structural alignment against the RCBS-PDB archive and 1623 natural TIM-barrel are obtained as output.

In order to have a non-redundant, high-quality collection, the 1623 structures were refined with PISCES [121]. This web-service was developed by the Dunbrack's group in 2003 at the Institute for Cancer Research in Philadelphia (USA) in order to cull proteins according to their structural resolution and sequence identity. Fasta sequences were uploaded on the web-site, and cut-offs of 80% for sequence identity and of 2 Å for structure resolution were set. With the help of PISCES, we obtained 228 high-quality and not-redundant structures; following visual inspection with Pymol, this number was reduced to 219.

Both collection, obtained using PDBeFOLD and PISCES and containing 1623 and 219 natural protein respectively, were analyzed for chain length. The PDBeFOLD collection will not be used for later analysis, and it is shown in Figure 2.1a only to confirm that the PISCES collection is well representing the whole dataset.

The two distributions are congruous with each other. The length range is between 172 and 1052 aa, with the highest number of representatives in the range of 250-300 aa. In the collection of 219 natural TIM-barrels (the only one considered from now on), the minimal length is represented by the structure with pdb code 1VKF, of 172 aa (Figure 2.1b), while the maximal length correspond to the structure 2FHF with 1052 aa (Figure 2.1c).

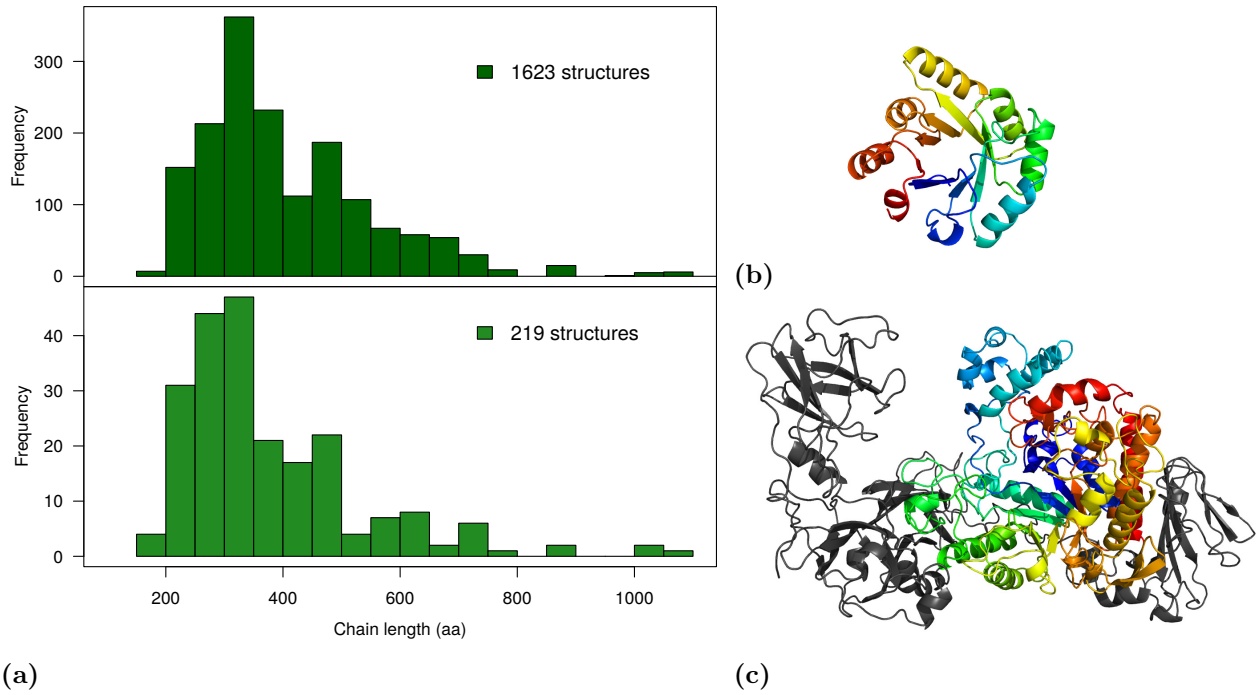


Figure 2.1: Collection of natural TIM-barrels

(**a**) Length distributions for the collections of 1623 and 219 natural TIM-barrels; (**b**) the shortest natural TIM-barrel of the collection, 172 aa (pdb ID: 1VKF) and (**c**) the longest one, 1052 aa (pdb ID: 2FHF).

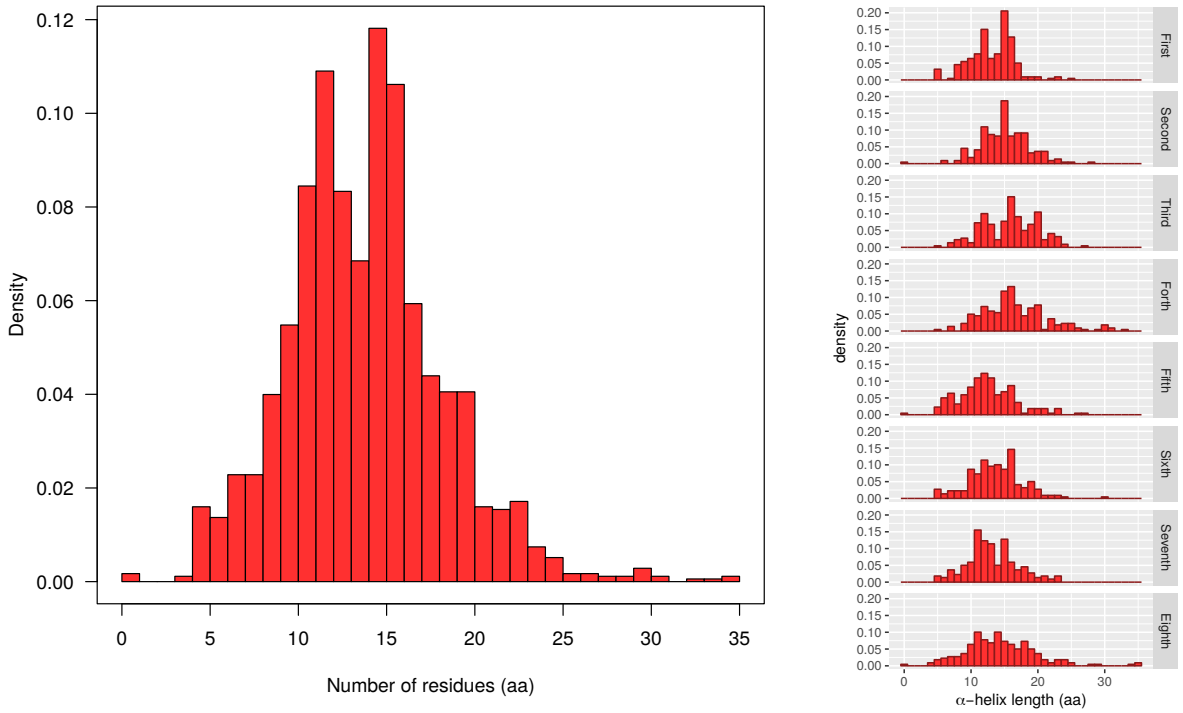
More than 60% of the sequence of 2FHF is not involved in the formation of the TIM-barrel fold: 392 aa are involved in the formation of 3 extra-domains at the N-term of the protein and 125 aa in the formation of one extra-domain at the C-term of the protein (highlighted in gray in Figure 2.1c).

Extra-domains are common in almost all the natural TIM-barrels and can be localized also in the loop regions. Their presence should be taken into account for the analysis of TIM-barrel features (i.e. the average element length), because they can affect the results.

2.1.2 Length distributions of α-helices

A plot showing the distribution of α-helices according to their length is shown in Figure 2.2a (the procedure is described in Section 3.1.2, page 169).

In the collection of 219 natural TIM-barrels, the only one considered, there are 3 proteins that lack one out of the eight α-helix that normally form the α-barrel (pdb IDs: 1H4P, 1TZZ and 2ZUV). In the collection, the minimal length for helices is 3 aa and the maximal is 34 aa. The highest populated lengths are 11 aa and 14-15 aa, showing that the distribution is not homogeneous. This 3-4 aa difference can be explained as a full turn of the helix, in order to keep the overall orientation of the N- and C- termini in direction of the loops; 1-2 extra aa would cause a misplacement of the N and C- termini in the helix



(a) **(b)**

Figure 2.2: α-helix length distribution

(**a**) Length distribution for all the helices of the collection that are involved in the TIM-barrel fold and (**b**) single distributions of the helices for each position in the TIM-barrel fold.

and longer loops would be required.

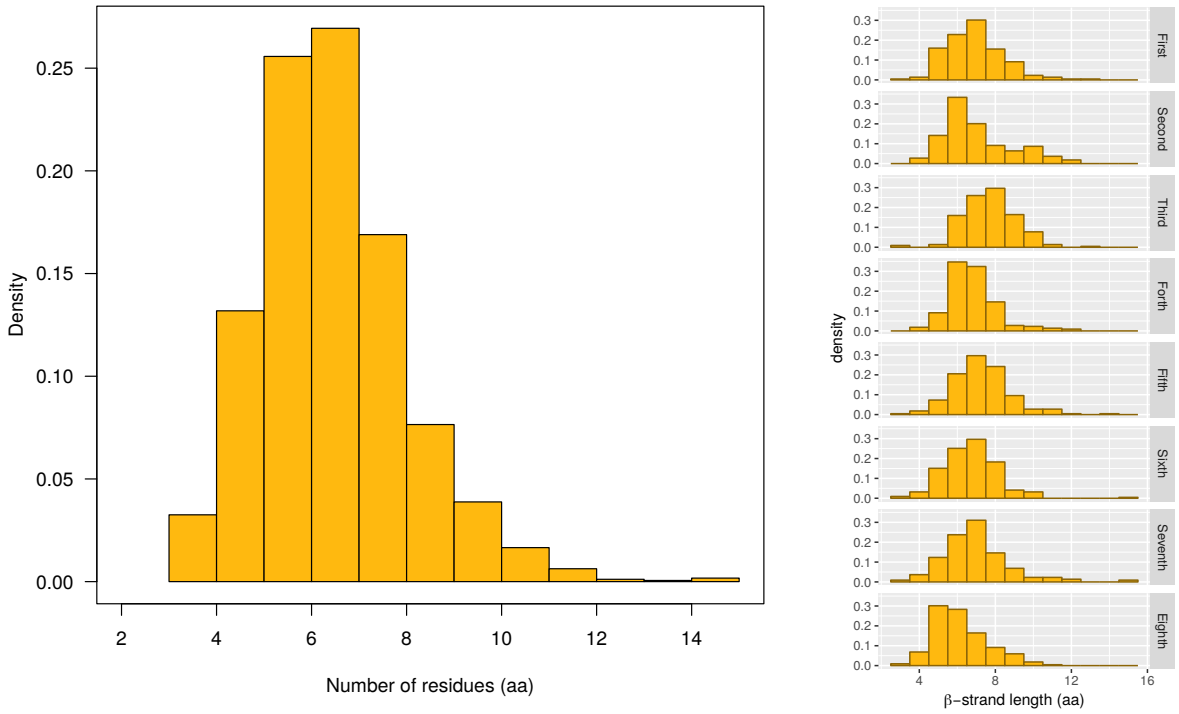
Length distributions at each position of the α-barrel are reported in Figure 2.2b. The general trend of main helix lengths at 11 aa and 14-15 aa is conserved and these three lengths together represent 20 to 47% of the whole population. This analysis further reveals that the 1^{st} and the 5^{th} helices are generally not longer than 17 aa. In contrast, the 4^{th} and the 8^{th} ones can be made of up to 35 aa. The theory of evolution of the TIM-barrel fold from two halves, described in Section 1.2.3, page 14, can be reflected in these distributions.

The length distribution of 8^{th} helix is wider than the other helices, and includes one missing helix, three helices with more than 30 aa and mini-helices of 3-4 aa that are not present in the other helix distributions.

2.1.3 Length distributions of β-strands

A plot showing the distribution of β-strands according to their length is shown in Figure 2.3a (the procedure is described in Section 3.1.2, page 169).

Contrary to the α-helices, in the β-strands there are no missing elements. The minimal



(a) **(b)**

Figure 2.3: β-strand length distribution

(**a**) Length distribution for all the strands of the collection that are involved in the TIM-barrel fold and (**b**) single distributions of the strands for each position in the TIM-barrel fold.

length is 3 aa and the maximal one is 14 aa. The highest populated lengths are 5 and 6 aa, representing together more than 50% of the population.

Figure 2.3b shows the length distributions for each individual β-strand of the β-barrel. The general behavior with the main length at 5-6 aa is conserved but odd strands show a maximum at $\geq$6 aa while the even ones have it at $\leq$6 aa. This trend can be explained taking into account the structural differences in the even and odd β-strands of the TIM-barrels, as reported in Section 1.2.1, page 11.

2.1.4 Length distributions of loops

In this work loops are considered as the connectivity unit between α-helices and β-strands involved in the TIM-barrel fold and they can include extra element of secondary structures.

Two kinds of loops are present in the TIM-barrel fold: from β-strand to α-helix (hereafter called C-term loop, because it is at the C-terminus of the β-strand), and from α-helix to β-strand (N-term loop). The first is mainly involved in function (i.e. active site), while the latter is important for the structure and stability of the protein, as discussed in

Section 1.2.1, page 11.

Plots showing the distributions of C-term and N-term loops according to their length are shown in Figs. 2.4b and 2.4a respectively. Their equivalent subgroups for position in the TIM-barrel are shown in Figs. 2.4d and 2.4c.

The minimal loop length is 1 aa and the maximal is 120 aa for C-term loops and 130 aa for N-term loops. The majority of the N-term loops is less than 10 aa long, and the outlier of 130 residues is exceptional and includes an entire extra-domain (pdb ID: 1ZFJA). The C-term loops are usually less than 50 aa long. This difference is typical in TIM-barrels: as described in Section 1.2.1, page 11, C-terminal loops are usually longer and involved in the function of the natural protein (i.e. enzymatic activity), while the N-term loops are shorter and important for structural stability.

In the N-term group, around 70% of the loops have a length of 1-3 aa, and this trend is strictly conserved at each of the seven possible positions in the barrel. On the contrary, the C-term group does not show conserved trends.

2.1.5 Rosetta total energy of natural TIM-barrels

Energy minimization (also called relax) of the 219 natural TIM-barrel proteins downloaded from RCBS-PDB was done with the Relax package of Rosetta [38] in order to remove bias and constraints present in the crystal structure (details in Section 3.1.3, page 170). The total energy score is calculated by Rosetta in terms of Rosetta Energy Units (REU), that are not related to the physical energy units kcal/-mol or kJ/mol and rely on an arbitrary scale.

As shown in Figure 2.5 with the natural protein 1A53 before (blue) and after (green) energy minimization, there are no relevant changes in the protein structure after relaxation. Only highly flexible regions (i.e. loops) can result slightly affected, with a small deviation of the main-chain compared to the original structure.

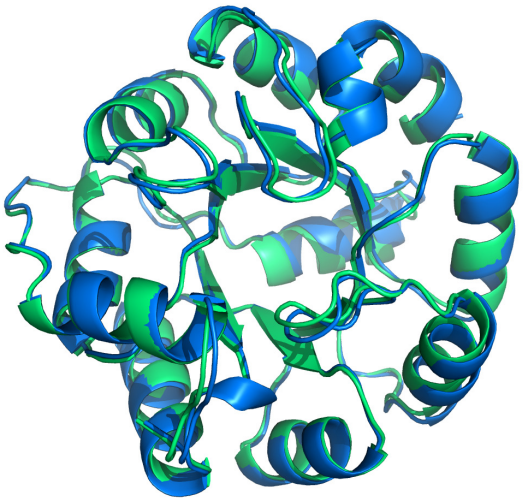


Figure 2.5: Energy minimization
The structure of the natural protein 1A53 before (blue) and after (green) energy minimization.

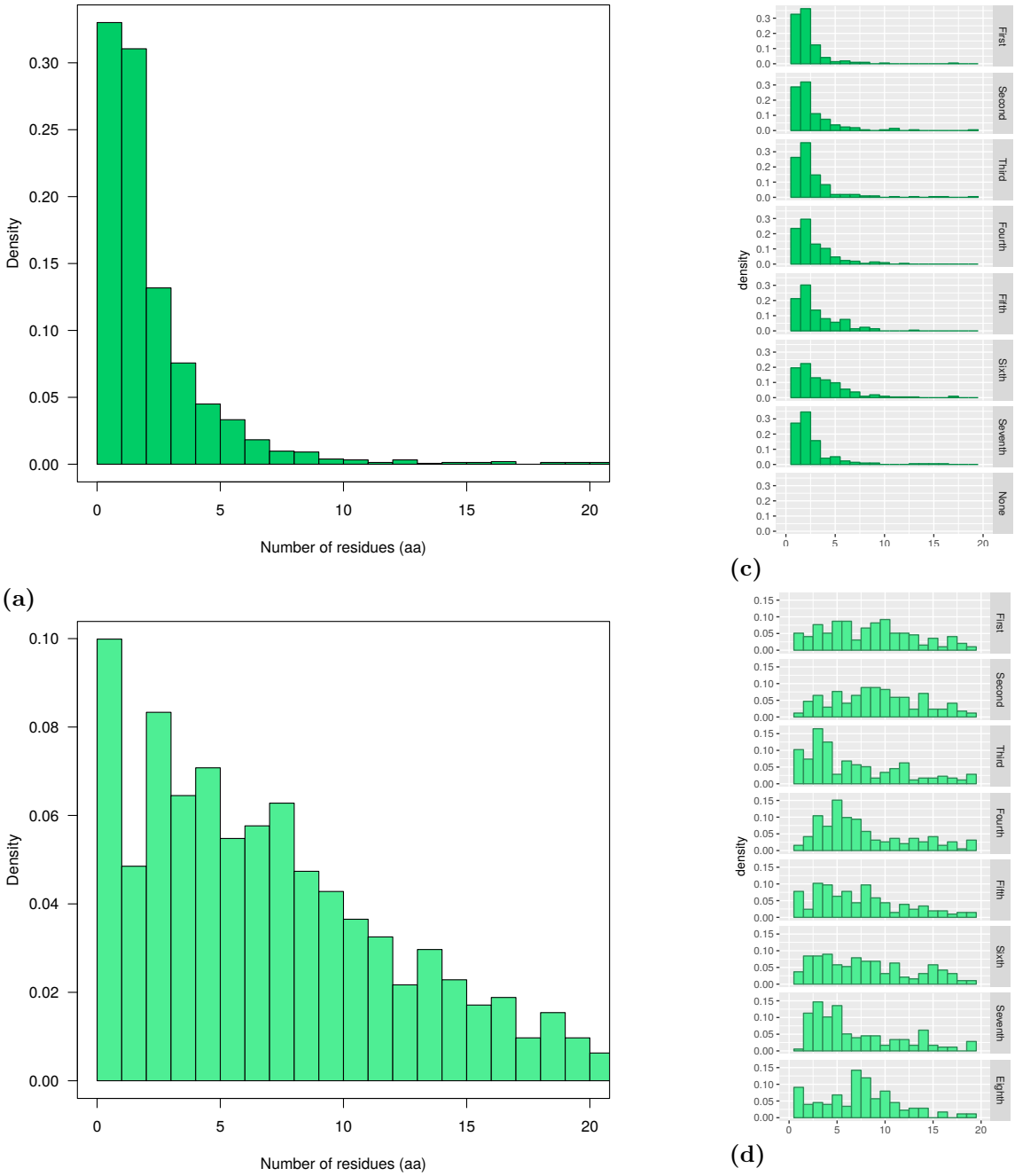


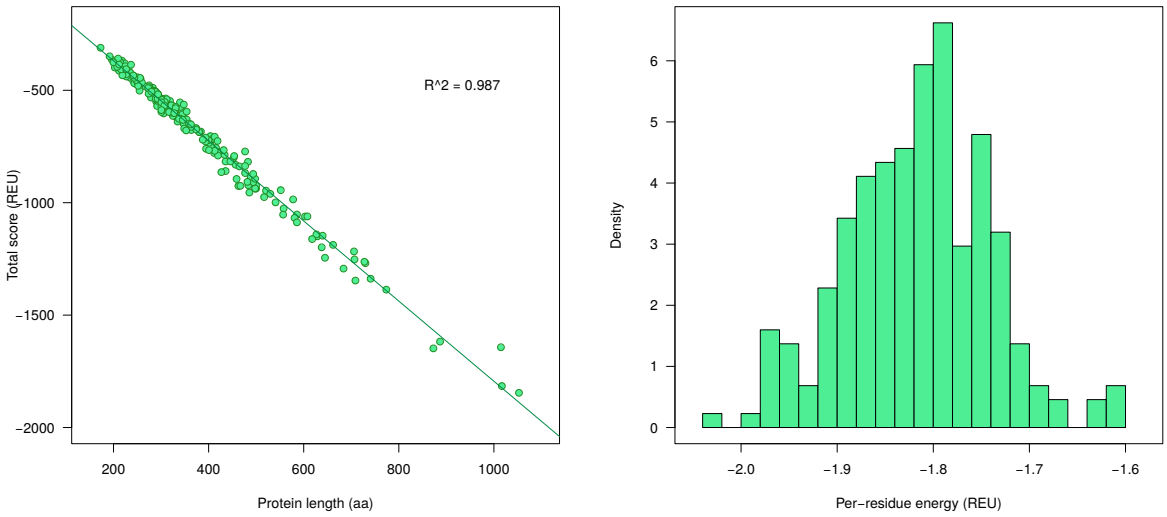
Figure 2.4: Loop length distribution

(**a** and **b**) Length distributions for N-term and C-term loops respectively; (**c** and **d**) their corresponding plots for individual position in the TIM-barrel fold. For graphical clearness, all graphs are showing only the first 20 aa in the scale. Full distributions are showed in Annex 6.3, page 240.

The total energy score for each natural TIM-barrel is calculated and plotted against the protein length (Figure 2.6a). There is a linear correlation between energy and protein

dimension: the bigger the protein the more negative is the score value.

Dividing the energy value by the amino acid length gives the average energy per residue for the collection of natural TIM-barrels, which is between -2.05 and -1.60 REU per residue (Figure 2.6b). Values bigger than -1.6 REU represent structures with poor stability, and this value will be used as cut-off value in the selection of the designed proteins. It is not clear whether it is possible to get values lower than -2.05 REU per residue and obtain super-stable variants, or if it represents a minimum limit. Thus, the range -2.05 to -1.60 REU per residue will be useful later on for the next step of design and validation of artificial TIM-barrels.



(a) **(b)**

Figure 2.6: Energy minimization on natural TIM-barrels

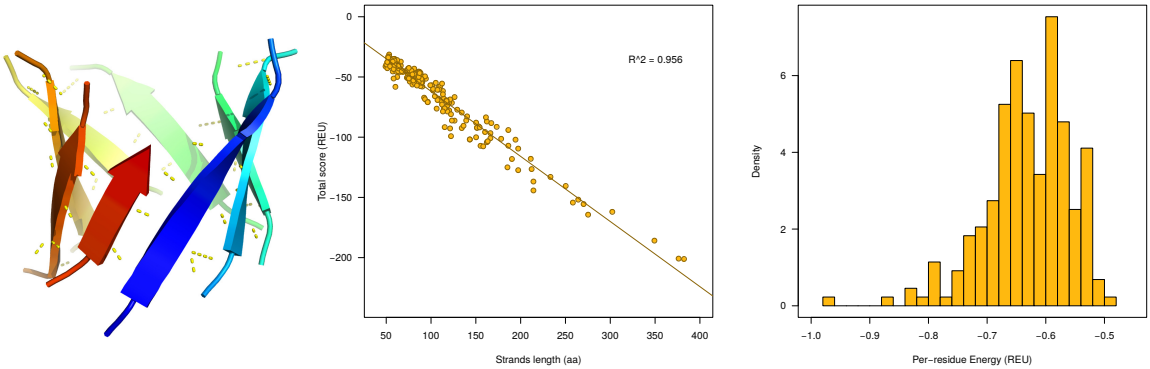
(**a**) Energy score versus protein length for each structure of the collection of natural TIM-barrels and (**b**) distribution plot of the per-residue energy.

2.1.6 Rosetta β-sheet energy of natural TIM-barrels

Analysis of the Rosetta energy score of the β-barrel is an important tool in the comparison of natural and artificial TIM-barrels. In nature, the eight strands form a single sheet that is closed on itself (forming the β-ring). If H-bonds are missing between the strands, the overall stability of the structure decreases. In particular, the H-bonds between the first and the last strands function as a zip that keeps the overall structure in a "closed" conformation (Figure 2.7a).

As described in Section 2.1.5, page 38 for the total energy, the energy value due to H-bonding between the backbone atoms in the β-strands was calculated following energy

minimization (referred as β-energy from now on). The number of residues of the β-strand conformation for each natural protein is plotted versus the β-energy (see Figure 2.7b).



(a) **(b)** **(c)**

Figure 2.7: β-energy of natural TIM-barrels

(**a**) β-barrel detail of the natural protein 1A53 with H-bonds highlighted, (**b**) β-energy score versus total β-length for each structure of the collection of natural TIM-barrels and (**c**) distribution plot of the per-residue β-energy.

The ratio of β-score versus length gives the average β-energy per residue for natural TIM-barrels, and it is shown in Figure 2.7c. The overall range in our collection is between -0.5 and -1.0 REU, with peaks at -0.6 and -0.65 REU. This range will be useful for the design and validation of artificial TIM-barrels.

2.1.7 Analysis of amino acid composition

The relative amount of each natural amino acid in the 219 TIM-barrel sequences was calculated. For example, the protein 1VKF (the smallest protein of the collection with 172 aa, shown in Figure 2.1b), contains 11% of valine, 10% of isoleucine, 9% of alanine, leucine and lysine, 8% of glutamic acid and glycine, 5% of aspartic acid, 4% of phenylalanine, arginine and serine, 3% of proline and threonine, 2% of methionine and asparagine, 1% of histidine, glutamine, thryptophan and tyrosine and no cysteine (see Section 3.1.4, page 171 for details).

Amino acid distributions across all the natural TIM-barrels are shown in Figure 2.8 and the ranges are reported in Table 2.1.

Out of the 219 sequences, 26 do not contain cysteines and 13 do not contain tryptophans. On the contrary, alanines, valines, leucines, and glycines are always found with a minimum of 3% of sequence coverage each. Valines and leucines are the predominant residues in β-strands [122] and it is intuitive their relevance in the TIM-barrel structure, but alanines and glycines are unexpected.

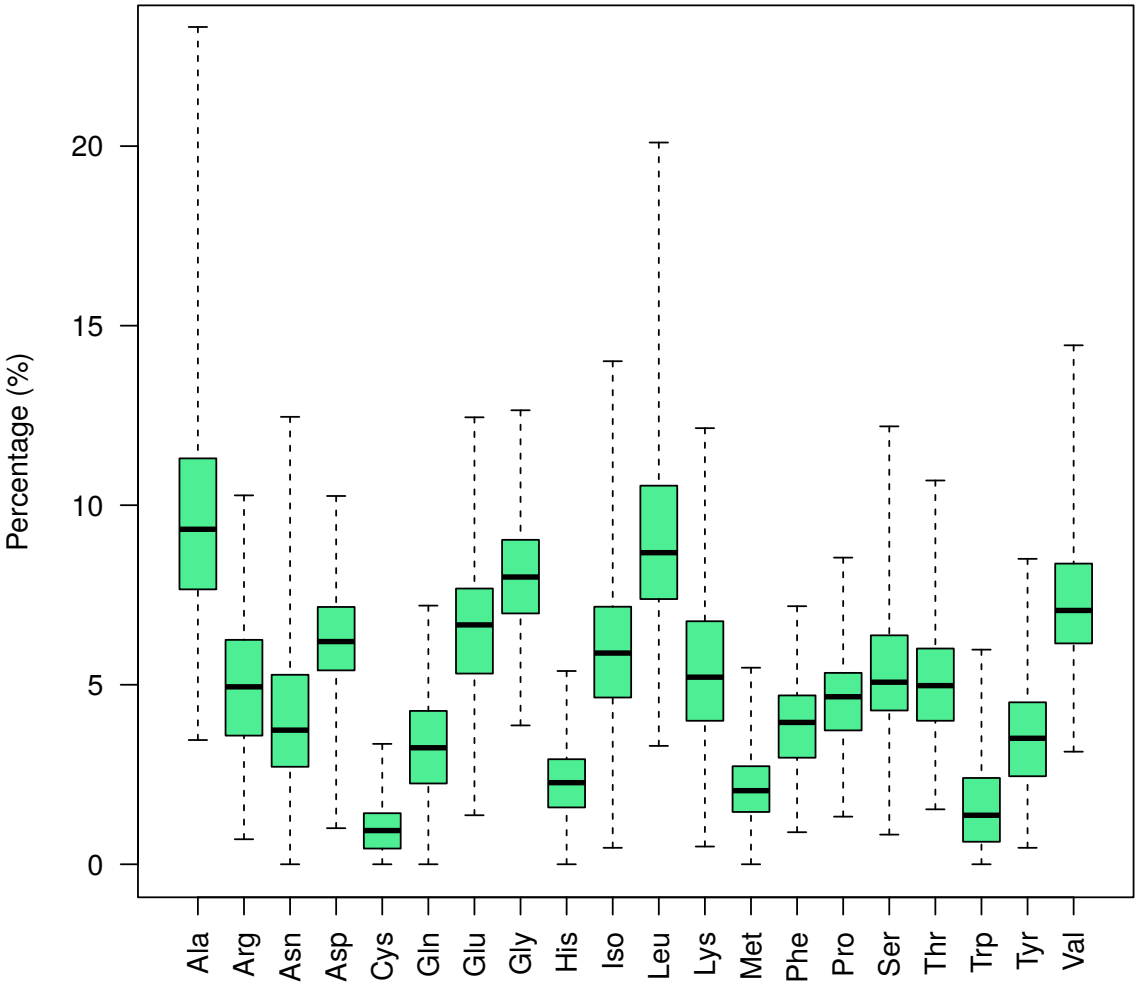


Figure 2.8: Amino acid composition of natural TIM-barrels
In each boxplot the black line inside the green bar is the median, the green bar represent 50% of the population and the upper and lower whiskers represent 25% of the population each.

IDs	Min %	Max %	IDs	Min %	Max %
Alanine	3.4	23.3	**Leucine**	3.3	20.1
Arginine	0.7	10.2	**Lysine**	0.5	12.1
Asparagine	0.0	12.4	**Methionine**	0.0	5.4
Aspartic acid	1.0	10.2	**Phenilalanine**	0.9	7.1
Cysteine	0.0	3.3	**Proline**	1.3	8.5
Glutammine	0.0	7.2	**Serine**	0.8	12.2
Glutamic acid	1.3	12.4	**Threonine**	1.5	10.6
Glycine	3.8	12.6	**Tryptophan**	0.0	5.9
Histidine	0.0	5.3	**Tyrosine**	0.4	8.5
Isoleucine	0.4	14.0	**Valine**	3.1	14.4

Table 2.1: Composition ranges in natural TIM-barrels

2.1.8 Analysis of amino acid properties

The distribution of amino acid in the collection of natural TIM-barrels is an useful information but it is not fully comprehensive. For a functional protein (i.e. enzymatic activity), the residues of the active site are highly conserved and their substitution is harmful in the majority of the cases [92]. On the contrary, outside the active site, sequences are less conserved and most residues can be substituted with others that have similar properties, such as size or charge [93]. For example a lysine can be exchanged with an arginine without changing the overall charge, and they can be consider equivalent. This trend is highly reflected in the TIM-barrel family, that have extremely low sequence identities (down to 15%) despite their common fold (see Chapter 1.2).

In order to detect hidden patterns due to equivalent residues, amino acids were organized in seven groups according to their properties: polar (D,E,H,K,S,N,Q,R,T,C), charged (D,E,K,R), positive charged (K,R), negative charged (D,E), small (A,C,G,S), apolar (W,Y,P,V,P,L,I,A,G,M) and aromatic (W,Y,F) (see Section 3.1.4 for details, page 171). The collection of 219 natural TIM-barrel was tested for each category composition and their distributions are shown in Figure 2.9 and in Table 2.2:

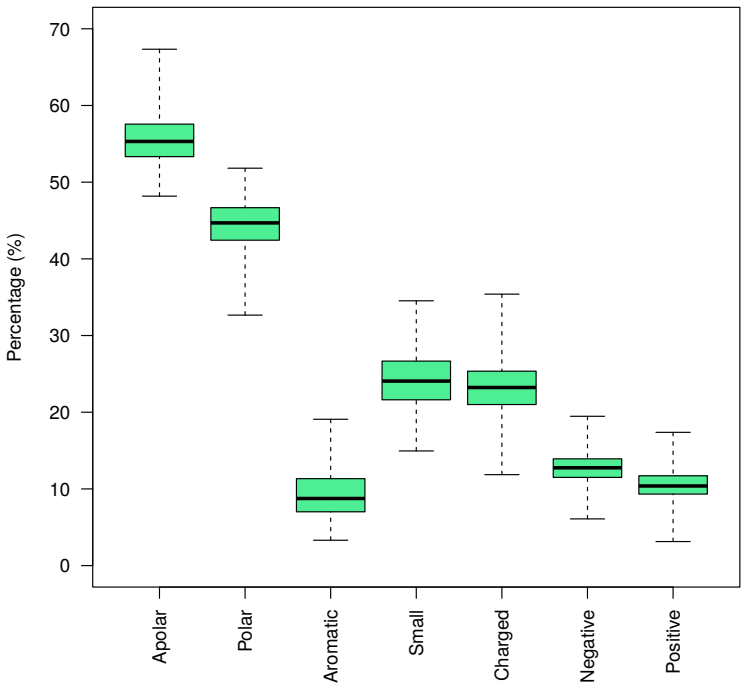


Figure 2.9: Composition by amino acid property of natural TIM-barrels

In each boxplot the black line inside the green bar is the median, the green bar represent 50% of the population and the upper and lower whiskers represent 25% of the population each.

The majority of the proteins in the collection have around 55% of apolar and 45% of polar residues. There is 24% of charged residues, with a slight excess of negatively (13%) versus positively (11%) charged ones. The aromatics represent around 8% of the total composition and the small residues 20%. This analysis will be used in the screening of artificial TIM-barrels, descibed in Section 2.3.5, page 58.

IDs	Min %	Max %	IDs	Min %	Max %
Polar	32.5	51.8	**Negative Charged**	6.0	19.4
Apolar	48.2	67.5	**Small**	14.9	34.5
Charged	11.8	35.4	**Aromatic**	3.3	19.0
Positive Charged	3.1	17.3			

Table 2.2: Amino acid property ranges in natural TIM-barrels

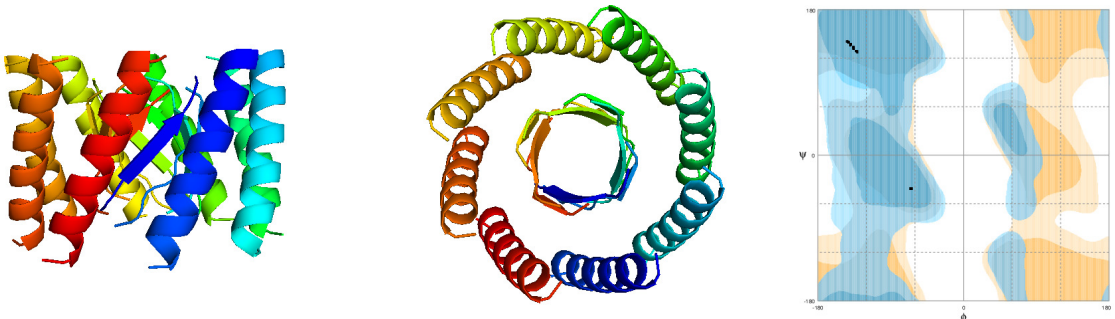
2.2 Protein design

This chapter describes the design from scratch of artificial TIM-barrel structures and sequences. Four steps are needed for the creation of more than 4000 backbone scaffolds. 28 out of them are chosen for sequence optimization in 10 cycles of design and minimization leading to the creation of more than 8000 final models and sequences.

2.2.1 Parametric backbone design

De novo design of scaffolds for artificial TIM-barrels was done with the BundleGridSampler package of Rosetta [38]. This program allows to design from scratch any secondary structure around a central bundle axis and is well suited for proteins that have circular symmetry, like TIM-barrels or helical bundles. Description of software, parameters and options is reported in Section 3.2.1, page 173.

The scaffold was designed by assembling 16 individual peptide fragments, 8 β-strands arranged in an internal β-barrel and 8 α-helices arranged in an external α-barrel, with a total length of 240 aa, all alanines (Figures 2.10a and 2.10b). Secondary structure elements are highly ideal: all residues belonging to the helices have exactly the same ϕ and ψ angles (-65°,-41°), whereas all residues belonging to strands have few different couples of angles: the ϕ values are between -131° and -143° and the ψ ones between 128° and 140° (Figure 2.10c). This small variation in dihedral angles is caused by the tilting of β-strands around the central axis.



(a) **(b)** **(c)**

Figure 2.10: Parametric Design with BundleGridSampler package of Rosetta (**a**) Front and (**b**) top view of the designed scaffold to mimic the TIM-barrel fold and (**c**) its Ramachandran Plot.

Due to the small size of alanine residues, no interactions are formed between the different secondary structure elements, and the Rosetta energy score is +14440.363 REU, which is outrageously high for a protein of this size (-450 REU was calculated from the

per residue distribution in Section 2.1.5, page 38). A more realistic energy score will be obtained following the next steps of design, i.e. alanine substitution in Section 2.2.2, page 46, loop closure in Section 2.2.3, page 48, and energy minimization in Section 2.2.4, page 49.

2.2.2 Alanine substitution

The substitution of alanine in the parametric structure (described in Section 2.2.1, page 45) is a necessary step in the creation of *de novo* scaffolds. Loop closure and energy minimization of the alanine scaffold will cause deformations of the whole structure because the programs are set to fill empty spaces in the core of a protein. The void space between α-helices and β-strands in the parametric scaffold is filled through the substitution of the small alanines with bigger residues, which side-chains form interactions between fragments, leading to increase in the stability of the structure before loop closure and energy minimization.

Alanine substitution was done with the Design package of Rosetta in 4 steps, one for each layer (i.e. core, boundaries, surface and loops, see Figure 2.11) in the structure. The attempt of substitution of all the alanines in one step resulted in a low variation at the sequence level, with just a few different combinations compared to the thousands possible. Parameters, options and instructions for substitution used are described in Section 3.2.2, page 174.

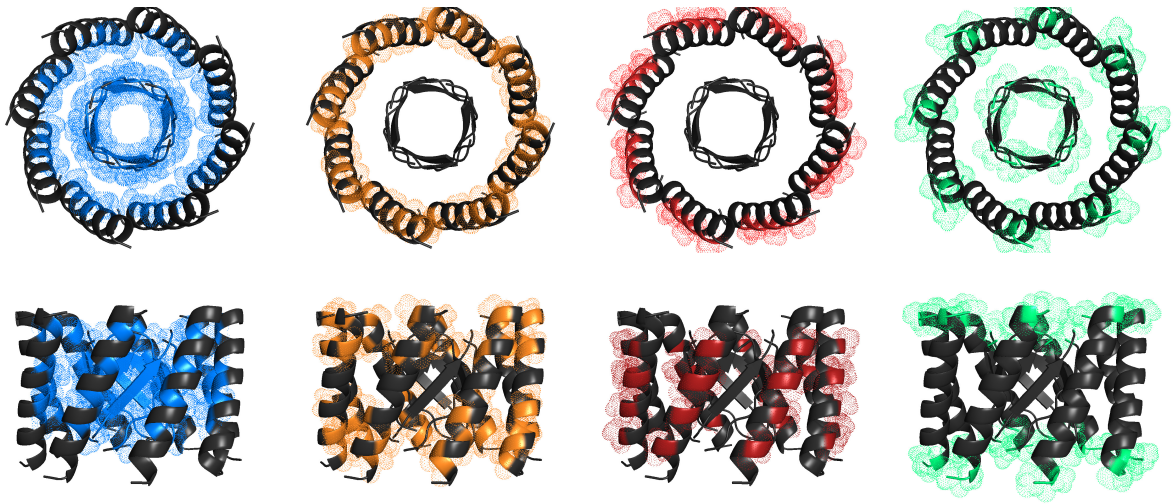


Figure 2.11: Protein layers

Top view and side view of the parametric structure by layers: the core in blue, the boundary in orange, the surface in red and the loop in green. All the 240 residues are attributed to one of the layers.

The first step of alanine substitution targeted all the amino acids that are involved in

the interface between α-helices and β-strands of the TIM-barrel (the hydrophobic ring), and those that point to the central axis of the barrel (in blue in Figure 2.11). The β-strand residues that points to the central axis were limited to be changed only to valine, isoleucine or leucine in order to keep the strand conformation, while those forming the hydrophobic ring are allowed to be changed in any apolar amino acid in order to increase sequence variety and to optimize the stability and the compactness of the structure.

200 outputs were produced and 129 out of them are non-redundant sequences (65%). The energy score improved by 3050 REU from the initial score of +14440 REU, with the worst model out of the 200 scoring +11388 REU and the best +11279. The first 8 best scoring models are selected for the next step.

The second step targeted 40 amino acids localized at the interface between the hydrophobic core and the hydrophilic surface of the protein (in orange in Figure 2.11). These boundary amino acids were allowed to be changed to both polar and apolar residues with the exception of alanine, proline, glycine and cysteine. 400 models were obtained as output, with no duplicates. The energy score improved by 10 REU from the previous step, resulting in range of +11289 and +11268 REU. The first 3 best scoring models for each initial input (a total of 24) were selected for the next step.

The third step targeted 40 amino acids that were exposed to the solvent on the protein surface (in red in Figure 2.11). They were allowed to be changed to any polar residues except glycine and cysteine. 1200 models were obtained (50 for each of the 24 inputs). Two sequences out of the 1200 resulted to be identical and one duplicate was discarded. The energy score improved by 13 REU, resulting in a range of values between +11254 and +11266 REU. 50 best scoring models were selected for the next step.

The fourth step targeted 64 amino acids that form the loop connections between secondary structure elements (in green in Figure 2.11). They were allowed to be changed to any residue with the exception of tryptophan, tyrosine, phenylalanine, glycine, cysteine and lysine. 4000 models were obtained and 32 out of them were discarded because duplicates. The resulting backbones structures are 3968.

The energy value increase from the previous step of 300 REU, resulting in a range of values between +11493 and +11586 REU. This is due to the fact that new residues partially overlap and cause repulsion between atoms, which gives an unfavorable contribution to the overall energy score. So far, neither the backbone nor the side-chains of the protein are been adjusted in order to form interactions (H-bonds, salt bridges or hydrophobic interactions) that will give a more favorable contribution to the score. The

following energy minimization step will remove these bias, but first the loops have to be connected to form a continuous polypeptide chain.

2.2.3 Loop closure

Loop closure to connect the 16 individual peptide fragments in a single chain was performed for each of the 3968 models with the Loopmodel package of Modeller [123]. 15 loops have to be created, 8 to connect β-strands to α-helices and 7 for the opposite direction. In all cases two residues of the β-strand and three of the α-helix were used to form the final loop (see details in Section 3.2.3, page 176).

Figures 2.12a and 2.12b show in green the residues that are involved before (a) and after (b) the loop formation step. There is no addition of amino acids and the rest of the fragments are not distorted, as shown in Figure 2.12c.

The Ramachandran plot (Figure 2.12d) of the single polypeptide is more dispersed compared to the one of the parametric scaffold (Figure 2.10c), but the majority of the residues of strands and helices are still displayed at the original coordinates because, so far, only the loop region was subjected to energy minimization, while the central body is still not relaxed. The energy score reflects this partial minimization, values improved of 5000 REU ranging from 6852 to 9273 REU, but it is still not close to expected scores.

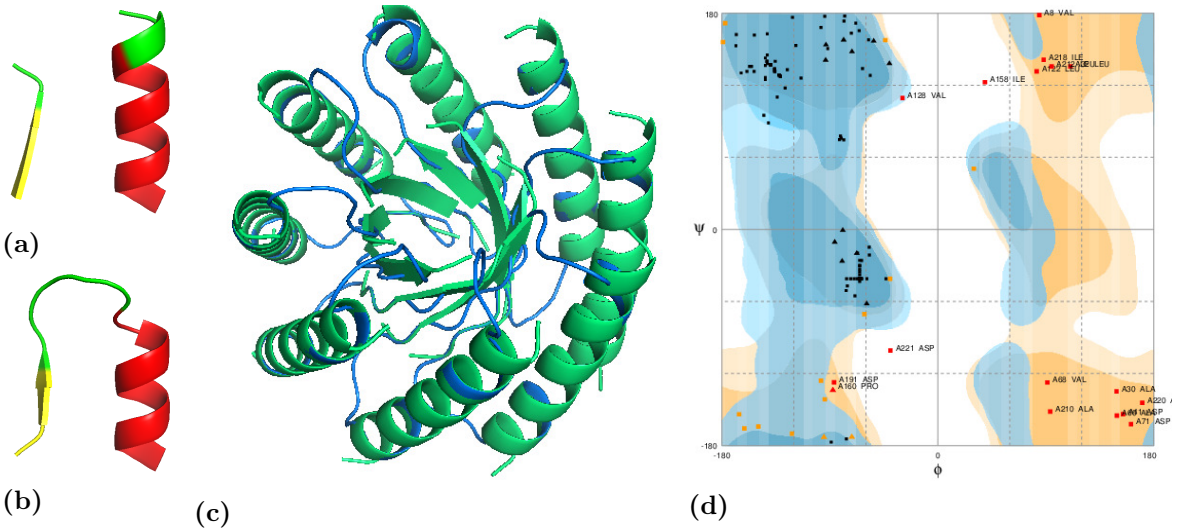
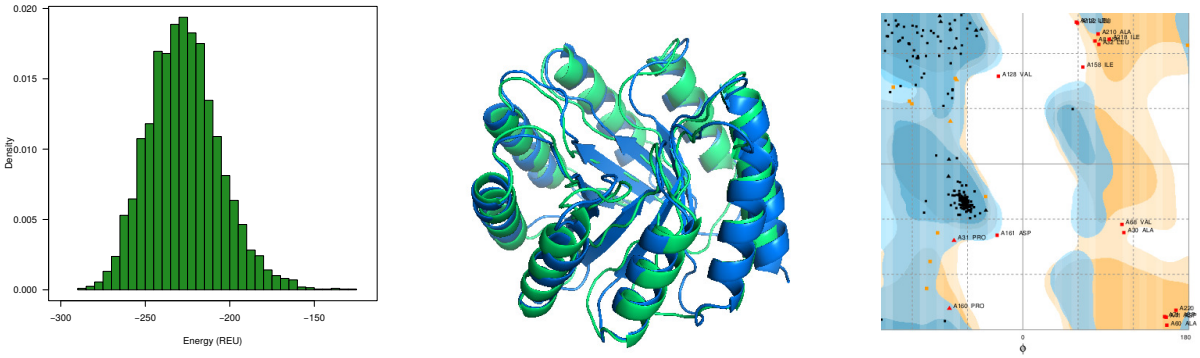


Figure 2.12: Loop Closure

(**a**) Example of a single loop before and (**b**) after the closure of the loop; (**c**) comparison of the whole model before (green) and after (blue) the closure and (**d**) Ramachandran plot of one of the scaffolds.

2.2.4 Energy minimization

Energy minimization (or relax) is an important step in protein modelling, especially for structures designed from scratch. Bias such as overlapping atoms or empty spaces in the models can be easily removed during the relaxation process, promoting the formation of interactions like H-bonds, salt bridges and Van der Waals clusters. Energy minimization was performed with the Relax package of Rosetta and details are reported in Section 3.2.4, page 176.



(a) **(b)** **(c)**

Figure 2.13: Energy Minimization

(**a**) Energy distribution of the 3968 models and (**b**) comparison between one model before (green) and after (blue) the relaxation process and (**c**) its Ramachandran plot.

Following energy minimization, all the models improved their Rosetta energy score by 7000 REU, reaching negative values in the range of -125 to -290 REU. Their distribution is shown in Figure 2.13a. The values are similar to those of small natural proteins, but remain above to those expected for a protein of 240 aa (in the range of -384 to -492 REU, as mentioned in Section 2.1.5, page 38). The score will be further improved by the sequence design step in Section 2.2.6, page 52.

The minimization process adjusts both the side-chain and the backbone atoms, and some of the conformational adjustments are visible in the main-chain (Figure 2.13b). The Ramachandran plot after minimization is more natural-like, with dispersed ϕ and ψ angles in α- and β-regions (Figure 2.13c).

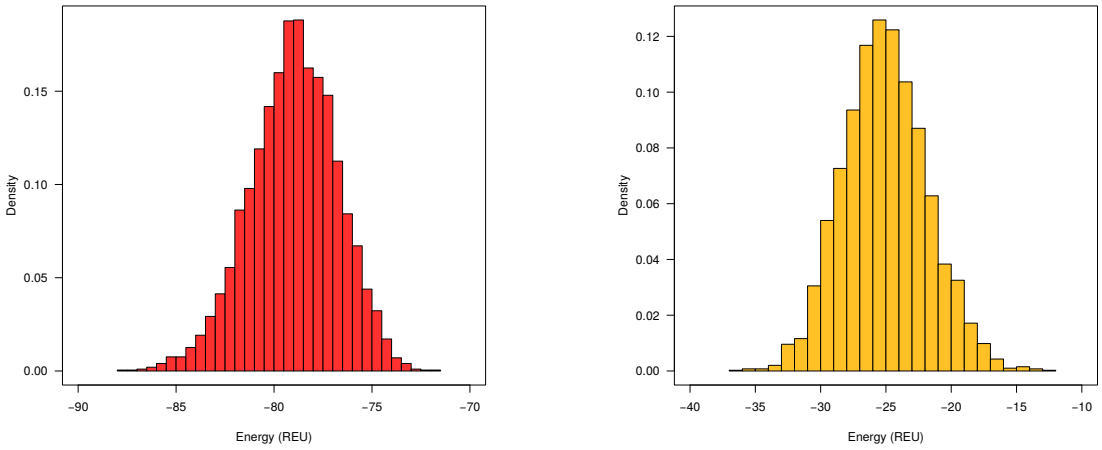
2.2.5 Backbone selection

The design of backbone structures that mimic the TIM-barrel-fold was performed with the programs Rosetta and Modeller and resulted in the creation of ∼4000 different scaffolds. Selection is necessary to find the best models for the following step of sequence

design. The selection was made according to structural features only (i.e. sequence information is not taken into account), in order to find models with the most stable α-helix and β-strand content.

The scoring function of Rosetta was used to rank the models on the basis of main-chain H-bonds values, for α-helices and β-strands individually. Bonds due to side-chain/side-chain or side-chain/main-chain interactions were not taken in account and the models were ranked according to the stability of their secondary structure elements only. For example, a model with an α-score of -35 REU contains more H-bonds in its helices compared to a model with score -15 REU, leading to a higher stability in the secondary structure elements.

The distribution of the α-backbone energy scores is shown in Figure 2.14, while the distribution of the β-backbone energy scores is shown in Figure 2.14b.



(a) **(b)**

Figure 2.14: Backbone energy scores

(**a**) Contribution to the energy score of the secondary structures for the 3968 scaffold for the helix (red) and (**b**) for the strand (yellow).

From the initial 3968 structures, the best 10% were selected according to their α-helix score. Of those 396, the best 10% were selected according to the β-strand rank. 11 out of the 39 best models were discarded because of their high sequence similarity, and the 28 remaining are shown in Figure 2.15. Although the overall shape is the same, in each model there are small differences that are due to the energy minimization step associated to different amino acid sequences. Differences are quite obvious at the level of the β-ring: some are perfectly spherical, others are more ovoid and others ones have a more triangular shape.

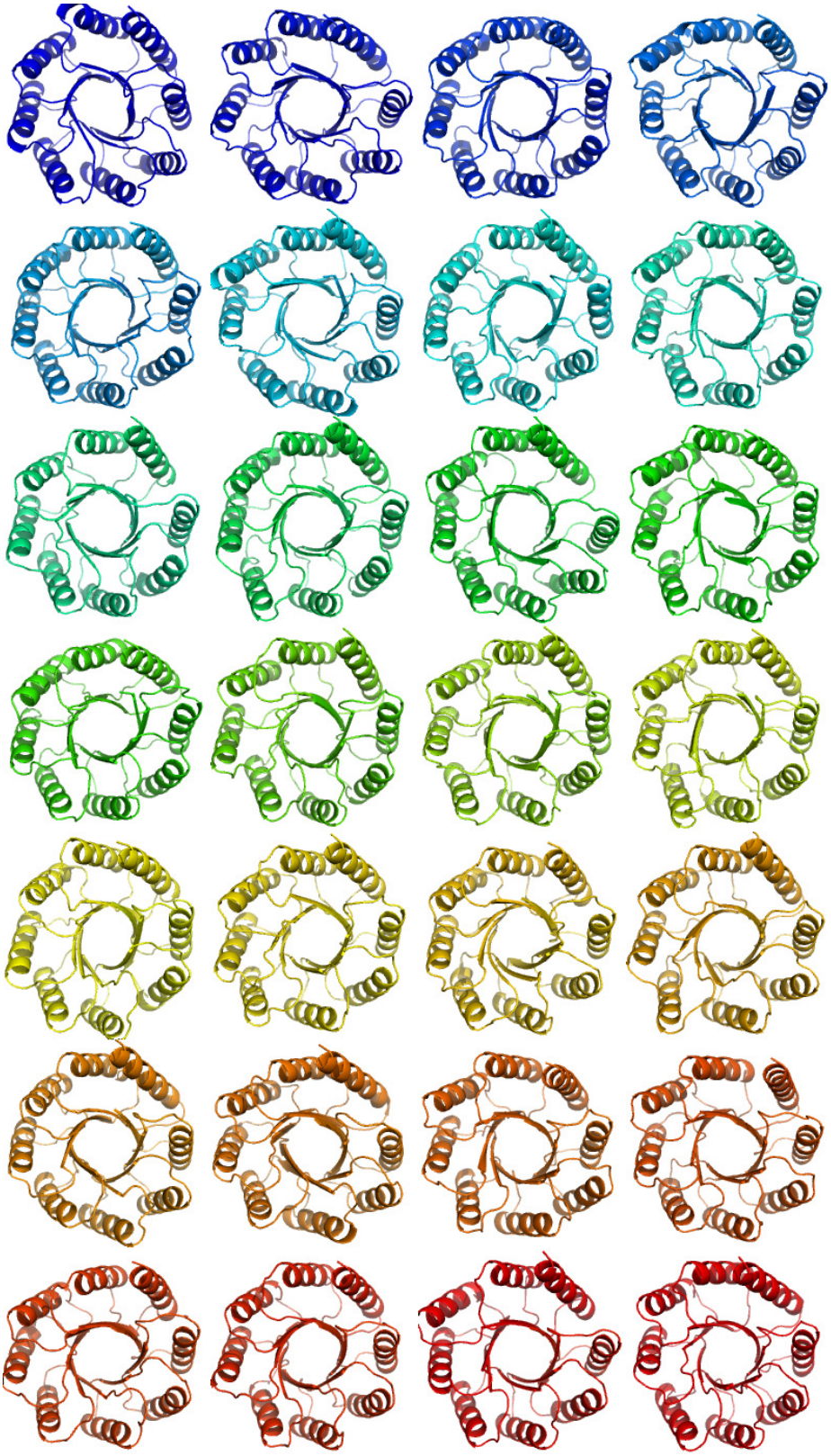


Figure 2.15: 28 selected backbone structures

2.2.6 Sequence design

The 28 selected backbone structures were subjected to 10 cycles of sequence design and energy minimization, using Rosetta. For a full description of the methods see Sections 3.2.5 and 6.5.7. In Figures 2.16 are reported the RMSD versus Total score for each of the designed sequences for cycle 1 (a), cycles 2-4 (b), cycles 5-7 (c) and cycles 8-10 (d).

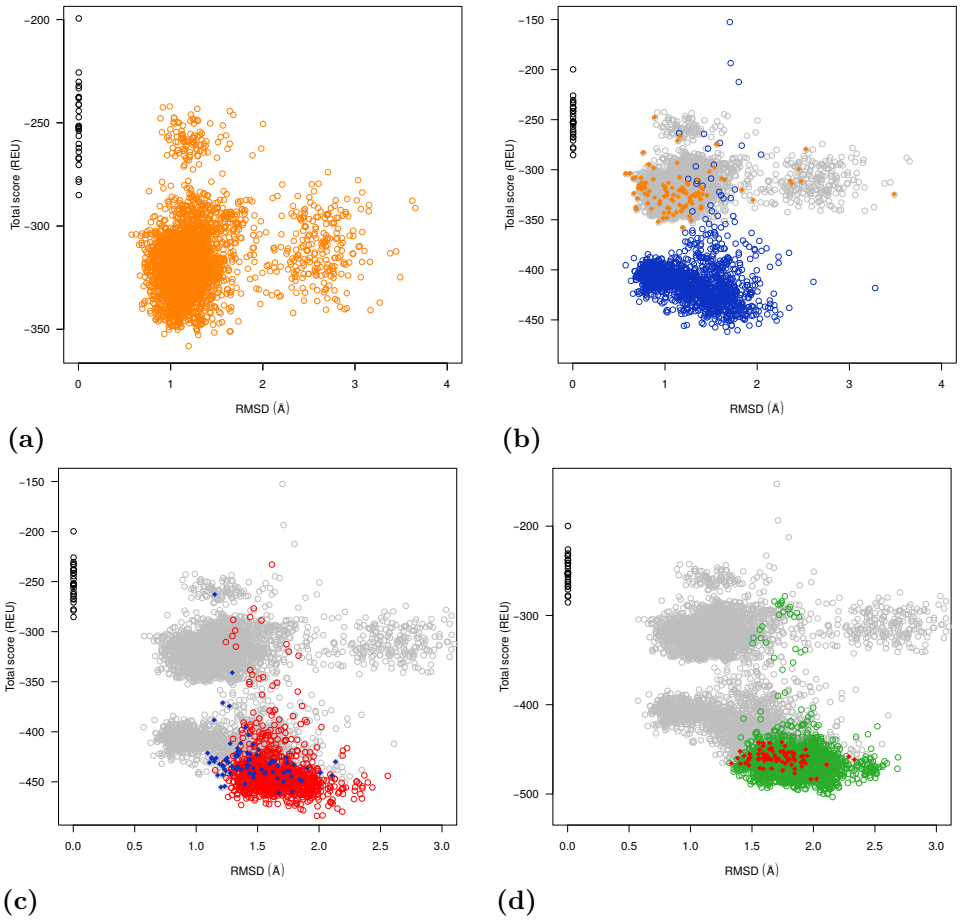


Figure 2.16: RMSD vs Energy in the sequence design cycles

(**a**) Initial 28 backbone structures (black) and output models (orange) of the first cycle of sequence design. (**b**) Input (orange) and output models (blue) for the 2^{nd} to 4^{th} cycles. (**c**) Input (blue) and output models (red) for the 5^{nd} to 7^{th} cycles. (**d**) Input (red) and output models (green) for the 8^{nd} to 10^{th} cycles. Gray circles are the models of the previous steps of design.

The 28 initial structures are used for comparison to calculate the RMSD of the designed models, and they are represented in Figure 2.16a as black circles, with RMSD equals to 0, and in Figure 2.17 as gray bars. Their energy range is between -199 and -284 REU, with the most populated value at -250 REU.

The first cycle targeted 96 aa belonging to the hydrophobic core of the fold in order to

improve the packing of the structure and increase the overall stability. The resulting 2800 outputs are represented in orange circles (Figure 2.16a) and bars (Figure 2.17), and show an improvement in the total energy, with a range between -357 and -241 REU and the more populated value at -310 REU. The RMSD is acceptable in the majority of the models with a minimum of 0.57 $\mathring{A}$ of C_α -deviation, due mainly to the side-chain substitution and backbone rearrangement. There are few outliers with RMSD above 4 $\mathring{A}$, up to a maximum of 15.68 $\mathring{A}$ (not shown in Figure 2.16a).

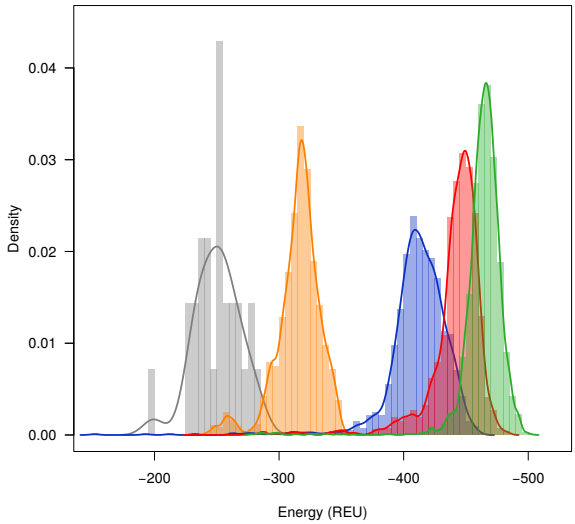


Figure 2.17: Total energy profiles of the sequence design

Total energy for the initial 28 backbone structures (gray) and for the output models sequence design: 1^{st} cycle in orange, 2^{nd} to 4^{th} cycles in blue, 5^{nd} to 7^{th} cycles in red and 8^{nd} to 10^{th} cycles in green.

The 2^{nd} to 4^{th} cycles targeted all the 240 aa of the artificial TIM-barrels, with restrictions that depended on the layer of the protein, apolar residues for the core, polar for the surface and all for boundary and loops. The 1060 outputs are shown as blue circles in Figure 2.16b and blue bars in Figure 2.17. The RMSD range from 0.57 to 3.27 $\mathring{A}$. The total energy improved in the majority of the models on average by ∼120 REU, but in few cases the total energy is worst than the score of the 28 not-optimized initial structures. The energy score range from -151 to -461 REU and the most populated value is at -400 REU.

For the majority of the models, the theoretic energy range per residue (see Section 2.1.5) is reached: for a protein of 240 residues the total energy should be between -492 and -384 REU.

The 5^{th} to 7^{th} cycles targeted all the 240 residues as described in the previous cycle. The 1070 outputs are shown as red circles in Figure 2.16c and red bars in Figure 2.17. The RMSD is in the range of 1.15-2.55 $\mathring{A}$. The total energy values slightly improved of

∼20 REU, ranging from -483 to -232 REU and the most populated value is -440 REU, showing that the major improvements in the packing are been reached in cycles 2^{nd} to 4^{th}. Now on the sequence design cycles will have as objective to increase the sequences variability.

The 8^{th} to 10^{th} cycles targeted all the 240 aa and there are no restrictions: any residue can change in any other, in order to mimic natural TIM-barrel in which, for example, polar residues are found in the internal β-ring (see Section 1.2, page 11). The 3217 outputs are shown as green circles in Figure 2.16c and green bars in Figure 2.17. The RMSD range from 1.31 to 2.86 $\AA$. The total energy values slightly improved of ∼20 REU, with a overall range of -277 to -502 REU, with the most populated value at -465 REU.

In Table 2.3 are reported all the steps of sequence design based on the initial family, with the number of selected inputs and of not-redundant outputs.

	1^{st} cycle		**2-3-4^{th} cycles**		**5-6-7^{th} cycles**		**8-9-10^{th} cycles**	
Family	**Inputs**	**Outputs**	**Inputs**	**Outputs**	**Inputs**	**Outputs**	**Inputs**	**Outputs**
1	1	100	4	40	4	40	2	69
2	1	100	4	40	4	40	3	130
3	1	100	4	40	4	40	2	134
4	1	100	3	30	2	20	2	83
5	1	100	4	40	3	30	4	141
6	1	100	4	40	2	20	2	65
7	1	100	4	40	3	30	1	65
8	1	100	4	40	3	30	1	56
9	1	100	4	40	4	40	4	113
10	1	100	3	30	3	30	4	142
11	1	100	4	40	2	20	2	76
12	1	100	4	40	4	40	2	88
13	1	100	4	40	3	30	2	79
14	1	100	4	40	4	40	4	126
15	1	100	4	40	3	30	1	58
16	1	100	4	40	3	30	3	134
17	1	100	4	40	3	30	3	139
18	1	100	3	30	3	30	4	140
19	1	100	3	30	4	40	3	140
20	1	100	4	40	3	30	3	101
21	1	100	4	40	3	30	3	134
22	1	100	4	40	3	30	4	129
23	1	100	4	40	3	30	3	142
24	1	100	4	40	4	40	4	134
25	1	100	4	40	2	20	3	151
26	1	100	4	40	2	20	4	145
27	1	100	3	30	4	40	3	138
28	1	100	3	40	3	40	4	146
Total	**28**	**2800**	**106**	**1060**	**88**	**880**	**80**	**3198**

Table 2.3: Resume of the sequence design

2.3 *In silico* validation

This chapter describes six methods used to select 10 models for experimental validation. The selection criteria are first tested on our collection of 219 natural TIM-barrels (see Section 2.1.1, page 33) in order to extract cut-off values for the selection of artificial TIM-barrels. Selection methods include the comparison of energy values, amino acid compositions, secondary structure predictions and molecular dynamics simulations. The 10 models chosen for experimental validation are shown and described at the end of the chapter.

2.3.1 Models selection

The previous chapter describes the creation of 7949 artificial and non-redundant sequences modeled on the TIM-barrel fold (Section 2.2.6, page 52). The 2800 outputs of the first cycle of sequence design (in orange in Figure 2.17), are not considered for experimental validation since their energy score is lower than -387 REU (the minimum energy score described in Section 2.1.5 for a protein of 240 aa, page 38). The remaining 5149 models, distributed among the 28 Families, are all included. The 28 Families are not equally represented, with a minimum number of 125 representatives for Family 06 and a maximum of 217 for Family 28 (see Table 2.4 in Section 2.3.8 for details, page 68).

2.3.2 Analysis of the total energy

The first step in the *in silico* validation is the analysis of the energy score. Energy values were calculated for each of the 5149 selected models with the Rosetta method used for natural TIM-barrel (see Section 3.1.3, page 170). The total energy score is directly correlated with protein length, as seen in Section 2.1.5, page 38, and a direct comparison of total energy values between artificial models and natural proteins can be misleading because of the length factor: natural proteins have different lengths while all the models are made up of 240 residues. The per-residue energy value is better suited for this analysis.

The range of per-residue energy for natural TIM-barrels is between -2.02 and -1.61 REU, with a peak centered at -1.8 REU (in yellow in Figure 2.18). The distribution of the artificial models is shown in blue, and their per-residue energy ranges from -2.09 to -0.63 REU (not shown in the picture). The peak is centered at -1.9 REU, showing that the majority of them have a lower per-residue energy than the natural proteins (and so

a higher per-residue stability). 112 models have values lower than the natural maximum of -2.09, and they are in theory more stable than natural proteins. Because there is no information about drawbacks in highly stable proteins, these models are kept in the validation process. On the contrary, models that have higher values than -1.61 REU are considered less stable. In order to be more strict in the selection of good variants, the cut-off limit is set at -1.7 REU, and 358 models are discarded. The remaining sequences are 4791, and details for each Family are reported in Table 2.4 in Section 2.3.8, page 68.

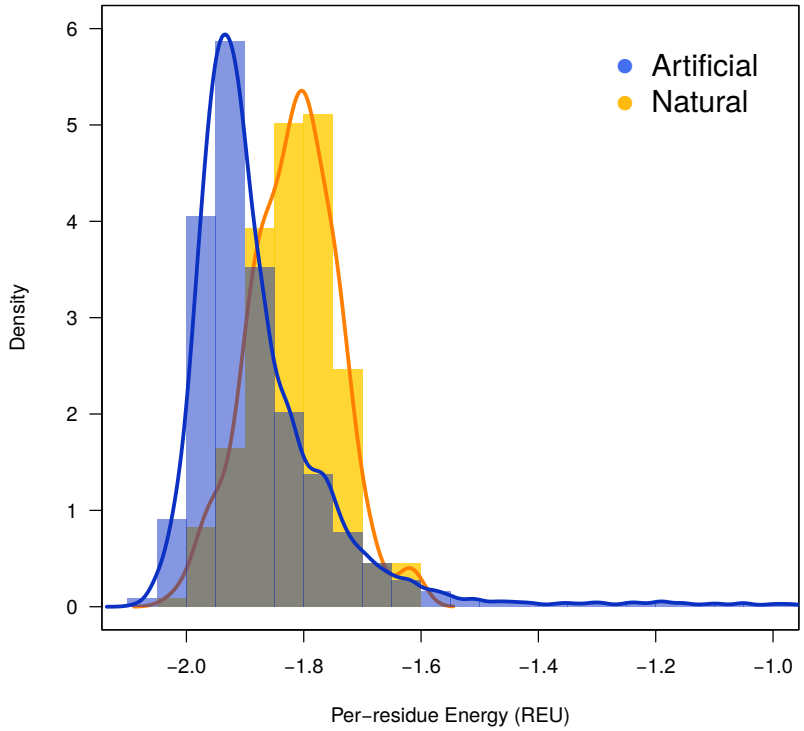


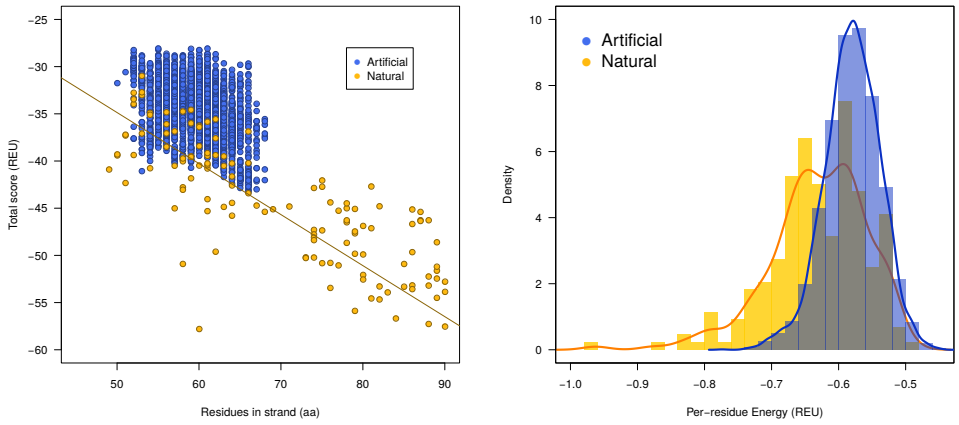
Figure 2.18: Per-residue energy scores of artificial TIM-barrels
Distributions of the per-residue energy scores for the 5149 artificial TIM-barrels (blue) and for the 219 natural ones (yellow).

2.3.3 Analysis of the β-energy

The second step in the *in silico* validation is the analysis of the energy score of the β-barrel, that takes in consideration only the H-bonds in the backbone structure of the β-strands. β-energy values were calculated for each of the 4791 remaining models with the method used for natural TIM-barrel (see Section 3.1.3, page 170).

The number of residues in β-strands versus the β-energy for both natural (yellow) and artificial (blue) TIM-barrels is shown in Figure 2.19a, and the per-residue β-energy distributions are in Figure 2.19b. For natural TIM-barrels the per-residue range of β-energies is between -0.96 and -0.49 REU, and for artificial ones it is between -0.77 and

-0.44 REU. The data show that on average artificial TIM-barrels have higher energies (and lower stability) in their β-strands compared to natural ones. This is in contrast with the total energy distribution shown in Section 2.3.2, page 55, in which the per-residue total energy is higher for the artificial group.



(a) **(b)**

Figure 2.19: Per-residue β-energy scores of artificial TIM-barrels
(**a**) Strand residues versus total β-energy for the 5149 artificial TIM-barrels (blue) and for the 219 natural ones (yellow), and (**b**) their per-residue β-energy distributions.

The results suggest that the H-bond network in the β-barrels is generally larger in natural TIM-barrels, but the H-bond network in α-helices or among side-chains is more extended in the artificial group, leading to a higher score in the total per-residue energy.

Despite that, almost all the models are included in the range of natural TIM-barrels, with only 67 outliers. As for the total energy selection, in order to be more restrictive, the cut-off value was set to -0.55 REU and a total of 1084 models are discarded. The remaining models for the next step of validation are 3707 (see Table 2.4 in Section 2.3.8 for details, page 68).

2.3.4 Analysis of amino acid composition

The third step in the *in silico* validation is the analysis of the amino acid composition. The 20 natural amino acids were counted in each of the 3707 models and their percentage was compared with the ranges that were obtained for natural TIM-barrels (Section 2.1.7, page 41). If all the amino acid counts are in the range, the model passes to the next step. If one or more of them is outside the range the model is discarded. Technical details for the method are described in Section 3.1.4, page 171, and the ranges of natural TIM-barrels are shown in Table 2.1 in Section 2.1.7, page 42.

This selection step discarded 2882 models, and the remaining ones are 825 (details are reported in Table 2.4 in Section 2.3.8, page 67).

The number of cysteines is out of range in all the discarded models, reaching up to 7.9% of sequence coverage instead of a maximum of 2.8%. This excess is a problem of the Rosetta software, that favors the cysteine due to its size and to its capacity to form H-bonds, but does not consider its reactivity and its danger in living cells.

The second residue out of range in 66% of the discarded models is the valine, underrepresented in all the cases with values lower than the minimal 3.1% of sequence coverage that is found in natural TIM-barrels.

The third residue out of range in 31% of the discarded models is the phenylalanine, overrepresented in all the cases with values higher than the maximal 6.6%. In the analysis of natural TIM-barrels this residue type is absent in 13 out of 219 structures, suggesting that it is not required for the TIM-barrel fold.

Methionines and glycines are both out of range in 18% of the discarded structures, mainly for overrepresentation. Lysines are overrepresented in 7% of the cases, and all the remaining residue types are out of range in less than 2% of the models, with the exception of alanines, asparagines, isoleucines and prolines that are in the correct range in all the tested models.

2.3.5 Analysis of amino acid properties

The fourth step in the *in silico* validation is the analysis of the artificial TIM-barrel composition on the basis of amino acid properties. As described in Section 2.1.8, page 43, amino acids can be grouped according to their basic properties, such as size or charge, and seven groups were selected and tested on natural TIM-barrels: polar, charged, positive, negative, small, apolar and aromatic amino acids (see Section 3.1.4 and Section 2.1.8 for details, page 171).

The same analysis was performed on the 825 remaining artificial models and 2 of them were discarded because out of range compared to natural TIM-barrels.

The discarded models belong to Family 25 and in both cases the apolar residues content is out of the maximum limit (67.5%). Details are reported in Table 2.4 in Section 2.3.8, page 68.

2.3.6 Secondary structure prediction

The fifth step in the *in silico* validation is the secondary structure prediction, in order to discard the sequences that are less-likely to fold into the correct secondary structure elements.

There are nowadays many software and web-services that can predict secondary structures starting from the amino acid sequences. They can be divided in two classes: the ones that use only the amino acid propensity and statistics of natural proteins for the prediction, and those that include also evolutionary information. In this second case, the target sequence is used on the one side for the propensity analysis and the statistic, and on the other side to find homologous sequences that have already been characterized for their secondary structure content. Results of the evolutionary predictors are highly efficient and the errors are minimal for natural proteins. However, the analysis of artificial models with this kind of predictors is affected since they lack of known homologous proteins.

Software that do not use evolutionary information are more prone to errors in the prediction of the secondary structure for natural proteins, but they are not affected by artificial sequences. For this reason, the screening of the artificial TIM-barrels is done with both kind of prediction software: SSpro [124], that is based on evolutionary information, and JPred4 [125] that is not.

In order to estimate the error of the predictions, fasta sequences of natural TIM-barrels were analyzed with both software, and the results were compared with the DSSP analysis of the corresponding pdb structures (see Section 3.3.1, page 179). All the residues with mismatching conformation are considered errors. The analysis of the error range in natural TIM-barrels will be useful to establish cut-off values for the prediction error, that is then used to screen the remaining 823 artificial sequences. A model is considered valid if it pass the cut-off value in both SSpro and JPred4 predictions.

The error distribution of the natural TIM-barrel sequences with SSpro and JPred4 are shown in Figures 2.20a and 2.20b, respectively.

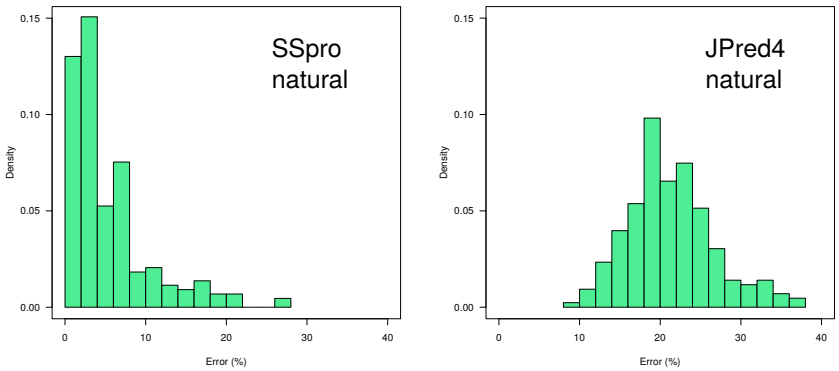
The error distributions of the two software are different: SSpro has a range between 0 and 28% of mismatching residues, with more than 90% of correct predictions in the majority of the proteins. The error range of JPred4 is between 8 and 37%, and the majority of the proteins have around 75% of correct predictions. This difference between SSpro and JPred4 shows the importance of evolutionary information in the quality of the prediction.

The cut-off value chosen for the SSpro predictor is 28%, its maximum value for natural TIM-barrels. The lack of evolutionary information in the analysis of artificial TIM-barrels will increase the error rate of the prediction, and a more restrictive cut-off value can be harmful, discarding potential good sequences. On the contrary, JPred4 is not influenced by artificial sequences and the chosen cut-off value is 25%, in order to be more restrictive

in the selection.

The results of the analysis of the artificial TIM-barrel structures are shown in Figure 2.21a for SSpro and in Figure 2.21b for JPred4. The SSpro predictions for the artificial sequences have a minimum of 10% and a maximum of 42% of mismatching residues. In comparison with the range 0-28% of the natural proteins, the importance of evolutionary information is evident. For the majority of the models, the error range in the artificial TIM-barrels prediction by JPred4 is comparable to the natural one; only 9 out of them have more than 37% of error.

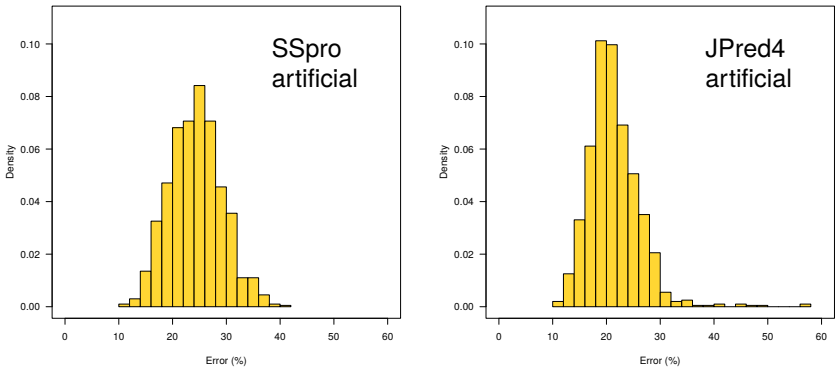
The cut-off values chosen for SSpro and JPred4 after the analysis of natural TIM-barrels are 28% and 25% respectively. The models that have higher values than the cut-off are 163 for SSpro and 116 for JPred4, with a total of 225 discarded and 598



(a) **(b)**

Figure 2.20: Secondary structure prediction of natural TIM-barrels

(**a**) Error distributions in the secondary structure prediction calculated for natural TIM-barrels with SSpro and (**b**) Jpred4.



(a) **(b)**

Figure 2.21: Secondary structure prediction of artificial TIM-barrels

(**a**) Error distributions in the secondary structure prediction calculated for artificial TIM-barrels with SSpro and (**b**) Jpred4.

remaining sequences. Details for each Family are reported in Table 2.4 in Section 2.3.8, page 68.

2.3.7 Molecular dynamics

In the fields of molecular biology and biochemistry, molecular dynamic (MD) is widely used to simulate the stability overtime of given macromolecules (proteins, DNA, RNA and lipids), their structural, conformational and functional characteristics, their interactions with solvent, ligands or other macromolecules and, in the case of proteins, their folding.

MD simulates atom motion of a system according to the Newton's second law: $F = ma$, in which F is the force, m the mass and a the acceleration. Because the force is correlated to the potential energy and the acceleration to the velocity, it is possible to calculate and simulate the trajectories that each atom of the system takes in a defined period of time.

The full-atom MD simulation of a protein is very accurate but it is slow due to the high number of atoms that are simulated. For example, a protein of 250 amino acids contains around 4000 atoms. A simulation of 50 ns requires 25000000 steps. This means that the number of trajectories that are calculated is 4000 atoms x 25000000 steps = 10^{11} trajectories. From the trajectories and the protein structure then we can obtain the 3D coordinates, the energies, the RMSD and so on. All these features however have a cost in term of calculation time: MD simulations are extremely accurate but very slow.

Moreover, in this example we did not mention that MD simulations can be done in explicit-solvent conditions. In this kind of simulations, water molecules (and optionally buffer, salt, detergent) are added around the protein to form a shell and both are simulated together. The simulation becomes more realistic because proteins are usually in solution, and more accurate since the effect of the solvent on the protein is taken into account. The advantages are enormous, however the simulation takes longer due to the presence of around 8000-9000 extra particles in the system. A 50 ns simulation for a protein of 250 aa in explicit solvent conditions (a total of 10000-11000 atoms), takes 5 days using the cluster VEGA of the Belgian Consortium des Équipements de Calcul Intensif (CECI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11.

Because of MD is time-consuming, it is not possible to test all the remaining 598 models. One model of each of the 28 Families was chosen as representative in order to obtain information on their stability and quality of the design. None of the models of

Family 11 passed the secondary structure prediction step, and the MD simulation was done on the best model of the previous step to evaluate its overall behavior. Family 11 will not be considered for experimental validation.

Additionally, four control proteins were chosen for the comparison with the artificial sequences: three are the positive ones and the remaining is the negative one. Among the positive control there are two natural TIM-barrel (1K77 and 4AAJ, of 260 and 200 aa, respectively), and sTIM-11, the artificial four-fold symmetrical TIM-barrel of 184 residues described in Section 1.4.2, page 30. The artificial OctaV (designed as TIM-barrel and folded in an $\alpha\beta\alpha$-sandwich, see details Section 1.3.3, page 22, and 1.3.5, page 24), is the negative control.

The 32 structures were simulated for 50 ns in explicit solvent with the GROMACS package [110], using the force-field is AMBER99SB [126] and the water-type is TIP3P [127]. AMBER99SB is one of the most common force-field for the simulation of soluble globular proteins. Moreover, it resulted the best one (among 10) in replicating the dynamic results obtained by NMR experiments for two globular proteins, ubiquitine and gb3 domain of protein G [128]. The choice of force-field and water-type was made based on the Details on the preparation of the structures, addition of water and ions, energy minimization and equilibration are described in Section 3.3.2, page 180.

The dynamics of 50 ns are shown in Figure 2.22a in green for the positive controls (1K77, 4AAJ and sTIM-11) and in red for the negative control (OctaV). RMSD distributions are shown in Figures 2.22b with the same color code.

The natural TIM-barrel 4AAJ (light green) is the most stable protein among the 4, with a RMSD distribution between 0.0765 and 0.1347 nm and the peak centered at 0.10 nm. 1K77 (dark green) has a RMSD distribution between 0.0828 and 0.1522 nm and the peak centered at 0.13 nm. Both proteins can be considered stable and in a rigid conformation. The artificial sTIM-11 (lime-green) is less stable over time compared to the two natural TIM-barrel, with a RMSD distribution between 0.0627 and 0.1883 nm and a peak centered at 0.17 nm, but there are not significant RMSD jumps during the course of the simulation.

The artificial model of OctaV (red) behaves differently from the positive controls. Already in the first nanoseconds, the RMSD jumps to values higher than 0.3 nm and the distribution ranges from 0.0939 to 0.4983 nm, with the peak centered at 0.42 nm. The model of the OctaV is known to be wrong, and the MD simulation is obviously able to predict it here. Note that it is not possible to follow conformational changes in 50 ns of simulation because theese take place on a μs timescale of. Drastic structural rearrangements in the first nanoseconds of the simulation could be an indication of the low

quality of the tested models.

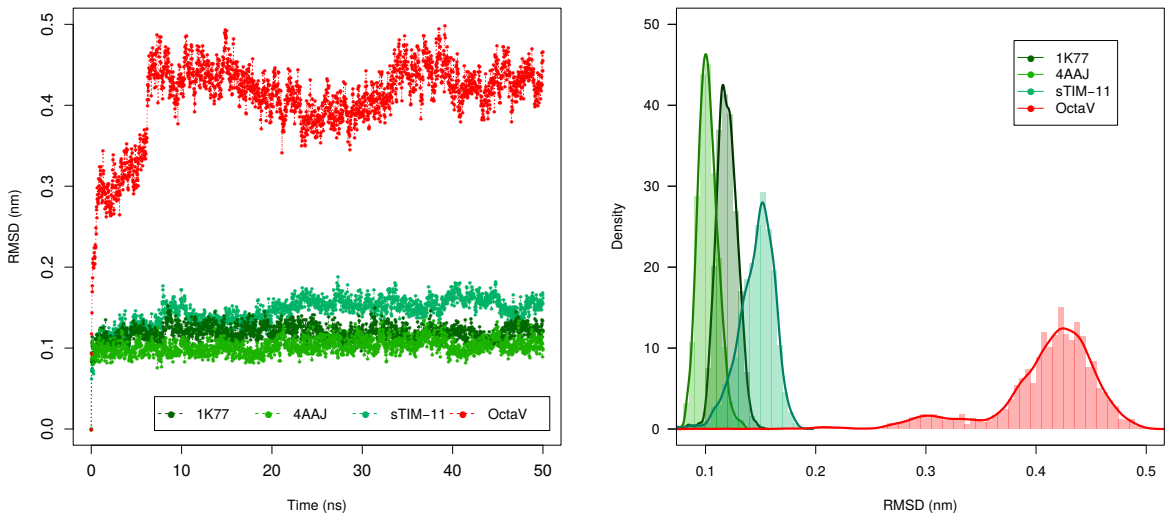
The RMSD vs Time and the RMSD distributions plots of the simulations of the 28 representatives are shown in Figures 2.23 to 2.50.
None of the tested models behave as well as the positive controls, but all of them perform better than OctaV, the negative control. The best Families are 06, 12, 16 and 26, with sharp distributions and the average RMSD lower than 0.2 nm. Distributions that are wider or that present multiple peaks are in general less preferable than single sharp peaks, because they reveal alternative protein conformations. One extreme example is Family 28, which shows a little shoulder below 0.2 nm, a peak centered at 0.22 nm and a second peak at 0.32 nm (Figure 2.50).

After analysis of the 28 simulations, 14 Families are discarded:

1. Families 19, 20, 27 and 03 are the worst behaving, with respectively 73.2%, 68.3%, 55.6%, 50.6% of RMSD values higher than 0.3 nm.

2. Families 01, 02, 14, 15, 18, 21 and 28 display multiple wider peaks in the region of 0.2 to 0.3 nm of RMSD, suggesting that they are having small re-arrangements and that they are not stable.

3. Families 10, 22 and 24 present a plateau before the peak that indicates a slow but constant re-arrangement of the structure in the first nanoseconds of simulation.



(a) **(b)**

Figure 2.22: Molecular dynamics of the control group

(**a**) RMSD vs Time and (**b**) RMSD distributions for the controls 1K77 (dark green), 4AAJ (light green), sTIM-11 (lime-green) and OctaV (red).

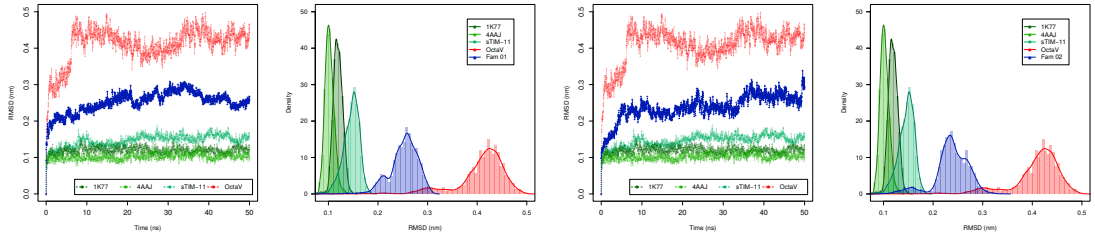


Figure 2.23: MD Family 01

Figure 2.24: MD Family 02

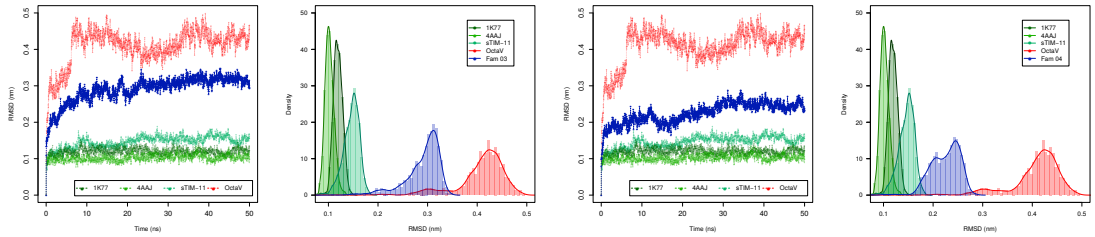


Figure 2.25: MD Family 03

Figure 2.26: MD Family 04

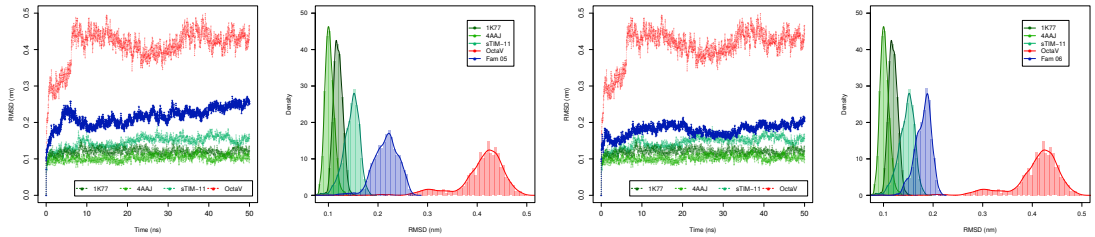


Figure 2.27: MD Family 05

Figure 2.28: MD Family 06

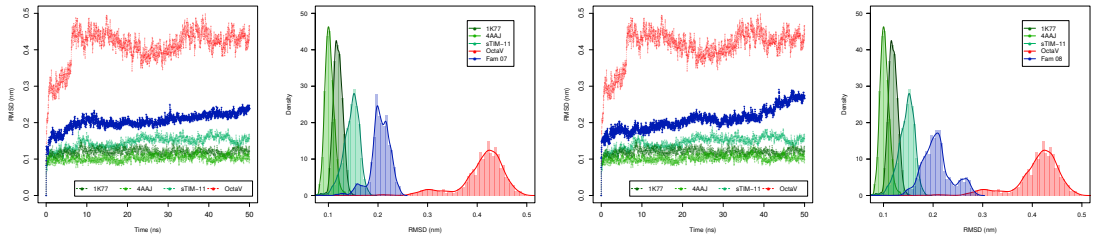


Figure 2.29: MD Family 07

Figure 2.30: MD Family 08

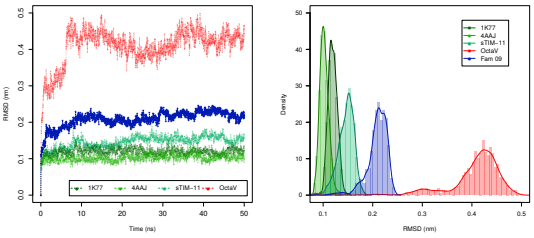


Figure 2.31: MD Family 09

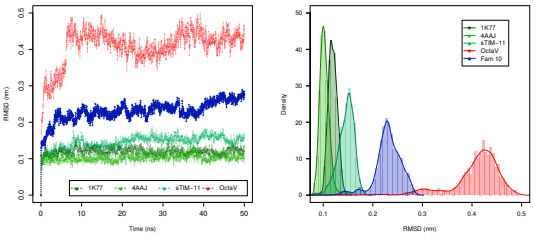


Figure 2.32: MD Family 10

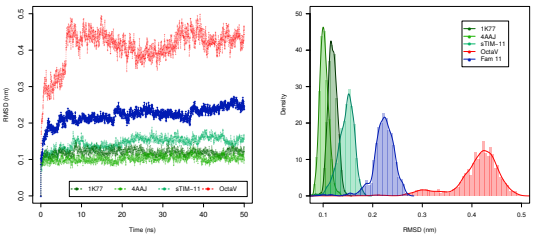


Figure 2.33: MD Family 11

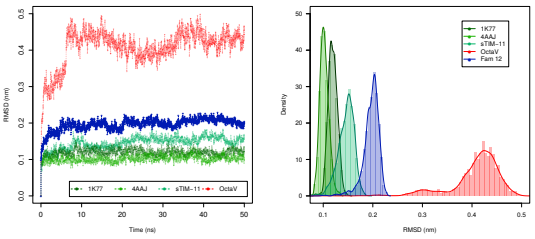


Figure 2.34: MD Family 12

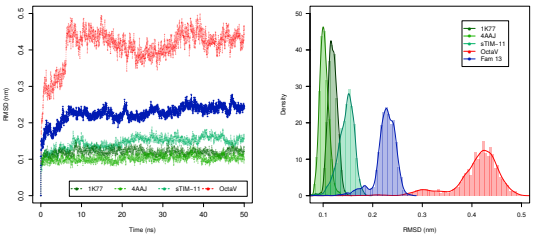


Figure 2.35: MD Family 13

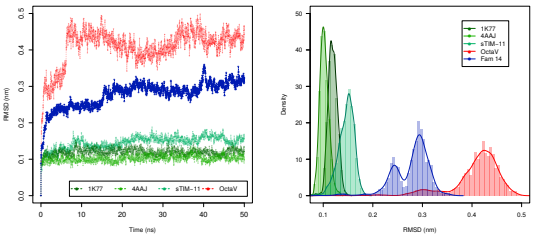


Figure 2.36: MD Family 14

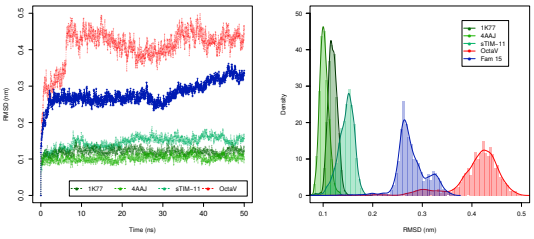


Figure 2.37: MD Family 15

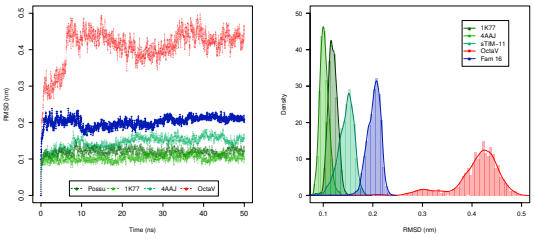


Figure 2.38: MD Family 16

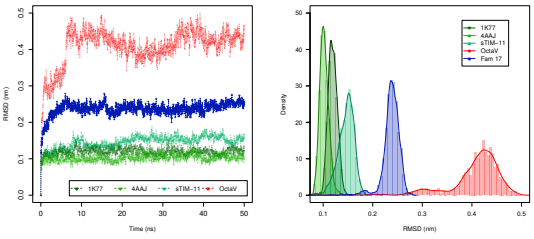


Figure 2.39: MD Family 17

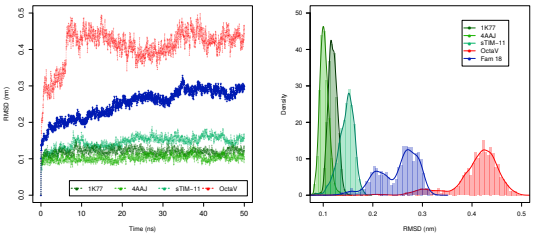


Figure 2.40: MD Family 18

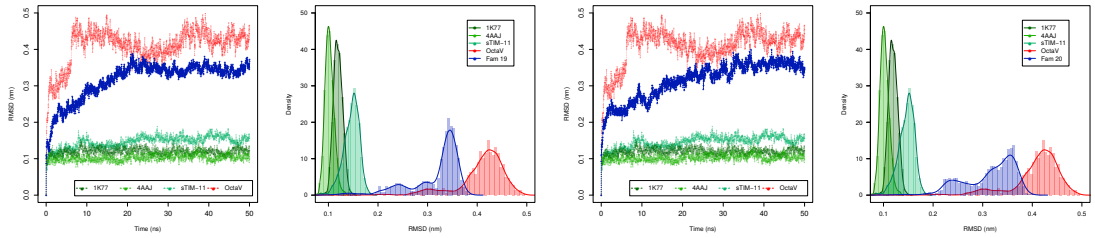


Figure 2.41: MD Family 19

Figure 2.42: MD Family 20

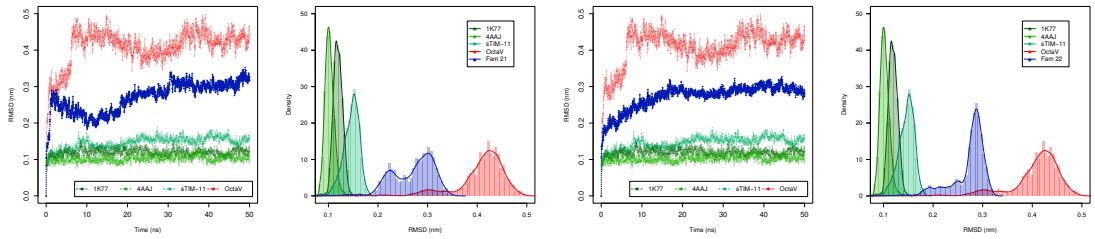


Figure 2.43: MD Family 21

Figure 2.44: MD Family 22

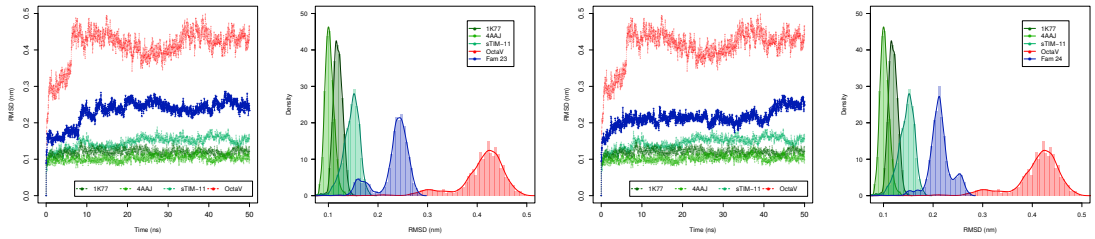


Figure 2.45: MD Family 23

Figure 2.46: MD Family 24

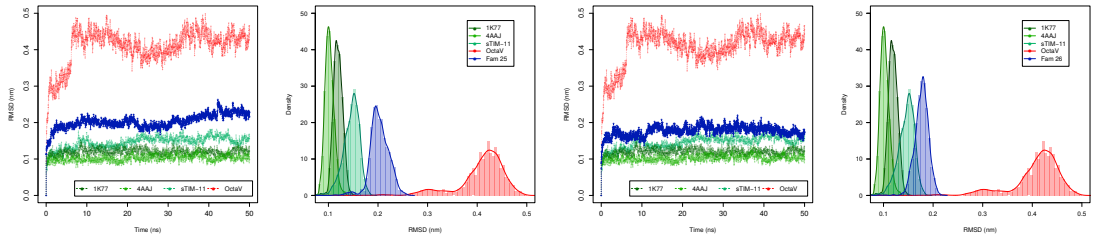


Figure 2.47: MD Family 25

Figure 2.48: MD Family 26

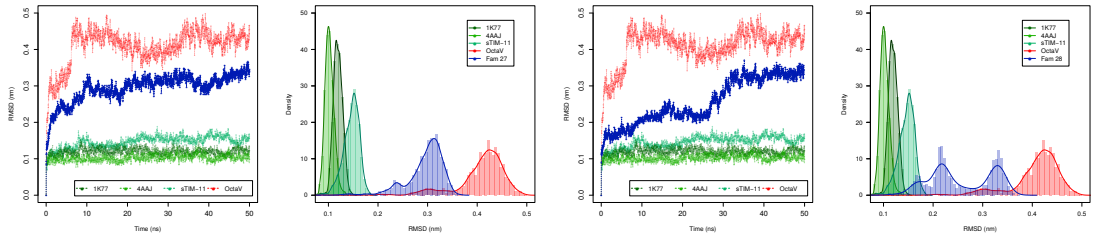


Figure 2.49: MD Family 27

Figure 2.50: MD Family 28

The number of remaining models for experimental validation is 355, and 14 Families out of 27 have been discarded. Details are reported in Table 2.4 in Section 2.3.8, page 68.

2.3.8 Summary of the *in silico* validation

A total of 5149 models, distributed among 28 Families, were selected. Each model was tested in comparison to our collection of natural TIM-barrels (described in Section 2.1.1), to identify the best models for experimental validation. The 6 steps of the *in silico* validation were:

1. Analysis of the total energy (Section 2.3.2, page 55)
2. Analysis of the β-energies (Section 2.3.3, page 56)
3. Analysis of amino acid composition (Section 2.3.4, page 57)
4. Analysis of amino acid properties (Section 2.3.5, page 58)
5. Secondary structure prediction (Section 2.3.6, page 58)
6. Molecular dynamic simulation (Section 2.3.7, page 61)

Table 2.4 summarizes the 6 validation steps for the 28 Families and give the number of successful models following each step.

The first step removed 331 models, with a minimum of 4.1% representatives in Family 28 and a maximum of 21.6% in Family 06. The second step reduced of 83% the population of Family 11, but it had no effect on Family 20 and 21. The discarded models at this step are 1084. The third step was the more efficient, with a minimum of 34.8% (Family 07) and a maximum of 99.5% (Family 28) of discarded models, a total of 2851. The forth step was the less efficient with just two models discarded, both belonging to Family 25. The fifth step removed 181 models. All the models of Families 22, 23 and 28 passed the secondary structure prediction test, whereas Family 11 was completely discarded. The sixth step excluded 14 Families and a total of 287 models.

More than 90% of the initial models were discarded after the six steps of *in silico* validation, leaving 355 sequences available for experimental validation.

Family	Models	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
01	149	127	81	6	6	5	0
02	210	196	195	36	36	32	0
03	214	210	174	36	36	32	0
04	133	123	107	14	14	9	9
05	211	202	161	32	32	20	20
06	125	98	55	4	4	3	3
07	135	117	89	58	58	40	40
08	126	104	61	8	8	7	7
09	193	176	64	4	4	3	3
10	202	197	139	18	18	16	0
11	136	124	21	2	2	0	0
12	168	160	126	18	18	13	13
13	149	136	132	36	36	31	31
14	206	192	191	55	55	10	0
15	128	115	53	2	2	1	0
16	204	188	112	52	52	35	35
17	209	206	109	42	42	34	34
18	200	182	153	36	36	24	0
19	210	198	193	33	33	28	0
20	171	169	169	22	22	20	0
21	204	186	186	30	30	19	0
22	199	181	48	5	5	5	0
23	212	207	204	44	44	44	44
24	214	200	181	13	13	7	0
25	211	200	172	87	85	71	71
26	205	199	137	78	78	45	45
27	208	190	187	53	53	27	0
28	217	208	207	1	1	1	0
Total	**5149**	**4791**	**3707**	**825**	**823**	**598**	**355**

Table 2.4: Summary of *in silico* validation

2.3.9 Model selection for experimental validation

Among the 355 models, 10 are selected for experimental validation from 10 different Families out of the remaining 13. The models are named OctaVII_01 to OctaVII_10 and they are described herefter.

OctaVII_01

The model is chosen among the 40 sequences that belong to Family 07 according to its Rosetta energy score (-473.823 REU), the best one in the group. Its β-energy per residue equals to -0.573 REU and its secondary structure prediction errors are 22.82% with JPred4 and 24.48% with SSpro. Its amino acid sequence is the following:

```
EFHIFIFGTTCNLDQYFIEAWKILMEAKDAHLGLGIQVEDQVIRYLFKKWHKLALEFELRGWISIFVYTTGDADALFREFLAFWLKVDQRCGAIALGGG
TGDLYNAVKKHLEDAKRSKAVHHALCVMLPPGPINDLFILLMILWELFRNGGGAIWIGVQSGGIKEMLELWIRIIKKGSESLLGVAVGGNSGDFDKAWE
IMLEILTKDSHANYAVGIIISNGRAKDKTKAWTLRFLKEQNS
```

The presence of aromatic residues is important for the experimental validation, and OctaVII_01 contains 9 tryptophans and 5 tyrosines. 3 cysteines are present in the structure and none of them is supposed to form a disulfide bridge.

Sequence alignment against the non-redundant protein database, performed with the program BLAST (see Section 3.3.3, page 181), produced 2 outputs: the first is a transcriptional regulator of the WhiB family [*Arthrobacter*] that has 36% of sequence identity in 29% of sequence coverage (around 25 residues out of 70) and the second is a sensory transduction histidine kinase [*Paenibacillus alvei*] that has 41% of identity in 18% of coverage (around 17 residues out of 43). Because sequence identities of the full protein are lower than 25% (respectively 10.4% and 7%), the sequence can be considered not related to natural sequences.

OctaVII_02

The second model is chosen among the 7 sequences that belong to Family 08 according to its Rosetta energy score (-464.922 REU), the best one in the group. Its β-energy per residue equals to -0.575 REU, slightly better than OctaVII_01, and its secondary structure prediction errors are 14.52% with JPred4 and 20.74% with SSpro, considerably low for an artificial sequence. Its amino acid sequence is the following:

```
KVMVVLFGKTKFAEKRFKDAMQIINDCDADGLAVMVAIFDTTGLELFKKAAELARDYSCGGMGLAMYGTQTDAKKVVAEIIKQLQNIDHDDVVCIVTGA
TDTALKIQEIAREMLEKADIRGGGLGITEQSGPLEKYARLAMANAKKFTYANFLFVIVISSGDKEDILLKLLEEWKKSGVAGGGVGIYGDDTAIVEHFK
KIVKIIAKLKCTNGIIMIMDSRGDFLELLRIFAEIAEKAQRA
```

OctaVII_02 contains 1 tryptophan, 5 tyrosines and 4 cysteines (none of them forming a disulfide bridge).

BLAST produced 2 outputs: an hypothetical protein BDEG_25319 [*Batrachochytrium dendrobatidis*], that has 43% of identity in 19% of sequence coverage (around 20 residues out of 45) and the chaperonin GroEL [*Methanosarcina mazei*], that has 42% of identity in 23% of coverage (around 23 residues out of 55). Because sequence identities of the full protein are lower than 25% (respectively 8.3% and 9.6%), the sequence can be considered not related to natural sequences.

OctaVII_03

The third model is chosen among the 35 sequences that belong to Family 16 according to its Rosetta energy score (-499.38 REU), the best one in the group. Its β-energy per residue equals to -0.592 REU, better than the previous ones, and its secondary structure prediction errors are 20.33% with JPred4 and 21.16% with SSpro. It is also one of the

best models according to molecular dynamics (Family 16). Its amino acid sequence is the following:

```
ILFIAFSCETTDNEKAFELAVKLVLDEQMEHIGIQVGGPGGPLEEAAAKFIKKMQLAKTSGQGFIVNFTSRDGNDWFEKARKLHQKSCHDNMVFLISVT
HTEALDLLEAWLRKLQKDKSSWMGVLFNHGNRNVEKAYQIAAEIFKKVLSLYCVWAILMTQGDIRDLAAKWAEQAANIKIWGCEIYLYNTNGDLEAIAR
ELAKIAKKYTGTFCGFGVVGAGEDLYNLNARLIKAAKEEMLN
```

OctaVII_03 contains 6 tryptophan, 6 tyrosines and 5 cysteines (none of them forming a disulfide bridge).

BLAST produced 2 outputs: a diguanylate cyclase [*Marinobacter persicus*] that has 44% of identity in 15% of sequence coverage (around 15 residues out of 36) and an hypothetical protein MYCFIDRAFT_41956, partial [*Pseudocercospora fijiensis CIRAD86*] that has 28% of identity in 68% of coverage (around 45 residues out of 163). Because sequence identities of the full protein are lower than 25%, (respectively 6.5% and 18.7%), the sequence can be considered not related to natural sequences.

OctaVII_04

The fourth model is chosen among the 44 sequences that belong to Family 23 according to its Rosetta energy score (-478.016 REU), the best one in the group. Its β-energy per residue equals to -0.634 REU, better than the previous ones, and its secondary structure prediction errors are 18.25% with JPred4 and 17.24% with SSpro. Its amino acid sequence is the following:

```
TIFVGLQGQETGADDKFKRIVEIVRALKSGECGVAVTLGTGDTEAQLEKWIRIAQKSECRSECIGVGGSEGDAEASWRKAAELHNKCDNSDSMLYSIAS
GSNKQEMFDRHLKAAEEHSKTLIAFFFEHDDTRADDKWLEFFLKLLNSSGGVIFTGIVASRGDVKNALHKWLEIAMKQKQGGWGVGINVSGDPVEEWWK
LILKFIKKYCGEQCAIFIVGTGSKMEKLLEKFAKELEKLLQA
```

OctaVII_04 contains 7 tryptophan, 2 tyrosines and 6 cysteines (none of them forming a disulfide bridge).

BLAST produced 1 output: a quinone-dependent dihydroorotate dehydrogenase [*Legionella moravica*] that has 31% of identity in 47% of sequence coverage (around 37 residues out of 112). Because sequence identity of the full protein is lower than 25%, (15.4%), the sequence can be considered not related to natural sequences.

OctaVII_05

The fifth model is chosen among the 45 sequences that belong to Family 26 according to its Rosetta energy score (-488.990 REU), the best one in the group. Its β-energy per residue equals to -0.586 REU and its secondary structure prediction errors are 18.25% with JPred4 and 22.40% with SSpro. It is also one of the best models according to molecular dynamics (Family 26). Its amino acid sequence is the following:

EGGIGFSGTGTANEKEWEKAREAVRKIDHEELFLIFIGCTTAERDEFKKFAEKAYKADIASFILAVGGHGTERKNYIEIALQIYLNLSVASNGWMVVGS
PDGFLDDFKWAVKRSIESDSKHLGLCLEGPNGDVEKAIREMLKMWQKASDGELMLGFVATSGNTLEILKIALEYFAKQTNHRACIFVKMSYGDIDNMAA
IIAKLINIADLGHRAEAYVGSGEYQEELLKEWIRRLKANLLK

OctaVII_05 contains 5 tryptophan, 7 tyrosines and 3 cysteines (none of them forming a disulfide bridge).

BLAST produced 1 output: a LacI family transcriptional regulator [*Celeribacter baekdonensis*] that has 51% of identity in 14% of sequence coverage (around 15 residues out of 33). Because sequence identity of the full protein is lower than 25%, (6.5%), the sequence can be considered not related to natural sequences.

OctaVII_06

The sixth model is chosen among the 3 sequences that belong to Family 06 according to its Rosetta energy score (-492.517 REU), the best one in the group. The model contains 2 cysteines in 2 different loop regions that are mutated to serines with Rosetta (see Section 3.3.4, page 182), with a new score of -496.148 REU. Its β-energy per residue equals to -0.522 REU and its secondary structure prediction errors are 18.67% with JPred4 and 21.16% with SSpro.

It is also one of the best models according to molecular dynamics (Family 06). Its amino acid sequence is the following:

TVVVLTYGHTSDFWKEMEKHLQELQKAGDAALEFGFIIHSGNLSEDLWWFVYLAKKYVTRSVALFFAGTGTKWEKEFRTALKILEMIGTTGSAFGFISG
NTVTDEWMRKAHAEFLKMREGKIHIGMEGNKGDEVELFKRALAEWLNAGKRANILFVARHKTEELKKAEEFIKMALKQQAISIALALNEDTGDALKVWA
EILKLLKKTKDGEFHALVIGTGTTAKKLLEIMRKMAIKMELG

OctaVII_06 contains 7 tryptophan and 3 tyrosines.

BLAST produced "No significant similarity found". The sequence can be considered not related to natural sequences.

OctaVII_07

The seventh model is chosen among the 13 sequences that belong to Family 12 according to its Rosetta energy score (-486.863 REU), the best one in the group. The model contains 6 cysteines in the loop region that are mutated to serines with Rosetta (see Section 3.3.4, page 182), without changes in the total score. Its β-energy per residue equals to -0.619 REU and its secondary structure prediction errors are 16.18% with JPred4 and 19.08% with SSpro.

It is also one of the best models according to molecular dynamics (Family 06). Its amino acid sequence is the following:

YFTIGHIRSTGAQDKYFAVALELILKSTGRDGAIIIIGAETKELKLAEEWMKRALKAETRITGLAIGGDTTNIDQVFLELWKIWLKITSTLSFFMIFASG
GDFKALLHKWLRLLEKWTDVDFGTGVVLTDTKESALFEEWLKELEKFQGKTGLVIIIAGDGNTRDALEEWLRKAIKASTGHLGVGIASSGKNARDYTQEA
IKLLRNTQSNDGALTVSGSDDRARDWLEIAIREAAKEALN

OctaVII_07 contains 8 tryptophan and 3 tyrosines.

BLAST produced 1 output: a hypothetical protein [*Endozoicomonas acroporae*] that has 33% of identity in 36% of sequence coverage (around 28 residues out of 86). Because sequence identity of the full protein is lower than 25%, (1.6%), the sequence can be considered not related to natural sequences.

OctaVII_08

The eighth model is chosen among the 31 sequences that belong to Family 13 according to its Rosetta energy score (-482.207 REU), the best one in the group. The model contains 6 cysteines in the loop region that are mutated to serines with Rosetta (see Section 3.3.4, page 182), with a new score of -478.687 REU. Its β-energy per residue equals to -0.606 REU and its secondary structure prediction errors are 21.57% with JPred4 and 24.89% with SSpro. Its amino acid sequence is the following:

KAAVGLAGQTKLADEIFKRIVQLLLEADTGGLGLVVAGNSKSLRDRFYEWARQAAEFKSGVINIGVDGDSGDMEAKFKEAFAIYLKFLYGALNWFLSDF
TKKQLELLRKFMELAMRADSTGMAVSFGKGTGDSDQIAKEIAKIWIDYTTQYGGLGVTYTDEHFLELLEKYLRIYAKTSTAELMIIIKTTDSDNSAIFE
TAIRILLKIVGPDLELIVIGSGDNMTKIIEKILRIAAKAEKT

OctaVII_08 contains 3 tryptophan and 8 tyrosines.

BLAST produced 1 output: a adenosine deaminase [*Streptomyces kasugaensis*] that has 49% of identity in 22% of sequence coverage (around 26 residues out of 53). Because sequence identity of the full protein is lower than 25%, (10.8%), the sequence can be considered not related to natural sequences.

OctaVII_09

The ninth model is chosen among the 9 sequences that belong to Family 04 according to its Rosetta energy score (-490.861 REU), the best one in the group. The model contains 5 cysteines in the loop region that are mutated to serines with Rosetta (see Section 3.3.4, page 182), with a new score of -486.798 REU. Its β-energy per residue equals to -0.630 REU and its secondary structure prediction errors are 19.91% with JPred4 and 22.40% with SSpro.

Its amino acid sequence is the following:

EIGVALVGGTTAMDKLFDELLKILQKLQTDSSMLGVVGFDSRDRALFQEWLKKAKKSDSALFAFGHGGSNGDEEKNYKEAQKDFLKIDAAYISNVTGSAK
GDTEKMFKEFVKIALKAEDRGAGFLFASGSGDLHRDWLEAMKEALKQTGQAIRFYISITNTQLQELLDKWFKESAKVTDGHSGVAILGYKGNQDKLFEEL
LQKIKKYAAGDGSLMIAIGGTGFRDRTAEHLKRMKKEDES

OctaVII_09 contains 3 tryptophan and 5 tyrosines. BLAST produced 1 output: a hypothetical protein [*Vitiosangium*] that has 22% of identity in 52% of sequence coverage (around 27 residues out of 124). Because sequence identity of the full protein is lower than 25%, (11.2%), the sequence can be considered not related to natural sequences.

OctaVII_10

The tenth model is chosen among the 20 sequences that belong to Family 05 according to its Rosetta energy score (-484.991 REU), the best one in the group. The model contains 6 cysteines in the loop region that are mutated to serines with Rosetta (see Section 3.3.4, page 182), with a new score of -483.617 REU. Its β-energy per residue equals to -0.576 REU and its secondary structure prediction errors are 18.67% with JPred4 and 24.48% with SSpro.

Its amino acid sequence is the following:

```
TFGIILAGTSTDREKAAKIIIELVEWANSGAWLLGFAEGGTEARDKLLEWAKEALKTTAEGFMMGFDGGSTRTLDEFEEFLKVAEWTKTSTIQVLVILRR
GESDKFAKEALRRLEESRSVGQGAGMMETGPGADEIFKKILYYAEVFTGDYFSLAIFAFSGTDAKQGEKWARLAAKSTEGVIVIGIELHSGRLRDGFHKL
RAIWEKLESRTGGLVIIGAGDDQKAEALEFLKEMEKHLKK
```

OctaVII_10 contains 6 tryptophans and 3 tyrosines.

BLAST produced 1 output: a hypothetical protein [*Vitiosangium*] that has 22% of identity in 52% of sequence coverage (around 27 residues out of 124). Because sequence identity of the full protein is lower than 25%, (11.2%), the sequence can be considered not related to natural sequences.

A summary of the characteristics of 10 models chosen for experimental validation is presented in Table 2.5, and their 3D structures is shown in Figure 2.51.

	Energy (REU)	β-energy (REU)	JPred4	SSpro	Trp	Tyr	Phe	Cys	Max ID
OctaVII_01	-473.823	-0.573	22.82 %	24.48 %	9	5	14	3	10.4 %
OctaVII_02	-464.922	-0.575	14.52 %	20.74 %	1	5	11	4	9.6 %
OctaVII_03	-499.38	-0.592	20.33 %	21.16 %	6	6	12	5	18.7 %
OctaVII_04	-478.016	-0.634	18.25 %	17.24 %	7	2	12	6	15.4 %
OctaVII_05	-488.99	-0.586	18.25 %	22.40 %	5	7	11	3	6.5 %
OctaVII_06	-496.148	-0.522	18.67 %	21.16 %	7	3	14	0	0.0 %
OctaVII_07	-486.863	-0.619	16.18 %	19.08 %	8	3	10	0	11.6 %
OctaVII_08	-478.687	-0.606	21.57 %	24.89 %	3	8	12	0	10.8 %
OctaVII_09	-486.798	-0.630	19.91 %	22.40 %	3	5	14	0	12.2 %
OctaVII_10	-483.617	-0.576	18.67 %	24.48 %	6	3	14	0	11.2 %

Table 2.5: Models for experimental validation

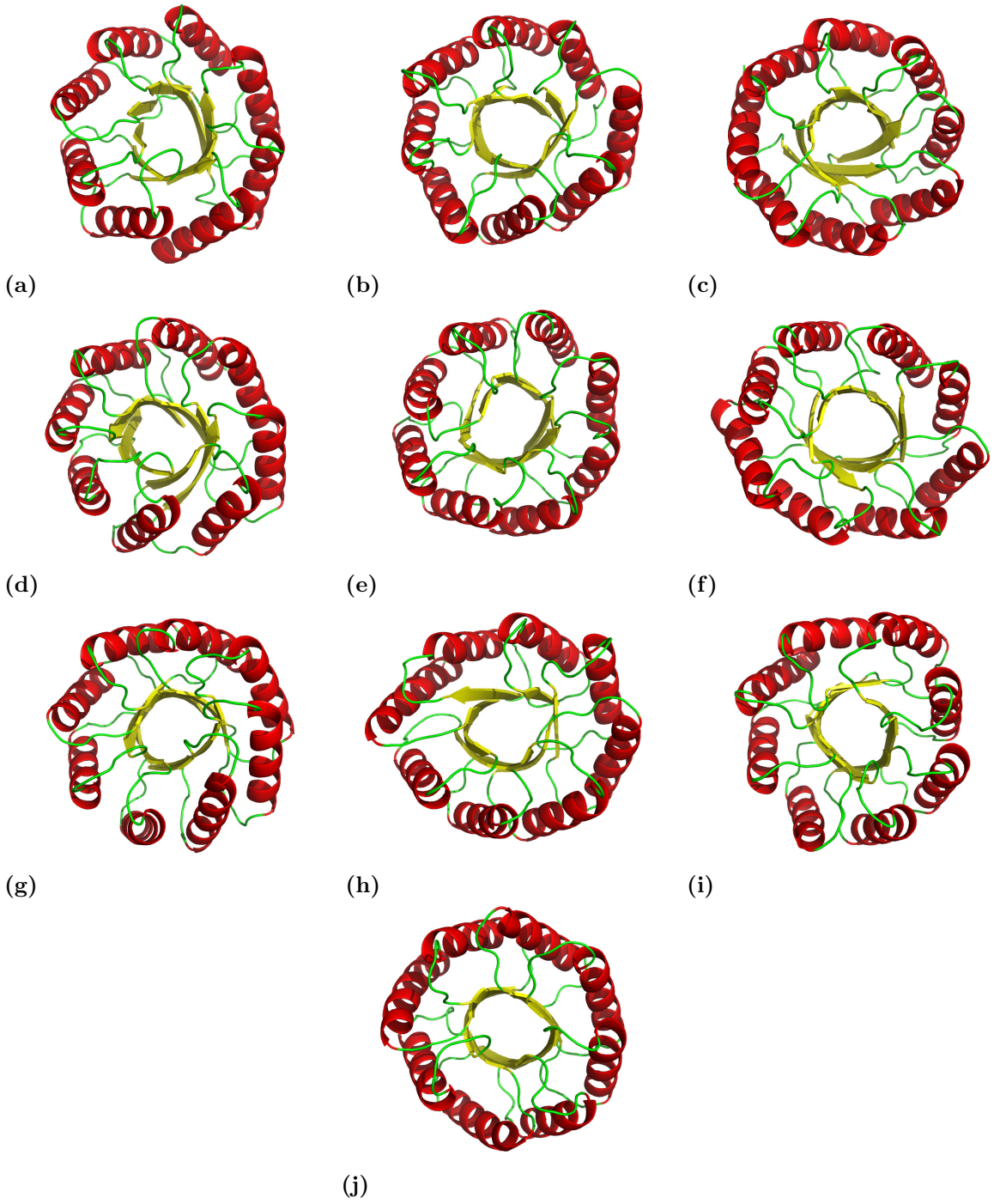


Figure 2.51: Selected OctaVIIs for experimental validation

(**a**) Models of OctaVII_01, (**b**) OctaVII_02, (**c**) OctaVII_03, (**d**) OctaVII_04, (**e**) OctaVII_05, (**f**) OctaVII_06, (**g**) OctaVII_07, (**h**) OctaVII_08, (**i**) OctaVII_09 and (**j**) OctaVII_10.

2.4 Experimental Validation

The experimental validation of the OctaVII proteins is described in this chapter. In addition to the 10 selected variants, 5 mutants have been produced and characterized. The first mutant, OctaVII_02 Y57Q, has been designed in our laboratory in order to solve a complication during the experimental validation of OctaVII_02 (see Section 2.4.2, page 83). The next three mutants, OctaVII_04 NoCys, OctaVII_04 WS and OctaVII_05 NoCys, have been designed following a collaboration with Dr. Wim Vranken at the Vrije Universiteit Brussel (VUB), Belgium. The goal of the collaboration was to find the best amino acid substitutions to improve the design according to the software developed in Vranken's lab (described in Section 2.4.6, page 99). The last mutant, OctaVII_09 WS, is the result of a collaboration with Dr. Jens Meiler at the Vanderbilt University, USA. Since software for protein modelling, including Rosetta, are constantly updated and improved, it was interesting to verify if the new version of the software could improve the OctaVII_09 design.

In the following sections we describe and discuss each of the 15 proteins individually. Some experiments have been conducted in parallel on several proteins, however, and there are cross-references and shared results. Moreover, the proteins are not produced and characterized at the same time: OctaVII_01 and OctaVII_02 are the first ones tested (June 2016) while OctaVII_09 WS is the last one (June 2018, at the really end of my PhD). For this reason the characterization of the last proteins is more poor compared to the first ones. The time-line in Figure 2.52 shows the dates when the genes were obtained after synthesis.

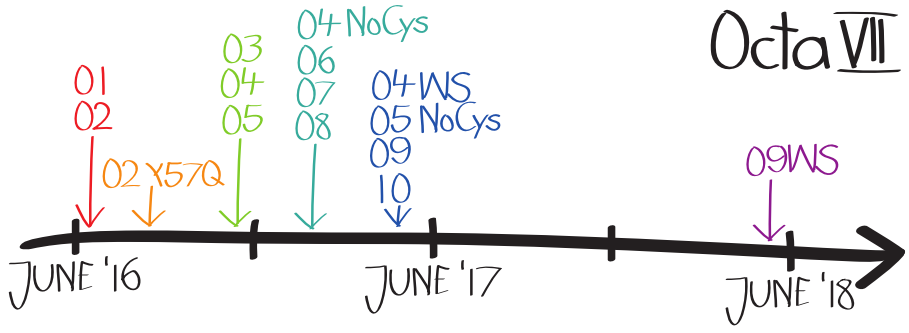


Figure 2.52: Time-line for experimental validation

The time-line shows the moment in which the genes for the different OctaVII proteins have been received at our lab. The representation was made for this thesis by Ruth Kellner.

Comparison among the OctaVIIs will be discussed in a recapitulation chapter at the end of the 15 individual sections.

2.4.1 OctaVII_01

OctaVII_01 is a 249 aa artificial protein that was chosen for experimental validation on the basis of its high Rosetta energy score among the models of Family 07 (details are given in Section 2.3.9, page 68). The gene was synthesized and inserted in the pET28a vector by the company IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet was resuspended in 40 μL of filtered mQ water and, after concentration measurement with Nanovue, a stock reserve was prepared with a final concentration of 50 ng/μL. Transformation of 50 μL of competent *E. coli* DG1 competent cells was done with 50 ng of OctaVII_01 DNA in order to replicate the plasmid for sequencing. The protocols for transformation, culture production, plasmid extraction and preparation of samples for sequencing are described in Sections 4.3.1, 4.3.2, 4.3.3 and 4.3.4, respectively, starting at page 193. Sequencing results confirmed the correct sequence of the OctaVII_01 gene (see Annex 6.4, page 241). The verified plasmids were used for a transformation in both *E. coli* BL21 (DE3) cells, to produce the protein, and DG1 cells to create as stock at -20°C.

Expression Trials

After transformation of BL21(DE3) competent cells with the OctaVII_01 plasmid, two colonies were selected from the plate and grown for 6 hours at 37°C. An aliquot was taken for the non induced (NI) sample, and 1 mM of IPTG was added to the rest of the culture. Aliquots were taken at 2 and 4 hours after induction (2h and 4h). SDS-PAGE analysis of the crude extract is shown in Figure 2.53.

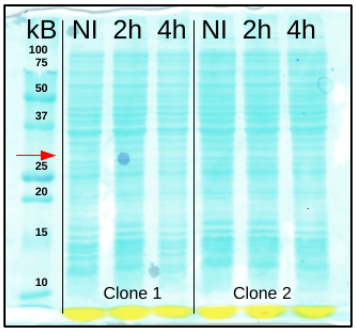
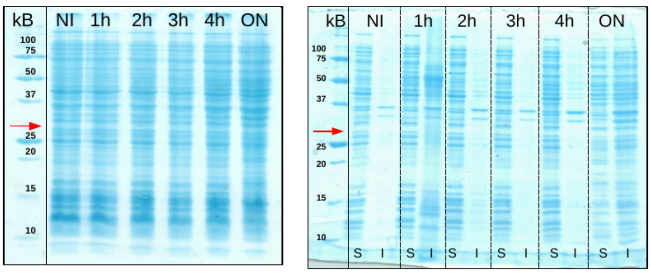


Figure 2.53: First expression trial of OctaVII_01

SDS-PAGE of the total extract for 2 different clones of OctaVII_01. Both clones have been induced at 37°C with IPTG 1 mM for 0, 2 and 4 hours (NI, 2h and 4h, respectively). The red arrow indicates the expected molecular mass for OctaVII_01 (27 kDa).

The theoretic molecular mass of OctaVII_01 is 28.3 kDa, however there are not significant bands in the induced samples, nor at that size nor in the rest of the gel.

A second trial for protein expression was done with a new colony of transformants. Induction was performed at 37°C with 1 mM IPTG for 0 (NI), 1, 2, 3, 4 hours and overnight (ON). For each condition, SDS-PAGE gels were prepared in order to show the protein expression in the total, soluble (S) and insoluble (I) fractions (Figure 2.54). As in the previous case, no evidence of expression is visible on the gel for any fraction.



(a) **(b)**

Figure 2.54: Second expression trial of OctaVII_01

(**a**) SDS-PAGE of the total fraction for 6 different conditions of induction at 37°C with IPTG 1 mM: 0, 1, 2, 3, 4 hours and overnight (NI, 1h, 2h, 3h, 4h and ON, respectively). (**b**) SDS-PAGE of the same samples after separation of the soluble fraction (S) and the insoluble fraction (I). The red arrows indicate the expected molecular mass for OctaVII_01 (27 kDa).

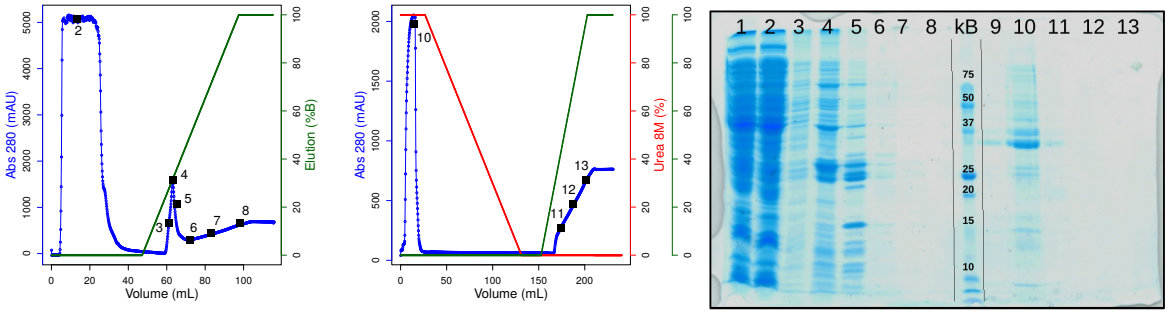
Purification

A third experiment to confirm that OctaVII_01 is not expressed in the cell is performed through purification on IMAC column. The principle at the basis of the experiment is that if OctaVII_01 is produced at minimal level (not visible in the SDS-PAGE), a large production (1 L) and purification might concentrate the protein enough to make it visible on SDS-PAGE.

The induction was done at 37°C for 4h, and cells were disrupted with 3 cycles of French Press. The soluble fraction was separated from the insoluble one through centrifugation and both fractions were loaded on a 5 mL HisTrap HP column. The purification profiles and the SDS-PAGE gels are shown in Figure 2.55. OctaVII_01 does not seem to be expressed, nor in the soluble fraction nor in the insoluble one, confirming the previous results.

Western blot

In order to detect a basal expression of the protein, a forth trial of expression was done following induction at 18°C using different concentrations of IPTG: 0, 10, 20, 50 and 100

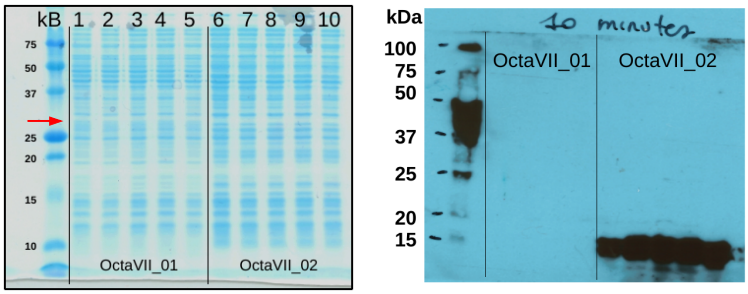


(a) (b) (c)

Figure 2.55: Purification of OctaVII_01

(**a**) Elution profiles of OctaVII_01 from the soluble fraction and (**b**) from the insoluble fraction. The blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the red line is the concentration of the denaturant buffer (urea 8 M). Black squares represent the fractions that are shown in (**c**), SDS-PAGE of the purifications from the soluble (lanes **1-8**) and the insoluble fractions (lanes **9-13**). Line **1,9** the input samples; **2,10** the flow-throughs; **3-8** and **11-13**, the elution peaks.

μM. The experiment was done in parallel with both OctaVII_01 and OctaVII_02, and the analysis was done with both SDS-PAGE (Figure 2.56a) and Western-Blot analysis using anti-HisTag antibodies (Figure 2.56b).



(a) (b)

Figure 2.56: Third expression trial of OctaVII_01

(**a**) SDS-PAGE of the total fraction for 5 different concentrations of IPTG for both OctaVII_01 (lanes **1-5**) and OctaVII_02 (lanes **6-10**). Lines **1,6**, non-induced; **2,7**, IPTG 10 μM; **3,8**, IPTG 20 μM; **4,9**, IPTG 50 μM and **5,10**, IPTG 100 μM. The red arrow indicates the expected molecular mass for OctaVII_01 (27 kDa). (**b**) Western Blot of the same samples after 10 minutes of film exposition, using anti HisTag antybodies.

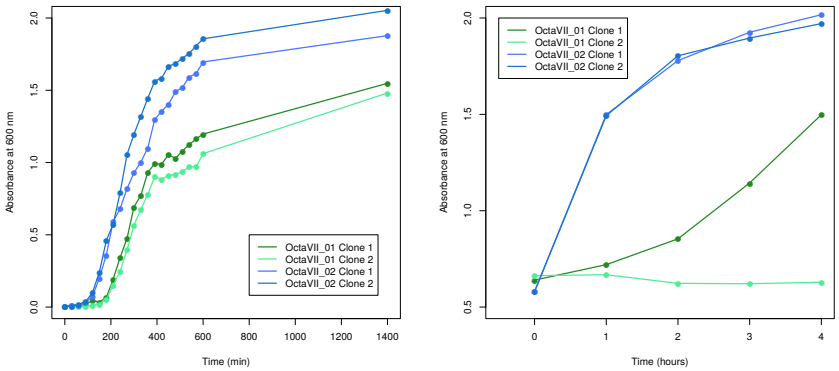
The SDS-PAGE gel with the expression of both OctaVII_01 and OctaVII_02 does not show significant bands for both proteins. However, as shown in Figure 2.56b, the antibody anti-HisTag recognizes both the tagged markers and the OctaVII_02 (its discussion will be done in the next chapter). Despite the long time of exposition there are no bands in the lanes of OctaVII_01. This experiment is a clear indication that the protein is not expressed in the cell, not even at basal level. One hypothesis that can explain the lack

of the protein in the crude extract of the cell culture is that the protein may be toxic for the cells even at low concentrations. In order to inquire this hypothesis, we analyzed the growth rate of the cells bearing the OctaVII_01 plasmid in comparison with cells bearing the OctaVII_02 plasmid.

Growth rate

A new transformation was done for both OctaVII_01 and OctaVII_02 in BL21(DE3) cells. Visual inspection of the agar plates revealed that colonies bearing the OctaVII_01 plasmid were smaller than those with the OctaVII_02 plasmid, suggesting that the protein may alter the metabolism of the cell.

For each protein, 2 colonies were selected and grown overnight in fresh LB-Kan medium. In the morning, cells were inoculated in fresh medium. To normalize the number of starting cells in each culture, the 4 pre-cultures were diluted to reach Abs_{600}=0.6 and the same volume was then injected in the fresh media. Their growth was monitored each 30 minutes through measurement of the absorbance at 600 nm, and the results are shown in Figure 2.57a.



(a) **(b)**

Figure 2.57: Growth rates for OctaVII_01 and OctaVII_02

(**a**) Time vs absorbance at 600 nm of 2 clones of OctaVII_01 (green) and 2 clones of OctaVII_02 (blue). Bacterial cells are grown in the same condition at 37°C, without induction. (**b**) Time vs absorbance at 600 nm for the same 4 clones (independent preparation) after induction with IPTG 1 mM.

The clones of OctaVII_01 have a longer lag-phase compared to the one of OctaVII_02 (around the double of time). The log-phase has the same rate for both proteins, but the deceleration phase for OctaVII_01 start around Abs_{600}=1.0, while OctaVII_02 one starts at Abs_{600}=1.5. After almost 24 hours, the absorbance of OctaVII_01 did not reach values over 1.7. This experimental data supports the hypothesis that OctaVII_01 may be partially toxic and may slow down the natural metabolism of the cell. The results were

obtained without induction of the expression of the protein, but the pET28a vector is known to be "leaky" and to allow a basal production of the target protein.

In order to verify what happens following induction, the same 4 clones were grown up in independent cultures to reach Abs_{600}=0.6, and were then induced with IPTG 1 mM. The absorbance of the induced samples was measured every hour and the result is shown in Figure 2.57b. The clones for OctaVII_02 have identical growth rates during induction, reaching Abs_{600}=2 in just 4 hours. The growth of the clones of OctaVII_01 is initially very moderated, and than vary depending on the clone: one is not growing at all during the 4h induction (expected behavior when proteins are toxic), but the other restarted to grow after around 2 hours, and reached Abs_{600}=1.5 in the remaining 2 hours.

Despite different trials, it was not possible to express the protein. The OctaVII_01 is discarded from further experimental validation.

2.4.2 OctaVII_02

OctaVII_02 has 250 residues and it is the best representative of Family 08 according to the Rosetta energy score (details in Section 2.3.9, page 68). The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

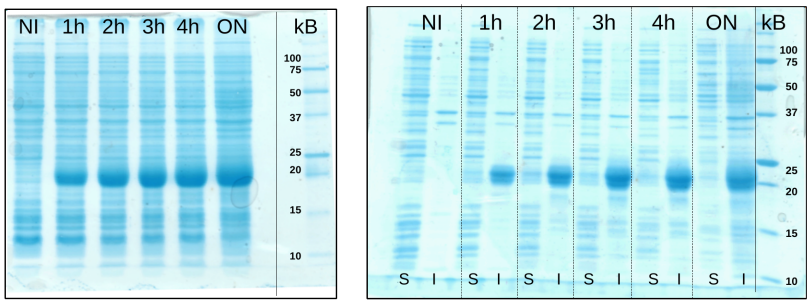
The DNA pellet of OctaVII_02 was manipulated for sequencing and storage as described for OctaVII_01 in Section 2.4.1. The results of the sequencing confirmed the correct sequence of the gene (see Annex 6.4, page 241), that was then used for a transformation in BL21 (DE3) cells for expression trials.

Expression Trials

Following transformation of *E. coli* BL21(DE3) competent cells with the OctaVII_02 plasmid, two colonies were selected from the plate and grown for 6 hours at 37°C. An aliquot was taken for the non induced (NI) sample, and 1 mM of IPTG was added to the rest of the culture. Aliquots were taken at 1, 2, 3, 4 and ∼15 hours after induction (1h, 2h, 3h, 4h and ON). SDS-PAGE analysis of the total, soluble (S) and insoluble (I) fractions is shown in Figure 2.58. It is clear that the protein is over-expressed and already visible on the gel 1 hour after the induction, and its production increases over-time. Analysis of the soluble (S) and insoluble (I) fractions of the crude extract shows that the protein is produced mainly in inclusion bodies (Figure 2.58b). The theoretic molecular mass of the protein is 27.5 kDa, but the bands appear at 20 kDa. It is not clear if the protein is degraded inside the cells to a lower molecular mass species after synthesis, or if it is produced directly at the wrong size. A western blot analysis was performed to answer that question.

Western blot

The western blot analysis shown in Figure 2.59 is the same as described for OctaVII_01 in Section 2.4.1, page 79. OctaVII_02 is not visible in the SDS-PAGE gel (Figure 2.59a), but it is clearly present in the western blot (Figure 2.59b), showing that it is expressed at low concentration of IPTG (0 to 100 μM) and low temperature (18°C). The molecular mass of the protein is 15 kDa and does not correspond to the theoretic one of 27.5 kDa. This truncation is surely located at the N-terminal of the protein because the 6x HisTag recognized by the antibody is at the C-terminal.

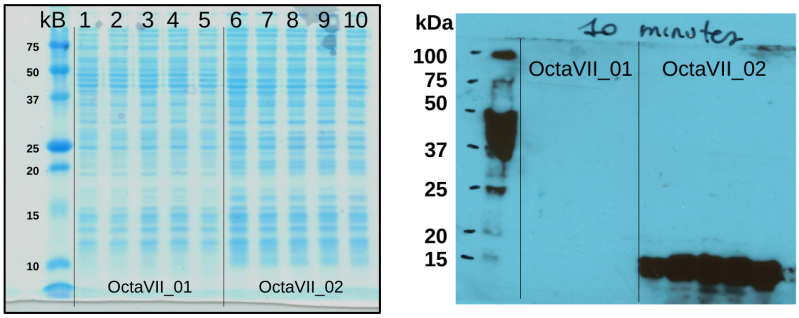


(a) (b)

Figure 2.58: Expression trial of OctaVII_02

(**a**) SDS-PAGE of the total fraction for 6 different conditions of induction of OctaVII_02 at 37°C with IPTG 1 mM: 0, 1, 2, 3, 4 hours and overnight (NI, 1h, 2h, 3h, 4h and ON, respectively). (**b**) SDS-PAGE of the same samples after separation of the soluble fraction (S) and the insoluble fraction (I).

The western blot also shows that the protein is not present as a single band but with at least two main populations at similar size, and few minor bands at lower molecular mass (very likely, degradation products).



(a) (b)

Figure 2.59: Western blot of OctaVII_02

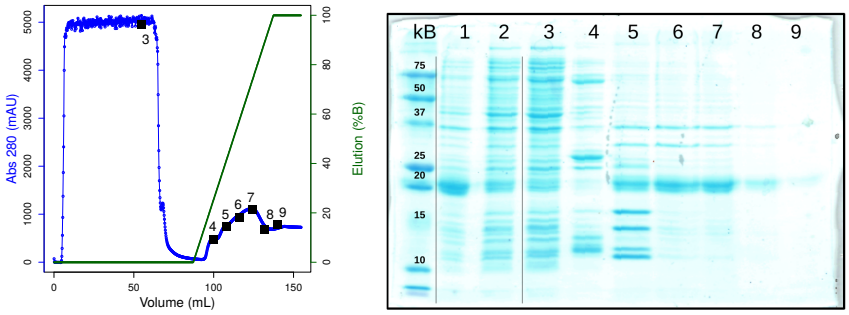
(**a**) SDS-PAGE of the total fraction for 5 different concentrations of IPTG for both OctaVII_01 (lanes **1-5**) and OctaVII_02 (lanes **6-10**). Lines **1,6**, non-induced; **2,7**, IPTG 10 μM; **3,8**, IPTG 20 μM; **4,9**, IPTG 50 μM and **5,10**, IPTG 100 μM. (**b**) Western Blot of the same samples after 10 minutes of film exposition.

Purification of the soluble fraction

Despite of the truncation and the double band of OctaVII_02, we decided to produce the protein. The cells of 1 L cultures were disrupted and the soluble fraction was separated from the insoluble one by centrifugation. Figure 2.60b shows the difference between the crude extract (total fraction obtained after cell disruption) and the soluble fraction (sample obtained after centrifugation of the crude extract). The band of OctaVII_02 is predominant in the crude extract (lane 1), while it is barely noticeable among the other

cellular proteins in the soluble fraction (lane 2).

After filtration on 0.22 μm filters, the soluble fraction of the crude extract was loaded on an HisTrap HP column for purification. The elution profile is shown in Figure 2.60a, and its analysis on SDS-PAGE is shown in Figure 2.60b in lanes 4 to 9.



(a) **(b)**

Figure 2.60: Purification of OctaVII_02, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_02: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the crude extract; **2** the input sample; **3** the flow-through; **4-9** the elution peak.

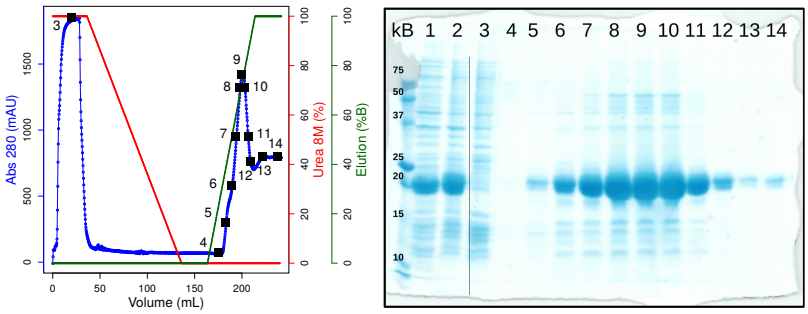
Most proteins of the crude extract did not interact with the column and were eluted in the flow-through (lane 3). Some contaminants are eluted in the first fractions following initiation of the imidazole gradient (lane 4), suggesting that their interaction with the Ni^{2+} matrix is weak and probably due to a single exposed histidine. Lines 5 to 9 present a band at 20 kDa that might correspond to OctaVII_02, but contaminant bands are present at higher and lower molecular mass, especially in lane 5. The molecular mass of 15 kDa that was shown in the western blot may just be an artifact due to the membrane transfer, but, despite the low quality of the gel resolution, it seems that two or more bands are overlapping at 20 kDa.

Purification of the insoluble fraction

The insoluble fraction of the crude extract of the 1 L culture was washed a couple of times prior to denaturation. The refolding of the protein was done in parallel with its purification: the unfolded protein in 8 M urea (see Section 4.7.4, page 202) was loaded in the HisTrap HP column and its 6x HisTag bounds tightly to the Ni^{2+} matrix. The concentration of urea was then progressively decreased with a gradient from 8 to 0 M (red line in Figure 2.61a). Under this condition, the protein can possibly refold and equilibrate in the standard buffer prior to elution with an imidazole buffer. This technique is called on-column refolding and is an efficient method to simultaneously refold and purify

proteins.

Inclusion bodies of OctaVII_02 were resuspended in a buffer with urea 8 M and left stirring overnight at room temperature. In the morning the solution of unfolded soluble proteins was centrifuged and filtered on 0.22 μm filters. Figure 2.61b, (lanes 1 and 2 shows no significant difference between the resuspended pellet and the filtrated solution.



(a) **(b)**

Figure 2.61: Purification of OctaVII_02, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_02: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the resuspended pellet; **2** the input sample; **3** the flow-through; **4-14** the elution peak.

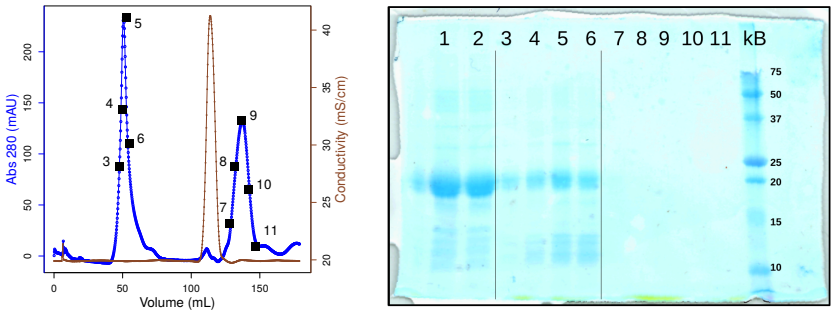
The unfolded OctaVII_02 was then loaded on the HisTrap HP column, refolded and eluted. The majority of the contaminants does not bind to the column and they are eluted in the flow-through (lane 3 in the SDS-PAGE gel). The refolding process (between 50 and 140 mL of the elution volume) does not affect the binding of the HisTag to the resin, as indicated by the low absorbance signal at 280 nm (blue line in Figure 2.61a). The elution of the protein starts at 150 mM of imidazole and present a sharp peak. The high absorbance signal after the elution peak is due to the increasing concentration of imidazole, that absorbs at 280 nm. The fractions on gels (lanes 4 to 14) shows that OctaVII_02 is the predominant band, although minor contaminants are visible at both higher and lower molecular masses. The next step in the analysis of OctaVII_02 is the evaluation of its oligomeric state through size exclusion chromatography.

Size Exclusion, Superdex75

The fractions containing the refolded OctaVII_02 were pooled in one sample that was centrifuged and filtered on 0.22 μm filters. The protein concentration dropped from 0.75 mg/mL to 0.25 mg/mL during this manipulation, suggesting that the refolded protein may not be fully soluble. However there are not differences in the intensity of the bands

on the gel before and after the filtration (lanes 1 and 2, respectively, in Figure 2.62b). This change in the concentration is probably bound to a decrease of imidazole in the solution due to precipitation. Pure imidazole does not absorbs at 280 nm, however impurities in the reagent may cause absorption. Indeed, the absorbance of the imidazole solution used during elution corresponds to 0.8 in Figure 2.61a. In order to avoid absorbance at 280 nm due to the impurities, only highly pure imidazole will be used for the purification of the following OctaVIIs.

The protein sample was loaded on a preparative Superdex-75 column for a size exclusion chromatography (see Table 4.13, page 206), and its elution profile is shown in Figure 2.62a. Two main peaks are visible according to absorbance measurements at 280 nm: the first one is centered at 50 mL and the second one at 140 mL. Fractions corresponding to both peaks were taken for SDS-PAGE analysis, shown in Figure 2.62b, lanes 3 to 11. The second peak (lanes 7-11) does not present protein bands, and thus the signal at 280 nm may be due to the imidazole molecules that absorbs in that range. The first peak presents the protein band and also contaminant species of low molecular mass, suggesting that the protein is interacting with them. The first elution peak corresponds to the void volume of the column, suggesting a molecular mass above 100 kDa. It is not clear if this high molecular mass is caused by interaction of OctaVII_02 with the contaminant in solution or with itself.



(a) (b)

Figure 2.62: Purification of OctaVII_02, size exclusion

(**a**) Size exclusion elution profile of the pool of the refolded OctaVII_02: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the pool of the refolded OctaVII_02; **2** the input sample; **3-6** the first elution peak; **7-11** the second elution peak.

So far OctaVII_02 is well expressed in inclusion bodies, it is truncated, not very soluble and apparently present in different isoforms and oligomerization states. The protein itself does not worth further characterization, but a last attempt we can do is to recover the full size of the protein.

N-sequencing

The sample with the refolded OctaVII_02 was used for N-sequencing (see Section 4.6.6 for the method description, page 200). The resulting sequence at the N-terminal of OctaVII_02 is SxGGM, that corresponds to the sequence SCGGM located at about 60 residues from the expected N-terminus. The truncation covers the first $\beta\alpha\beta\alpha$ unit and the theoretic molecular mass of the truncated form correspond to 20.9 kDa, that is in good agreement with the bands on the gels at ∼20 kDa.

The reasons at the basis of the truncation in OctaVII_02 are not clear, and after sequencing of the plasmid we could exclude errors in the DNA sequence of the gene. Two other possibilities have been considered: 1- an error during the production of the protein, in which the translation machinery missed the first starting codon, ATG, but recognized another one downstream the sequence, or 2- a post-translational cleavage due to endogenous proteases.

We could exclude the first hypothesis because the truncated version of OctaVII_02 does not begin with a methionine, and the closest one is 23 residues upstream the truncation region. It is very unlikely that the truncation of the protein is due to an error during the translation of the protein.

The second hypothesis was assessed through analysis of the sequence cleavage sites. Two predictors for protein cleavage were used: PeptideCutter [129] and PROSPER [130]. The first predictor found 2 proteases that cut the protein between Y59 and S60, the first residue of OctaVII_02: chymotrypsin-like proteases (that cleave the C-term of F, W, Y, M and L residues), and Proteinase K (that cleaves the C-term of aromatic residues). The second program predicted a cleavage by a serine proteases.

Ehrmann group at the University of Duisburg-Essen has a public web-site that resume all the proteases of *E. coli*:

https://www.uni-due.de/zmb/members/ehrmann/e-coli-proteases/.

Among all the proteases, the ATP-dependent Clp proteolytic subunit may be responsible for the truncation of OctaVII_02: it is localized in the cytoplasm of *E. coli* cells, it has a chymotrypsin-like activity (cleavage after F, W, Y, M and L residues) and it plays a major role in the degradation of misfolded protein (like the OctaVII_02 before compartmentalization in inclusion bodies).

In order to understand if the reason of the truncation is a cleavage by endogenous proteases of *E. coli*, and if it is possible to recover the full length of the protein, we decided to produce a variant of OctaVII_02 in which the recognition site of the putative protease (Y59) is mutated to a glutamine (OctaVII_02 Y59Q).

2.4.3 OctaVII_02 Y59Q

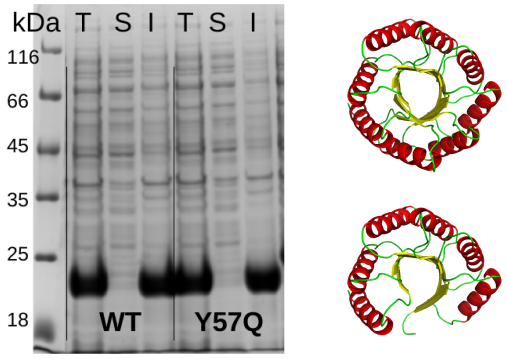
Site-directed mutagenesis and sequencing

The single point mutation in the gene of OctaVII_02 was done by PCR with the kit *QuickChange II XL Site-Directed Mutagenesis* from Agilent Technologies. The sequence of the two primers, Y57Q forward and Y57Q reverse, for the mutagenesis are reported in Section 4.1.3, page 184.

Transformation of DG1 competent cells and plasmid replication were done in order to verify by sequencing the presence of the correct mutation. 4 out of 5 clones contained the good sequence which was used for stock preparation and protein expression.

Expression Trials

After transformation of *E. coli* BL21(DE3) competent cells with both OctaVII_02 wild-type and OctaVII_02 Y59Q, two cultures were prepared to verify the protein size of the Y59Q mutant. The SDS-PAGE gel is shown in Figure 2.63a, and it clearly shows that the single point mutation does not prevent the putative cleavage observed with OctaVII_02. Both OctaVII_02 and OctaVII_02 Y59Q are discarded from further characterization because they lack 60 out of 250 residues (Figure 2.63b).



(a) (b)

Figure 2.63: Expression trial of OctaVII_02 Y59Q

(**a**) SDS-PAGE of the total (T), soluble (S) and insoluble (I) fractions of OctaVII_02 (WT) and OctaVII_02 Y59Q. (**b**) Models of the theoretic full-length protein (top) and of its truncated version (bottom).

2.4.4 OctaVII_03

OctaVII_03 has 250 residues and it is the best representative of Family 16 according to the Rosetta energy score (details in Section 2.3.9, page 68). It is also the best representative of all the models according to molecular dynamics (details in Section 3.3.2, page 180). The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_03 was manipulated for sequencing and storage as described for OctaVII_01 in Section 2.4.1. The results of the sequencing confirmed the correct sequence of the OctaVII_03 gene (see Annex 6.4, page 241), that was then used for a transformation in BL21 (DE3) cells for expression trials.

Expression Trials

After transformation, a colony was selected from the plate to grow for 8 hours at 37°C. An aliquot was taken for the non induced (NI) sample, and 1 mM of IPTG was added to the rest of the culture for an overnight induction at 37°C. The overnight sample was used for analysis of the total (T), the soluble (S) and the insoluble (I) fractions. SDS-PAGE analysis for the 4 conditions (NI, T, S and I) is shown in Figure 2.64. The protein is over-expressed and mainly produced in inclusion bodies. The theoretic molecular mass of the protein is 28.1 kDa, but the bands appears at 23 kDa, suggesting that OctaVII_03 might also be truncated. Since our efforts to recover the full length of OctaVII_02 did not brought significant results (see Section 2.4.3, page 89), OctaVII_03 is discarded without further analysis.

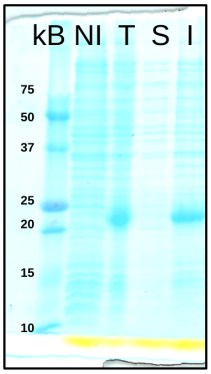


Figure 2.64: Expression trial of OctaVII_03

SDS-PAGE of OctaVII_03 after induction with 1 mM IPTG overnight. From left to right: the non-induced sample (NI), the total fraction (T), the soluble fraction (S) and the insoluble fraction (I).

2.4.5 OctaVII_04

OctaVII_04 has 250 residues and it is the best representative of Family 23 according to the Rosetta energy score (details in Section 2.3.9, page 68). The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_04 was manipulated for sequencing and storage as described for OctaVII_01 in Section 2.4.1, page 77. The results of the sequencing confirmed the correct sequence of the OctaVII_04 gene (see Annex 6.4, page 241), that was then used for a transformation in BL21 (DE3) cells for expression trials.

Expression Trials

After transformation of *E. coli* BL21(DE3) competent cells with the OctaVII_04 plasmid, a colony was selected from the plate to grow for 8 hours at 37°C. An aliquot was taken for the non induced (NI) sample, and 1 mM of IPTG was added to the rest of the culture for an overnight induction at 37°C. SDS-PAGE analysis of the total (T), soluble (S), and insoluble (I) fractions is shown in Figure 2.65. It is clear that the protein is over-expressed and mainly produced in inclusion bodies. The theoretic molecular mass of the protein is 27.9 kDa, and the bands appear at the expected position in the gel. Expression trials were performed also at 18°C overnight with the same results (data not shown).

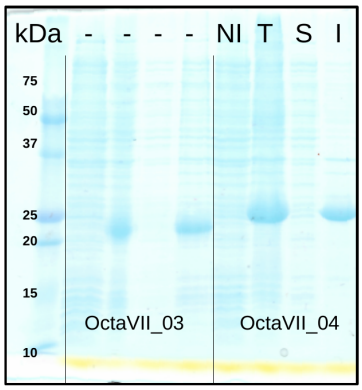


Figure 2.65: Expression trial of OctaVII_04

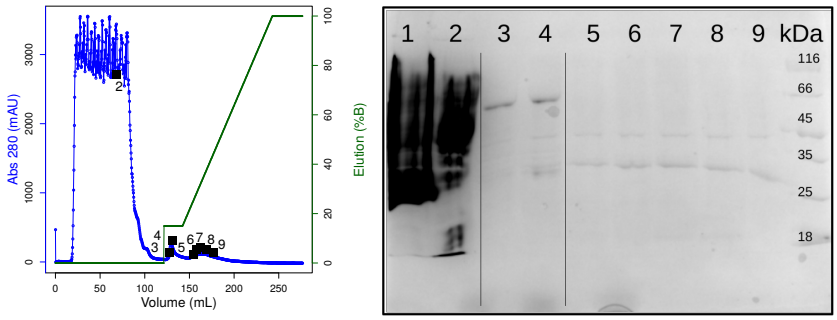
SDS-PAGE of OctaVII_03 (already analyzed in Section 2.4.4, page 91) and OctaVII_04, after induction with 1 mM IPTG overnight. From left to right: the non-induced sample (NI), the total fraction (T), the soluble fraction (S) and the insoluble fraction (I).

Purification of the soluble fraction

Cells of 1 L culture with overnight induction at 37°C were disrupted. Soluble and insoluble fractions were separated by centrifugation. After filtration on 0.22 μm filters, the soluble fraction of the crude extract was loaded on an HisTrap HP column for purification. The elution profile is shown in Figure 2.66a, and its analysis on SDS-PAGE is shown in Figure 2.66b in lanes 2 to 9.

Almost all the proteins of the sample do not interact with the column and are eluted in the flow-through (lane 2). Some contaminants at high molecular mass are eluted with imidazole 75 mM in the first fractions (lanes 3 and 4). Lanes 5 to 9 correspond to a band at 28 kDa that might represent the OctaVII_04.

As for the purification of the soluble fraction of OctaVII_02, the overall amount of OctaVII_04 is extremely low for an over-expression. Its presence in the soluble fraction may be due other factors rather than spontaneous folding in the cytoplasm of the cells.



(a) (b)

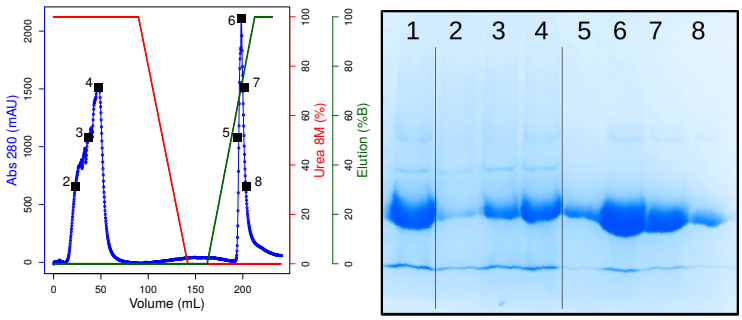
Figure 2.66: Purification of OctaVII_04, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_04: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2** the flow-through; **3-4** the first elution peak at imidazole 75 mM; **5-9** the second elution peak.

Purification of the insoluble fraction

The insoluble fraction of the crude extract of the 1 L culture was washed a couple of times prior to denaturation. As for OctaVII_02, the refolding of the protein was performed in parallel with its purification. The unfolded OctaVII_04 in 8 M urea was loaded on the HisTrap HP column, refolded and eluted (Figure 2.67a). In the input solution OctaVII_04 is the predominant band, however few contaminants are visible in the SDS-PAGE (lane 1 in Figure 2.67b). The contaminants and part of OctaVII_04 are eluted in the flow-through

(lanes 2-4). The presence of OctaVII_04 in the flow-through is due to simple saturation of the column.



(a) (b)

Figure 2.67: Purification of OctaVII_04, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_04: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-4** the flow-through; **5-8** the elution peak.

The refolding process (between 100 and 150 mL of the elution volume) does not affect the binding of the protein to the matrix because there are no changes in the absorbance at 280 nm (blue line in Figure 2.67a). The elution of the protein starts at 250 mM of imidazole and present a sharp peak. The fractions on gels (lanes 5 to 8) shows that OctaVII_04 is the predominant band and that it is of high purity (> 95%).

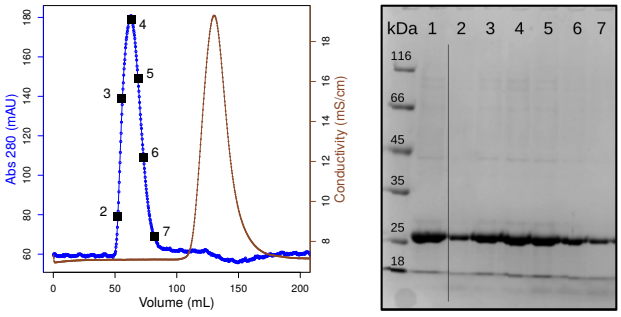
Desalting

In order to remove the imidazole from the protein sample and prevent any precipitation, a desalting chromatography was done just after the elution of the refolded OctaVII_04. The fractions containing the protein were pooled together, the sample was centrifuged, filtered on 0.22 μm filters and loaded onto the desalting column. The elution profile is shown in Figure 2.68a.

The peak of elution of the protein (from 50 to 100 mL in the elution volume) is well separated from the peak in conductivity (110 to 160 mL), due to the imidazole. The SDS-PAGE in Figure 2.68b shows the presence of few contaminants at 45 kDa, but OctaVII_04 is always the predominant band.

Biophysical characterization

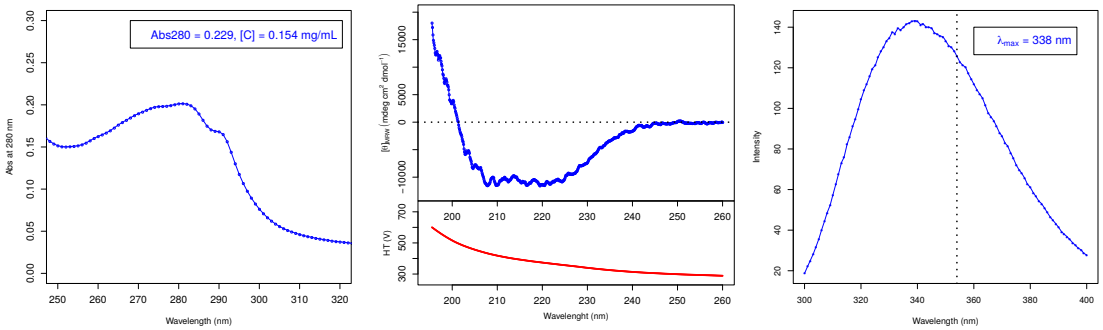
The fractions containing OctaVII_04 were pooled together and centrifuged at 20000 rpm for 20 mins. The sample was filtered on 0.22 μm filters and an absorption spectrum was recorded (Figure 2.69a).



(a) (b)

Figure 2.68: Purification of OctaVII_04, desalting

(**a**) Elution profile of the desalting of the pool of the refolded OctaVII_04: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the input sample; **2-7** the elution peak.



(a) (b) (c)

Figure 2.69: Biophysical characterization of OctaVII_04

(**a**) Absorption spectrum of the pool of OctaVII_04 after desalting. Abs_{280} is used to calculate the protein concentration. (**b**) CD spectrum of the protein (top) and high-tension (bottom); the dotted line indicates the baseline at $[\Theta]$=0. (**c**) Emission fluorescence spectrum of the protein. The dotted line at 354 nm indicates the expected maximum for unfolded proteins.

The Abs_{260}/Abs_{280} ratio is 0.82 and indicates that there is no significant contamination by nucleic acids; the shoulder at 290 nm is typical of tryptophan and the Abs_{280} is 0.23. The concentration of the protein was calculated with the Equation 4.3, and equals to 0.154 mg/mL. The molar extinction coefficient (ε_m) is reported in Table 4.16 in the method section 4.8.1, page 209.

The sample was then diluted to 0.1 mg/mL and 0.01 mg/mL for analysis by circular dichroism (CD) and fluorescence, respectively. Reliable CD signal could be measured down to 195.6 nm before saturation of the photo-multiplicator. The CD spectrum shows the presence of α-helices, with minima around 222 and 208 nm. The analysis of the spectrum by CDpro, using the program CDSSTR, is compared with the DSSP analysis obtained from the model structure of OctaVII_04. Both are shown in Table 2.6.

	% α-Helix	% β-Strand	% Turn	% Unstruc
CDpro (CDSSTR)	34.9	17.4	19.5	27.9
DSSP	47.9	18.7	16.6	16.6

Table 2.6: Secondary structures content, OctaVII_04

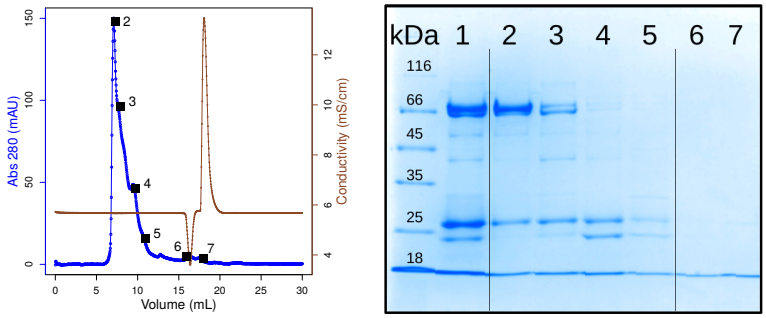
The percentages of β-strands and turns obtained from the experimental data are in good agreement with the model ones, 18% each. However, the α-helix content is decreased of 13% compared to the model.

Analysis of the tertiary structure was done by fluorescence and is reported in Figure 2.69c. The maximum in intensity is at 338 nm and indicates that the aromatic residues are not exposed to the solvent. The protein have a compact 3D structure, but it is not clear if it is well-packed or in a molten globule state.

The protein was subjected to filtration steps in order to reach a concentration > 2 mg/mL. Unfortunately, the majority of the protein precipitated after just few cycles. It was not possible to reach a concentration greater than 0.2 mg/mL in any trial done with OctaVII_04.

Size Exclusion, Superdex75

A new sample of the refolded OctaVII_04 was loaded on a Superdex-75 column in order to inquire its oligomerization state. The sample was less pure than the previous one: contaminant bands are visible on SDS-PAGE gel at 66 kDa and 23 kDa (lane 1 in Figure 2.70b).



(a) **(b)**

Figure 2.70: Purification of OctaVII_04, size exclusion

(**a**) Size exclusion elution profile of the pool of the refolded OctaVII_04: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the input sample; **2-5** the first elution peak; **6-7** the second elution peak.

The elution profile of OctaVII_04 is shown in Figure 2.70a, in which a main peak with multiples shoulders is visible. The peak is centered at 7.9 mL, corresponding to the void volume of the column. This result suggests that the majority of the protein is in a higher oligomerization state than 75 kDa, as for OctaVII_02. The SDS-PAGE gel shows that a proteins at 66 kDa and the OctaVII_04 co-elute in the main peak (lanes 2 and 3). One of the shoulders is centered at 10 mL, corresponding to 60 kDa. In this fraction the protein can be in a dimeric form, but the SDS-PAGE gel shows that OctaVII_04 is co-eluting with the protein at 23 kDa.

It is not clear if OctaVII_04 is interacting with the contaminant bands in solution, or if it is forming homo-oligomers on its own. The presence of 6 free cysteines in the sequence of OctaVII_04 may be a reason for its high oligomerization state. It can possibly be related also to the low solubility of the protein in solution. We decided to discard OctaVII_04 from further analysis and to create a mutant without cysteins in order to obtain a monomeric form.

2.4.6 OctaVII_04 NoCys

OctaVII_04 NoCys is a mutant of OctaVII_04 that does not contain cysteine. The substitution of the 6 original cysteines was done in collaboration with the laboratory of Dr. Wim Vranken at the Vrije Universiteit Brussel (VUB), Belgium. He and his group developed a methodology based on the Hidden Markov model (named HMM) that is able to score a protein starting from its sequence and to suggest beneficial mutations that may improve the score. Rosetta does the same, but it requires a starting 3D structure. HMM was trained on databases of natural proteins that furnish statistical results, that were then used for the analysis of the protein sequence. The HMM methodology is very promising, because it can be used with all the proteins, natural or artificial, that do not have a solved structure. In this collaboration, all the work related to the HMM software is done by Gabriele Orlando, a PhD student in Vranken's Lab.

The OctaVII_04 sequence was analyzed with HMM, that suggested the beneficial mutations for the substitution of the 6 cysteines: three of them have one single option of mutation (C61I, C89L and C210A), while the remaining three has two options each (C34 T or V, C65 S or V and C214 T or M). All the possible combinations are 8. We tested each of them with both HMM (that uses as input the amino acid sequence) and Rosetta (that uses as input a 3D model structure). The preparation of the 8 models is described in Section 3.3.4. The scores of HMM and Rosetta are reported in Table 2.7. Their normalized values (against OctaVII_04) are plotted in Figure 2.71.

Name	Mutations	HMM score	Rosetta score
OctaVII_04	C34, C61, C65, C89, C210, C214	27.62	-478.016
Mutant_1	C34V, C61I, C65S, C89L, C210A, C214T	26.01	-479.136
Mutant_2	C34V, C61I, C65S, C89L, C210A, C214M	27.82	-476.161
Mutant_3	C34V, C61I, C65V, C89L, C210A, C214T	29.47	-472.391
Mutant_4	C34V, C61I, C65V, C89L, C210A, C214M	30.02	-473.415
Mutant_5	C34T, C61I, C65S, C89L, C210A, C214T	29.67	-474.525
Mutant_6	C34T, C61I, C65S, C89L, C210A, C214M	28.22	-477.103
Mutant_7	C34T, C61I, C65V, C89L, C210A, C214T	29.94	-474.011
Mutant_8	C34T, C61I, C65V, C89L, C210A, C214M	30.65	-470.749

Table 2.7: Design of OctaVII_04 NoCys

The HMM scores are all improving compared to the score of OctaVII_04 with the exception of Mutant_1. The Rosetta scores behaves in the opposite way: the best Rosetta score is of Mutant_1. Compared to OctaVII_04 it is the only one with a lower score, while the other 7 variants has an higher total energy (so, a lower stability). Mutant_8 has 7.2

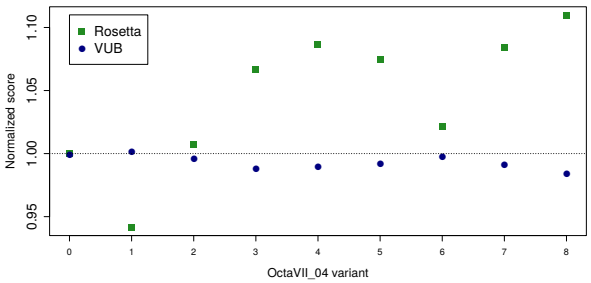


Figure 2.71: Cysteines substitution of OctaVII_04

Normalized scores for the 8 mutants of OctaVII_04 according to HMM (blue) and Rosetta (green) scores. The numbers 1 to 8 on the x-axis represent the 8 mutants and 0 is the original OctaVII_04.

REU of difference compared to the original, but this difference is not significant (1.5% of the total energy).

We chose Mutant_8 among the 8 variants, that has the best score according to HMM. It is renamed OctaVII_04 NoCys and it has 250 residues. The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

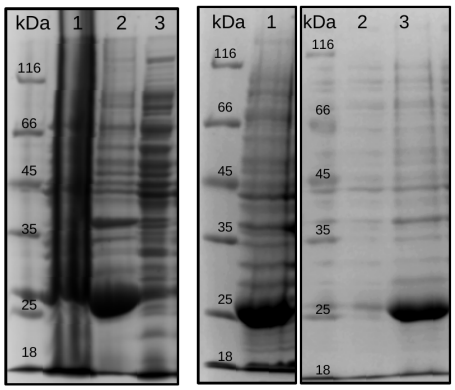
The DNA pellet of OctaVII_04 NoCys was manipulated for sequencing and storage as described for OctaVII_01 in Section 2.4.1, page 77. The results of the sequencing confirmed the correct sequence of the OctaVII_04 NoCys gene (see Annex 6.4, page 241), that was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials.

Expression Trials

Expression trials of OctaVII_04 NoCys were performed with IPTG 1 mM in two conditions: overnight at 18°C and 4 hours at 37°C. In both cases the protein was highly produced and represents the major band in the total fraction (lines 1 in Figure 2.72). The protein was mainly produced in inclusion bodies (insoluble fraction), but the gel suggests that a small amount might be present also in the soluble fraction. For this reason, we tried to purify the protein from the soluble fraction with the HisTrap HP column.

Purification of the soluble fraction

The protein was produced in 1 L of cell culture with 4 hours of induction at 37°C. The crude extract was obtained after 3 cycles of disruption and it is shown in 2 different concentrations in Figure 2.73b (lanes 1 and 2). The soluble fraction was separated from the pellet via centrifugation and filtered on 0.22 μm filters. The input sample (lane 3)

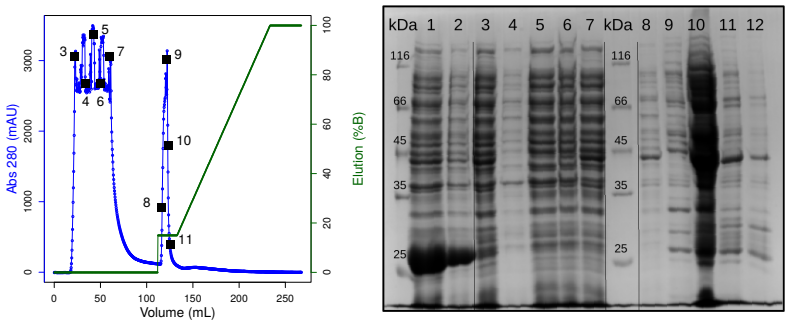


(a) **(b)**

Figure 2.72: Expression trials of OctaVII_04 NoCys, soluble fraction

(**a**) Expression trials of OctaVII_04 NoCys at 18°C overnight: lane **1** total, **2** insoluble and **3** soluble fractions. (**b**) Expression trials of OctaVII_04 NoCys at 37°C for 4 hours: **lane 1** total, **2** soluble and **3** insoluble fractions.

was loaded on the HisTrap HP column, and the protein was purified as described in Section 4.7.3, page 201. The chromatogram of the purification is shown in Figure 2.73a. The elution peak is visible at 75 mM of imidazole. This concentration is too low for the 6x HisTag, that normally eluates at 300 mM of imidazole. The peak is composed by contaminants, as shown in the SDS-PAGE in lanes 8 to 12, and OctaVII_04 NoCys is not produced in the soluble fraction. This result may suggest that the small elution peak that was obtained after purification of the soluble fraction is not due to spontaneous folding.



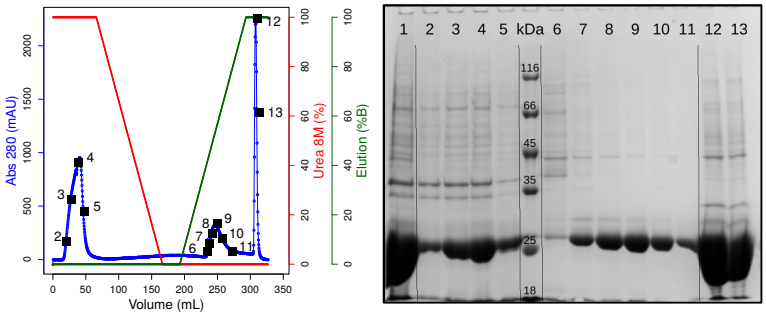
(a) **(b)**

Figure 2.73: Purification of OctaVII_04 NoCys, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_04 NoCys: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lanes **1-2**, the crude extract, 4 μL and 1 μL, respectively; **3** the input sample; **4-7** the flow-through; **8-12** the elution peak.

Purification of the insoluble fraction

The insoluble fraction (pellet) produced in 1 L of cell culture was washed and re-suspended at room temperature overnight in a buffer containing urea 8 M. The solution was centrifuged at 20000 rpm for 20 min and the supernatant was filtered on 0.22 μm filters. This input sample was loaded on the HisTrap HP column for purification and refolding (described in Section 4.7.4, page 202) and it is shown in Figure 2.74b in lane 1. OctaVII_04 NoCys is the predominant protein, but contaminants are visible at higher molecular mass. OctaVII_04 NoCys is present in the flow-through (lanes 2 to 5), suggesting that the column reached saturation. There are two elution peaks: the first one is between 200 and 300 mM of imidazole, which is the optimal concentration for the elution of HisTagged proteins, and the second is at 500 mM of imidazole. It is not clear why part of the protein elutes at higher concentration of imidazole. The purification was repeated 4 times with the same results. The second peak is more concentrated than the first one, and it contains more contaminants compared to the first one (in particular the band at 33 kDa that is not visible in the first peak). However, OctaVII_04 NoCys is always the main band. The two peaks are individually pooled and manipulated for comparison.



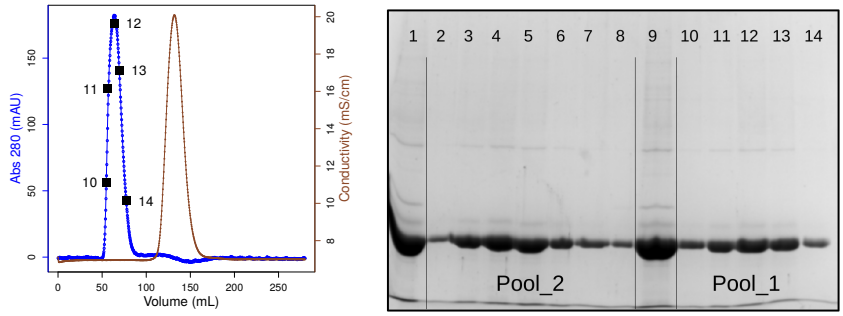
(a) **(b)**

Figure 2.74: Purification of OctaVII_04 NoCys, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_04 NoCys: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-5** the flow-through; **6-11** the first elution peak and **12-13** the second elution peak.

Desalting

For each of the two elution peaks of the refolded OctaVII_04 NoCys, a pool was created (Pool 1 and Pool 2, respectively). Both were centrifuged and filtered prior to loading onto the desalting column. The elution profile is shown only for Pool 1 (Figure 2.75a), but the SDS-PAGE gel shows the elution fractions for both samples.



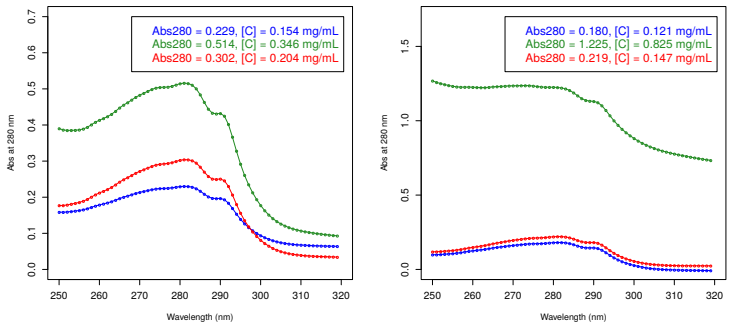
(a) **(b)**

Figure 2.75: Purification of OctaVII_04 NoCys, desalting

(**a**) Elution profile of the desalting of Pool 1 of the refolded OctaVII_04: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the input sample of Pool 2 and **2-8** its elution peak; lane **9** the input sample of Pool 1 and **10-14** its elution peak.

Biophysical characterization

After desalting, Pool 1 and Pool 2 were centrifuged in order to remove possible aggregates. Absorbance spectra were recorded in order to quantify the protein and they are shown in Figure 2.76.



(a) **(b)**

Figure 2.76: Concentration trials of OctaVII_04 NoCys

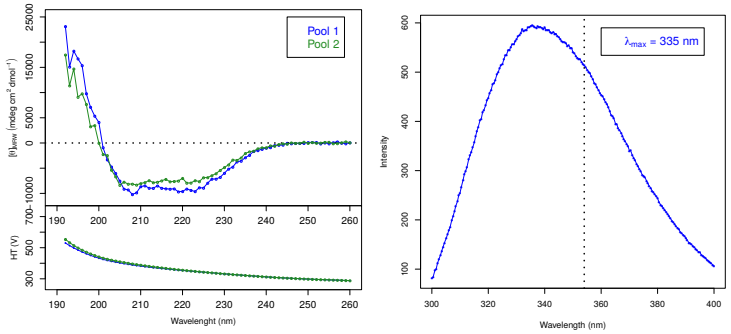
(**a**) Absorption spectrum of Pool 1 of OctaVII_04 NoCys after desalting (blue), after concentration (green) and after centrifugation and filtration (red). Abs_{280} is used to calculate the protein concentration. (**b**) The same analysis for Pool 2 of OctaVII_04 NoCys.

Pool 1 appears to have aggregates in solution, because the signal above 310 nm is not zero (blue line in Figure 2.76a). After 2 steps of concentration by ultrafiltration (green line), the aggregates are visible in solution, and a centrifugation step at high speed is necessary to remove them. The solution is then filtered and the protein concentration resulted 0.2 mg/mL (red line). This concentration is similar to the one obtained for OctaVII_04. Pool 2 does not seem to have aggregates after desalting (blue line in Figure 2.76b), anyway they show up after 4 cycles of concentration by ultrafiltration (green

line). Centrifugation at high speed and filtration removed the aggregates, but also the majority of the protein from the solution, and the final concentration of the sample is 0.15 mg/mL.

The concentrations of Pool 1 and Pool 2 of OctaVII_04 NoCys are low, but enough to perform far-UV circular dichroism and fluorescence measurements.

Figure 2.77a shows the CD spectra of Pool 1 and Pool 2 of OctaVII_04 NoCys. Both show a significative fraction of α-helical secondary structures, as indicated by the two minima at 222 nm and 208 nm. The spectre are, however, not identical. The analysis of the 2 datasets with CDpro, using the program CDSSTR, confirmed the presence of differences between the two pools (Table 2.8).



(a) **(b)**

Figure 2.77: Biophysical characterization of OctaVII_04 NoCys

(**a**) CD spectra of Pool 1 in blue and Pool 2 in green of OctaVII_04 NoCys (top) and high-tension (bottom); the dotted line indicates the baseline at [Θ]=0. (**b**) Emission fluorescence spectrum of Pool 1 of OctaVII_04 NoCys; the dotted line at 354 nm indicates the theoretic maximum for unfolded proteins.

Pool 1 contains 33.7% of helix, while Pool 2 only 22.3%. However, the strand content is lower in Pool 1 (18.3%) and higher in Pool 2 (26.8%). These results may suggest that there are two populations of OctaVII_04 NoCys. This might be due to the refolding on-column, however we did not inquire deeper the situation: both population do not reach the expected content of helix predicted by DSSP (47%). The content of the strands is however in good agreement between Pool 1 and DSSP, around 18.5%. Only Pool 1 was so used to analysis of the tertiary structure by fluorescence, as shown in Figure 2.77b.

	% α-Helix	% β-Strand	% Turn	% Unstruc
CDpro (CDSSTR) Pool 1	33.7	18.3	19.0	28.6
CDpro (CDSSTR) Pool 2	22.3	26.8	21.6	29.0
DSSP	47.9	18.7	16.6	16.6

Table 2.8: Secondary structures content, OctaVII_04 NoCys

The maximum intensity is centered at 335 nm, suggesting that the aromatic residues are localized in the core of the protein and that they are not exposed to the solvent. Attempts to concentrate the protein for near-UV CD analysis (up to 2 mg/mL) led to protein precipitation. The maximum concentration that was reached without evidences of aggregates was 0.35 mg/mL. Because of its poor solubility and its tendency to form aggregates, OctaVII_04 NoCys is discarded from further analysis.

2.4.7 OctaVII_04 WS

OctaVII_04 WS (Weak Spots) is a mutant of OctaVII_04 NoCys, designed in collaboration with the laboratory of Wim Vranken at the Vrije Universiteit Brussel (VUB), in Belgium. The goal of the collaboration was to improve the sequence of OctaVII_04 NoCys to fold in a TIM-barrel with the use of HMM methodology. However, the methodology needed to be adapted in order to recognize the TIM-barrel fold among unknown sequences. To do so, the software was trained with a database of TIM-barrel proteins only (described in section **HMM training**). This new methodology (named HMM-TIM) was then used to analyze the OctaVII_04 NoCys sequence and discriminate the amino acids that are less favorable to form a TIM-barrel fold. Once these "weak spots" were defined, HMM-TIM was used to predict suitable amino acid substitutions (see section **HMM-TIM predictions**). As mentioned for the cysteine substitution of OctaVII_04, all the computational work bound to the HMM training and the HMM-TIM predictions was done by Gabriele Orlando, a PhD student in Vranken's Lab.

HMM training

In order to train the HMM mothodology to recognize the TIM-barrel fold among unknown sequences, a structure-based multiple sequence alignment (MSA) of 100 known TIM-barrels was generated based on information from the Protein Data Bank. Then, every amino acid was encoded with a set of features that describes their physical-chemical behavior. 16 features with less then 0.4 of pairwise Pearson's correlation coefficient were selected from the Aaindex (www.genome.jp/aaindex). To this, the DynaMine predicted backbone dynamics were added, resulting in a total of 17 features (reported in Annex 6.6). With this information, Vranken's Lab built a logistic profile Hidden Markov Model (lHMM) based on the MSA information (named HMM-TIM) and trained it to distinguish TIM-barrel sequences from other folds.

In order to evaluate the prediction capability of HMM-TIM, Vranken's Lab performed three blind tests with datasets of random folds, including the TIM-barrel one. The datasets were prepared in order to contain representatives for all kinds of proteins: only-α, only-β and α-β-proteins. Particular attention was paid to include in each dataset representatives for the α-β-α sandwich and the Rossmann-like folds which have very similar secondary structure content as the TIM-barrel fold. They then used the trained software, to discriminate between TIM-barrels and other folds in the blind datasets. The results were analyzed with the area under the ROC curve (AUC) method, shown in Figure 2.78. They ranged from 92 to 96%, indicating that the biophysical characteristics enabled a good separation between TIM barrel-forming and non-TIM barrel forming sequences.

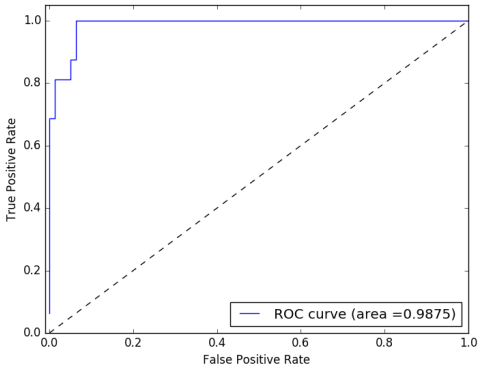


Figure 2.78: HMM-TIM blind test

Analysis of one blind test of HMM-TIM according to the area under the ROC curve method. This plot is used to evaluate the capability of the machine learning program to discriminate among true positive and false positive predictions. The dashed diagonal line indicates random predictions, while the blue line indicates the HMM-TIM predictions. Since the blue curve is in the upper left-hand corner, the software has a low rate of false positive predictions and an high-rate of true positive predictions. This means that the predictions of the software are very accurate.

HMM-TIM Predictions

Following the training and the blind test, HMM-TIM was used to predict which amino acid in OctaVII_04 NoCys sequence was less favorable to form a TIM-barrel fold. Results are shown in Figure 2.79.

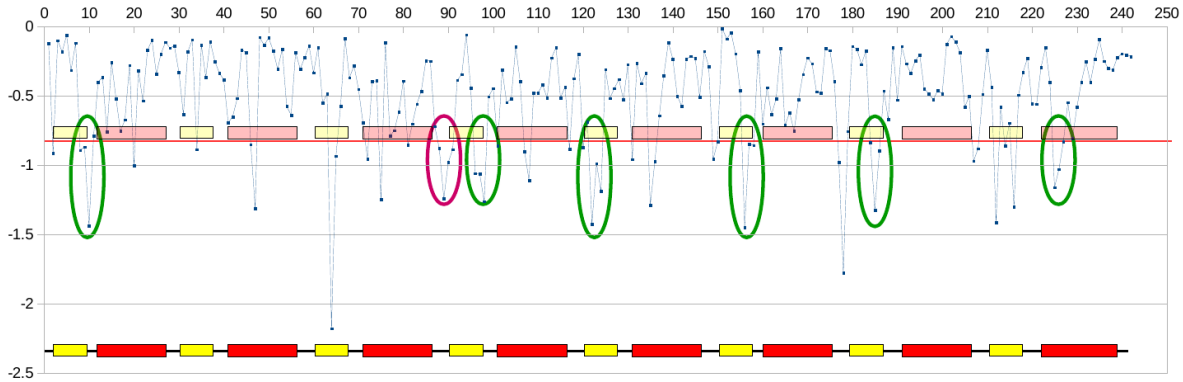


Figure 2.79: Analysis of the sequence of OctaVII_04 NoCys

Blue dots HMM-TIM score for each residue in the OctaVII_04 NoCys sequence. **Red and yellow bars** representations of the secondary structures of the protein: helices in red and strands in yellow. **Green and purple circles** cluster of 3 and 4 consecutive residues, respectively, that are less-likely to fold in a TIM-barrel structure.

The x-axis corresponds to the amino acid sequence of OctaVII_04 NoCys, whereas the score of the software is given on the y-axis. Values close to 0 indicate that the residues are very-likely to fold in a TIM-barrel, while values that are lower than -0.8 are less likely to

form a TIM-barrel. A schematic representation of the helices (red) and strands (yellow) of the protein is reported in full colors at the bottom of the graph and in soft colors at the level of the cut-off value (-0.8). 49 out of 240 residues have a lower score than -0.8. Among them, 22 were selected since they form clusters of consecutive "bad" residues, or weak spots. Green and purple circles indicate the 6 clusters of 3 consecutive residues and the single cluster of 4 residues, respectively.

Following the analysis of the OctaVII_04 NoCys sequence, HMM-TIM was then used to find the best amino acid substitution for each of the 22 mutations. Two out of 22 residues have only one suggested amino acid, while the remaining 20 have 2 suggested residues. All the possible combinations are 2^{20} (more than 1 million of possibilities) and it is not possible to compute them all at once. To simplify the computations, each cluster is modeled independently (52 models) and the best combinations of mutations are then combined in the final sequence. All the suggested mutations are reported by cluster in Table 2.9 together with the scores obtained with both the HMM-TIM software and Rosetta. In general, the mutations are considered "synonymous": the residues are substituted with other amino acids with the same properties (dimension, charge, polarity). For instance, leucine 8 is exchanged with an isoleucine or a valine, that share the same dimension and hydrophobicity; tyrosine 97 is exchanged with the aromatic residues phenylalanine and tryptophan; glycine 185 is exchanged with the small residues alanine and serine. "Non-synonymous" examples are asparagine 91, that is exchanged in one case with a bigger and charged residue, lysine, and in the second case with a small and uncharged glycine, and glutamic acid 226 that is changed in one case in a positive charged residue, arginine, and in the other case with a neutral glutamine.

Name	Mutations	HMM-TIM score	Rosetta score
OctaVII_04 NoCys		31.50	-470.74
Cl1_1	L8V, Q9D, G10N	31.77	-464.97
Cl1_2	L8V, Q9D, G10S	32.11	-464.77
Cl1_3	L8V, Q9N, G10N	32.26	-461.38
Cl1_4	L8V, Q9N, G10S	32.56	-467.31
Cl1_5	L8I, Q9D, G10N	31.60	-464.90
Cl1_6	L8I, Q9D, G10S	31.82	-464.90
Cl1_7	L8I, Q9N, G10N	32.02	-466.14
Cl1_8	L8I, Q9N, G10S	32.24	-466.42
Cl2_1	K88E, L89I, D90N, N91K	30.54	-464.28
Cl2_2	K88E, L89I, D90N, N91G	31.37	-463.43
Cl2_3	K88E, L89I, D90E, N91K	31.09	-457.85
Cl2_4	K88E, L89I, D90E, N91G	31.71	-462.77
Cl2_5	K88N, L89I, D90N, N91K	30.49	-464.13

Name	Mutations	HMM-TIM score	Rosetta score
Cl2_6	K88N, L89I, D90N, N91G	31.37	-463.88
Cl2_7	K88N, L89I, D90E, N91K	30.66	-463.18
Cl2_8	K88N, L89I, D90E, N91G	31.41	-461.59
Cl3_1	L96I, Y97W, S98E	32.80	-461.49
Cl3_2	L96I, Y97W, S98N	33.30	-461.35
Cl3_3	L96I, Y97F, S98E	34.15	-467.45
Cl3_4	L96I, Y97F, S98N	34.62	-467.53
Cl3_5	L96V, Y97W, S98E	32.17	-461.03
Cl3_6	L96V, Y97W, S98N	32.67	-461.25
Cl3_7	L96V, Y97F, S98E	33.93	-463.91
Cl3_8	L96V, Y97F, S98N	34.40	-464.75
Cl4_1	T122S, L123M, I124V	31.15	-465.90
Cl4_2	T122S, L123M, I124L	31.37	-466.56
Cl4_3	T122S, L123V, I124V	29.62	-466.33
Cl4_4	T122S, L123V, I124L	29.86	-470.23
Cl5_1	T156A, G157S, I158V	32.25	-463.36
Cl5_2	T156A, G157S, I158F	32.62	-461.50
Cl5_3	T156A, G157N, I158V	31.78	-461.82
Cl5_4	T156A, G157N, I158F	32.20	-460.54
Cl5_5	T156V, G157S, I158V	33.28	-465.65
Cl5_6	T156V, G157S, I158F	33.57	-461.66
Cl5_7	T156V, G157N, I158V	32.95	-451.29
Cl5_8	T156V, G157N, I158F	33.25	-457.11
Cl6_1	W184Y, G185S, V186L	31.01	-468.28
Cl6_2	W184Y, G185S, V186I	31.43	-469.42
Cl6_3	W184Y, G185A, V186L	30.70	-459.06
Cl6_4	W184Y, G185A, V186I	31.11	-471.95
Cl6_5	W184F, G185S, V186L	31.04	-467.83
Cl6_6	W184F, G185S, V186I	31.46	-468.17
Cl6_7	W184F, G185A, V186L	30.80	-467.07
Cl6_8	W184F, G185A, V186I	31.20	-471.75
Cl7_1	M225I, E226R, K227R	33.13	-467.35
Cl7_2	M225I, E226R, K227E	33.20	-466.70
Cl7_3	M225I, E226Q, K227R	32.96	-467.86
Cl7_4	M225I, E226Q, K227E	32.97	-469.00
Cl7_5	M225V, E226R, K227R	32.53	-467.45
Cl7_6	M225V, E226R, K227E	32.61	-467.00
Cl7_7	M225V, E226Q, K227R	32.38	-470.08
Cl7_8	M225V, E226Q, K227E	32.33	-470.58

Table 2.9: Mutations in the clusters of OctaVII_04 NoCys

The result of the mutations in Clusters 1, 3, 5 and 7 is always positive according to

HMM-TIM. The best combinations are: Cl1_4 (L8I, Q9N, G10S), Cl3_4 (L96I, Y97F, S98N), Cl5_6 (T156V, G157S, I158F) and Cl7_4 (M225I, E226Q, K227E).

Clusters 2, 4 and 6 on the contrary have all the combinations with a lower score compared to OctaVII_04 NoCys, with the exception of Cl2_4 (K88E, L89I, D90E, N91G). The results according to Rosetta score are all higher than OctaVII_04 NoCys (and so less favorable), with the exception of Cl6_4 and Cl6_8. The higher change in the overall energy is of 4.1% with Cl5_7.

The best combination of mutations for each cluster is then combined with the others in one model, OctaVII_04 WS (Weak Spots). The 22 mutations are: L8V, Q9N, G10S, K88E, L89I, D90E, N91G, L96I, Y97F, S98N, T122S, L123V, I124V, T156V, G157S, I158F, W184F, G185S, V186I, M225I, E226R and K227E. The Rosetta score of OctaVII_04 WS drop from -470.75 to -435.34 REU. Its gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

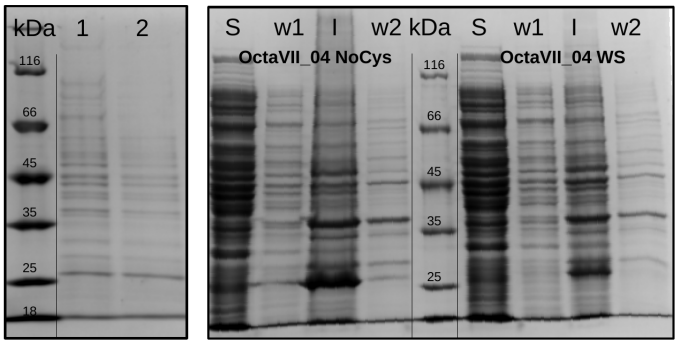
The DNA pellet of OctaVII_04 WS was manipulated for sequencing and storage as described for OctaVII_01 in Section 2.4.1, page 77. The results of the sequencing confirmed the correct sequence of the OctaVII_04 WS gene (see Annex 6.4, page 241), that was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials.

Expression Trials

Expression trials for OctaVII_04 WS were performed in two conditions: overnight induction at 18°C (data not shown) and 4 hours induction at 37°C (Figure 2.80a). Although the SDS-PAGE gel was clearly underloaded, a band is present at 27 kDa for both clones (lanes 1 and 2). A small culture of both OctaVII_04 NoCys and OctaVII_04 WS was prepared in order to compare the crude extracts, shown in Figure 2.80b. In both cases, the protein is not visible in the soluble fraction (S). Experiments on OctaVII_04 NoCys already confirmed the absence of the protein in the soluble fraction, but this was not confirmed yet for OctaVII_04 WS (next section). A predominant band at the correct size is visible for both proteins in the insoluble fraction (I), which contains also many contaminants at various molecular masses. Despite two washing of the inclusion bodies (w1 and w2), the pellet remained highly contaminated.

Purification of the soluble fraction

In order to verify the presence of OctaVII_04 WS in the soluble fraction, 1 L of cell culture was produced and disrupted (the crude extract is shown in lane 1 of Figure 2.81b).

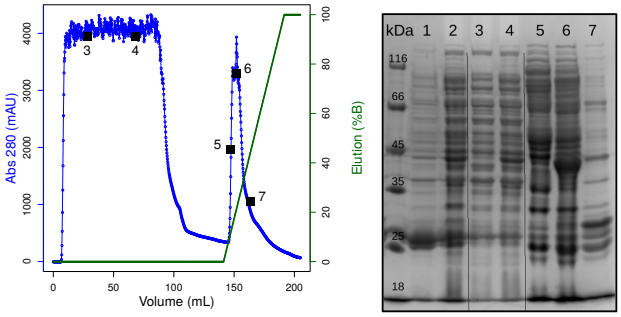


(a) (b)

Figure 2.80: Expression trials of OctaVII_04 WS

(**a**) Expression trials of two clones of OctaVII_04 WS at 37°C for 4 hours; (**b**) Crude extracts of OctaVII_04 NoCys (left) and OctaVII_04 WS (right): S is the soluble fraction, I the insoluble one and w1 and w2 are the washing steps.

The sample was centrifuged and the supernatant filtered and loaded on the HisTrap HP column (the input sample is shown in lane 2). The majority of the proteins was eluted in the flow-through, as visible in the chromatogram (Figure 2.81a) and in lanes 3 and 4 of the SDS-PAGE gel. A peak of elution appears at 50 mM of imidazole, that is a low concentration for the elution of an HisTagged protein (250 mM imidazole). As shown on gel in lanes 5 to 7, the peak of elution is composed almost entirely by contaminants at high molecular mass, and OctaVII_04 WS does not seem to be present in the soluble fraction of the crude extract.



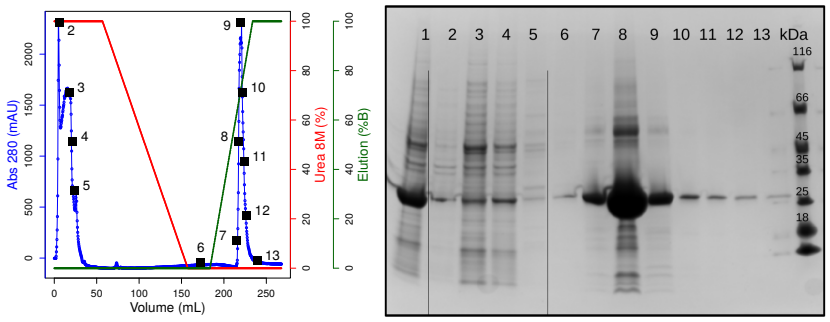
(a) (b)

Figure 2.81: Purification of OctaVII_04 WS, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_04 WS: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the crude extract; **2** the input sample; **3-4** the flow-through; **5-7** the elution peak.

Purification of the insoluble fraction

The pellet obtained from the culture of 1 L was washed twice in order to decrease the amount of contaminants. It was then resuspended overnight at room temperature in a buffer containing urea 8 M. The sample was then centrifuged, filtered and loaded on HisTrap HP columns for the refolding and the purification. The input sample is shown in lane 1 of Figure 2.82b. OctaVII_04 WS is the predominant band, but contaminants are present at high and low molecular masses. The chromatogram is shown in Figure 2.82a, many contaminant protein are eluted in the flow-through (lanes 2 to 5 in the SDS-PAGE). The refolding of the protein from 60 mL to 160 mL does not cause elution of the proteins since there is no signal at 280 nm. The elution of the protein started at around 250 mM of imidazole, which was the expected concentration for the dissociation of the 6x HisTag from the matrix of the column. The elution peak is shown in lanes 6 to 13 of the SDS-PAGE gel and the presence of contaminant is visible in lane 8.



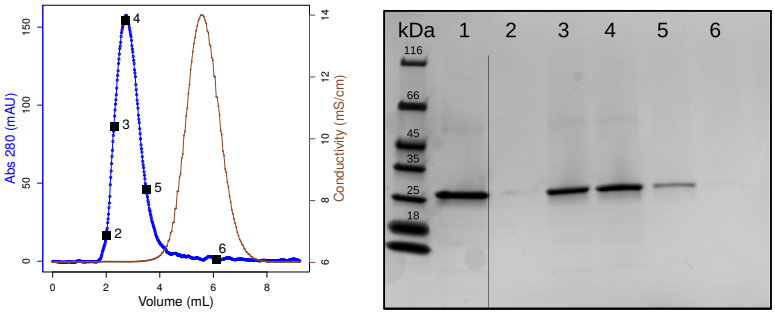
(a) (b)

Figure 2.82: Purification of OctaVII_04 WS, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_04 WS: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-5** the flow-through; **6-13** the elution peak.

Desalting

The fractions collected during the elution of OctaVII_04 WS after the refolding on column were pooled together and centrifuged at 20000 rpm in order to remove possible aggregates. The supernatant was filtered and loaded on a desalting column in order to remove the imidazole from the protein sample. The input sample is shown in Figure 2.83b in lane 1. The chromatogram of the purification is shown in Figure 2.83a, where the separation of the protein (blue line) and the imidazole (brown line) is visible. The SDS-PAGE in Figure 2.83b shows that OctAVII_04 WS is predominant, anyway there are contaminants at ∼48 kDa.



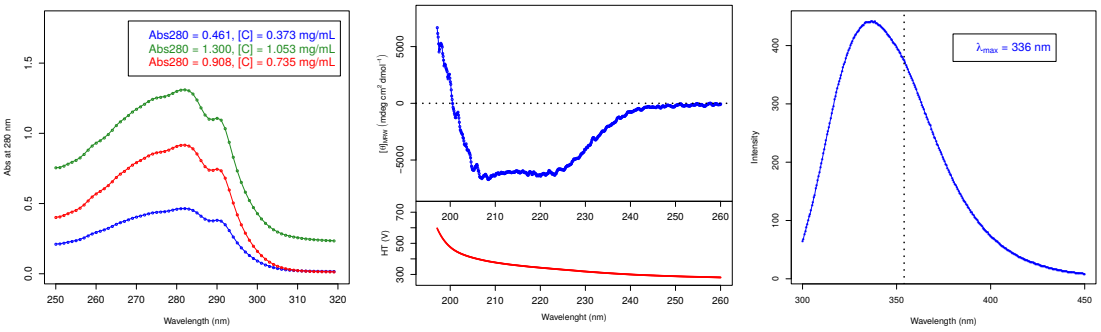
(a) **(b)**

Figure 2.83: Purification of OctaVII_04 WS, desalting

(a) Elution profile of the desalting of the pool of the refolded OctaVII_04 WS: the blue line is the absorbance at 280 nm and the brown line is the conductivity. **(b)** SDS-PAGE of the desalting: lane **1** is the input sample, **2-6** the elution peak.

Biophysical characterization

The fractions containing OctaVII_04 WS after desalting were pooled together, centrifuged in order to remove possible aggregates and analyzed by absorbance spectrometry (line blue in Figure 2.84a). The sample was then concentrated by ultrafiltration (green line) and centrifuged at high speed and filtered (red line). The values at Abs_{280} are used to calculate protein concentration, that is 0.73 mg/mL, a much higher value compared to the concentrations of OctaVII_04 and OctaVII_04 NoCys.



(a) **(b)** **(c)**

Figure 2.84: Biophysical characterization of OctaVII_04 WS

(**a**) Absorption spectrum of OctaVII_04 WS after desalting (blue), after concentration (green) and after centrifugation and filtration (red). Abs_{280} is used to calculate the protein concentration. (**b**) CD spectra of the protein (top) and high-tension (bottom); the dotted line indicates the baseline at $[\Theta]$=0. (**c**) Emission fluorescence spectrum of OctaVII_04 WS; the dotted line at 354 nm indicates the theoretic maximum for unfolded proteins.

The sample was used for analysis by far-UV circular dichroism and fluorescence measurements (Figure 2.84). The CD signal presents minima at 222 nm and 208 nm, as expected for a protein with high helical content. Analysis of the spectrum by CDpro,

using the program CDSSTR, indicates that the sample contains 34.4% of helix and 30.3% of strands. These results are not in agreement with the DSSP analysis of the model, that it is supposed to contain 48% of helix and 20% of strands. However, the content of helix in OctaVII_04 WS is similar to the one of OctaVII_04 NoCys (33.7%). On the contrary, the content of strands and of unstructured regions are 15% higher and 15% lower, respectively, in the mutant compared to the wild-type. These results suggests that in OctaVII_04 WS, part of the unstructured regions of OctaVII_04 NoCys folds mainly in β-strands, passing from 13.3% to 30.3%.

	% α-Helix	**% β-Strand**	**% Turn**	**% Unstruc**
CDpro (CDSSTR)	34.4	30.3	16.0	13.1
DSSP	48.3	20.4	15.4	15.8

Table 2.10: Secondary structures content, OctaVII_04 WS

Analysis of the tertiary structure of OctaVII_04 WS (see Figure 2.84c), showed a maximum in intensity centered at 336 nm, suggesting that the protein is folded and that the aromatics are not exposed to the solvent.

As for the previous OctaVIIs, OctaVII_04 WS was concentrated by ultrafiltration in order to perform near-UV CD analysis. However, the protein precipitated at a concentration lower than 1 mg/mL. Also OctaVII_04 WS is discarded for its low solubility.

2.4.8 OctaVII_05

OctaVII_05 has 249 residues and it is the best representative of Family 26 according to the Rosetta energy score (details in Section 2.3.9, page 68). The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_05 was manipulated for sequencing and storage as described for OctaVII_01 in Section 2.4.1, page 77. The results of the sequencing confirmed the correct sequence of the OctaVII_05 gene (see Annex 6.4, page 241), that was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials.

Expression Trials

Expression trials of OctaVII_05 were done in two conditions: overnight induction at 18°C and 4 hours induction at 37°C (Figure 2.85). The SDS-PAGE gel shows the total (T), the insoluble (I) and the soluble (S) fractions of the crude extract after sonication. OctaVII_05 is visible in the total fraction and in the pellet of the condition at 37°C, but not in the condition at 18°C. Multiple trials were done with different clones and transformations, but the protein was never expressed at lower temperature. The reasons of this particular phenomenon were not further inquired and large volume production of OctaVII_05 cultures was performed at 37°C only.

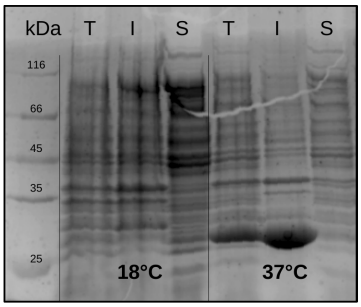


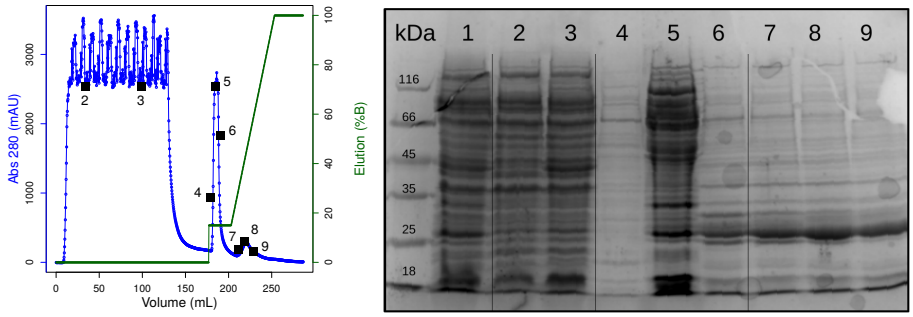
Figure 2.85: Expression trials of OctaVII_05

(**a**) Expression trials of OctaVII_05 with overnight induction at 18°C (left) and at 37°C for 4 hours (right); for each condition the total (T), the soluble (S) and the insoluble (I) fractions of the crude extract are reported.

Purification of the soluble fraction

The soluble fraction of the crude extract of 1 L cell culture of OctaVII_05 was filtered and loaded on HisTrap HP column for purification. The input sample is shown in lane

1 of Figure 2.86b. A small band is visible at ∼27 kDa, the theoretic molecular mass of OctaVII_05. The chromatogram of the purification is shown in Figure 2.86a. The majority of the proteins elutes in the flow-through, but two peaks are visible during the imidazole gradient: a first and higher one at 75 mM of imidazole and a second and smaller one at the beginning of the gradient from 15% of imidazole to 100%. Fractions of both peaks were loaded on SDS-PAGE gels and are shown in lanes 4 to 6 and 7 to 8, respectively. The first peak is mainly composed of contaminants at higher and lower molecular mass, but a quite intense band is visible at the correct size for OctaVII_05. The second peak contains mainly OctaVII_05, but some contaminants are still present.



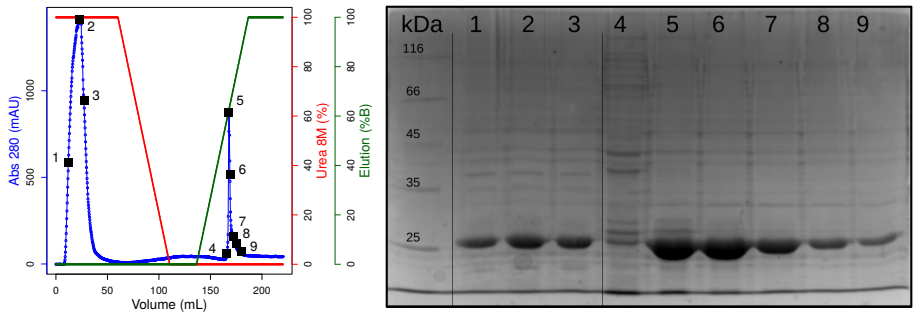
(a) (b)

Figure 2.86: Purification of OctaVII_05, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_05: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-3** the flow-through **4-6** a first elution peak at imidazole 75 mM; **7-9** a second elution peak.

Purification of the insoluble fraction

The pellet of the crude extract obtained from 1 L of cell culture was washed twice prior to resuspension in the buffer urea 8 M. The resuspension of the inclusion bodies in a buffer containing urea 8 M was performed at room temperature overnight and the sample is then centrifuged and filtered. The refolding and the purification of the protein is done with an HisTrap HP column and the chromatogram is shown in Figure 2.87a. Part of the protein was eluted in the flow-through (lanes 1-3 in Figure 2.87b), suggesting that the column reached saturation. The protein was not eluted during the refolding step (60 to 110 mL of elution volume), but during the imidazole gradient at a concentration of ∼ 300 mM. The protein resulted highly pure in the SDS-PAGE gel (lanes 4 to 9).



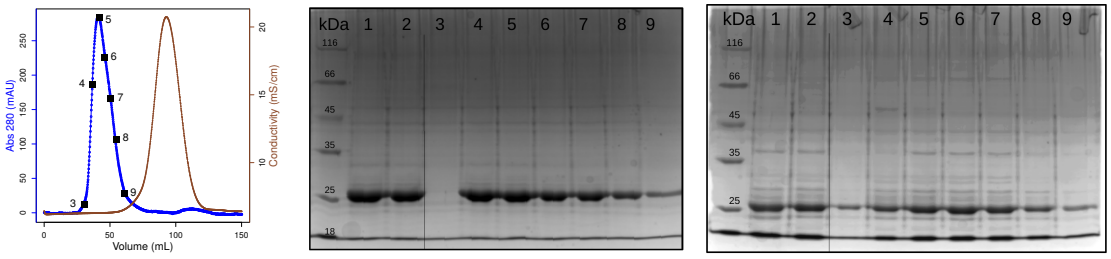
(a) (b)

Figure 2.87: Purification of OctaVII_05, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_05: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lanes **1-3** there is the flow-through and in lanes **4-9** the elution peak.

Desalting

The fractions containing OctaVII_05 after purification of the soluble fraction were pooled together and termed *Pool Sol*, lane 1 in Figure 2.88c, and the ones obtained by refolding of the inclusion bodies were pooled together and termed *Pool Insol*, lane 1 of Figure 2.88b. Both samples were centrifuged and filtered (lanes 2). The samples were loaded on the desalting column for removal of imidazole and the chromatogram relative to this step is shown for *Pool Insol* only, in Figure 2.88a. Lanes 3 to 9 in both SDS-PAGE gels show the fractions obtained in the elution peaks.



(a) (b) (c)

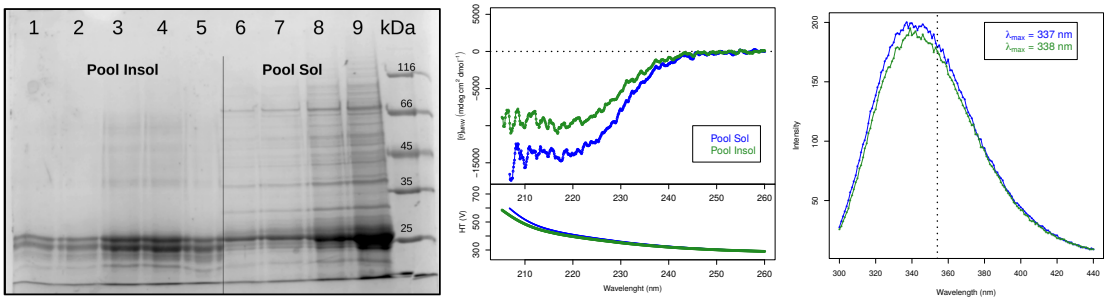
Figure 2.88: Purification of OctaVII_05, desalting

(**a**) Elution profile of the desalting of the pool of OctaVII_05 *Pool Insol*: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), its SDS-PAGE gel. (**c**), SDS-PAGE of the desalting of *Pool Sol*. For both samples, lanes **1** are the input samples before centrifugation; **2** the inputs after centrifugation and filtration; **3-9** the elution peaks.

Biophysical characterization

After desalting, the fractions were pooled together in *Pool Insol* and in *Pool Sol*, lanes 1 and 6, respectively, in Figure 2.89a). *Pool Insol* showed signs of degradation: two bands are visible at around 27 kDa and 25 kDa. The time interval between the desalting (Figure 2.88b) and the SDS-PAGE (Figure 2.89a) was of three days only, suggesting that the protein is prone to degradation. Both samples were concentrated by 3 steps of ultrafiltration (lanes 7-9 and 2-4, respectively), and centrifuged at high speed in order to remove aggregates. They were then diluted at 0.1 mg/mL and 0.01 mg/mL for CD and fluorescence analysis, respectively.

CD spectra for both OctaVII_05 *Pool Sol* and *Pool Insol* are shown in Figure 2.89b. The data-points collected when the high-tension is higher than 600 V are discarded. OctaVII_05 *Pool Sol* results to have signal at 222 nm of around -15000 mdeg cm^2 mol^{-1}, which is quite unlikely. It is possible that the high number of contaminants in the sample (lane 9 in Figure 2.89a), may affect the calculation of the concentration of the protein, and thus affect the calculation for the residue molar ellipticity. However, both samples show the typical signal for helices at 222 nm. The spectra are not reaching wavelengths lower than 205 nm and it is not possible to calculate the percentages of secondary structures. The samples are probably contaminated by low amounts of DTT or imidazole that disturb the CD analysis.



(a) **(b)** **(c)**

Figure 2.89: Biophysical characterization of OctaVII_05

(**a**) Concentration steps for *Pool Insol* and *Pool Sol* of OctaVII_05: lanes **1,6** the pool after desalting; **2,7**, first step of concentration by ultrafiltration; **3,8**, second step; **4,9**, third step and **5**, sample after centrifugation and filtration. (**b**) CD spectra of the two protein samples (top) and high-tension (bottom); the dotted line indicates the baseline at [Θ]=0. (**c**) Emission fluorescence spectra of the protein samples. The dotted line at 354 nm indicates the theoretic maximum for unfolded proteins.

Analysis of the tertiary structure of both samples was done by fluorescence (Figure 2.89c). The two spectra are very similar, with just a small difference in the maximum intensity that is higher in the soluble OctaVII_05. Again this difference may just be caused by an error in the quantification of the protein due to the presence of many contaminants.

The two λ_{max} are centered at 337 and 338 nm, respectively, indicating that the aromatic residues of the proteins are not exposed to the solvent.

Disulfide interactions

The presence of OctaVII_05 in the soluble fraction of the cell extract is really promising, but we do not know if the protein spontaneously folds in low percentages in the cytoplasm or if there are other reasons that can explain this behavior. OctaVII_05 bears 3 free cysteines and a possible explanation for its presence in the soluble fraction may be the formation of disulfide bonds with soluble proteins of the cell. In order to verify if the protein is covalently bound to soluble proteins we performed a simple test on both the OctaVII_05 obtained from the soluble fraction (*Pool Sol*) and the refolded one (*Pool Insol*). The purified proteins were centrifuged at high speed in order to precipitate possible aggregates. The supernatant was separated from the pellet, which was then resuspended in the same volume. For each pool of OctaVII_05, three samples were taken for SDS-PAGE analysis: the first, from the supernatant, was treated with a loading buffer that does not contain β-mercaptoethanol (lanes 1 in Figure 2.90). The other two were taken from the supernatant and from the resuspended pellet and were treated with the reducing agent (lanes 2 and 3, respectively).

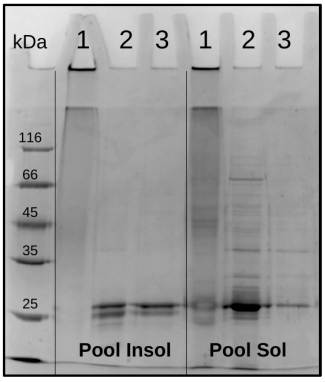


Figure 2.90: Disulfide bonds

Analysis of disufide bond formation for the *Pool Sol* and the *Pool Insol* of OctaVII_05. Lanes **1** the supernatant without β-mercaptoethanol, **2** the supernatant with β-mercaptoethanol and **3** the resuspended pellet with β-mercaptoethanol.

From the SDS-PAGE it is clear that both pools of OctaVII_05 are in a higher oligomerization state due to disulfide bonds since the samples without β-mercaptoethanol are not even entering into the gel (dark band at the beginning of the well in lanes 1). The same sample with the reducing agent clearly shows the presence of the protein at the correct molecular mass. In the case of the soluble OctaVII_05 also all the contaminant are visible

in lane 2. This experiment seems to confirm that the presence of OctaVII_05 in the soluble fraction of the crude extract is due to covalent binding with soluble protein. These proteins seems also to prevent the precipitation of OctaVII_05 since its amount in the resuspended pellet (lane 3) is much lower in comparison to the refolded OctaVII_05.

In normal conditions of growth, the cytoplasm should be in a constant reduced state thanks to the thioredoxin and the glutaredoxin systems, that use NADPH as source of reducing power [131]. However, the production of OctaVII_05 is a stressful condition for the cell: first, the cell is forced by external induction to overproduce the protein (which is not even useful for the cellular metabolism!). This over-expression is so intense that usually after just one hour of induction the target protein is the main band on the SDS-PAGE. Second, OctaVII_05 has 3 free cysteines that have to be reduced during the over-expression and perhaps synthesized in response to the over-expression. Third, the protein is actively segregated in inclusion bodies just after translation. The energetic resources of the cell may drastically diminish during the induction, and it may be that the NADPH levels in the cytoplasm are too low to avoid the formation of disulfide bonds between OctaVII_05 and endogenous proteins. Moreover, this covalent interaction between proteins seems to be inaccessible to reducing agents such as DDT, freshly added in all the buffers for protein extraction and purification.

Free cysteines are present in the first 5 OctaVIIs, from OctaVII_01 to OctaVII_05. OctaVII_01 and OctaVII_03 did not show this behavior because the first was never expressed and the second was discarded prior to purification. OctaVII_02 is purified in high amount from the soluble fraction (Figure 2.60), and, in a much lower amount, OctaVII_04 also is purified from the soluble fraction (Figure 2.66). These results suggest that OctaVII_02 and OctaVII_04 may share the same phenomena found in OctaVII_05, however we did not tested it. To exclude the hypothesis of spontaneous folding of OctaVII_05 in the soluble fraction of the protein, and possibly avoid aggregation due to disulfide formation, we created a mutant without cysteins of OctaVII_05, called OctaVII_05 NoCys.

2.4.9 OctaVII_05 NoCys

OctaVII_05 NoCys is the mutant of OctaVII_05 that contains no cysteines. As for the couple OctaVII_04 and OctaVII_04 NoCys, the choice of residues to replace the 3 original cysteines of OctaVII_05 was done in collaboration with the laboratory of Dr. Wim Vranken at the VUB. The HMM methodology, (described in Section 2.4.6, page 99), suggested one mutation for C40 (threonine), and two possible mutations for C126 (serine or alanine), and C184 (threonine or valine). As for OctaVII_04, we created the four possible combinations and we tested each of them with both HMM (that uses the amino acid sequence as input) and Rosetta (that uses a 3D model structure as input). The preparation of the 4 models is described in Section 3.3.4, page 182. The scores of HMM and Rosetta are reported in Table 2.11 and the normalized values (against OctaVII_04) are plotted in Figure 2.91.

Name	Mutations	HMM score	Rosetta score
OctaVII_05	C40T, C126, C184	31.33	-488.99
Mutant_1	C40T, C126S, C184T	31.02	-488.31
Mutant_2	C40T, C126S, C184V	31.85	-486.68
Mutant_3	C40T, C126A, C184T	31.04	-487.92
Mutant_4	C40T, C126A, C184V	31.89	-489.64

Table 2.11: Design of OctaVII_05 NoCys

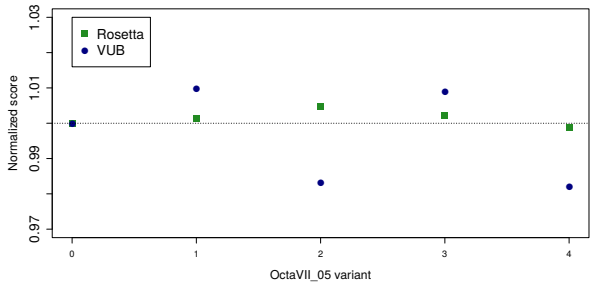


Figure 2.91: Cysteines substitution of OctaVII_05

Normalized scores for the 4 mutants of OctaVII_05 according to HMM (blue) and Rosetta (green) scores. The numbers 1 to 4 on the x-axis represent the 4 mutants and 0 is the original OctaVII_05.

The scores of OctaVII_05 improved in 2 combinations out of the 4 (Mutant_2 and Mutant_4) according to HMM, and in 1 out of 4 (Mutant_4) according to Rosetta. Among the 4 variants, we chosed Mutant_4, that has the best score according to both software. It is renamed OctaVII_05 NoCys and it has 250 residues. The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_05 NoCys was manipulated for sequencing and storage as described for OctaVII_01 in Section 2.4.1, page 77. The results of the sequencing confirmed the correct sequence of the OctaVII_05 NoCys gene (see Annex 6.4, page 241), that was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials.

Expression Trials

Expression trials for OctaVII_05 NoCys were done in two conditions: induction at 37°C for 4 hours and at 18°C overnight. As for OctaVII_05, the protein was highly expressed at 37°C (lane 1 in Figure 2.92), but not at 18°C (lane 4). The protein was produced in inclusion bodies (lane 3), but it is not clear if a small amount is present in the soluble fraction. As for the other OctaVIIs, we searched for the target protein in the soluble fraction.

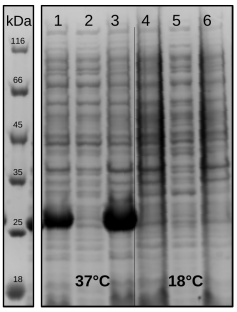
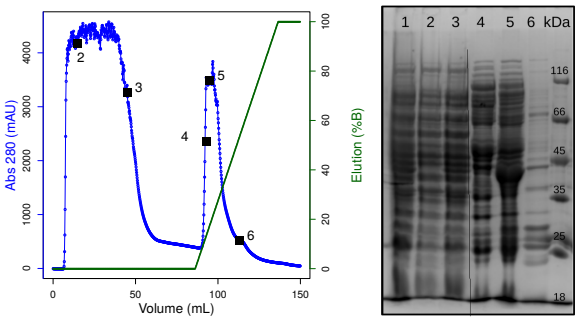


Figure 2.92: Expression trials of OctaVII_05 NoCys

Expression trials of OctaVII_05 NoCys at 37°C with 4 hours induction (left) and at 18°C overnight (right). Lanes **1,4** show the total fraction of the crude extract, **2,5** the soluble one and **3,6** the insoluble one.

Purification of the soluble fraction

OctaVII_05 NoCys was produced in 1 L of cell culture. After cell disruption the crude extract was centrifuged in order to separate the surnatant (soluble fraction) from the inclusion bodies (insoluble fraction), and filtered. The input sample (lane 1 in Figure 2.93b) was loaded on HisTrap HP columns for purification. Its chromatogram is reported in Figure 2.93a. The majority of the proteins are not binding the column and are eluted in the flow-through (lanes 2-3). The elution peak is eluted at ∼100 mM of imidazole, which is a low concentration for the dissociation of the 6x HisTag (250 mM). The peak is composed by contaminants at all molecular masses as shown in lanes 4-6 of the SDS-PAGE gel. OctaVII_05 NoCys is not present in the elution peak.



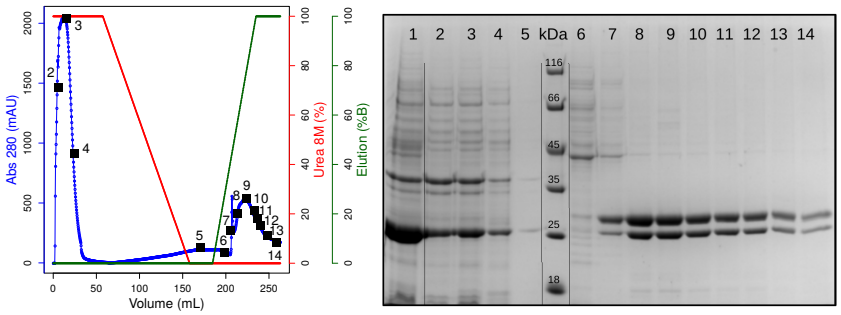
(a) (b)

Figure 2.93: Purification of OctaVII_05 NoCys, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_05 NoCys: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-3** the flow-through and **4-6** the eluction peak.

Purification of the insoluble fraction

The pellet (inclusion bodies) obtained from 1 L of cell culture was washed twice and dissolved in urea 8 M at room temperature overnight. The solution was centrifuged, filtered and loaded on HisTrap HP column for refolding and purification (chromatogram in Figure 2.94a).



(a) (b)

Figure 2.94: Purification of OctaVII_05 NoCys, insoluble fraction

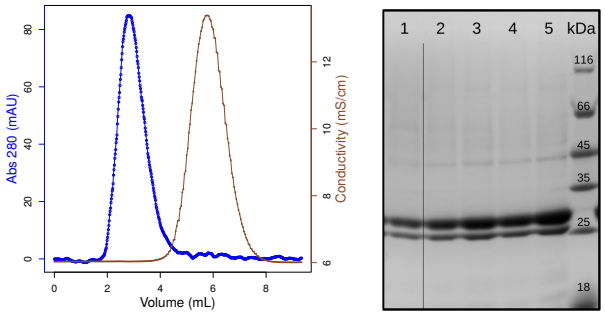
(**a**) Elution profiles of the insoluble fraction of OctaVII_05 NoCys: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-4** the flow-through; **5** a post-refolding fraction **6-14** the elution peaks.

The input sample is shown in lane 1 of Figure 2.94b: OctaVII_05 NoCys is the main band on the gel, but contaminants are present especially at high molecular mass. The majority of them is eluted in the flow-through (lanes 2-3), together with part of the protein, suggesting that the column reached saturation. The refolding of the protein (60 to 160 mL) did not cause dissociation of the HisTag from the matrix, and its elution

presents 2 peaks: a sharp one at 150 mM of imidazole that is composed by contaminants (lane 6 on the SDS-PAGE gel), and a second and wider peak (lanes 7 to 14). The SDS-PAGE for the elution shows two bands, one at the size of OctaVII_05 NoCys and the second at higher molecular mass. This is presumably still OctaVII_05 NoCys due to its high concentration. Also OctaVII_05 shows a double band after refolding and desalting, but it is not clear why it happens.

Desalting

The pool of the refolded OctaVII_05 NoCys was centrifuged and filtered prior to loading on the desalting column. The chromatogram of the purification is shown in Figure 2.95a. After the purification the fractions containing OctaVII_05 NoCys were pooled together (lane 1 in Figure 2.95b) and concentrated through 4 steps of ultrafiltration with Amicon filters. Each step was done by centrifugation at 4000 rpm for 5 minutes and they are shown in lanes 2-5 in the SDS-PAGE gel. After the third step of concentration, the sample was centrifuged at 14000 rpm for 20 minutes in order to remove the aggregates that were visible in the sample (lane 4) . The supernatant was then concentrated a last time and filtered on 0.22 μm filters. The double band of OctaVII_05 NoCys is still visible through all the steps of concentration.



(a) **(b)**

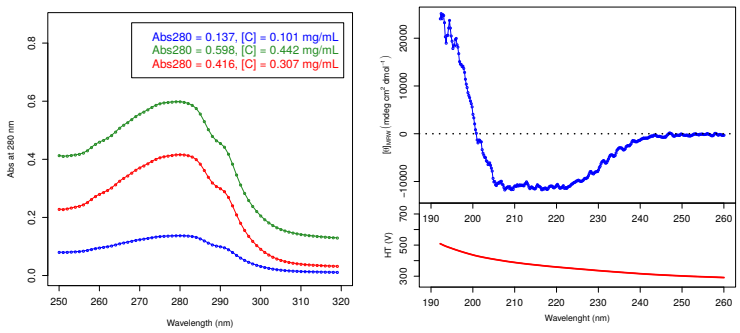
Figure 2.95: Purification of OctaVII_05 NoCys, desalting

(**a**) Elution profile of the desalting of the pool of the refolded OctaVII_05 NoCys: the blue line is the absorbance at 280 nm and the brown line is the conductivity; (**b**), the SDS-PAGE gel of the concentration steps of the protein. Lane **1** the pool after desalting; **2** first concentration step; **3** second step; **4** third step followed by centrifugation and **5** forth step of concentration followed by filtration.

Biophysical characterization

The outcome of the concentration steps described in the previous section are shown in Figure 2.96a: the blue line is the pool after desalting, the green line the third step of

concentration by ultrafiltration (lane 4 in the SDS-PAGE) and the red line the pool after centrifugation at high speed and filtration. The final concentration is 0.3 mg/mL.



(a) **(b)**

Figure 2.96: Biophysical characterization of OctaVII_05 NoCys

(**a**) Absorption spectrum of OctaVII_04 WS5 NoCys after desalting (blue), after concentration (green) and after centrifugation and filtration (red). Abs_{280} is used to calculate the protein concentration. (**b**) CD spectra of the protein (top) and high-tension (bottom); the dotted line indicates the baseline at $[\Theta]$=0.

The analysis by CD is shown in Figure 2.96b, where the typical signal of the helix is present with minima at 222 nm and 208 nm. The analysis of the spectrum by CDpro, using the program CDSSTR, indicates a content of 40% helix and 17% strands, which is more in agreement with the DSSP analysis compared to the other OctaVIIs: 48% helix and 19% strands. These results, however, should be carefully taken into account since the sample present a double band of proteins.

	% α-Helix	% β-Strand	% Turn	% Unstruc
CDpro (CDSSTR)	39.6	17.2	17.6	25.2
DSSP	47.9	18.7	17.0	17.5

Table 2.12: Secondary structures content, OctaVII_05 NoCys

The attempts to concentrate the protein for near-UV CD analysis however were not successful. As for the previous OctaVIIs, the protein did not reach 1 mg/mL of concentration without the formation of aggregates that were visible by eye. The low solubility and the presence of a double band after refolding are the reason why OctaVII_05 NoCys is discarded from further analysis.

2.4.10 OctaVII_06

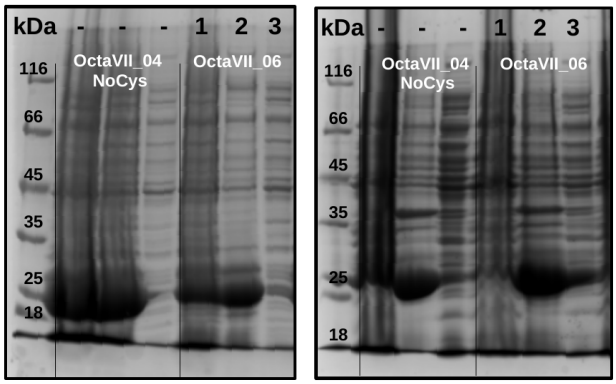
OctaVII_06 has 250 residues and it is the best representative of Family 06 according to the Rosetta energy score (details in Section 2.3.9, page 68). It is also one of the best model according to molecular dynamics. The gene was synthesized by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_06 (gBlock®), was digested with the restriction enzymes *NcoI* and *XhoI* in parallel with an empty pET28a vector (see Section 4.2.2, page 190). After purification of both DNA (Section 4.2.3, page 191), the gene was inserted in the vector by ligation with the T4 ligase (Section 4.2.5, page 191). The plasmid was used for transformation in *E. coli* DG1 competent cells for sequencing. The results of the sequencing confirmed the correct sequence of the OctaVII_06 gene in 1 out of 4 clones. The plasmid bearing the good sequence was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials and in DG1 competent cell for stock preparation.

Expression Trials

Expression trials for OctaVII_06 were done in two conditions: induction with IPTG 1 mM at 37°C for 4 hours (Figure 2.97a) and at 18°C overnight (Figure 2.97b).



(a) **(b)**

Figure 2.97: Expression trials of OctaVII_06

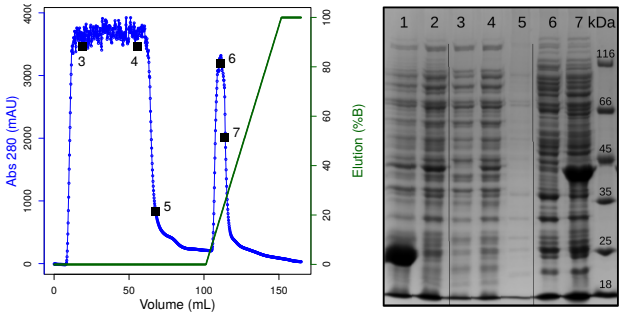
(**a**) Expression trials of OctaVII_04 NoCys (already discussed in Section 2.4.6) and of OctaVII_06 at 37°C for 4 hours and (**b**), at 18°C overnight: lanes **1** total fraction, **2** insoluble fraction and **3** soluble fraction.

In both cases the protein is highly expressed in inclusion bodies (lanes 2). The low quality of the gel does not allow us to understand if part of the protein is present in the

soluble fraction. So, as for the previous OctaVIIs, we tried to purify the protein from a bigger volume of culture.

Purification of the soluble fraction

OctaVII_06 was produced in 1 L of culture with overnight induction at 18°C. The crude extract obtained after disruption (lane 1 in Figure 2.98b) was centrifuged, filtered and loaded on the HisTrap HP column. The chromatogram of the purification is shown in Figure 2.98a and the SDS-PAGE in Figure 2.98b. The majority of the proteins are eluted in the flow-through (lanes 3 to 5) and the remaining one that bound the column are eluted at low concentration of imidazole. As shown in the SDS-PAGE, the elution peak is composed only by contaminants, and OctaVII_06 is not produced in the soluble fraction.



(a) **(b)**

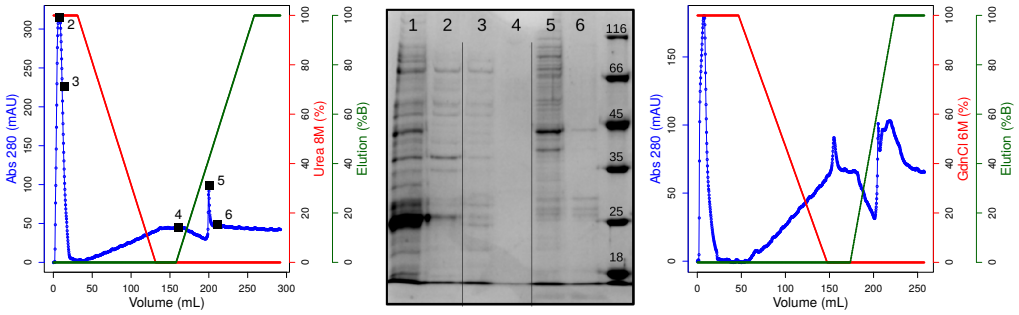
Figure 2.98: Purification of OctaVII_06, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_06: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the crude extract; **2** the input sample; **3-5** the flow-through; **6-7** the elution peak.

Purification of the insoluble fraction

The pellet of 1 L culture was washed twice and resuspended in buffer with urea 8 M in order to dissolve the inclusion bodies. The solution was stirred overnight at room temperature, and centrifuged the day after. Differently from the other OctaVIIs, the inclusion bodies of OctaVII_06 did not completely dissolve in urea and a significant amount of pellet was present after centrifugation. The sample in urea before and after centrifugation is shown in lanes 1 and 2, respectively, in Figure 2.99b. The supernatant was anyway filtered and loaded on the HisTrap HP column but only few contaminants were eluted (Figure 2.99a). The increase in absorbance during the refolding step is due to the buffer exchange. It is present in all the purification from the insoluble fraction, but here is more

noticeable because the absorbance scale is short (from 0 mAU to 300 mAU). The denaturation of the inclusion bodies in urea 8 M was tried twice, and we obtained the same negative results. A last trial to dissolve the pellet was done with a buffer containing 6 M guanidinium chloride (GdmCl) as denaturing agent instead of urea and it was possible to purify a small amount of protein. The chromatogram of the purification of the sample in GdmCl is shown in Figure 2.99c. Since the GdmCl and the SDS-PAGE technique are incompatible, the gel of the second purification is not shown.



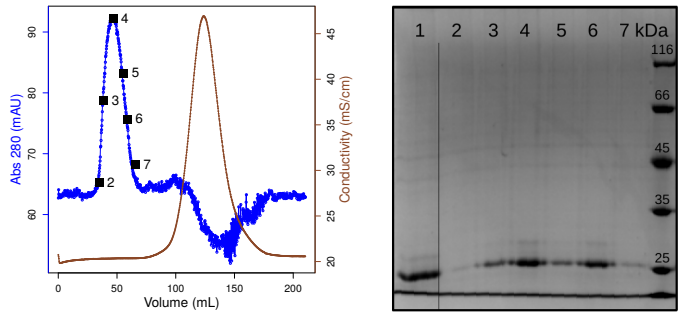
(a) **(b)** **(c)**

Figure 2.99: Purification of OctaVII_06, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_06 dissolved in a buffer with urea 8 M: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-4** the flow-through; **5-6** the elution peak. (**c**) Elution profiles of the insoluble fraction of OctaVII_06 dissolved in a buffer with GdnCl 6 M.

Desalting

The fractions collected after the purification of the OctaVII_06 in GdmCl were pooled together and centrifuged in order to remove possible aggregates. The sample was then filtered (lane 1 in Figure 2.100b) and loaded on the desalting column to remove the imidazole. The elution peak is shown in the chromatogram (Figure 2.100a) and in lanes 2 to 7 of the SDS-PAGE gel.



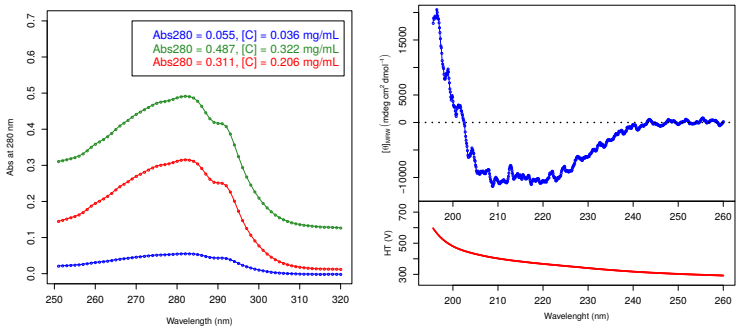
(a) (b)

Figure 2.100: Purification of OctaVII_06, desalting

(**a**) Elution profile of the desalting of the pool of the refolded OctaVII_06: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the input sample; **2-7** the elution peak.

Biophysical characterization

The fractions containing the refolded OctaVII_06 were pooled together and centrifuged in order to remove possible aggregates. The sample was used to measure by absorbance the protein concentration, which is extremely low: 0.036 mg/mL (blue line in Figure 2.101a). 8 steps of concentration by ultrafiltration were done until precipitates were visible in solution (green line). The sample was centrifuged at high speed and filtered (red line). The final concentration was 0.2 mg/mL, enough for analysis by far-UV circular dichroism. The CD signal is shown in Figure 2.101b, and presents the usual curve for proteins containing helices.



(a) (b)

Figure 2.101: Biophysical characterization of OctaVII_06

(**a**) Absorption spectrum of OctaVII_06 after desalting (blue), after concentration (green) and after centrifugation and filtration (red). Abs_{280} is used to calculate the protein concentration. (**b**) CD spectra of the protein (top) and high-tension (bottom); the dotted line indicates the baseline at $[\Theta]$=0.

The analysis by CDpro, using the program CDSSTR, is shown in Table 2.13, and as for the previous OctaVIIs there is not an agreement in the content of secondary structures

between the experimental protein and its model. Although the strand content is similar (20.8% and 17.5%, respectively), the helix one is different of about 10% (37.8% and 48.3%).

	% α-Helix	**% β-Strand**	**% Turn**	**% Unstruc**
CDpro (CDSSTR)	37.8	20.8	15.7	25.4
DSSP	48.3	17.5	15.0	19.7

Table 2.13: Secondary structures content, OctaVII_06

The protein is also poorly soluble because attempts to reach higher concentration ended with its precipitation. For these results, also OctaVII_06 is discarded from further characterization.

2.4.11 OctaVII_07

OctaVII_07 has 250 residues and it is the best representative of Family 12 according to the Rosetta energy score (details in Section 2.3.9, page 68). The gene was synthesized by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_07 (gBlock®), was manipulated for digestion, ligation, transformation and sequencing as described for OctaVII_06 in Section 2.4.10, page 129. The results of the sequencing confirmed the correct sequence of the OctaVII_07 gene in 1 clone out of 9. The plasmid bearing the good sequence was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials and in DG1 competent cell for stock preparation.

Expression Trials

Expression trials for OctaVII_07 were done in two conditions: induction for 4h at 37°C (Figure 2.102) and overnight induction at 18°C (data not shown). In both cases there is over-expression of the protein, but the molecular mass seems to be lower than expected: 25 kDa versus the theoretic 27.8 kDa.

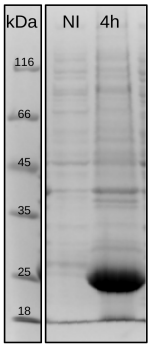


Figure 2.102: Expression trial of OctaVII_07

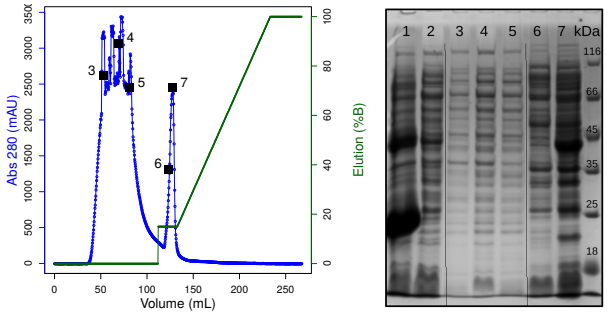
SDS-PAGE of OctaVII_07 before and after induction at 37°C. The non induced sample (NI) is compared with the sample induced for 4 hours (4h).

Purification of the soluble fraction

OctaVII_07 was produced in 1 L culture with induction at 18°C overnight. After cell disruption, the crude extract was centrifuged in order to separate the soluble fraction from the insoluble one. In lanes 1 and 2 of Figure 2.103b the crude extract is shown before and after the centrifugation step. OctaVII_07 is highly produced but mainly in

inclusion bodies. Again, its size appears to be also lower than 25 kDa, suggesting that it is truncated as OctaVII_02 and OctaVII_03. However, the soluble fraction of the protein was loaded on the HisTrap HP column for purification (the chromatogram is shown in Figure 2.103a). The majority of the proteins were eluted in the flow-through (lanes 3 to 5 in the SDS-PAGE), and the few ones that bound the column were eluted at 75 mM of imidazole, a concentration too low for the dissociation of the HisTag from the matrix. The SDS-PAGE gel confirmed that the elution peak is composed only by contaminants of the crude extract and the protein is not visible in the elution fractions.

OctaVII_07, as OctaVII_02 and OctaVII_03, is discarded from further analysis for its low molecular mass.



(a) **(b)**

Figure 2.103: Purification of OctaVII_07, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_07: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the crude extract; **2** the input sample; **3-5** the flow-through; **6-7** the elution peak.

2.4.12 OctaVII_08

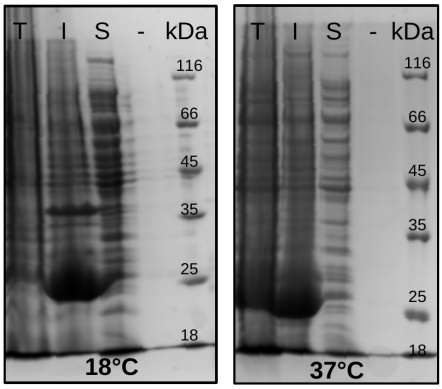
OctaVII_08 has 250 residues and it is the best representative of Family 13 according to the Rosetta energy score (details in Section 2.3.9, page 68). The gene was synthesized by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_08 (gBlock®), was manipulated for digestion, ligation, transformation and sequencing as described for OctaVII_06 in Section 2.4.10, page 129. The results of the sequencing confirmed the correct sequence of the OctaVII_08 gene in 3 clones out of 4. The plasmid bearing the good sequence was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials and in DG1 competent cell for stock preparation.

Expression Trials

Expression trials for OctaVII_08 were done in 2 conditions: induction at 18°C overnight (Figure 2.104a) and at 37°C for 4 hours (Figure 2.104b). In both cases the protein is highly expressed in inclusion bodies (I). The low quality of the gel do not allow us to understand if part of the protein is present in the soluble fraction, however we will test it in the next section.



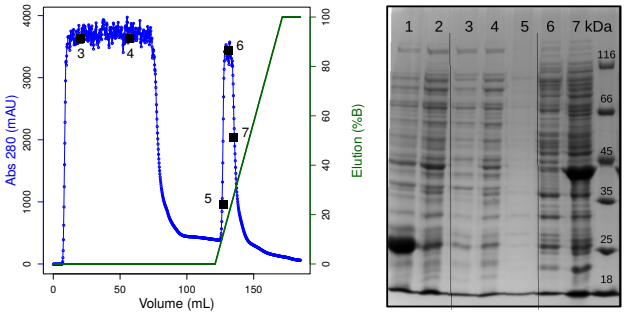
(a) **(b)**

Figure 2.104: Expression trials of OctaVII_08

(**a**) Expression trials of OctaVII_08 at 18°C overnight and (**b**) at 37°C for 4 hours: total fraction (T), insoluble fraction (I) and soluble fraction (S).

Purification of the soluble fraction

OctaVII_08 was produced in 1 L of culture with induction at 37°C for 4 hours. The crude extract obtained after disruption (lane 1 in Figure 2.105b) was centrifuged, filtered and loaded on the HisTrap HP column. The input sample (soluble fraction) is shown in lane 2 of the SDS-PAGE gel, and the chromatogram of the purification in Figure 2.105a. The majority of the proteins are eluted in the flow-through (lanes 3 to 5) and the remaining ones that bound the column are eluted at low concentration of imidazole. As shown in the SDS-PAGE, the elution peak is composed only by contaminants, and OctaVII_08 is not produced in the soluble fraction.



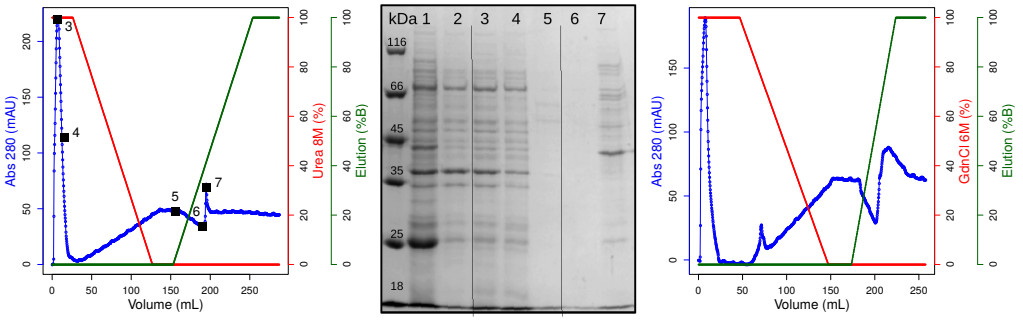
(a) **(b)**

Figure 2.105: Purification of OctaVII_08, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_08: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the crude extract; **2** the input sample; **3-5** the flow-through; **6-7** the elution peak.

Purification of the insoluble fraction

The pellet of 1 L culture was washed twice and resuspended in buffer with urea 8 M overnight at room temperature. After centrifugation, a consistent pellet was found in the tube, and as for OctaVII_06, OctaVII_08 did not completely dissolve in urea. The sample in urea before and after centrifugation is shown in lanes 1 and 2, respectively, in Figure 2.106b. The supernatant was anyway filtered and loaded on the HisTrap HP column but only few contaminants were eluted (Figure 2.106a). As for OctaVII_06, the pellet was dissolved in a buffer containing GdmCl 6 M and loaded again in the HisTrap column. Finally, a small amount of protein was purified. The chromatogram of the purification of the sample in GdmCl is shown in Figure 2.106c. Since GdmCl and the SDS-PAGE technique are incompatible, the gel of the second purification is not shown.



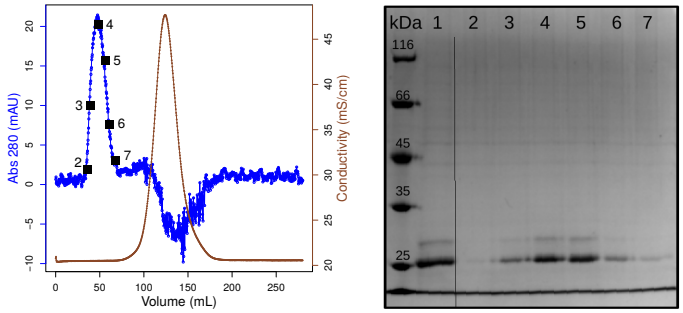
(a) **(b)** **(c)**

Figure 2.106: Purification of OctaVII_08, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_08 dissolved in a buffer with urea 8 M: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8M or GdnCl 6M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the sample before centrifugation; **2**, the input sample after centrifugation; **3-5** the flow-through; **6-7** the elution peak. (**c**) Elution profiles of the insoluble fraction of OctaVII_08 dissolved in a buffer with GdnCl 6 M.

Desalting

The fractions collected after the purification of the OctaVII_08 in GdmCl were pooled together and centrifuged in order to remove possible aggregates. The sample was then filtered (lane 1 in Figure 2.107b) and loaded on the desalting column to remove the imidazole. The elution peak is shown in the chromatogram (Figure 2.107a) and in lanes 2 to 7 of the SDS-PAGE gel. As for OctaVII_06, the negative peak centered at 140 mL is noticeable only because the scale of absorbance is short (20 mAU maximum), but it is always present in the desalting chromatogram.



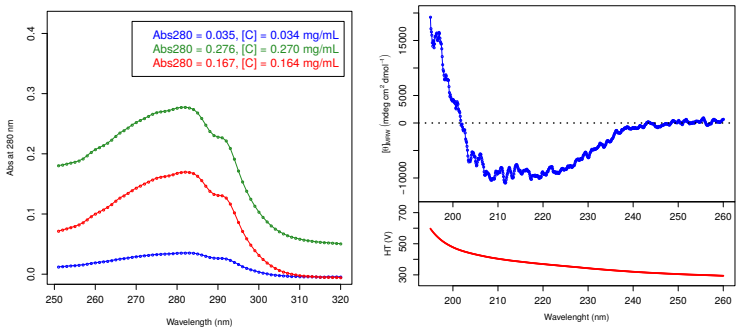
(a) **(b)**

Figure 2.107: Purification of OctaVII_08, desalting

(**a**) Elution profile of the desalting of the pool of the refolded OctaVII_08: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the input sample; **2-7** the elution peak.

Biophysical characterization

The fractions containing the refolded OctaVII_08 were pooled together and centrifuged in order to remove possible aggregates. The protein concentration was measured (blue line in Figure 2.108a), and it resulted low, as for OctaVII_06. Steps of concentration by ultrafiltration lead to the formation of aggregates (green line), that are precipitated by centrifugation at high speed and filtration (red line). OctaVII_08 reached a final concentration of 0.16 mg/mL that was enough for analysis by far-UV circular dichroism (Figure 2.108b). The spectrum is typical of a protein containing helices and its analysis by CDpro, using the program CDSSTR, indicates that the sample contains 37.2% of helices and 22% of strands. As for the previous OctaVIIs, the CD results are not in agreement with the DSSP analysis on the model of OctaVII_08 that should contain 48.7% of helices and 17.5% of strands.



(a) **(b)**

Figure 2.108: Biophysical characterization of OctaVII_08

(**a**) Absorption spectrum of OctaVII_08 after desalting (blue), after concentration (green) and after centrifugation and filtration (red). Abs_{280} is used to calculate the protein concentration. (**b**) CD spectra of the protein (top) and high-tension (bottom); the dotted line indicates the baseline at $[\Theta]$=0.

	% α-Helix	% β-Strand	% Turn	% Unstruc
CDpro (CDSSTR)	37.2	22.0	16.6	23.6
DSSP	48.7	17.5	14.6	19.1

Table 2.14: Secondary structures content, OctaVII_08

As for all the previous proteins, we were not able to reach a concentration higher that 1 mg/mL, and also OctaVII_08 is discarded from further analysis.

2.4.13 OctaVII_09

OctaVII_09 has 250 residues and it is the best representative of Family 23 according to the Rosetta energy score (details in Section 2.3.9, page 68). The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_09 (gBlock®), was manipulated for digestion, ligation, transformation and sequencing as described for OctaVII_06 in Section 2.4.10, page 129. The results of the sequencing confirmed the correct sequence of the OctaVII_09 gene in 1 clone out of 2. The plasmid bearing the good sequence was then used for transformation in *E. coli* BL21 (DE3) cells for expression trials and in DG1 competent cell for stock preparation.

Expression trials of OctaVII_09

Expression trials for OctaVII_09 were performed in two conditions: induction at 37°C for 4 hours (Figure 2.109) and at 18°C overnight (not shown). In both cases the protein is highly expressed in inclusion bodies (I). The low quality of the gel did not allow us to understand if part of the protein was present in the soluble fraction, however we looked for the target protein in the soluble fraction.

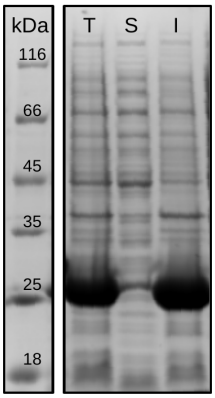


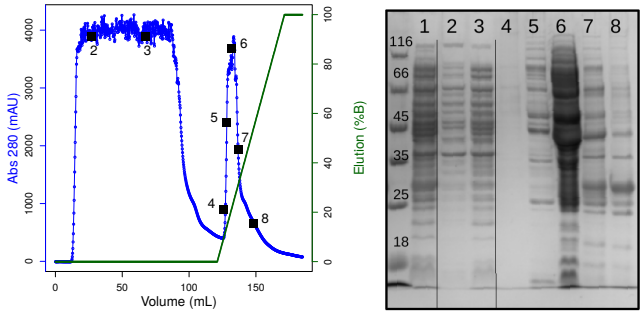
Figure 2.109: Expression trials of OctaVII_09

SDS-PAGE shows the expression of OctaVII_09 at 37°C for 4 hours for the fractions: total (T), soluble (S) and insoluble (I).

Purification of OctaVII_09, soluble fraction

OctaVII_09 was produced in 1 L of culture. The soluble fraction was separated from the insoluble one by centrifugation. After filtration the sample was loaded onto HisTrap

HP column for purification (line 1 of Figure 2.110b). The chromatogram is shown in Figure 2.110a. The flow-through is shown in lanes 2 and 3 and the elution peak in lanes 4 to 8. The peak is composed only by contaminants that elute at low concentration of imidazole (50 mM).



(a) (b)

Figure 2.110: Purification of OctaVII_09, soluble fraction

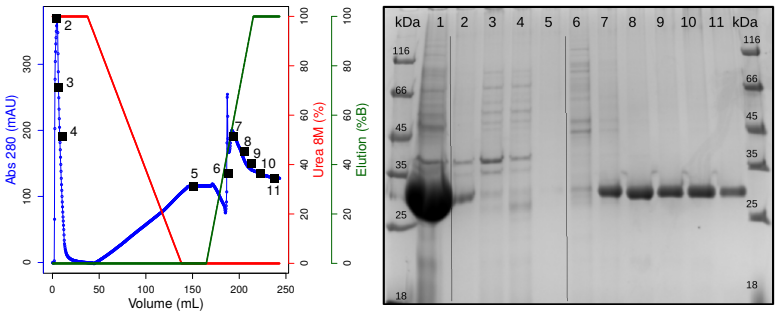
(**a**) Elution profiles of the soluble fraction of OctaVII_09: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-3** the flow-through and **4-8** the elution peak at imidazole 50-200 mM.

Purification of OctaVII_09, insoluble fraction

The insoluble fraction of the 1 L culture was washed twice and dissolved in buffer containing 8 M urea. The sample was centrifuged, filtered and loaded on the HisTrap HP column for refolding and purification. The input sample is visible in lane 1 of Figure 2.111b, OctaVII_09 is the main band but few contaminants are visible at higher molecular mass. The chromatogram of the purification is shown in Figure 2.111a: the protein completely binds the matrix and almost all the contaminant are removed with the flow-through (lanes 2-4). The refolding did not caused elution of OctaVII_09 (lane 5), that is then eluted starting from imidazole 200 mM. Except for the first sharp peak that contains contaminants, the protein seems to be the only band in lanes 7 to 11 of the gel.

Purification of OctaVII_09, desalting

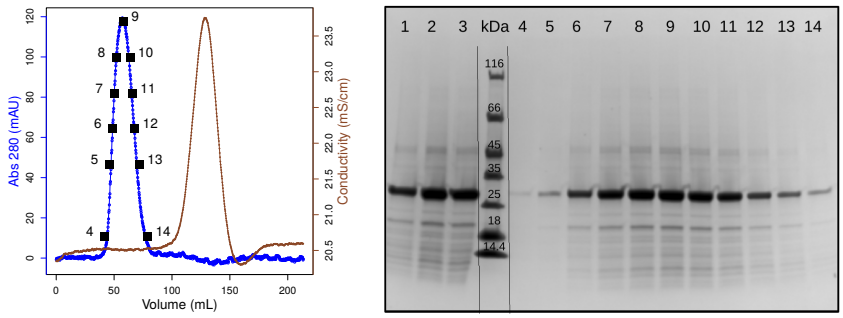
The fractions containing the refolded OctaVII_09 were pooled together and concentrated by ultracentrifugation in Amicon filters. After concentration the protein was centrifuged at 20000 rpm and filtered. The samples shown in Figure 2.112b were isolated before concentration (lane 1), after ultrafiltration (lane 2), and after centrifugation and filtration to remove aggregates (lane 3). The filtrated sample was loaded onto the desalting column and the chromatogram of the purification is shown in Figure 2.112a.



(a) (b)

Figure 2.111: Purification of OctaVII_09, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_09: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-4** the flow-through; **5** a fraction after the refolding and **6-11** the elution peak.



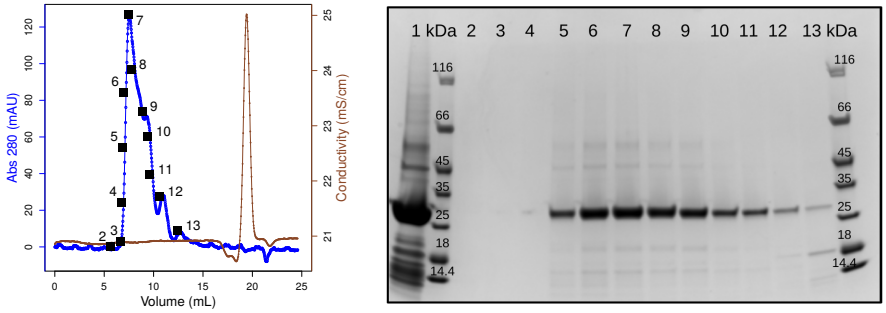
(a) (b)

Figure 2.112: Purification of OctaVII_09, desalting

(**a**) Elution profile of the desalting of the pool of the refolded OctaVII_09: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the pool of refolded OctaVII_09; **2** the same pool after concentration by ultrafiltration; **3** the input sample and **4-14** the elution peak.

Purification of OctaVII_09, size exclusion

The refolded protein was loaded on a size exclusion column to verify its oligomerization state, and the chromatogram is shown in Figure 2.113a. The calibration of the column was reported in Section 4.7.7, page 208. A main peak is eluted in the void volume, suggesting that OctaVII_09 has a molecular mass higher to 75 kDa. Two shoulders are visible at 10 mL and at 11.5 mL of elution volume, but analysis on gel (Figure 2.113b) shows that there are no differences among any fraction of the entire peak. The peak at 11.4 mL may however be composed by the monomer of OctaVII_09: the molecular mass that is calculated from the calibration curve is 30.2 kDa, that is in agreement with the size of the protein.



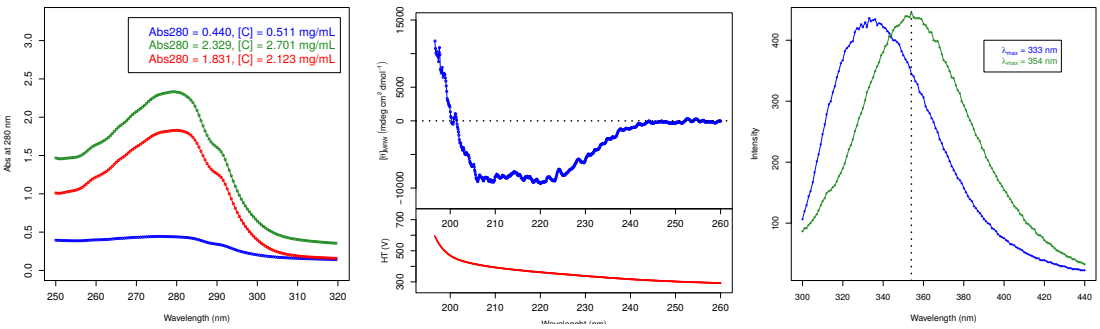
(a) **(b)**

Figure 2.113: Purification of OctaVII_09, size exclusion

(**a**) Size exclusion elution profile of the pool of the refolded OctaVII_09: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the input sample; **2-13** the elution peak.

Biophysical characterization

The refolded and desalted OctaVII_09 was concentrated by ultrafiltration. The protein concentration was measured by absorbance spectrometry in three conditions: before concentration (blue line in Figure 2.114a), after 5 cycles of ultrafiltration (green line), and after centrifugation at high speed to remove possible aggregates (red line).



(a) **(b)** **(c)**

Figure 2.114: Biophysical characterization of OctaVII_09

(**a**) Absorption spectrum of OctaVII_09 after desalting (blue), after concentration (green) and after centrifugation and filtration (red). Abs_{280} is used to calculate the protein concentration. (**b**) CD spectra of the protein (top) and high-tension (bottom); the dotted line indicates the baseline at $[\Theta]=0$. (**c**) Emission fluorescence spectra of OctaVII_09 in standard conditions and in GdmCl 6 M; the dotted line at 354 nm indicates the theoretic maximum for unfolded proteins.

For the first time in this work, the protein was stable in solution at a concentration higher than 1 mg/mL! The final concentration reached in this experiment was 2.1 mg/mL, but in later trials we were able to reach 4 mg/mL.

The sample was diluted for analysis by far-UV circular dichroism (Figure 2.114b) and

by intrinsic fluorescence (Figure 2.114c) measurements.

The CD signal is very similar to the ones obtained for the other OctaVIIs, presenting two minima at 208 nm and 222 nm, typical of proteins containing α-helices. The analysis of the spectrum by CDPro, with the program CDSSTR, indicates that the sample contains 30% of helix, 22.5% of strands and 20% of turns. The comparison of these values with the one obtained from the DSSP analysis on the model shows similar results with the other OctaVIIs: although the strand content is generally in agreement, the helix content in the sample is at least 10% lower than the expected one. In this case, the model has 47.5% of helices while the experimental sample contains only 30% of them.

	% α-Helix	% β-Strand	% Turn	% Unstruc
CDpro (CDSSTR)	29.2	22.5	20.0	27.8
DSSP	47.5	19.2	15.8	17.5

Table 2.15: Secondary structures content, OctaVII_09

For fluorescence analysis, the refolded protein was diluted in both standard and denaturing buffers (GdmCl 6 M) (Figure 2.114c). The λ_{max} is 333 nm for the native protein and 354 nm for the unfolded one. This 20 nm shift clearly indicates that in the native conformation the aromatic residues are not exposed to the solvent and that the protein is not in a molten globule state.

Thanks to the higher concentration of OctaVII_09 compared to the other artificial proteins, we were able to perform additional analysis, such as chemical unfolding, thermal unfolding and near-UV CD.

The chemical unfolding was followed by both CD and fluorescence. The protein was diluted in different concentrations of denaturing buffer using the pipetting workstation at the Robotein Platform (the robot).

For the CD analysis, only the denaturing agent GdmCl was used (urea is not suitable for the technique). The robot prepared 48 samples ranging from 0 to 5.5 M of GdmCl. The CD signal at 222 nm was then measured for each of the 48 samples for 60 seconds, and its average value was plotted against the concentration of denaturing agent (Figure 2.115a).

The signal at 222 nm is stable up to 2 M of GdmCl, suggesting that the folding of the protein is not affected below this concentration. The unfolding transition occurs between 2 M and 5 M and the signal stabilizes after 5 M of GdmCl. It is not clear why there is a big "jump" in the curve at 2 M GdmCl. Since OctaVII_09 is present in solution in a high oligomeric form, the jump in the transition may indicate dissociation

to a monomeric form. The same samples were recorded in the same conditions 15 days after the first measurement and the jump was not present anymore (data not shown). This evidence may be suggest that samples with less than 2 M GdmCl require more time of equilibration compared to samples with an high concentration of denaturant. This experiment was performed only once. However, the transition does not correspond to a cooperative unfolding one. CD spectra were recorded every 0.5 M of denaturing agent, and they are shown in Figure 2.115b.

The same 48 samples were used for fluorescence analysis, and their intensities at 318 nm are reported versus the concentration of GdmCl in Figure 2.115c. At that wavelength there is the major difference in the folded (0 M GdmCl) and unfolded (5.5 M) curves. In this experiment the signal is stable up to 1.5 M GdmCl and the transition is observed between 1.5 M and 5 M of denaturing agent, again showing a non-cooperative behavior. The same analysis was done on 48 samples of OctaVII_09 in different concentration of urea, ranging from 0 to 8 M (Figure 2.115e). The intensity at 318 nm is stable until 3 M of denaturant. The transition is again non-cooperative and spans from 3 M to 8 M of urea, suggesting that the protein is not completely unfolded at that concentration.

All the experiments of chemical unfolding shows that the protein is moderately resistant to the chemical agents, with stable signals up to 2 M GdmCl and 3 M of urea. The transition curves however are spanned on a large range of concentrations and do not show the typical sigmoidal curve of a cooperative unfolding.

For the thermal unfolding of OctaVII_09, the sample was heated from 25°C to 96°C and then cooled back to 25°C. Changes in the tertiary and secondary structures were followed by fluorescence (Figure 2.116a) and by far-UV CD (Figure 2.117a), respectively.

In the fluorescence analysis, the unfolding transition is shown in blue (from 25°C to 96°C). Although the intensity is decreasing proportionally to the temperature, there is not the typical transition of the thermal unfolding. The green line in the figure is the refolding curve (from 96°C to 25°C). Again the transition is not present, but the intensity increases proportionally to the decrease of temperature. These results may suggest that the protein recover some of its compactness after thermal unfolding. Regular fluorescence spectra are recorded before and after the unfolding and are shown in Figure 2.116b with the same color code. The native protein (blue line) has the λ_{max} centered at 338 suggesting that the 3D structure is compacted. After heating and cooling, the sample does not recover the same initial intensity (green line), and the λ_{max} is centered at 344 nm. There are two possible explanations for this result: in one case, the proteins recovered part of their folding but in a less compacted structure than the native one; in the second case, only a fraction of the sample recover the native folding, while the remaining fraction is

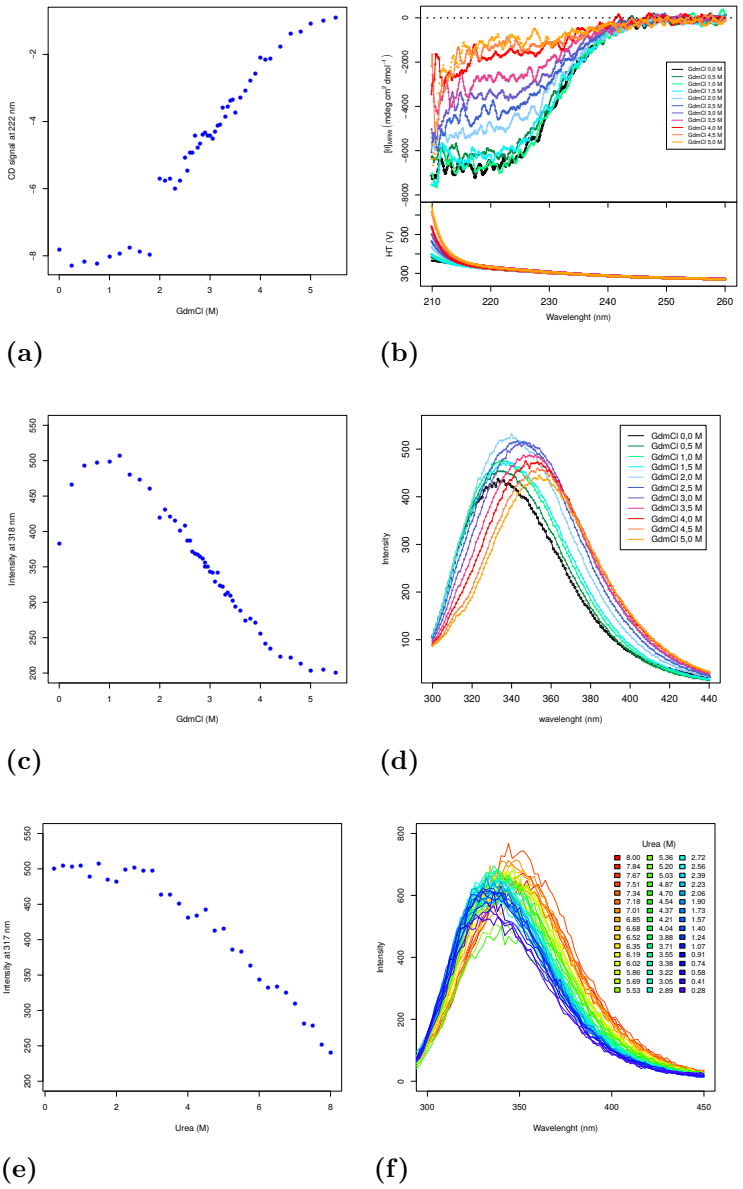
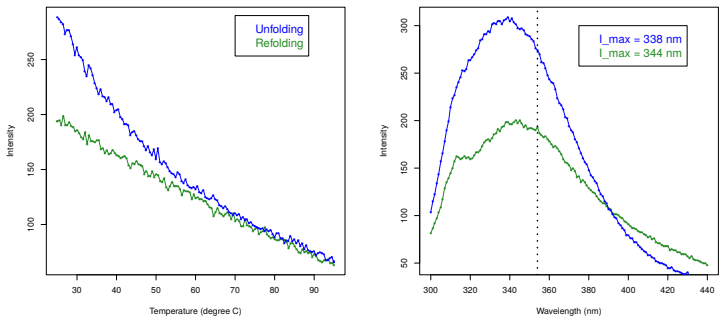


Figure 2.115: Chemical unfolding of OctaVII_09

(**a**) Chemical unfolding of OctaVII_09 in GdmCl followed by CD, and (**b**) their CD spectra. (**c**) Chemical unfolding of OctaVII_09 in GdmCl followed by fluorescence, and (**d**) their emission fluorescence spectra. (**e**) Chemical unfolding of OctaVII_09 in urea followed by fluorescence, and (**f**) their emission fluorescence spectra.

in an unfolded state. Although we do not know which case corresponds to our results, in both cases a partial re-folding is obtained.

Thermal unfolding of OctaVII_09 was followed also by CD at 222 nm, and the signal is shown in blue in Figure 2.117a. As for the fluorescence experiment, no transition is visible during the thermal unfolding. In order to force the unfolding of the protein, we repeated



(a) **(b)**

Figure 2.116: Thermal unfolding of OctaVII_09 by fluorescence

(**a**) Thermal unfolding (blue line) and refolding (green line) followed at 319 nm. (**b**) Emission fluorescence spectra obtained before (blue) and after (green) the thermal unfolding. The dotted line at 354 nm indicates the theoretic maximum for unfolded proteins.

the experiment with OctaVII_09 in 1, 2 and 3 M of GdmCl (green, red and orange curve, respectively). The unfolding transition is not visible in any of the 3 additional conditions. In particular the protein behaves in the same way for the conditions with 0, 1 and 2 M of denaturant, with a per-residue molar ellipticity of -8000 mdeg cm^2 mol^{-1} at 25°C and of -6000 mdeg cm^2 mol^{-1} at 90°C. The condition with 3 M GdmCl is affected by the presence of the denaturant already at 25°C, with a value of -6000 mdeg cm^2 mol^{-1} that decrease to around -4000 mdeg cm^2 mol^{-1} at 90°C. This result is in agreement with the chemical unfolding experiment, shown in Figure 2.115a, which indicates that the secondary structures of the protein start to unfold with a concentration of GdmCl higher than 2 M.

For each of the 4 conditions, 3 CD spectra were recorded: at 25°C before the thermal unfolding, at 96°C and at 25°C after cooling of the samples. The results are shown in Figures 2.117b to 2.117e with the same color code. In the conditions with 0, 1 and 2 M of GdmCl the protein recovers part of its signal after the cooling, suggesting that the process is reversible. For the condition with 3 M of GdmCl the reversibility of the denaturation is less evident, but in all the cases the protein still contains defined secondary structures, also at 90°C. The protein with 0 M of denaturant was also successfully tested for unfolding and reversibility after incubation at 90°C for 60 minutes, suggesting that OctaVII_09 is thermostable and that its unfolding is reversible.

Finally, OctaVII_09 was tested by near-UV circular dichroism. The protein was tested in 3 different concentration: 0.8, 1.2 and 1.9 mg/mL, and the curves are shown in Figure 2.118.

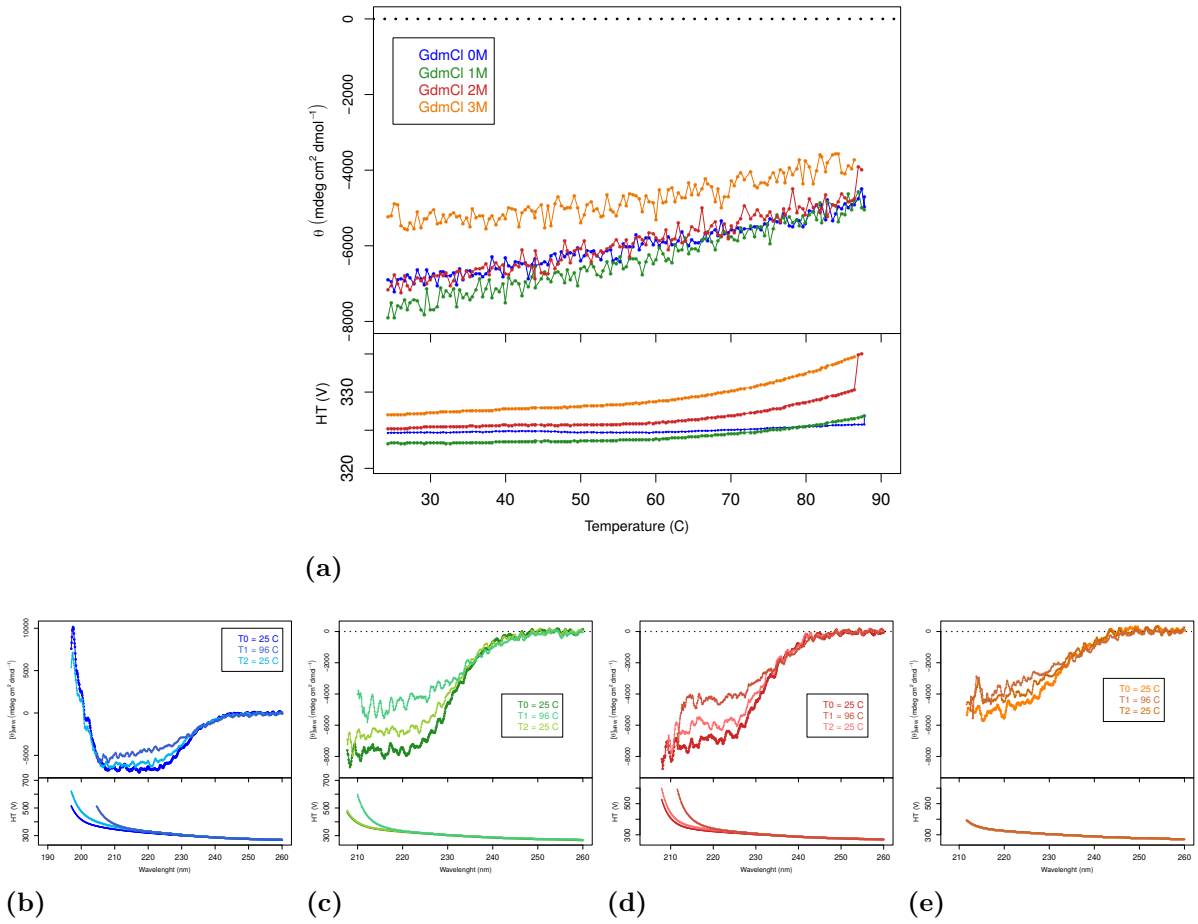


Figure 2.117: Thermal unfolding of OctaVII_09 by circular dichroism

(**a**) Thermal unfolding followed at 222 nm for OctaVII_09 in 4 different conditions: standard buffer (blue line), GdmCl 1M (green line), GdmCl 2 M (red line) and GdmCl 3 M (orange line). (**b-e**) CD spectra of the 4 samples at 25°C before unfolding, at 96°C and at 25°C after cooling of the sample. The color scheme is the same of the thermal unfolding.

The near-UV signal is present and it is concentration dependent for the concentrations of 0.8 and 1.2 mg/mL. The concentration of 1.9 mg/mL still shows a similar signal, but it is not concentration dependent compared to the other two conditions. This can be explained by the fact that OctaVII_09 forms oligomers at high concentration, and their presence in solution may affect the near-UV CD signal.

Crystallization trials

OctaVII_09 was subjected to 2 trials of crystallization, each of them consisting in 480 different conditions. The first one was done with the protein in phosphate buffer at a concentration of 2 mg/mL. After ∼12 months of incubation, only one condition formed

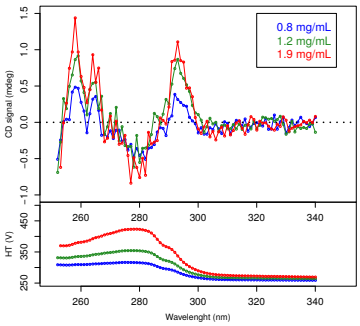


Figure 2.118: Near-UV CD of OctaVII_09

Near-UV CD signal of OctaVII_09 in three different concentrations (top) and high-tension profile (bottom).

crystals (20% PEG 1000, 100 mM Sodium acetate pH 4.5, 200 mM Zinc acetate). The crystals have not been tested for X-ray diffraction yet, so we do not know if they are formed by the protein or by salt.

The second trial of crystallization was done with the protein in a Tris buffer with a concentration of 4 mg/mL. After 8 months of incubation there is no evidence of crystal formation.

2.4.14 OctaVII_09 WS

OctaVII_09 WS (Weak Spot) is a mutant of OctaVII_09 that was designed in collaboration with Jens Meiler at the Vanderbilt University in Nashville, USA. OctaVII_09 was designed in 2015 with the Rosetta software, version 3.5. Over the years many improvements have been done on the software and in particular on its scoring function. In 2018, we tested the model of OctaVII_09 with the new scoring function of Rosetta, version 3.8, and we decided to improve its design according to the suggestions of the software. All the computational work described hereafter in the design of OctaVII_09 WS was done by Samuel Schmitz, a graduate student in the laboratory of Jens Meiler.

In silico saturated mutagenesis

The pdb structure of OctaVII_09 was tested with Rosetta version 3.5 (used in 2015) and with Rosetta version 3.8 (2018). The energy score obtained with the old version is -484 REU, while the one obtained with the new version is -777 REU. The two scores are not comparable, because they are calculated with different scoring functions. However, in order to see if there are mutations that may improve the energy score of OctaVII_09, the model was subjected to *in silico* saturated mutagenesis with the recent version of Rosetta: the 240 residues in the sequence were individually substituted with each of the remaining 19 amino acids (for a total of 4560 models). 50 out of the 4560 single point mutations improved the design and the energy score of OctaVII_09 (Figure 2.119). 27 out of the 50 improved the score of at least 0.5 REU, 16 out of them of at least 1.0 REU, and the mutation S93L alone improved the score by 3.81 REU.

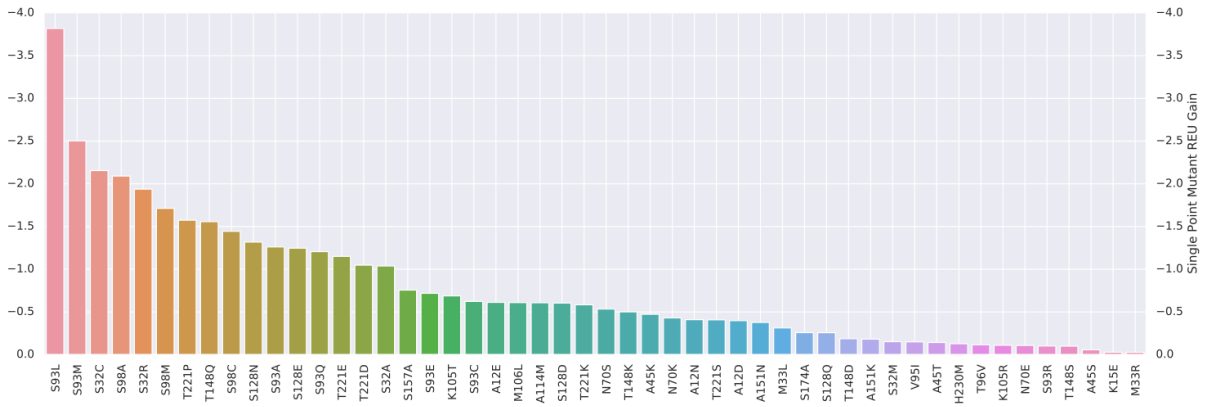
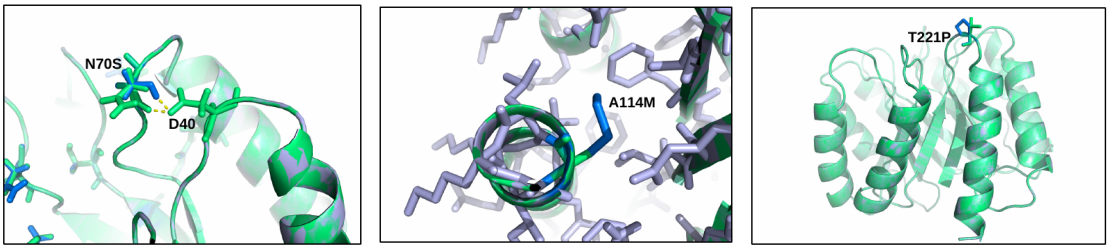


Figure 2.119: *In silico* saturated mutagenesis of OctaVII_09

The 50 best single point mutations of OctaVII_09 after saturated mutagenesis. The mutations are reported on the x-axis, while the Rosetta energy gain is plotted on the y-axis.

All the mutations were analyzed at structural level in order to understand their contribution to the total energy score. For example the mutation N70S (Figure 2.120a) improves the score by formation of an H-bond with the residue D40. This interaction may improve the folding and the stability of the protein because it connects residues that are localized on two different loops. Another example is the mutation A114M (Figure 2.120b), that fills an empty space at the interface between two helices and a strand, improving the packing. A last example is T221P (Figure 2.120c), that is localized in the last loop of the TIM-barrel. This mutation stabilizes the kink of the loop and supports the alignment of the last helix to the β-sheet.

14 mutations were combined in the design of OctaVII_09 WS: 4 were localized in the loops, 6 on the helices (4 exposed to the solvent and 2 buried in the protein core) and 4 on the strands. The 14 mutations improved the score (calculated with Rosetta, version 3.8), from -777 REU to -796 REU, a total of 19 REU (more than 1 REU per mutation!).



(a) **(b)** **(c)**

Figure 2.120: Single point mutations of OctaVII_09

Structural details of the single mutations: (**a**) N70S, (**b**) A114M and (**c**) T221P. The wild-type residues is colored in green and the mutation in blue. The single poin mutations do not introduce structural changes.

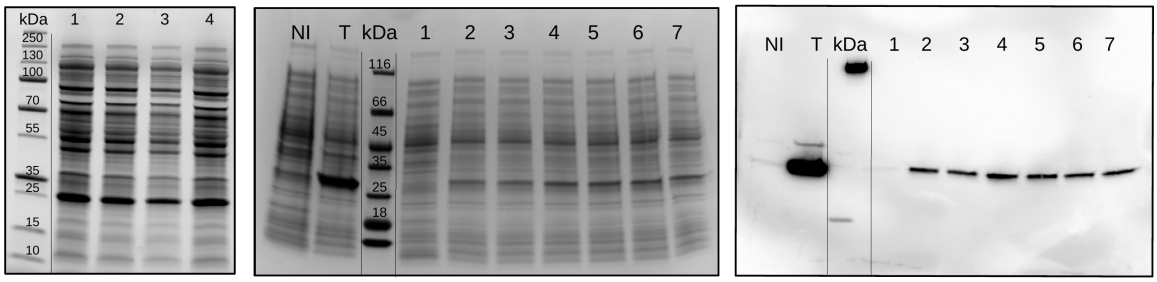
The 14-residue mutant is called OctaVII_09 WS (Weak Spot) and it has 250 residues. Its gene was synthesized by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_09 WS (gBlock®), was manipulated for digestion, ligation, transformation and sequencing as described for OctaVII_06 in Section 2.4.10, page 129. The results of the sequencing confirmed the correct sequence of the OctaVII_09 WS gene in 1 clones out of 4. The plasmid bearing the good sequence was then used for transformation in *E. coli* BL21 (DE3) cells for expression trials and in DG1 competent cell for stock preparation.

Expression Trials

Expression trials for OctaVII_09 WS were done with 4 different clones at 37°C, with induction with IPTG 1 mM for 4 hours (Figure 2.121a). All of them expressed the protein at a molecular mass of ∼27 kDa. The protein expression was also tested at 18°C with different concentrations of IPTG: 0, 5, 10, 50, 100, 500 and 1000 μM. Figure 2.121b shows the total fractions of the non-induced (NI) and the overnight (T) samples, and the protein is clearly expressed at ∼27 kDa. The samples with different concentrations of IPTG were sonicated, and the soluble fraction was separated from the pellet by centrifugation. Lanes 1 to 7 of the SDS-PAGE show the soluble fractions only. A band corresponding to the molecular mass of OctaVII_09 WS is present in all the fractions except the one without IPTG (lane 1). The intensity of the band increases between 0 and 50 μM of IPTG, however it remains constant between 50 and 1000 μM. In order to verify if the band is effectively due to OctaVII_09 WS, a western blot anti-HisTag was performed on the same samples, (shown in Figure 2.121c). The band clearly corresponds to OctaVII_09 WS that is expressed in the soluble fraction of the cell.



(a) **(b)** **(c)**

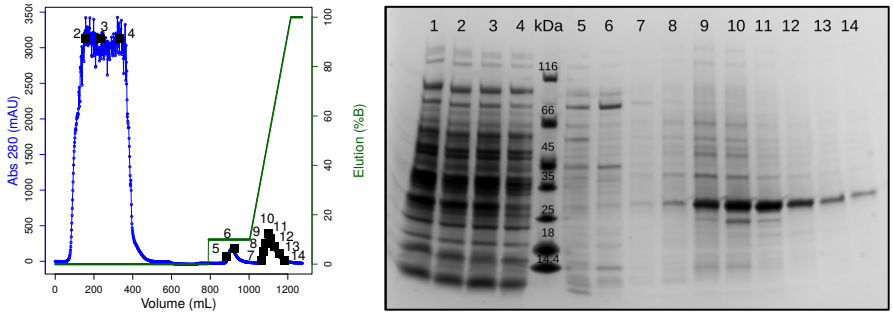
Figure 2.121: Expression of OctaVII_09 WS

(**a**) Expression trial for OctaVII_09 WS at 37°C; the 4 lanes show the total fraction of 4 different clones. (**b**) Total fractions of the non-induced sample (NI) and of the overnight sample (T) at 18°C, and expression trials with different concentration of IPTG: lane **1**, IPTG 0 μM; **2**, IPTG 5 μM; **3**, IPTG 10 μM; **4**, IPTG 50 μM; **5**, IPTG 100 μM; **6**, IPTG 500 μM and **7**, IPTG 1000 μM. (**c**) western blot anti-HisTag of the same fractions described in (b).

Purification of OctaVII_09 WS, soluble fraction

The results obtained with the expression trials at both 37°C and 18°C, suggested that OctaVII_09 WS is partially expressed in the soluble fraction. OctaVII_02, OctaVII_04 and OctaVII_05 were found in the soluble fractions, but the experiments with OctaVII_05 (see Section 2.4.8, page 117) suggested that this is due to the interaction of the free cysteines with endogenous proteins rather than spontaneous folding. OctaVII_09 WS does not contain cysteines, so we are convinced that the 14 mutations that differentiate

OctaVII_09 from OctaVII_09 WS improved its solubility in the cell. In order to purify the protein from the soluble fraction, 1 L of culture was produced. The crude extract was obtained after 4 cycles of disruption, and the soluble fraction was separated from the pellet through centrifugation. The soluble fraction was filtered and loaded on HisTrap HP column for purification (Figure 2.122a). The input sample is visible in lane 1 of Figure 2.122b; a consistent band at ∼27 kDa seems to be present. The majority of the proteins are eluted in the flow-through (lanes 2 to 4). Two peaks of elution are visible on the chromatogram: the first is at 75 mM of imidazole and it is composed only by contaminants (lanes 5 and 6 in the gel), while the second one is eluted at 100 mM of imidazole. A consistent band is present at the expected molecular mass for OctaVII_09 WS (lanes 7 to 14).



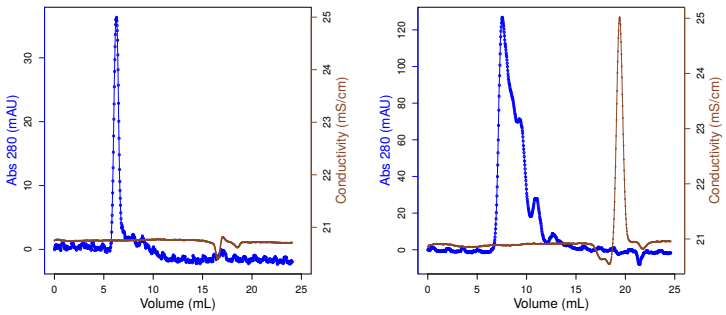
(a) **(b)**

Figure 2.122: Purification of OctaVII_09 WS, soluble fraction

(**a**) Elution profiles of the soluble fraction of OctaVII_09 WS: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-4** the flow-through; **5-6** the first elution peak and **7-14** the second elution peak.

Purification of OctaVII_09 WS, size exclusion

The fractions containing OctaVII_09 WS, obtained from the purification of the soluble fractions were pooled together and centrifuged in order to remove aggregates. An aliquote of the pool was then filtered and loaded on the Superdex-75 column for a size exclusion chromatography. The chromatogram is reported in Figure 2.123a; Figure 2.123b shows the one obtained with refolded OctaVII_09 for comparison. The initial concentration of the two pools was different, however it seems clear that OctaVII_09 WS is eluted in a single peak that correspond to the void volume of the column (7.4 mL). The 2 shoulders that are present in the refolded OctaVII_09 and that may correspond to the monomer and the dimer of the protein are completely absent in the chromatogram of OctaVII_09 WS. This results suggest that OctaVII_09 WS is in an higher oligomerization state.



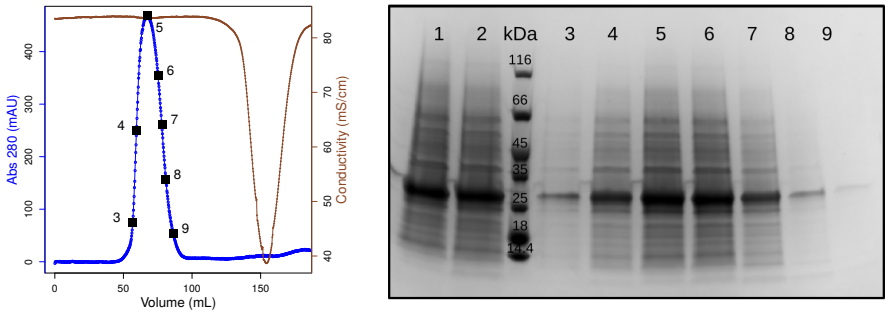
(a) **(b)**

Figure 2.123: Purification of OctaVII_09 WS, desalting

(**a**) Size exclusion chromatography of OctaVII_09 WS from the soluble fraction and (**b**), of the refolded OctaVII_09. The blue line is the absorbance at 280 nm and the brown line is the conductivity.

Purification of OctaVII_09 WS, desalting

Despite the apparently high oligomerization state of OctaVII_09 WS, we decided to load the sample in the desalting column in order to remove the imidazole. The pool obtained after purification is shown in lane 1 of Figure 2.124b. Many contaminants at higher and lower molecular mass are present, and they were not removed after centrifugation and filtration (lane 2). The sample was loaded on the desalting column and the chromatogram is presented in Figure 2.124a. The conductivity curve (brown) forms a negative peak because of the buffer exchange, the Tris buffer has a higher conductivity than the phosphate one. The fractions of the elution peak are shown in lanes 3 to 9 of the SDS-PAGE.



(a) **(b)**

Figure 2.124: Purification of OctaVII_09 WS, desalting

(**a**) Elution profile of the desalting of the pool of the refolded OctaVII_09 WS: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the pool after desalting; **2** the input sample and **3-9** the elution peak.

Purification of OctaVII_09 WS, boiling

The sample obtained from the desalting purification was highly contaminated and it was necessary to perform an extra step of purification prior to possible crystallization. Since OctaVII_09 resulted to be thermostable at high temperature, we attempted to remove some contaminant by increasing the temperature of the sample. For this experiment, 9 samples of 100 μL of OctaVII_09 WS were used. Except for the untreated sample (lane 1 in Figure 2.125), the remaining 8 were incubated at 95°C for 1, 5, 10, 20, 30, 40, 50 and 60 minutes, and then centrifuged at high speed to remove possible aggregates (lanes 2 to 9, respectively). The SDS-PAGE shows that OctaVII_09 WS is thermostable as OctaVII_09 and it does not precipitate despite 60 minutes at high temperature. The contaminants at high molecular mass are on the contrary precipitating just after 5 minutes of incubation at high temperature. However, the contaminants at lower molecular mass seems as thermostable as OctaVII_09 WS and it was not possible to remove them from the solution.

The pool containing OctaVII_09 WS was then incubated at 60°C for 90 minutes and then centrifuged at high speed.

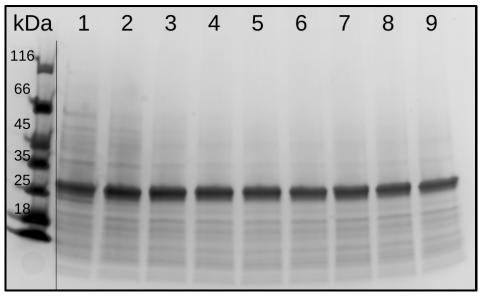


Figure 2.125: Purification of OctaVII_09 WS, boiling

SDS-PAGE of the purification of OctaVII_09 WS by heat: the same sample is boiled at 95°C for 0, 1, 5, 10, 20, 30, 40, 50 and 60 minutes (Lanes **1-9**, respectively).

Crystallization trials

OctaVII_09 WS was tested for crystallization trial in 480 different conditions. The protein was in a phosphate buffer at a concentration of 2.3 mg/mL. After 5 months of incubation, 4 conditions presented crystals. The first one is the same that gave crystals for OctaVII_09: 20% PEG 1000, 100 mM Sodium acetate pH 4.5, 200 mM Zinc acetate. The remaining three are: 1- 28% PEG 400, 100 mM HEPES sodium pH 7.5, 200 mM $CaCl_2$; 2- 20% 2-propanol, 100 mM Sodium acetate pH 4.6, 200 mM $CaCl_2$; and 3- 30% PEG 400, 100 nM TRIS-HCl pH 8.5, 200 mM $MgCl_2$. Crystals have not been tested for X-ray diffraction, so we do not know yet if they are formed by the protein or by salts.

2.4.15 OctaVII_10

OctaVII_10 has 250 residues and it is the best representative of Family 05 according to the Rosetta energy score (details in Section 2.3.9, page 68). The gene was synthesized and inserted in the pET28a plasmid by IDT, that shipped it as dry pellet.

Sequencing

The DNA pellet of OctaVII_10 (gBlock®), was manipulated for digestion, ligation, transformation and sequencing as described for OctaVII_06 in Section 2.4.10, page 129. The results of the sequencing confirmed the correct sequence of the OctaVII_10 gene in 1 clone out of 10. The plasmid bearing the good sequence was then used for a transformation in *E. coli* BL21 (DE3) cells for expression trials and in DG1 competent cell for stock preparation.

Expression Trials

Expression trials of OctaVII_10 were performed in two conditions: overnight induction at 18°C (Figure 2.126) and 4 hours induction at 37°C (not shown). OctaVII_10 is visible in the total fraction at the correct size.

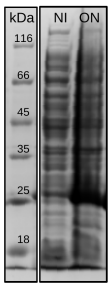


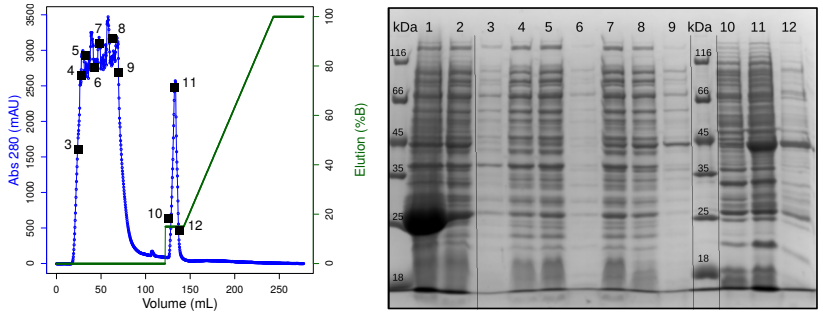
Figure 2.126: Expression trials of OctaVII_10

Expression trials of OctaVII_10: the non-induced sample is on the left (NI) and the sample with overnight induction at 18°C is on the right (ON).

Purification of the soluble fraction

OctaVII_10 was produced in 1 L culture with induction at 37°C for 4 hours. The crude extract (lane 1 in Figure 2.127b) was centrifuged and filtered. The majority of the target protein was removed from the soluble fraction (lane 2). The chromatogram in Figure 2.127a shows that most proteins were eluted in the flow-through, whereas a minor fraction only was eluted in a peak at 75 mM of imidazole. The SDS-PAGE confirms that

OctaVII_10 is not produced in the soluble fraction because the elution peak is composed of contaminants only.



(a) **(b)**

Figure 2.127: Purification of OctaVII_10, soluble fraction

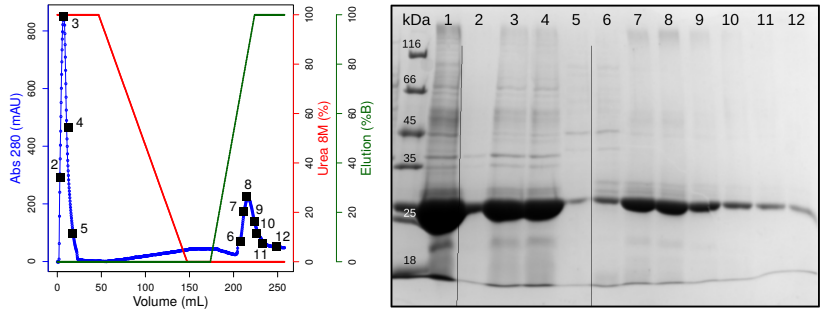
(**a**) Elution profiles of the soluble fraction of OctaVII_10: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the crude extract, **2** the input sample; **3-9** the flow-through and **10-12** the elution peak.

Purification of the insoluble fraction

The pellet collected after centrifugation of the crude extract of OctaVII_10 was washed twice and dissolved in the buffer containing 8 M urea. Following centrifugation, the supernatant was filtered and loaded in the HisTrap HP column for refolding and purification. The chromatogram is shown in Figure 2.128a and the input sample in lane 1 of Figure 2.128b. The flow-through contains the target protein, suggesting that the column reached saturation. The refolding did not cause unbinding of the protein from the matrix, that is then eluted at 250 mM of imidazole buffer. The fractions of the elution peak (lanes 6 to 12 of the SDS-PAGE) are highly pure. Few contaminants of 28 to 35 kDa are visible in the gel, however OctaVII_10 is always the predominant band.

Desalting

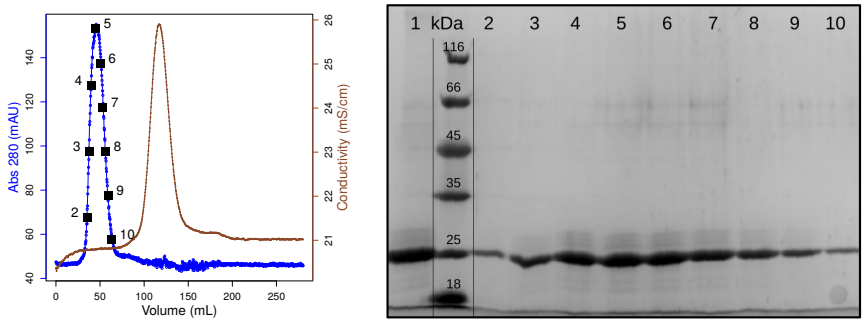
The fractions containing the refolded OctaVII_10 were pooled together, centrifuged and filtered before desalting. The chromatogram is shown in Figure 2.129a, the input sample is shown in lane 1 of Figure 2.129b and the elution peak in lanes 2 to 10.



(a) (b)

Figure 2.128: Purification of OctaVII_10, insoluble fraction

(**a**) Elution profiles of the insoluble fraction of OctaVII_10: the blue line is the absorbance at 280 nm, the green line is the concentration of the elution buffer (Imidazole 0.5 M), the red line is the concentration of the denaturing buffer (Urea 8 M) and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. In lane **1** there is the input sample; **2-5** the flow-through and **6-12** the elution peak.



(a) (b)

Figure 2.129: Purification of OctaVII_10, desalting

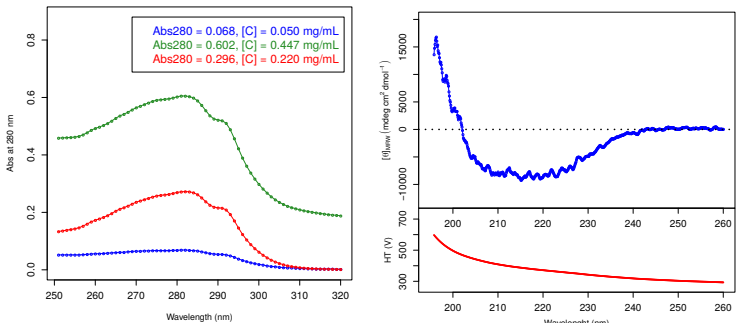
(**a**) Elution profile of the desalting of the pool of the refolded OctaVII_10: the blue line is the absorbance at 280 nm, the brown line is the conductivity and the black squares represent the fractions that are shown in (**b**), the SDS-PAGE gel. Lane **1** the input sample; **2-10** the elution peak.

Biophysical characterization

The desalted sample of OctaVII_10 was analyzed by absorbance spectroscopy under 3 conditions: before concentration (blue line in Figure 2.130a), after ultrafiltration (green line), and after centrifugation to remove the aggregates (red line).

The final concentration of the protein corresponds to 0.22 mg/mL. Trials to bring the protein to a higher concentration failed, and lead to precipitation of the sample.

However, it was possible to record a far-UV CD spectrum, shown in Figure 2.130b. The percentages of secondary structures were calculated from the spectrum with CDpro, using the program CDSSTR. The results are compared with the predicted percentages calculated by DSSP on the 3D model in Table 2.16.



(a) **(b)**

Figure 2.130: Biophysical characterization of OctaVII_05 NoCys

(**a**) Absorption spectrum of OctaVII_10 after desalting (blue), after concentration (green) and after centrifugation and filtration (red). Abs_{280} is used to calculate the protein concentration. (**b**) CD spectra of the protein (top) and high-tension (bottom); the dotted line indicates the baseline at $[\Theta]$=0.

	% α-Helix	% β-Strand	% Turn	% Unstruc
CDpro (CDSSTR)	22.9	28	20.5	28.4
DSSP	47.5	19.2	15.8	17.5

Table 2.16: Secondary structures content, OctaVII_10

The percentages from CDpro and DSSP are highly different: the helix content is 25% lower of the expected one, while the strand one is 10% higher. It is quite unlikely that the protein is folded in the correct structure. OctaVII_10 is discarded from further characterization due to its low solubility and to the results of circular dichroism.

2.4.16 Comparison of OctaVIIs

The results obtained for the 15 artificial proteins designed and produced in this thesis project are summarized in Table 2.17.

	01	02	02 YQ	03	04	04 NC	04 WS	05	05 NC	06	07	08	09	09 WS	10
Expr. 37°C	x	V	V	V	V	V	V	V	V	V	V	V	V	V	V
Expr. 18°C	x	V	V	V	V	V	V	x	x	V	V	V	V	V	V
Mass	-	x	x	x	V	V	V	V	V	V	x	V	V	V	V
Purif. Soluble	-	V	-	-	V	x	x	V	x	x	x	x	x	V	x
Purif. Pellet	-	V	-	-	V	V	V	V	V	V	-	V	V	-	V
Far-UV CD	-	-	-	-	V	V	V	V	V	V	-	V	V	-	V
Fluorescence	-	-	-	-	V	V	V	V	-	-	-	-	V	-	-
mg/mL	-	-	-	-	0.2	0.2	0.7	0.2	0.3	0.2	-	0.1	4	2	0.2
Crystallization	-	-	-	-	-	-	-	-	-	-	-	-	V	V	-

Table 2.17: Experimental validation of the 15 OctaVIIs

Codes: "V" = succesfull experiment; "x" = unsuccesfull experiment and "-" = not performed experiment.

Expression Trials (Expr. 37°C and Expr. 18°C in Table 2.17)

Among the 15 proteins, only OctaVII_01 could not be expressed at all, neither at 37°C nor at 18°C. We did not further inquire the reasons why the protein was not expressed under our experimental conditions since this was not the goal of this work. However, similar observation was reported by Winther and co-workers in their publication on the re-design of the thioredoxin (described in the Introduction, Section 1.1.2, page 7) [31]. Thus, among the 48 models chosen for experimental validation, 16 were not expressed. Considering that: 1- thioredoxin is ∼110 residues, less than half of the size of the OctaVIIs, and 2- their design was done starting from natural backbones, we can be happy of our yield in which only 1 protein out of 10 is not expressed (vs. 1 out of 3 with thioredoxin).

The remaining 14 proteins could all be expressed at 37°C and 18°C, with the exception of OctaVII_05 and its mutant OctaVII_05 NoCys that were not expressed at lower temperature. Since there is no difference in the plasmid sequences of all constructs, those exceptions rely only on the gene sequence of OctaVII_05 and OctaVII_05 NoCys. We did not inquire further the reasons why this phenomenon occurs and there are no similar results in the literature. Indeed, temperature changes during growth and induction of *E. coli* are widely suggested to tune the expression and the folding of recombinant proteins.

Molecular Mass (MM in Table 2.17)

Among the 10 artificial proteins, 3 are truncated: OctaVII_02, OctaVII_03 and OctaVII_07. Attempts to recover the full length of OctaVII_02 with the mutation Y57Q

failed. Many examples of protein truncation in *E. coli* can be found in the literature [132]. In a review about recombinant protein expression in *E. coli*, Rosano and Ceccarelli [133] pointed out that this phenomenon may be due to codon bias: depletion of low-abundant tRNA causes amino acid misincorporation and/or truncation during the protein production. However, we exclude this hypothesis for two reasons: first, the gene sequences were designed with a codon optimization for the production in *E. coli*, and second, tRNA depletion affects mainly the C-terminal part of the expressed protein, while the truncations in the OctaVII proteins are all located at the N-terminus.

An alternative explanation for protein truncation is the presence in the gene of an internal ribosome binding site (internal RBS), which alters the starting point for protein translation [134, 135]. Whitaker and co-workers [136] demonstrated that eukaryotic genes expressed in *E.coli* are more prone to contain internal RBS compared to prokaryotic ones. This is due to the fact that bacterial genes have been subjected to negative selection pressure against RBS, while eukaryotic genes have not. Since artificial sequences are not subjected to evolutionary selection pressure, they have higher probability to present internal RBS that causes N-terminal truncations. We did not inquire further for the presence alternative ribosome binding sites in the sequences of OctaVII_02, OctaVII_03 and OctaVII_07.

Purfication from the soluble fraction (Purif. Soluble in Table 2.17)

12 proteins have been tested for purification from the soluble fraction, using affinity chromatography on a HisTrap HP column. Only 4 of them could be observed on SDS-PAGE following purification: OctaVII_02, OctaVII_04 and OctaVII_05 and OctaVII_09 WS. In all cases, however, most or all of the protein of interest was found in inclusion bodies.

Three variants, OctaVII_02, OctaVII_04 and OctaVII_05, contain free cysteines that may interact with endogenous proteins. This interaction may shift part of the protein production from the insoluble fraction (inclusion bodies) to the soluble one. This hypothesis is not fully demonstrated, but in Section 2.4.8, page 117, we showed how OctaVII_05, purified from the soluble fraction, forms covalent bond with endogenous proteins that are broken only in presence of both reducing and denaturing agents (β-mercaptoethanol and SDS). The fact that the mutants without cysteines of OctaVII_04 and OctaVII_05 are not observed in the soluble fraction also supports this hypothesis.

The situation is different for OctaVII_09 WS, which is also partially found in the soluble fraction. This protein is a 14-residue mutant of OctaVII_09. Both contains no cysteines, but OctaVII_09 is produced only in inclusion bodies. The 14 mutations are

responsible for the improved solubility of the protein upon expression, but we do not know if they contribute as a whole or not. However, this protein is the only one out of 15 that, at least in part, spontaneously folds in the soluble fraction of the cell.

In the work of Winther [31], 9 designs out of 48 were partially expressed in the soluble fraction, and 3 out of them could be successfully purified. These results are in good agreement with the ones presented in this work: 1 protein out of 15 is obtained from the soluble fraction.

Other results found in the literature are less detailed for a comparison. For example, in the work of Huang (reported in Section 1.4.2, page 30), 22 proteins out of 22 are highly expressed in *E. coli*, without showing troubles as non-expression, truncation or insolubility [117].

Purification from the insoluble fraction (Purif. Pellet in Table 2.17)

10 OctaVIIs out of 15 were successfully purified from the insoluble fraction of the cellular extract. Solubilization of inclusion bodies 8 M in urea is a standard procedure for refolding of insoluble proteins, however it did not work with two of our artificial proteins, OctaVII_06 and OctaVII_08. The use of guadinidium chloride, a more powerful chaotropic agent than urea, led to a partial dissolution of the inclusion bodies. The amount of protein obtained by extraction with GdmCl was, however, much lower than that obtained with urea for the other OctaVIIs. The fact that inclusion bodies do not dissolve in urea 8 M is not reported in literature, however there is a post on the forum of ResearchGate that mention a similar result:

www.researchgate.net/post/Problem_with_protein_purification_not_soluble_in_8M_urea

Unfortunately, nobody had an explanation for this phenomenon. It may be that the inclusion bodies of OctaVII_06 and OctaVII_08 are too much insoluble for dissolution in urea.

The observation that urea may not be the best chaotropic agent to dissolve inclusion bodies is also supported by the chemical unfolding experiment performed on the refolded OctaVII_09 in Section 2.4.13, page 141. The unfolding transition followed by fluorescence (see Figure 2.115e), does not reach a plateau at high concentration of urea, indicating that the protein is not fully unfolded. It is possible that, during solubilization, inclusion bodies were not completely dissolved and that this contributed to the aggregation issues of OctaVII_09.

Secondary structures (Far-UV CD in Table 2.17)

All the OctaVIIs that were analyzed by far-UV circular dichroism shows presence of secondary structures: in particular clear minima at 208 and 222 nm revealed a significant content in α-helical content. In all cases, however, this was lower (23 to 39%) than expected (46 to 49%)

In contrast, the percentage of β-strands is in general consistent with the expectation (17-21%). Only OctaVII_04 WS (30%) and OctaVII_10 (28%) are out of the range. OctaVII_04 WS has 22 mutations of difference compared to OctaVII_04 NoCys. The mutations seem to increase the percentage of β-strands from 18% to 30%. The content of helices and of turns remains the same, while the unstructured regions decrease from 29% to 19%. The 22 mutations are probably involved in the folding of portions of the protein that were previously unstructured.

OctaVII_10 is unlikely to be folded as expected, in fact its content of secondary structures is 10% higher for the strands (28% instead of 19%), and 25% lower for the helices (23% instead of 48%) compared to the model ones.

OctaVII_09 is the best model according to solubility and thermostability, but its helical content is less in agreement with the model than the other OctaVIIs. In fact, the expected percentage is 47.5% and the experimental one is 29%.

The remaining six OctaVIIs have more or less the same content in the strands (17.2% to 22.0%), in the helices (33.7% to 39.6%) and in turn regions (15.7% to 19.5%). The similarity in their secondary structure content may be due to the fact that the proteins descend from the same structural ancestor (the parametric structure described in Section 2.2.1). In fact, their 3D models are similar, and the percentages extracted with DSSP spans in short ranges: 17.0%-20.5% for the strands, 46.6%-48.7% for the helices and 14.1%-19.5% for the turns. The content of strands and turns is in good agreement with the expectation, while the helical content is always higher in the model of by least 10%.

Remarkably, the percentages of secondary structures obtained for most proteins are in good agreement with those obtained for the natural TIM-barrel of *Thermotoga marittima*: 18% of strands, 36% of helices and 19% of turns (unpublished data of Matagne's Lab). We can not know if the proteins are folded in a TIM-barrel fold based only on the CD results. As reported for OctaV.1 in Section 1.3.5, page 24, the helix and strand contents may be in the range of natural TIM-barrels, but their folding may not correspond to the desired one.

Tertiary structures (Fluorescence in Table 2.17)

Fluorescence emission spectra were recorded for 5 variants. The expected λ_{max} for completely unfolded proteins is at 355 nm, due to full exposure of the tryptophan indole groups (shown with OctaVII_09 in Figure 2.114c). All of the 5 OctaVIIs have the λ_{max} in the range of 333-338 nm, indicating that the structure is compact and that the aromatics are not fully exposed to the solvent. Unfortunately, only one protein out of the 5 could be concentrated up to 2 mg/mL, the concentration requested for the analysis by near-UV circular dichroism. In this case, a significant CD signal (Figure 2.118) demonstrated the occurence of stable tertiary contents.

Solubility (mg/mL in Table 2.17)

10 out of 15 OctaVIIs have been subjected to cycles of ultrafiltration in order to increase their concentration. However, the majority of them aggregated and precipitated already in the first steps and their final concentration was always < 1 mg/mL.

The low solubility of artificial proteins is quite common, especially in the Octarellin history (see Chapter 1.3). In the work of Winther [31] on the re-design of thioredoxin, the majority of the protein have low solubility. Out of the 48 proteins they produced, only 9 were tested for purification. Among these, 7 were discarded due to precipitation on column, low stability in solution and low solubility. Also the two remaining proteins, successfully isolated by size exclusion chromatography in a monomeric form, showed low solubility: one reached a concentration of $\sim$0.35 mg/mL, the other was more soluble, but only at low pH. In the work of Huang [117] on the symmetric TIM-barrel, there is no evidence for poor solubility of their models.

Interestingly, the 22 mutations that distinguish OctaVII_04 NoCys and OctaVII_04 WS improved the solubility from 0.20 mg/mL to 0.74 mg/mL. 10 out of the 22 mutations are localized in loop regions, 9 in β-strands and the remaining 3 in the last α-helix. We did not inquired further which residues are involved in the improved solubility of OctaVII_04 WS, however the improvement is significant. OctaVII_09 and its mutant OctaVII_09 WS are the only proteins that passed the limit of 1 mg/mL, reaching 4 and 2 mg/mL, respectively.

Oligomerization and aggregation (in Table 2.17)

All the 10 refolded proteins showed signs of oligomerization or aggregation. In particular, all the proteins except OctaVII_09 and its mutant have low solubility. They precipitate at low concentration, and aggregates were visible in solution. OctaVII_04, Oc-

taVII_04 NoCys, OctaVII_04 WS, OctaVII_05, OctaVII_05 NoCys and OctaVII_09 have been tested by dynamic light scattering (DLS) and showed the presence of aggregates (data not presented in this work). However, for OctaVII_09, the aggregates were stable in solution for months, and the protein could be concentrated at 4 mg/mL without precipitation. This stability overtime and upon concentration may be due to an oligomerization process more than aggregation, however we did not inquired further.

In the history of the Octarellins, there are evidences of aggregation in solution only for OctaV and OctaVI. Following refolding, OctaV is present in solution in both a monomeric and an aggregated form [4], that are then separated by size-exclusion chromatography. Instead, OctaVI is refolded in a monomeric form only, but it forms aggregates when heated above 74°C.

In the work of Winther [31] on the re-design of the thioredoxin, the authors stated that most of the protein they tried to purify did not elute from the column, probably because of on-column aggregation, but there are no evidence to support their theory.

Among the 4 Symmetrins designed by Nagarajan [116], only one resulted present in solution in a high oligomeric state (Symmetrin-3). It is interesting to notice that its T_m is ∼20°C higher than the one of the other Symmetrins (63°C versus 44°C). High thermostability is often associated to high oligomerization state [137], and this may be the case also for OctaVII_09, that is both thermostable and oligomeric.

Crystallization (Table 2.17)

OctaVII_09 and its mutant OctaVII_09 WS are the only proteins that underwent crystallization trials, since they were concentrated to 4 mg/mL and 2 mg/mL, respectively. One crystal was obtained after a year of incubation for OctaVII_09, while 4 crystals were obtained for OctaVII_09 WS after only 5 months. Interestingly, one condition of crystallization is common in both proteins. This result seems promising, however the 5 crystals have not been tested for X-ray diffraction yet. It is possible that they are composed only by salts. The diffraction experiment is planned and soon we will obtain an answer. If the crystals are formed by the protein, we have a stock sample to repeat the crystallization and possibly to determine the 3D structure of the proteins.

The Weak Spot mutant

Two of the 15 proteins, OctaVII_04 WS and OctaVII_09 WS, are mutant that have been designed in order to improve the properties of OctaVII_04 NoCys and OctaVII_09, respectively.

OctaVII_04 WS has been created in collaboration with the laboratory of Wim Vranken. The software they used to analyze OctaVII_04 NoCys was supposed to improve its folding to form a TIM-barrel structure. However, its secondary structure content, as estimated from the analsis of far-UV CD data, is not in agreement with that of the model, nor with the results obtained with the natural TIM-barrel of *Thermatoga maritima*. In fact, it presents 30% of β-strands instead of 20%. The biophysical results seem to suggest that the 22-mutations present in OctaVII_04 WS do not really fold into a TIM-barrel, but it is impossible to determinate it without resolution of the structure.

However, the mutations contributed to increase the solubility of the protein after refolding of 3-4 times. Moreover, according to the results obtained by far-UV CD, 12% of the unstructured regions of OctaVII_04 NoCys resulted structured in OctaVII_04 WS. These results are impressive if we consider that the HMM-TIM methodology did not use direct structural information of the protein.

On the contrary, OctaVII_09 WS is the mutant of OctaVII_09, designed in collaboration with the laboratory of Jens Meiler, using structural informations. The 14-mutations led to an improved solubility after expression, in fact OctaVII_09 WS is partially produced in the soluble fraction.

In the history of the Octarellins, the expression of OctaV was shifted from the inclusion bodies to the soluble fraction thanks to directed evolution. We obtained the same results in OctaVII_09 WS thanks to a rational analysis of the protein structure (and an improved software). This is remarkable because we were able to mimic at computational level the directed evolution process. This means reduced experimental time (one month *in silico* versus 6 at the bench), and reduced costs.

There is only a single drawback of the 14-mutation of OctaVII_09 WS: the low percentage of monomeric protein in solution found by size-exclusion with OctaVII_09 is completely lost in the mutant, that is present in solution only in an high oligomerization state.

Chapter 3

Computational Protocols

3.1 Natural TIM-barrel analysis

3.1.1 Collection of natural TIM-barrels

85 pdb structures were downloaded from DATE (web-link, page 231). After visual inspection with Pymol [118], 33 out of them were discarded because presenting one or more extra domains in their structure. The presence of other folds, rather then the TIM-barrel one, will affect the next step of structural alignment. In fact, alternative domains might be found in proteins that do not contain the TIM-barrel fold, leading to a mixed collections of TIM-barrels and other folds. For this reason the 33 multi-domains structures were discarded. The remaining 52 were individually uploaded to the PDBeFOLD web-service [119], (web-link, page 231). 1623 different natural TIM-barrels were obtained after structural alignment with PDBeFOLD. Fasta sequences were loaded on PISCES web-service [121] (web-link, page 231), and were refined according to sequence identity, lower than 80%, and to structure resolution, lower than 2$\mathring{A}$. The 229 output sequences were reduced to 219 after visual inspection with Pymol. The 10 discarded proteins did not contained the TIM-barrel fold in their structure. For a description of the software, see Annexes 6.1, page 231, and for the list of 219 pdb IDs see Annex 6.5.4, page 247. The total length of the proteins was extracted from the PISCES output.

3.1.2 Secondary structure assignment

For each of the 219 pdb files, N- and C-terminal domains not involved in the TIM-barrel fold (extra-domains) were deleted from the pdb with Pymol. Secondary structure assignment was done with the program DSSP [138]; the command line is reported in Annex 6.5.1.

Information about α-helices (H,G and I in the DSSP output) and β-strands (E) was extracted and the secondary structures not involved in the TIM-barrel fold (extra-elements), were discarded. For each natural structure, the element number in the TIM-barrel fold (i.e. 1^{st} strand, 1^{st} helix, 2^{nd} strand, etc.) and the residue numbers at the beginning and at the end of the secondary structures were assigned. The length of each element was calculated by subtraction of the residue numbers.

Loops were defined in this work as the connection between β-strands and α-helices involved in the TIM-barrel fold, and they can include extra-domains and individual secondary structures. The calculation of loop length was done by subtraction of the residue numbers.

3.1.3 Energy scores in natural TIM-barrels

The 219 natural TIM-barrels were subjected to energy minimization (or relax) with the Relax package of Rosetta [38], (version 2015.19.57819_bundle). The energy function used by rosetta is Talaris2013. It is possible to modify the energy function and adapt it to specific system, however in this work we used the default one that is preferable for soluble globular proteins. The command line is reported in Annex 6.5.2, page 245. The program produced as outputs the relaxed pdb and a file with the corresponding energy scores.

Total energy values were extracted after minimization (the `total_score` column in the score file) and plotted against amino acid length of the structures (see Section 2.1.5, page 38). The R-square value was calculated with R-language [139].

Energy values that depends only on H-bonds between the strands backbones (and not on side-chain/side-chain and side-chain/main-chain interactions), were automatically calculated in the score file and extracted after minimization (the `hbond_lr_bb` column in the score file). They will be defined as β-energies here on. The number of residues belonging to β-strands were extracted with DSSP and plotted against the β-energy values (see Section 2.1.6, page 40)

3.1.4 Amino acid composition

The 219 natural TIM-barrels were analyzed for their amino acid content. Each of the 20 natural amino acids was counted in each protein sequence and the resulting value was transformed in percentage according to the protein length. The range of percentages in the collection of natural TIM-barrel for each natural amino acid is shown in Figure 2.1.

A similar analysis was done by grouping amino acids according top the following properties (property groups):

1. **Small:** Alanine (A), Cysteine (C), Glycine (G), Serine (S)
2. **Polar:** Aspartic acid (D), Glutamic Acid (E), Histidine (H), Lysine (K), Serine (S), Asparagine (N), Glutamine (Q), Arginine (R), Threonine (T), Cysteine (C)
3. **Charged:** Aspartic acid (D), Glutamic Acid (E), Lysine (K), Arginine (R)
4. **Positive:** Lysine (K), Arginine (R)
5. **Negative:** Aspartic acid (D), Glutamic Acid (E)
6. **Apolar:** Tryptophan (W), Tyrosine (Y), Phenylalanine (F), Valine (V), Proline (P), Leucine (L), Isoleucine (I), Alanine (A), Glycine (G), Methionine (M)
7. **Aromatics:** Tryptophan (W), Tyrosine (Y), Phenylalanine (F)

3.2 Protein design

3.2.1 Parametric design

The design of the artificial TIM-barrel backbones can be done in two ways: from scratch using geometrical information extracted from natural TIM-barrels (parametric design), or through homology modelling of multiple existing TIM-barrel structures (comparative modelling). Both techniques were preliminary tested and only the first one resulted optimal for the design. The comparative design had three major problems. The first is that it was not possible to perform multiple structure alignment on all the 219 natural TIM-barrels structures that have been described in the previous chapter. We were able to obtain valid alignments using only 30 proteins out of the 219, decreasing dramatically the heterogeneity of our dataset. The second problem is bound to the characteristic of natural TIM-barrels to have extra-structures and extra-domains in their loops. Since they are not conserved, they negatively affect the comparative modelling and output structures lacked multiple loop regions. The third problem was related to the characteristic of natural TIM-barrels to have a low sequence identity (in certain cases also less than 5%). It was really hard to find a consensus sequence in agreement with all the 30 structures used as input. This led to outputs that collapsed on themself at the first energy minimization. The comparative modelling is an excellent technique when the starting inputs have high similarity in both sequence and structure, but it became poorly accurate when they are more variable. For this reason we decided to design our backbones with the parametric method that is described hereafter.

The parametric design of the artificial TIM-barrel backbones was performed with the package BundleGridSampler of the Rosetta software (version 2015.19.57819_bundle). This program required 8 parameters to design each element, 5 of them were common to both α-helices and β-strands and 3 were for strand only. The common parameters for each element were: the radius (`r0`), the length (`helix_length`), the localization in space around the bundle axis (`delta_omega0`), the orientation around the internal axis (`delta_omega1`) and the inclination (`omega0`). Parameters for β-strands only were: the strand definition to differentiate from the helix (`crick_params_file`); the orientation of N- to C-terminus that is opposite to the helix one (`invert`), and the vertical shift of the elements in order to align them with the α-barrel (`delta_t`).

The radius values used for the design of the β- and of the α-barrels were calculated from natural TIM-barrel structures with DeepView [140], and set 7.5 $\mathring{A}$ and 17.4 $\mathring{A}$, re-

spectively. The chosen length of the β-strands was 9 aa: 5 aa for the strand itself plus 2 aa at the N-term and 2 aa at the C-term in order to form the loop connection in the following steps of design (see Loop Closure, in Sections 3.2.3, page 176). The chosen length of the α-helices was 21 aa: 16 aa for the helix itself plus 3 aa at the N-term and 2 aa at the C-term. The localization in space of the first β-strand was set at 0.0 rad on the β-circumference with `r0`=7.5 $\mathring{A}$, and the following strands were positioned at intervals of 0.78539816 rad (that equals to $1/8^{th}$ of circumference in radians). In a similar way, the first α-helix was set at 5.890486230 rad on the α-circumference with `r0`=17.4 $\mathring{A}$ and the following were positioned at intervals of 0.78539816 rad. The orientation and the inclination of the secondary structure elements were found by comparison with natural TIM-barrel. For helices, the orientation around the helical axis was set at 3.14 rad and the inclination at 0.04 rad. For the strands, the inclination value was set at 0.32 rad, while the orientation value differed for even and odd elements: respectively 4.31 rad and 1.57 rad. The vertical shift values was set 0.25 aa for even strands and 1 aa for odd ones.

All these parameters were inserted in the PARAMETER file, one of the 3 input files necessary to run the program. The second input was a FASTA file to assign side-chains to each residue (in our case, all the residues will be Ala), and the last one was the OPTION file, in order to define the number of output (one in our case), the energy function for the scoring (talaris2014), and other standard option required by the software. The command line and the 3 input files are reported in Annex 6.5.3, page 245.

3.2.2 Alanine substitution

The alanine substitution was performed with the Design package of Rosetta software (version 2015.19.57819_bundle). The program required as inputs a PARAMETER file, an OPTION file, a pdb file and a RESFILE that contains the instructions for the amino acid substitution. Description of the RESFILE usage is reported in Section 6.1, page 231.

Alanine substitution was divided into 4 steps, according to the structure layer: core, boundary, surface and loop. This separation was necessary to create multiple combinations of amino acids and increase the sequence variability; trials with a single step of substitution resulted in low sequence differentiation. The substitution instructions correspond to the overall characteristics of globular proteins, in which the hydrophobic core is mainly composed by apolar residues, the protein surface by polar amino acids and the boundary regions by a mix of both. Loops can be composed by both apolar and polar residues depending on their position, so they can be considered as “boundaries”. How-

ever, in this work, we separated loops from "rigid" boundaries (belonging to structured α-helices or β-strands) in order to avoid larger residues in the loop (W,F,Y and K), and to avoid flexible residues in the rigid structure (G,A).

All the steps shared the same PARAMETER and OPTION files while the RESFILE and the input pdb were changing. RESFILEs for the alanine substitution are shown in Section 6.5.4, page 247.

The first step of alanine substitution targeted the hydrophobic core of the protein: 7 residues of each β-strand and 5 of each α-helix that faces the internal β-barrel, for a total of 96 out of 240 aa. The input pdb was the output of the parametric design described in Section 2.2.1, page 45. 200 outputs were produced and ranked by energy score by Rosetta. The substitution instructions (RESFILE 1) were:

- 4 residues on each β-strand that face the center of the barrel, can be substituted only with valine (V), isoleucine (I) or leucine (L).

- 3 residues on each β-strand that face the α-barrel, can be substituted with any apolar amino acid with the exception of glycine (G) and cysteine (C).

- 5 residues on each α-helix that face the β-barrel can be substituted with any apolar amino acid with the exception of glycine (G) and cysteine (C).

The second step targeted the boundary amino acids between the hydrophobic core and the hydrophilic surface of the protein, 40 out of 240 aa. 8 input pdb were chosen between the best of the previous step for the energy score and, for each of them, 50 outputs were produced and ranked. The substitution instructions (RESFILE 2) were:

- 5 residues on each α-helix, can be substituted with any amino acid with the exception of alanine (A), prolines(P), glycine (G) and cysteine (C).

The third step targeted the amino acids on surface of the protein, 40 out of 240 aa. 24 input pdb were chosen between the best of the previous step for the energy score and, for each of them, 50 outputs were requested. The substitution instructions (RESFILE 3) were:

- 5 residues on each α-helix can be substituted with any polar amino acid with the exception of glycine (G) and cysteine (C).

The fourth step targets the amino acids that will form the loops of the protein, 64 out of 240 aa. 50 input pdb structures are chosen between the best of the previous step for the energy score and, for each of them, 80 outputs are requested. The substitution instructions (RESFILE 4) are:

- 2 residues on each β-strand and 6 on each α-helix can be substituted with any amino acid with the exception of glycine (G), cysteine (C), tryptophan (W), tyrosine (Y), lysine (K) and phenylalanine (F).

After each step, the models were checked for sequence duplicates (redundant outputs were discarded), and for energy score. At the end of the substitution, a total of 3968 different models were created.

Command lines and PARAMETER, OPTION and RESFILE files are reported in Annex 6.5.4, page 247.

3.2.3 Loop closure

Loop closure was performed with the package Loopmodel of Modeller. The program formed the loop connection bringing the selected residues of the first and of the second chain close enough in space to form a peptide bond. Two residues were selected from the strand, at both N- and C-termini, and three from the helix. Out of this 5, one was the boundary residue with the β-strand, one with the α-helix and the central ones are the final loop.

For each of the 3968 models an instruction file was prepared (details in Annex 6.5.5, page 255). Instructions included the input pdb name, the method to use (loopmodel), the residue number at the beginning and at the end of the connection and the number of outputs (one). In our case the first loop was formed from aa 8 to 12; the second loop from aa 27 to 31, etc. The residues that were involved in the loop formation were shortly minimized by the program, but not the α-helices and β-strands.

3.2.4 Energy minimization

Energy minimization of the 3968 models was performed with the package Relax of the Rosetta software (version 2015.19.57819_bundle). The program required as input only the pdb files of interest and produced as outputs the relaxed pdb and a file with the corresponding energy scores.

The option `relax:constrain_relax_to_start_coords` was used in order to restrict the relaxation to the initial coordinates of the structure.

3.2.5 Sequence design

28 backbone structures were selected among the 3968 models (see Section 2.2.6, page 52) and were subjected to 10 cycles of sequence design and energy minimization with the packages Design and Relax of Rosetta, (version 2015.19.57819_bundle).

The first cycle of design targeted the 96 aa that form the hydrophobic core of the proteins (see Section 3.2.2 for details, page 174), and allowed them to be changed in any apolar residue. For each input structure, 100 outputs were requested, for a total of 2800 structures. After energy minimization, the models were checked for sequence redundancy and ranked by both total energy and RMSD against the initial structure. 106 were selected based on the best score in both categories, at least 3 models from each of the 28 initial families.

2^{nd}-4^{th} cycles targeted all the 240 aa of the selected 106 models. The 96 residues of the core layer were allowed to be changed only in apolar aa, the 40 residues of the surface layer only in polar aa and the 104 aa of boundaries and loops layers in any residue. For each input, 10 outputs were requested. After 3 cycles of design and energy minimization, the models were checked for sequence redundancy, ranked for total energy and for the RMSD against the initial structures, and 88 best candidates were selected, at least 2 for each of the 28 input structure.

5^{nd}-7^{th} cycles targeted all the 240 aa of the selected 88 models, with the same restrictions described in the previous paragraph. For each input, 10 to 15 outputs were requested. After 3 cycles of design and energy minimization, the models were checked for sequence redundancy, ranked for total energy and for the RMSD against the initial structures, and 80 models were selected for each of the input structure.

8^{nd}-10^{th} cycles targeted all the 240 aa of the selected 80 models and allowed them to be changed in any other residue without restrictions. For each input, 35 outputs were requested.
Command lines and PARAMETER, OPTIONS and RESFILE files containing the instruction for each layer of the proteins are reported in Section 6.5.7, page 256.

3.3 *In silico* validation

For experimental validation, 10 out of 7949 models were selected in 6 consecutive steps by *in silico* validation. Each step was first applied to the natural TIM-barrel collection (Section 2.1.1, page 33) in order to establish cut-off values that were then applied to the artificial models. The first 4 steps of validation were already described in the previous sections: analysis of the total energy (Step 1) and β-energy (Step 2) in Section 3.1.3, page 170; analysis of amino acid composition (Step 3) and properties (Step 4) in Section 3.1.4, page 171. Secondary structure prediction (Step 5) and molecular dynamic simulations (Step 6) are described in this chapter.

3.3.1 Secondary structure prediction

Secondary structure predictions were performed with SSpro and Jpred4. For both programs, the predictions are based on the amino acid propensity to form α-helix and β-strands, but SSpro also integrates evolutionary information that increases the precision of the prediction for natural proteins.

SSpro was installed and run locally. Fasta sequences were the input files and each residue of the protein was assigned to one of three structural elements: C for coil (unstructured regions), H for α-helix and E for β-strand. For example, the SSpro output of the natural protein 1A53 is:

```
>1a53A
CCCCCHHHHHHHHHHHCCCCCCCCCCCCCCCCHHHHHHHHHHCCCCCEEEEECCECCCCCECCCCHHHHHHHHCCCCCEEEEECCCCCCCCCCHHHHHHHHHHCC
CCEEEECCCCCCHHHHHHHHHHCCCEEEEEHHHCCHHHHHHHHHHHHHCCCCCEEEECCHHHHHHHHHCCCCEEEEECECCCCCCECHHHHHHHHHCCCCCCEE
EEECCCCCCHHHHHHHHHCCCCEEEECHHHHHCCCHHHHHHC
```

JPred4 is a web-service, and the fasta sequences were uploaded to the web-site (the web-link is reported in Annex 6.1, page 231). Outputs included both the amino acid sequence and the secondary structure prediction, and differently from SSpro, unstructured regions were defined with the dash symbol (-) instead of "C":

```
>1a53A
PRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAEYKRKSPSGLDVERDPIEYSKFMERYAVGLSILTEEKYFNGSYETLRKIASSVS
---HHHHHHHHHHHHHHHHHHHHHH------HHHHHHHH-------EEEEEE------------HHHHHHHHH----EEEEE---------HHHHHHHHHH--

IPILMKDFIVKESQIDDAYNLGADTVLLIVKILTERELESLLEYARSYGMEPLIEINDENDLLDIALRIGARFIGINSRDLETLEINKENQRKLISMIPSNVVK
--EEE-------HHHHHHHHH---HHHHHHHH---HHHHHHHHHHHH----EEEEE--HHHHHHHHH----EEEEE------------HHHHHHHH-----EE

VAESGISERNEIEELRKLGVNAFLIGSSLMRNPEKIKEFIL
EEE-----HHHHHHHHH----EEEEE--------HHEEE--
```

Output files of both SSpro and JPred4 were modified afterwards to annotate the unstructured regions with the identifier "L", and then compared with the DSSP outputs

(that use the "L" identifier). JPred4 and SSpro assigned the secondary structure elements based on the amino acid sequence, while DSSP assigned them based on the 3D coordinates of pdb structure files (See Section 3.1.2, page 169). If the residue has the same conformation in the comparison between SSpro and DSSP or between JPred4 and DSSP the prediction is considered valid, otherwise not.

The pool of natural TIM-barrels was analyzed and for each software a cut-off value was chosen for the evaluation of the artificial sequences.

3.3.2 Molecular dynamics

Molecular dynamic simulations were done with GROMACS. The procedure consisted in 7 steps: the first six for the preparation of the protein structure and its environment, and the last one for the proper simulation. Command lines and PARAMETER files for each step are reported in Annex 6.5.8, page 275.

1. **Generate the topology:** pdb structure files (downloaded from RCBS-PDB or created with Rosetta) cannot be used as input files by GROMACS directly, they have first to be converted into the GROMACS format (topology files), that includes information on atom type, charges, bonds, angles, dihedral angles, force-fields and water-type. The program `pdb2gmx`, included in the GROMACS package, was used to transform the pdb files in topology files. The force-field in all the simulations was AMBER99SB, and the water-type was TIP3P.

2. **Add periodic boundaries:** the program `editconf` of the GROMACS package was used to create a "box" around the protein in order to define the dimensions of the simulation. The choice of the box geometry is important, because the number of atoms in the box directly influences the time of the simulation. For this reason a dodecahedron shape was chosen, which has 30% less of volume compared to the cubic shape. The box was created at least at 1 nm distance from the protein atoms.

3. **Add water molecules:** all simulations were performed in an aqueous environment. Water molecules were added with the package `solvate` of GROMACS arranged in multiple layers around the surface of the protein.

4. **Add ions:** ions were added to the system to mimic the effect of buffer solutions and to balance the charges present on the protein surface. Na^+ and Cl^- atoms were added to the system (protein + water) with the package `genion` of GROMACS, to

mimic a final concentration of 150 mM NaCl, which will be used in experimental conditions.

5. **Energy minimization:** to remove overlapping atoms, clashes or empty regions which may have been created in the previous steps, energy minimization was run on all the atoms of the system (protein, water and ion). The package for energy minimization was `mdrun`, included in GROMACS.

6. **Temperature and pressure coupling:** after energy minimization the protein structure is ready for the simulation, however the solvent is still not completely equilibrated and oriented in the correct way. Simulations in these conditions can collapse the system and 2 steps of equilibration are recommended: first temperature and then pressure coupling. In both cases the protein structure was restricted in its position while the solvent molecules and ions were free to move around the protein. When temperature was reached, pressure was applied to the system in order to reach the proper density. As for the energy minimization, the package `mdrun` was used.

7. **Molecular dynamic simulation:** after preparation of the structure, addition of water molecules and ions, energy minimization and equilibration, the system is ready for the simulation. It lasted 50 ns, with $25x10^6$ steps of 2 fs each. Atom coordinates, velocities and energies were saved every 10000 fs, for a total of 5000 data-points. The package used for the simulation was `mdrun`.

Molecular dynamics is time-consuming and it is not possible to simulate all the 598 models that are left after the first 5 steps of *in silico* validation. For each of the 28 Families, the best model in terms of Rosetta energy scores was selected for the simulation. Two natural TIM-barrels (1K77 of 260 aa and 4AAJ of 200 aa) and the model of the artificial TIM-barrel described in Section 1.4.2, page 30, (sTIM-11 of 184 aa) were simulated as positive controls. The model of Octarellin V.1, described in Section 1.3.5, page 24, was simulated as negative control.

3.3.3 Sequence Alignment

The 10 models that were chosen for experimental validation (see Section 2.3.9, page 68) were tested for sequence alignment against a not-redundant database of natural protein sequences with BLAST (web-link, page 231).

Fasta sequences were pasted into the "Enter Query Sequence" box and "Non-redundant protein sequences (nr)" was selected in the "Database" drop down menu. None of the

other options were selected and the “BLAST” button was chosen to start the alignment.

3.3.4 Cysteine removal

Cysteine removal was done with Rosetta, a round of Design, to substitute the side-chain atoms, followed by a round of Energy Minimization, to relax the structure. The procedure is the same that was used in the Sequence Design chapter (Section 2.2.6, page 52). OctaVII_04 contained 6, OctaII_05 contained 3, OctaVII_06 contained 2, OctaVII_09 contained 5 and OctaVII_07, OctaVII_08 and OctaVII_10 contained 6 cysteines each. Command lines and OPTION, PARAMETERS and RESFILE files are reported in Annex 6.5.9, page 279.

Chapter 4

Material and Methods

4.1 Materials

4.1.1 Chemicals and consumables

Chemicals have been purchased from the following suppliers: Carl Roth, Filter Company, Merck Chemicals, MP Biomedicals, Thermo Fisher Scientific and VWR. Consumables have been purchased from the following suppliers: BD Medical Technology, Filter Company, Rocc, Sarstedt and VWR.

4.1.2 Commercial kits, enzymes and buffers

Commercial kits, enzymes and buffers are reported in Table 4.1.

Name of kit or enzyme	**Supplier**	**Method**
NcoI, 10000 units/mL	NEB	DNA Digestion
XhoI, 20000 units/mL	NEB	DNA Digestion
Buffer SmartCut, 10x	NEB	DNA Digestion
NucleoSpin® Gel and PCR Clean-up	Macherey-Nagel	DNA purification
T4-DNA Ligase	Bioke	Ligation
T4-DNA Ligase Buffer	Bioke	Ligation
GeneJET Plasmid Miniprep Kit	Thermo Fisher Scientific	Plasmid purification
NucleoBond® Xtra Midi/Maxi	Macherey-Nagel	Plasmid purification
QuickChange II XL Site-Directed Mutagenesis Kit	Agilent Technologies	Mutagenesis
Midori Green Advance	Nippon Genetics Europe	Agarose electrophoresis
Gel Loading Dye Purple 6x	NEB	Agarose electrophoresis
O'GeneRuler 1kb DNA ladder	Thermo Scientific	Agarose electrophoresis
Benzonase, 250000 units/mL	Merck	Protein extraction
Precision Plus™ Unstained Protein Standards	Bio-Rad	SDS-PAGE
Unstained Protein Molecular Weight Marker	Thermo Scientific	SDS-PAGE

Name of kit or enzyme	Supplier	Method
His-Tag Antibody HRP Conjugate kit	Merck Chemicals	Western-Blot
Clarity™ Western ECL substrate	Bio-Rad	Western-Blot
Gel Filtration Calibration Kit	GE Healthcare	Protein purification

Table 4.1: Commercial kits, enzymes and buffers

4.1.3 Primers

Primers for sequencing and for *in-situ* mutagenesis are reported in Table 4.2.

Primer Name	Length	Sequence	Method
T7 prom	20 nt	TAATACGACTCACTATAGGG	Sequencing
T7 term	19 nt	CTAGTTATTGCTCAGCGGT	Sequencing
Y57Q forward	39 nt	CAGAGCTGGCACGTGAT**CAG**AGCTGTGGTGGTATGGGTC	Mutagenesis
Y57Q reverse	39 nt	GACCCATACCACCACAGCT**CTG**ATCACGTGCCAGCTCTG	Mutagenesis

Table 4.2: Oligonucleotides

The Y57Q pair of primers are used to introduce a site point mutation in OctaVII_02. The codon "TAC", that codifies for a tyrosine (Y), is changed in "CAG", that codifies for a glutamine (Q). This codon is reported in bold to highlight the losation of the mismatching nucleotides.

4.1.4 Bacterial strains

Commercial *E. coli* competent cells are reported in Table 4.3.

Strain	Genotype	Supplier	Method
BL21 (DE3)	fhuA2 ompT gal γ DE3 [dcm] ΔhsdS	NEB	Protein production
DG1	mcrA ΔM15 ΔlacX74 recA1 araD139 Δ(ara-leu)7697	Eurogentec	Plasmid replication

Table 4.3: *E. coli* strains

4.1.5 Growth media

Media for bacterial growth are reported in Table 4.4. All media were autoclaved prior to use.

Medium	Components	Concentration	Method
LB broth	Tryptone	10.0 g/L	Bacterial growth
	Yeast Extract	5.0 g/L	
	NaCl	10.0 g/L	
LB agar	Tryptone	10.0 g/L	Plating
	Yeast Extract	5.0 g/L	
	NaCl	10.0 g/L	
	Agar	15.0 g/L	
2XYT 2% Glucose	Tryptone	16.0 g/L	Fermentation
	Yeast Extract	10.0 g/L	
	NaCl	5.0 g/L	
	Glucose	20.0 g/L	
SOC	Tryptone	20.0 g/L	Transformation
	Yeast Extract	5.0 g/L	
	NaCl	0.6 g/L	
	Glucose	4.0 g/L	
	KCl	1.9 g/L	
	$MgCl_2$	2.0 g/L	
	$MgSO_4$	2.5 g/L	

Table 4.4: Growth media

4.1.6 Kanamycin

Kanamycin stock was prepared dissolving 1 g of kanamycin-sulphate powder in 20 mL of mQ water and filtering the solution with 0.22 μm filters. Aliquots of 1.5 mL were prepared and frozen for later use. The kanamycin stock of 50 mg/mL was diluted to a final concentration of 50 μg/mL.

4.1.7 IPTG

Isopropyl-3-thiogalactopyranoside (IPTG) stock was prepared dissolving 4.76 g of IPTG powder in 20 mL of mQ water and filtering the solution with 0.22 μm filters. Aliquots of 1.5 mL were prepared and frozen for later use. The IPTG stock of 1 M was diluted to a final concentration of 1 mM.

4.1.8 Buffers

Buffers were prepared with fresh mQ water and their pH was adjusted with hydrochloric acid (HCl, 37%) and sodium hydroxide (NaOH, 5 N). Buffers are listed in Table 4.5.

Name	Components	Concentration
Wash buffer, pH=7	Na_2HPO_4	50 mM
	NaCl	50 mM
Sample buffer 1, pH=8	Na_2HPO_4	50 mM
	NaCl	150 mM
	DTT	5 mM
Sample buffer 2, pH=7	Na_2HPO_4	50 mM
	NaCl	150 mM
Sample buffer 3, pH=8	Na_2HPO_4	50 mM
	NaCl	150 mM
Elution buffer 1, IMAC, pH=8	Na_2HPO_4	50 mM
	NaCl	150 mM
	Imidazole	500 mM
	DTT	5 mM
Elution buffer 2, IMAC, pH=7	Na_2HPO_4	50 mM
	NaCl	150 mM
	Imidazole	500 mM
Elution buffer 3, IMAC, pH=8	Na_2HPO_4	50 mM
	NaCl	150 mM
	Imidazole	500 mM
Denaturing buffer 1, urea, pH=8	Na_2HPO_4	50 mM
	NaCl	150 mM
	Urea	8 M
	DTT	5 mM
Denaturing buffer 2, urea, pH=7	Na_2HPO_4	50 mM
	NaCl	150 mM
	Urea	8 M
Denaturing buffer 3, urea, pH=8	Na_2HPO_4	50 mM
	NaCl	150 mM
	Urea	8 M
Denaturing buffer 1, GdmCl, pH=7	Na_2HPO_4	50 mM
	NaCl	150 mM
	GdmCl	6 M
Denaturing buffer 2, GdmCl, pH=8	Na_2HPO_4	50 mM
	NaCl	150 mM
	GdmCl	6 M
TAE buffer	Tris-acetate	40 mM
	EDTA	1 mM
SDS-PAGE, TGS buffer	Tris-HCl	50 mM
	SDS	1%
	Glycine	385 mM

Name	Components	Concentration
SDS-PAGE, Loading Blue 4x, pH=6.8	Tris-HCl	200 mM
	SDS	8%
	Glycerol	25%
	β-mercaptoethanol	5%
	Bromophenol Blue	1%
Transfer buffer, pH=8.0	Tris	25 mM
	Glycine	192 mM
	Methanol	25%
TBS buffer, pH=7.5	Tris-HCl	10 mM
	NaCl	500 mM
TBSTT, pH=7.5	Tris-HCl	20 mM
	NaCl	500 mM
	TritonX-100	0.2%
	Tween-20	0.05%

Table 4.5: Buffers

4.2 Vector construction

4.2.1 Gene and vector design

The vector chosen for *E. coli* protein production is pET28a by Novagen, that carries a T7lac promoter, a kanamycin resistance gene and a 6-His Tag that can be added at both the N- or/and the C- terminal parts of the protein. The restriction sites chosen for gene insertion are *NcoI* at the 5' and *XhoI* at the 3' of the gene. The target sequence of the *NcoI* restriction site is CCATGG. It contains the starting codon *ATG* that codifies for the initial methionine and it forces the use of a triplet *GXX* in the second amino acid position. Valine, alanine, aspartic acid, glutamic acid and glycine are the residues codified by the *GXX* triplet. OctaVII_01, OctaVII_05 and OctaVII_09 had allowed residues and only the methionine was added at the N-terminal part of the sequence. The remaining OctaVIIs were not compatible with the restriction site sequence and two residues were added at the N-term of the proteins: methionine and glycine.

The XhoI restriction site was necessary in order to insert the 6-His Tag coding sequence at the 3' part of the gene sequence. It involved the addition of 2 codons for leucine and glutamic acid (LE) between the gene and the 6x His-tag coding sequence. In total, 8 residues are added at the C-terminal part of the protein. In Table 4.6 there is a summary of the genes design and in Figure 4.1 there is an example of the final construct.

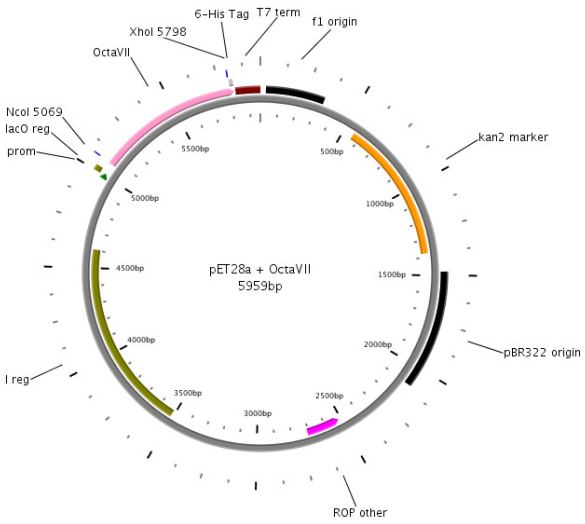


Figure 4.1: pET28a-OctaVII vector

Representation of the final construct pET28a-OctaVII. In evidence there are the sequences for sequencing (T7-prom and T7-term), restriction sites for the insertion of the OctaVII genes (NcoI and XhoI), the 6x His-tag coding sequence and the kanamycin resistance gene. The picture is realized with PlasMapper [141].

The final amino acid sequences of the 10 OctaVII were retro-translated to nucleotide sequences with the program "Codon Optimization Tool" of Integrated DNA Technologies (IDT), the company that synthesizes the gene fragments. The program is available at: www.idtdna.com/CodonOpt, and the amino acid sequences were pasted in the "Single Entry" box. The options "Amino Acids", "gBlocks Gene Fragments" and "*Escherichia coli* K12" were chosen for the queries "Sequence Type", "Product Type" and "Organism", respectively.

Gene sequences for OctaVII_01 to OctaVII_05 were synthesized and directly cloned in the pET28a vector by IDT, while OctaVII_06 to OctaVII_10 were synthesized by IDT as DNA fragments of 750 nb, called gBlocks®, and cloned in vector by us (see Section 4.2.2, page 190). In each gBlock®, the gene sequence of the OctaVII is 723 or 726 nb long. The remaining 24 nb are flanking region with the exact sequence of the pET28a vector, 12 nt upstream the *NcoI* and 12 nt downstream the *XhoI* restriction sites. Gene sequences for each OctaVII are reported in Annex 6.4, page 241.

	N-term aa	C-term aa	Total Length	Order type
OctaVII_01	M-	-LEHHHHHH	249	plasmid
OctaVII_02	MG-	-LEHHHHHH	250	plasmid
OctaVII_03	MG-	-LEHHHHHH	250	plasmid
OctaVII_04	MG-	-LEHHHHHH	250	plasmid
OctaVII_05	M-	-LEHHHHHH	249	plasmid
OctaVII_06	MG-	-LEHHHHHH	250	gBlocks®
OctaVII_07	MG-	-LEHHHHHH	250	gBlocks®
OctaVII_08	MG-	-LEHHHHHH	250	gBlocks®
OctaVII_09	M-	-LEHHHHHH	249	gBlocks®
OctaVII_10	MG-	-LEHHHHHH	250	gBlocks®

Table 4.6: Post-design modifications to the protein sequence

4.2.2 Digestion with *NcoI* and *XhoI*

gBlocks® with 500 ng of DNA fragments for the genes of OctaVII_06 to _10 were shipped by IDT as dry pellet. Centrifugation at 4000 rpm was done for 5 seconds to collect the DNA at the bottom of the tube, that was then resuspended in 20 μL of autoclaved and filtered mQ water. The final concentration of the gBlocks® solution was 25 ng/μL.

The 5 gene fragments and the pET28a plasmid were digested with *NcoI* and *XhoI* as first step in the vector construction. pET28a vectors were available in the stock of the

lab. The reaction mix (for both gBlocks® and plasmid) is reported in Table 4.7. The reaction tubes were incubated at 37°C for 1 hour.

	gBlocks®	**pET28a**
Buffer SmartCut (10x)	5 μL	5 μL
NcoI (10 U/μL)	1 μL	1 μL
XhoI (10 U/μL)	1 μL	1 μL
DNA	125 ng	1 ug
Total	50 μL	50 μL

Table 4.7: Digestion mixes

4.2.3 DNA clean-up

The digested gBlocks® fragments and pET28a plasmids were purified from the restriction enzymes using the *NucleoSpin® Gel and PCR Clean-up Kit* by Macherey-Nagel according to the protocol of the supplier. DNA was eluted in autoclaved and filtered mQ water.

4.2.4 DNA quantification

DNA final concentrations were measured with the NanoVue Plus instrument by GE Healthcare using 4 μL of DNA solution. The concentrations of the gene fragments are reported in Table 4.8.

4.2.5 Ligation

The ligation was performed with the T4 Ligase enzyme in the T4 Ligase Buffer and the 20 μL reaction mixes are reported in Table 4.8. Sample tubes were left at 4°C overnight and the reactions were completed at room temperature for 30 min in the morning, following an home made protocol that is widely used in the laboratory.

	[DNA]	**Insert**	**pET28a**	**Buffer**	**T4 Ligase**	**mQ water**
OctaVII_06	28.5 ng/μL	3 μL	1.25 μL	2 μL	1μL	12.75μL
OctaVII_07	17.0 ng/μL	5 μL	1.25 μL	2 μL	1μL	10.75μL
OctaVII_08	27.5 ng/μL	3 μL	1.25 μL	2 μL	1μL	12.75μL
OctaVII_09	23.9 ng/μL	4 μL	1.25 μL	2 μL	1μL	11.75μL
OctaVII_10	22.0 ng/μL	4 μL	1.25 μL	2 μL	1μL	11.75μL
pET28a	76.5 ng/μL	/	/	/	/	/

Table 4.8: DNA concentrations and ligation mixes

4.2.6 DNA agarose gel electrophoresis

Agarose gels were prepared melting in a microwave 1 g of agarose in 100 mL of TAE buffer (see Table 4.5). 10 μL of Midori Green Advance 10000x from Nippon Genetics Europe were added to the solution, and then poured in the electrophoresis chamber with a 6-wells comb inserted to generate the loading wells. DNA samples were mixed with Gel Loading Dye, Purple 6x by NEB, and loaded in the wells. 5 μL of O'GeneRuler markers by Thermo Fisher were added in a separated well. The voltage was set at 110 V and the electrophoresis lasted 30 min. DNA bands could be detected with a Gel Doc EZ Imager machine by Bio-Rad.

4.2.7 Site-directed mutagenesis

Site-directed mutagenesis was performed with the kit *QuickChange II XL Site-Directed Mutagenesis* from Agilent Technologies. The primers Y57Q forward and reverse were designed following the instruction of the supplier and are reported in Table 4.2 in the Material chapter.

The procedure for the site-directed mutagenesis is described in the supplier protocol. After transformation, different clones were subjected to sequencing in order to confirm the presence of the mutation (see Section 4.3.4, page 193).

4.3 Transformation and sequencing

4.3.1 Transformation

Transformation was performed with all the 10 pET28a-OctaVII vectors in both BL21(DE3) and DG1 *E.coli* competent cells. Aliquots of 50 μL of competent cells were mixed with 10-100 ng of pET28a-OctaVII vector and incubated in ice for 30 min. Cells were then heat-shocked at 42°C for 30 s and immediately transfered in ice for 5 min. 200 μL of room-temperature SOC medium was added to the competent cells and the mixtures were incubated at 37° for 60 min in shaking condition (250 rpm). 100 μL of cells culture were then spread onto selection plates (LB-Kan) and incubated overnight at 37°C.

4.3.2 Plasmid replication

E. coli DG1 transformants were used to produce the plasmid for sequencing and storage. Colony picking of 2 to 4 colonies for each OctaVII was done in 5 mL of fresh LB-Kan medium. The culture was grown for ∼8 hours at 37°C in shacking conditions (200 rpm). For a mini-preparation, 100 μL of culture were transferred into 5 mL of fresh LB-Kan medium. For a maxi-preparation, 4 mL of culture were transferred into 200 mL of fresh LB-Kan medium. In both cases, the flask were kept overnight at 37°C in shacking conditions (200 rpm).

4.3.3 Mini and maxi prep

Extraction of the plasmid from *E.coli* DG1 cells was done with two different kits: *GeneJET Plasmid Miniprep Kit* by Thermo Fisher Scientific for 5 mL cultures, and *NucleoBond® Xtra Midi/Maxi* by Macherey-Nagel for 200 mL cultures. In both cases the elution of the plasmid was done in filtered and autoclaved mQ water. Plasmid concentrations were measured as described in Section 4.2.4, page 191.

4.3.4 Sequencing

Sequencing of the OctaVII genes was done by the platform *Sanger Cycle Sequencing* of the GIGA-Genomics department at the University of Liège. Each plasmid was sequenced twice, in the forward direction with T7 prom primer and in the reverse direction with the T7 term primer. Each reaction mix was composed by 10 μL of primer 5 μM and by 10 μL of template pET28a-OctaVII 40-50 ng/μL.

4.4 Protein expression trials and production

4.4.1 Protein expression trials

After gene sequencing, the correct sequences were selected and used for transformation of BL21(DE3) *E.coli* cells (see Section 4.3.4 for details, page 193). Expression trials were done considering different induction times: 1h, 2h, 3h and 4h at 37°C or overnight at 18°C. Colonies were picked from the plates and grown for ∼8 hours at 37°C in shacking conditions (200 rpm). Cultures were done with a dilution 1:100 in 25 ml of fresh LB-Kan medium in the same conditions. Induction with IPTG 1 mM was done at 37°C for 1h, 2h, 3h or 4h or at 18°C overnight. At the end of the induction, cultures were diluted to reach Abs_{600}=0.6. Aliquots of 1 mL were centrifuged 5 min at 6000 rpm and the supernatant was discarded.

4.4.2 Sonication

The cell pellets of the expression trials were resuspended in 300 μL of Wash buffer 2 (see Section 4.1.8, page 185) and the proteins were extracted by sonication with the machine BioRuptor Plus by Diagenode. Each sample underwent to 10 cycles of 1 min (30 s of sonication and 30 s of pause) at 4°C. Aliquots of the crude extract (or total fraction, T) were centrifuged at 14000 rpm for 20 min at 4°C to produce the soluble fraction (S) and the insoluble fraction (I). Inclusion bodies were collected in the insoluble fraction. Analysis of the expression trials was performed by SDS-PAGE, see Section 4.5.1, page 195 for details.

4.5 Blotting techniques

4.5.1 SDS-PAGE

Sodium Dodecyl Sulphate Poly Acrylamide Gel Electrophoresis (SDS-PAGE) was used to separate proteins according to their molecular mass.

Gel preparation

During this project, both commercial and home-made polyacrylamide gels were used. Commercial polyacrylamide gels (Mini-PROTEAN® TGX™ Precasted Gel by Bio-Rad) presented a gradient from 4% of acrylamide in the top of the gel to 20% at the bottom. Home-made poly-acrylamide gels were composed of 2 parts, the stacking gel for the loading of the samples with 4% of acrylamide, and the running gel for the separation of the protein with 12% of acrylamide. 2 gels were prepared with the following reagents:

	Running Gel (12%)	**Stacking Gel 4%**
mQ water	3.4 mL	3.2 mL
Tris (1.5 M pH=8.8 / 0.5 M pH=6.8) buffers	2.5 mL	1.5 mL
Acrylamide / Bis-acrylamide	4 mL	1.2 mL
SDS 10%	100 μL	60 μL
Ammonium Persulfate 10%	100 μL	60 μL
TEMED	10 μL	6 μL

Table 4.9: Acrylamide gels preparation

Sample preparation

Sample solutions were mixed with SDS-PAGE Loading buffer (see Table 4.5) and heated at 95°C for 5 minutes. After a short spin-down, 20 or 10 μL of sample were loaded in the wells of the gel, depending on their capacity. 5 μL of marker (*Precision Plus™Unstained Protein Standards* by Bio-Rad or *Unstained Protein Molecular Weight Marker* by Thermo Scientific) were added in the gel wells.

Electrophoresis settings

Prior to sample loading, the gels were mounted in the Bio-Rad chamber for SDS-PAGE. The central chamber was filled with fresh SDS-PAGE TGS-buffer. After sample and marker loading, the chamber was closed and electrophoresis was performed for commercial gels at 120 V for 30 min. Hand-made gels were run at 100 V for the first 30 min

and at 150 V for the remaining time.

Gel staining

Gels were colored by immersion in the commercial InstantBlue buffer by Expedeon for a minimum of 15 min. The gels did not need decoloration steps and the gels were scanned with the Gel Doc EZ Imager instrument by Bio-Rad.

4.5.2 Western-Blot

Western blots were done on unstained SDS-PAGE with the His-Tag Antibody HRP Conjugate kit by Merck Chemicals that targets the histidine tag present at the C-terminal of the artificial proteins.

Protein transfer

Protein samples were loaded on SDS-PAGE gels and run as described in Section 4.5.1, page 195, but they were not stained at the end of the electrophoresis. Instead, the proteins in the acrylamide gel were transfered to a nitrocellulose membrane with the Trans-Blot Turbo machine by Bio-Rad.

The transfer pack (or blotting sandwich) was composed of 4 layers: the bottom ion reservoir, the nitrocellulose membrane, the acrylamide gel and the top ion reservoir. Each reservoir was composed by 3 Waterman papers soaked in Transfer buffer (see Table 4.5 for the composition, page 187). The nitrocellulose membrane was activated by a short immersion in ethanol 100% and equilibration in the Transfer buffer. The blotting sandwich was inserted into the cassette, in which the bottom and top trays were the anode and the cathode, respectively.

The program "Mixed MW" was chosen among the pre-programmed protocols of the Bio-Rad machine since it is well-suited for the transfer of a broad range of molecular masses (5-150 kDa).

Immuno-blotting

The nitrocellulose membrane was washed twice for 10 min with 15 mL 1x TBS buffer (see Table 4.5, page 187). It was incubated overnight in the blocking solution provided by the His-Tag Antibody HRP Conjugate kit. It was then washed twice for 10 min in 20 mL TBSTT buffer and one more time for 10 min in 15 mL TBS buffer. The antibody included in the kit was diluted 1:1500 in the blocking solution and it was incubated for 1

hour with the membrane. Three wash of 10 min were then performed, the first two in 20 mL TBSTT and the last one in 15 mL TBS.

Chemiluminescent detection

The substrate for chemiluminescent detection was prepared immediately before use with the Clarity™ Western ECL substrate kit by Bio-Rad. 1 mL of Luminol/Enhancer was mixed with 1 mL of Stable Peroxide Solution. The membrane was incubated for 1 min in the substrate solution. The chemiluminescence was detected with an ImageQuant LAS4000 instrument by GE Healthcare after 1 to 10 min exposure.

4.6 Protein production

4.6.1 Protein production in flasks

Colony picking

Individual colonies were picked with a sterile tip from the selection plate and were transferred to 5 mL of fresh LB-Kan medium. Culture was grown at 37°C for 8 to 15 hours in shacking conditions (200 rpm).

Pre-culture

The pre-culture was prepared starting from the 5 mL of the colony picking culture with a dilution of 1:100 in fresh LB-Kan medium. The pre-culture was grown at 37°C for 8 to 15 hours in shacking conditions (200 rpm).

Culture

The culture was prepared with a dilution 1:100 of the pre-culture in 2-4 L of fresh LB-Kan. The culture was grown at 37°C for 8h or at 18°C overnight in shacking condition (200 rpm). Induction was done with a final concentration of IPTG of 1 mM at 18°C overnight, or at 37°C for 1-4 hours.

4.6.2 Protein production by fermentation

Colony picking and pre-culture preparation were performed as described in the previous ection. The minimal volume for the pre-culture preparation was 500 mL.

Cultures of 10 L were prepared in New Brunswick™ BioFlo® 415 fermenters. The medium 2XYT 2% Glucose was autoclaved directly in the fermenter and the kanamycin was added together with the pre-culture. The culture was grown for 8h at 37°C, and induction was done overnight at 18°C with 1 mM IPTG.

4.6.3 Cell harvesting

Separation of the cells from the medium was done by centrifugation at 6000 rpm for 10 minutes at 4°C. The pellet was then resuspended in Wash buffer 2 (see Section 4.1.8, page 185) and stirred for 30 min at 4°C. Centrifugation was performed in the same conditions to re-collect the cells that were weighted for the next step of cell disruption.

4.6.4 Cell disruption

Each gram of pellet was resuspended in 4 mL of Sample buffer (1, 2 or 3 depending on the protein, see Section 4.1.8 for details, page 185). For each liter of culture, 4 μL of Benzonase were added to the resuspension. The sample was loaded in the Emulciflex-C3 homogenizer by ATA Scientific, and 3 to 4 cycles of disruption were performed. Centrifugation of the crude extract at 20000 rpm for 20 min at 4°C caused the separation of the soluble fraction (supernatant) from the insoluble fraction (inclusion bodies).

4.6.5 Inclusion bodies preparation

Inclusion bodies contains many soluble contaminants, and 2 washing cycles with the Wash buffer (see Section 4.1.8, page 185) were recommended. The pellet was collected by centrifugation at 20000 rpm for 20 min at 4°C and frozen for later use.

4.6.6 Protein N-sequencing

Sequencing of the N-terminus of OctaVII_02 was done at the University of Liège by Nicole Otthiers. The instrument is a "Precise Protein Sequencing System" produced by Perkin Elmer Corporation. The method is based on the sequential degradation of proteins, developped in the '50s by Pher Edman. The N-term of the first amino acid was first labeled, then cleaved and extracted from the protein solution for identification.

4.7 Protein purification

4.7.1 Sample and buffer preparation

Prior to any type of purification, the sample was centrifuged at 20000 rpm for 20 min at 4°C and the supernatant was filtered on 0.22 μm filters. All the buffers used for protein purification were filtered on 0.22 μm filters.

4.7.2 System preparation

Purifications were performed with a ÄKTA-Explorer chromatography system by GE Healthcare. Prior to any type of purification, the inlets, the system and the columns were washed with mQ water in order to remove the EtOH 20%. After equilibration in mQ water, inlets were washed with the correct buffer solution and the column is equilibrated with the loading buffer for 1 to 10 CV.

At the end of the purification, the samples were analyzed by SDS-PAGE as described in Section 4.5.1, page 195.

4.7.3 IMAC for soluble fraction

Column type

The purification of the soluble fraction of the crude extract was done with 2x 5 mL HisTrap HP columns by GE Healthcare. The flow rate and pressure setting were 5 mL/min and 0.3 MPa respectively. After each purification the columns were cleaned and stripped according to the protocol of the supplier.

Program Setting

The purification of the soluble fraction of the crude extract by IMAC was done in 6 steps: 1- short equilibration, 2- sample loading, 3- wash, 4- first step of elution at 15% of Imidazole buffer (75 mM), 5- second step of elution by gradient from 15 to 100% of imidazole buffer (0.5 M) and 6- post-elution step in which the imidazole buffer is kept at 100% for 5 CV. The program settings for each step are shown in Table 4.10 and a schematic gradient profile is shown in Figure 4.2a.

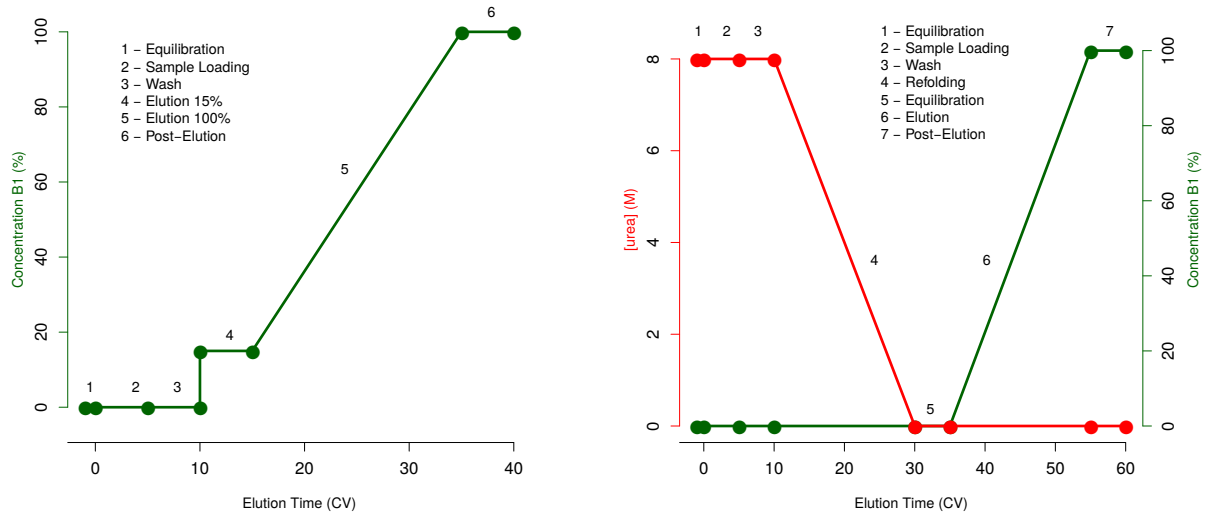
Three inlets were used in this purification: A11 for the equilibration buffer (Sample buffer), B1 for the elution buffer (Elution buffer, IMAC) and A12 for the sample loading.

Step	Variable	Setting
General setting	Column	HisTrap HP 5 mL
	FlowRate Equil	5 mL/min
	Column PressureLimit	0.3 MPa
	Wavelength 1	280 nm
	Wavelength 2	260 nm
	Averaging Time UV	5.12 sec
	BufferValve A1 Inlet	A11
1 Column Equilibration	Compensation Volume	8 mL
	Equilibrate with	1 CV
	FlowRate WashOut	5 mL/min
	Flowthrough FracSize	4 mL
2 Sample Loading	SampleInlet	A12
	Injection Flow rate	5 mL/min
	Sample Volume	XXX mL
3 Wash	Complete Flow rate	5 mL/min
	Complete Sample Load	20 mL
	Wash column with	5 CV
4 Elution with 15% imidazole	1 ConcB Step	15 %B
	1 Fraction Size	4 mL
	1 PeakFraction Size	0 mL
	1 Length of Step	4 CV
5 Gradient of elution	2 Fraction Size	4 mL
	2 PeakFraction Size	0 mL
	2 Target ConcB	100 %B
	2 Length of Gradient	20 base
6 Post-Elution	3 ConcB Step	100 %B
	3 Fraction Size	4 mL
	3 PeakFraction Size	0 mL
	3 Length of Step	5 CV
	Gradient Delay	8 mL

Table 4.10: Settings for soluble fraction purification by IMAC

4.7.4 IMAC for insoluble fraction and refolding

Prior to purification, inclusion bodies were dissolved in a denaturing buffer overnight at room temperature. Thanks to the 6x HisTag at the C-term of the proteins it was possible to perform both the refolding and the purification in one single step (called on column refolding). Elution profiles of IMAC purification from the soluble and from the insoluble fractions are shown in Figure 4.2 as example.



(a) **(b)**

Figure 4.2: Example of IMAC elution profiles

(**a**) Purification of the soluble fraction and (**b**) purification and refolding in column of the insoluble fraction by IMAC chromatography.

Column type

Purification and refolding were done with 2x 5 mL HisTrap HP columns by GE Healthcare. The flow rate and pressure setting were 5 mL/min and 0.3 MPa respectively. After each purification the column was cleaned and stripped according to the protocol of the supplier.

Program Setting

The purification and refolding of the inclusion bodies by IMAC was done in 7 steps: 1- short equilibration, 2- sample loading, 3- wash, 4- refolding, 5- short equilibration, 6- elution and 7- post-elution step in which the imidazole buffer is kept at 100% for 5 CV. The program settings for each step are shown in Table 4.11 and a schematic gradient profile is shown in Figure 4.2b.

Four inlets were used in this purification: A11 for the denaturation buffer (Denaturing buffer urea or GdmCl based), B1 for the refolding buffer (Sample buffer), A12 for the elution buffer (Elution buffer, IMAC) and A13 for the sample loading.

Step	Variable	Setting
General setting	Column	HisTrap HP 5 mL
	FlowRate Equil	5 mL/min
	Column PressureLimit	0.3 MPa
	Wavelength 1	280 nm
	Wavelength 2	260 nm
	Averaging Time UV	5.12 sec
	BufferValve A1 Inlet	A11
1 Column Equilibration	Compensation Volume	8 mL
	Equilibrate with	1 CV
	FlowRate WashOut	5 mL/min
	Flowthrough FracSize	4 mL
2 Sample Loading	SampleInlet	A13
	Injection Flow rate	5 mL/min
	Sample Volume	XXX mL
3 Wash	Complete Flow rate	5 mL/min
	Complete Sample Load	20 mL
	Wash column with	5 CV
4 Refolding	InletValve	A11
	2 Fraction Size	4 mL
	PeakFraction Size	0 mL
	Target ConcB	100 %B
	Length of Gradient	20 base
5 Equilibration	1 ConcB Step	100 %B
	Fraction Size	4 mL
	PeakFraction Size	0 mL
	Length of Step	5 CV
6 Gradient elution	InletValve	A12
	2 Fraction Size	4 mL
	PeakFraction Size	0 mL
	Target ConcB	0 %B
	Length of Gradient	20 base
7 Post-Elution	InletValve	A12
	3 ConcB Step	0 %B
	Fraction Size	4 mL
	PeakFraction Size	0 mL
	Length of Step	5 CV
	Gradient Delay	8 mL

Table 4.11: Settings for insoluble fraction purification by IMAC

4.7.5 Desalting

After purification by IMAC, a step of desalting was necessary to remove imidazole from the protein sample.

Column type

Desalting was done with a 135 mL Sephadex G-25 column by GE Healthcare. The flow rate and pressure setting were 5 mL/min and 0.3 MPa respectively.

Program Setting

Desalting of the imidazole sample was done in 3 steps: 1- short equilibration, 2- loading of the sample and 3- elution. The program settings for each step are shown in Table 4.12.

The inlets A11 was used for the Sample buffer. The sample was injected by a loop, with a maximum volume of 20 mL.

Step	Variable	Setting
General setting	Column	SephadexG25 26/40 135 mL
	FlowRate Equil	5 mL/min
	Column PressureLimit	0.3 MPa
	Wavelength 1	280 nm
	Wavelength 2	260 nm
	Averaging Time UV	0.01 sec
	BufferValve A1 Inlet	A11
1 Equilibration	Compensation Volume	8 mL
	Equilibrate with	0.1 CV
	FlowRate WashOut	5 mL/min
2 Sample Loading	Injection Flow rate	5 mL/min
	Empty loop with	30 mL
3 Elution	Elution Flow rate	5 mL/min
	Eluate Frac Size	4 mL
	Length of Elution	1.5 CV

Table 4.12: Program setting for desalting

4.7.6 Size exclusion: Superdex75

Size exclusion chromatography (SEC) was performed in two ways. For OctaVII_02 was used a preparative column, that allows bigger volumes to be loaded and that contributes to the purification of the protein. For the following proteins we decided to use an analytical column, that uses smaller volumes of sample and that is faster compared to the preparative one.

Column type

The two columns are: a preparative 120 mL Superdex 75 16/60 column and an analytical 24 mL Superdex-75 10/300GL column, both by GE Healthcare. The flow rate and pressure setting for the preparative column were 1 mL/min and 0.5 MPa, while for the analytical one were 0.5 mL/min and 1.8 MPa.

Program Setting

In both cases, size exclusion was done in 3 steps: 1- short equilibration, 2- loading of the sample and 3- elution. The program settings for each step are shown in Table 4.13 for the preparative column and in Table 4.14 for the analytical one. One inlets was used in this purification: A11 for the buffer. The sample was injected by a loop, with a maximum volume of 5 mL for the preparative column and of 500 μL for the analytical one.

Step	Variable	Setting
General setting	Column	Superdex-75 16/60GL 120 mL
	FlowRate Equil	1 mL/min
	Column PressureLimit	0.5 MPa
	Wavelength 1	280 nm
	Wavelength 2	260 nm
	Averaging Time UV	5.12 sec
	BufferValve A1 Inlet	A11
1 Equilibration	Compensation Volume	8 mL
	Equilibrate with	0.1 CV
	FlowRate WashOut	1 mL/min
2 Sample Loading	Injection Flow rate	1 mL/min
	Empty loop with	5 mL
3 Elution	Elution Flow rate	1 mL/min
	Eluate Frac Size	2 mL
	Length of Elution	1 CV

Table 4.13: Settings for preparative SEC

Step	Variable	Setting
General setting	Column	Superdex-75 10/300GL 23 mL
	FlowRate Equil	0.5 mL/min
	Column PressureLimit	1.8 MPa
	Wavelength 1	280 nm
	Wavelength 2	260 nm
	Averaging Time UV	5.12 sec
	BufferValve A1 Inlet	A11
1 Equilibration	Compensation Volume	8 mL
	Equilibrate with	0.1 CV
	FlowRate WashOut	0.5 mL/min
2 Sample Loading	Injection Flow rate	0.5 mL/min
	Empty loop with	500 μL
3 Elution	Elution Flow rate	0.5 mL/min
	Eluate Frac Size	1 mL
	Length of Elution	1 CV

Table 4.14: Settings for analytical SEC

Column calibration

Calibration was done only for the analytical column with the standards: Aprotinin, RNAse A, Ovalbumin and Conalbumin of the *Gel Filtration Calibration Kit* by GE Healthcare, and the Trypsin Inhibitor by Sigma. Three mixes were prepared:

- **Mix A:** Aprotinin 100 μL, Ovalbumin 100 μL, Sample buffer 100 μL
- **Mix B:** RNase A 100 μL, Conalbumin 100 μL, Sample buffer 100 μL
- **Mix C:** Trypsin inhibitor 100 μL, Sample buffer 200 μL

The three mixes were filtered on 0.22 μm filters and 100 μL were loaded in the Superdex-75 column. The elution profile is shown in Figure 4.3a. Molecular weight of the standards and their elution volumes are reported in Table 4.15.

Standard	Molecular weight	Elution volume
Aprotinin	6500 Da	15.22 mL
RNAse A	13700 Da	13.46 mL
Trypsin inhibitor	21000 Da	11.94 mL
Ovalbumin	44000 Da	10.21 mL
Conalbumin	75000 Da	9.45 mL

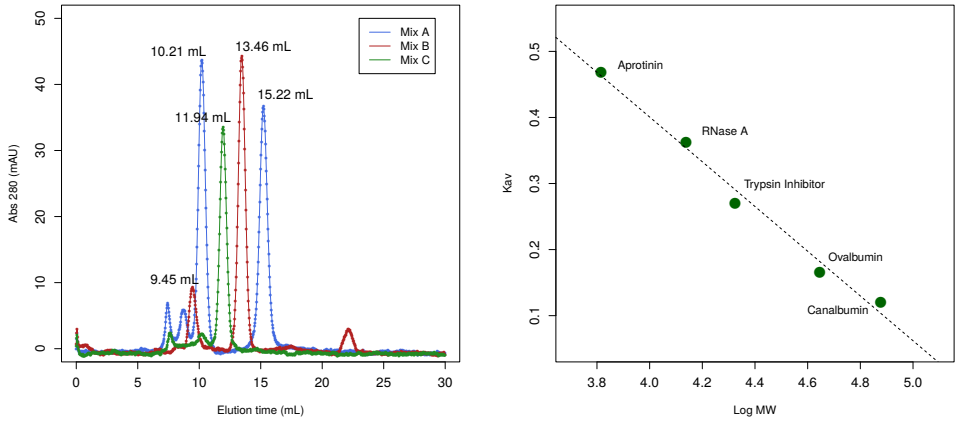
Table 4.15: Calibration standards

The calibration curve was calculated using the equation 4.1, in which K_{av} is the gel-

phase distribution coefficient, V_e is the elution volume (reported in Table 4.15), V_c is the geometric column volume (24 mL) and V_0 is the void volume (7.43 mL).

$$K_{av} = \frac{V_e - V_0}{V_c - V_0} \tag{4.1}$$

The calibration K_{av} versus Log molecular weight was plotted in Figure 4.3b.



(a) **(b)**

Figure 4.3: Superdex-75 calibration

(**a**) Elution profile of the standard proteins for calibration of Superdex-75 column and (**b**) plot of the molecular weight versus the gel-phase distribution coefficient (calibration curve).

4.7.7 Concentration

Protein concentration was performed by ultrafiltration with Amicon® Ultra centrifugal filters by Merck's Millipore Ltd with cut-off of 10 kDa and 15 mL of capacity. The sample solution was loaded onto the pre-equilibrated membrane and centrifuged at 4000 rpm at 4°C in 5 min cycles. When the final concentration was reached, the sample solution was centrifuged at 14000 rpm for 30 min at 4°C and filtered on 0.22 μm filters.

4.8 Biophysical Characterization

4.8.1 Absorbance

Absorption spectroscopy is a common technique to evaluate the concentration of a protein sample. At 280 nm aromatic residues (tryptophan and tyrosine) absorb part of the incident light, proportionally to the protein concentration. The Beer-Lambert law (Equation 4.2) explains the correlation between absorbance and protein concentration:

$$A_{280} = d \cdot \varepsilon \cdot C \tag{4.2}$$

where A_{280} is the absorbance at 280 nm, d is the path length (cm), ε is the molar extinction coefficient ($M^{-1}cm^{-1}$) and C the molar concentration of the protein sample (M). The protein concentration can be extracted in mg/mL with the following equation:

$$\frac{mg}{mL} = \frac{A_{280} \cdot MM}{d \cdot \varepsilon} \tag{4.3}$$

The molecular mass (MM) and the theoretic ε can be calculated from the amino acid sequence of the protein. For all the OctaVIIs, the theoretic extinction coefficient (ε), the molecular mass (MM) and the theoretic isoelectric point (pI) were calculated with ProtParam by ExPASy (see Table 4.16 for details).

	Molecular Mass (Da)	**Theoretic ε ($M^{-1}cm^{-1}$)**	**Theoretic pI**
OctaVII_01	28296.93	56950	6.90
OctaVII_02	27504.13	12950	6.46
OctaVII_03	28181.49	41940	6.60
OctaVII_04	27920.87	41480	6.27
OctaVII_04_NC	27930.87	41480	6.27
OctaVII_04_WS	27936.77	34490	6.02
OctaVII_05	28013.08	37930	6.01
OctaVII_05_NC	27974.98	37930	6.01
OctaVII_06	28410.09	42970	8.84
OctaVII_07	27888.90	48470	6.32
OctaVII_08	27811.14	28420	6.18
OctaVII_09	27773.58	23950	6.81
OctaVII_09_WS	27760.62	22460	6.21
OctaVII_10	27802.73	37470	5.90

Table 4.16: OctaVIIs: MW, ε and pI

Absorbance spectra were recorded with a JASCO V-630 spectrophotometer. The wavelength range was 230-350 nm, and data were collected at intervals of 1 nm with a scan speed of 100 nm/min at room temperature. The spectrum of the buffer (blanc) was subtracted to the sample spectrum before calculation of the protein concentration.

4.8.2 Far UV-CD

Circular dichroism (CD) in the far-UV region is a technique that allows evaluation of the secondary structure content in a protein sample. The principle of the technique is based on the fact that asymmetric molecules (i.e. proteins) may absorb in a different extend left- and right-handed circularly polarized light. The CD signal is the result of the difference in absorption of the two polarized lights, as shown in Equation 4.4.

$$\Delta A = A_R - A_L \tag{4.4}$$

where ΔA is the difference in absorbance (or CD signal) and A_R and A_L are the absorbances of right-handed and the left-handed circularly polarized lights respectively. The Beer-Lambert law, shown in Equation 4.5, is at the basis of the circular dichroism spectroscopy:

$$A = \varepsilon \cdot d \cdot C \tag{4.5}$$

where A is the absorbance, ε the molar extinction coefficient, d the path length and C the sample concentration. The CD signal is related to the Beer-Lambert law by the following equation:

$$\Delta A = (\varepsilon_R - \varepsilon_L) \cdot d \cdot C = \Delta\varepsilon \cdot d \cdot C \tag{4.6}$$

where ε_R and ε_R are the molar extinction coefficients for right-handed and left-handed circularly polarized lights respectively, and $\Delta\varepsilon$ is their difference.

For historical reasons, CD is usually expressed in ellipticity (Θ) that is related to ΔA with the equation:

$$\Theta = 32.98 \cdot \Delta A \tag{4.7}$$

The ellipticity can be transformed in molar ellipticity ($[\Theta]$) according to the following equation:

$$[\Theta] = \frac{\Theta \cdot 100 \cdot MM}{C \cdot d} \tag{4.8}$$

where MM is the molecular weight of the protein. The CD signal can also be expressed as mean residue ellipticity ($[\Theta]_{MRW}$):

$$[\Theta]_{MRW} = \frac{[\Theta]}{N_{aa} - 1} \tag{4.9}$$

where (N_{aa} - 1) is the number of peptide bonds in the protein sample.

CD spectra were recorded with a JASCO J-810 spectropolarimeter at 20°C, with a wavelength range between 260 and 185 nm. Protein samples were pipetted in 1mm pathlength quartz Suprasil cell (Hellma) at a concentration between 0.05 and 0.15 mg/mL and maximum volume = 300 μL. Four scans (10 nm/min, 1 nm bandwidth, 0.2 nm data pitch and 1 s DIT) were averaged. The high-tension voltage (HT) was recorded in parallel to the CD spectrum. If the HT signal was higher than 600 V, the CD signal was not reliable and those data-point were discarded. The spectrum of the buffer (blanc) was subtracted to the protein spectrum and the CD signal was transformed in mean residue ellipticity according to the following equation:

$$[\Theta]_{MRW} = \frac{\Theta \cdot 100 \cdot MM}{C \cdot d \cdot N_{aa}} \tag{4.10}$$

4.8.3 Intrinsic Fluorescence

Aromatic amino acids (tryptophan and tyrosine) of proteins can absorb light at 280 nm (excitation wavelength) and release energy through fluorescent emission. Depending on the environment that surrounds the aromatic residues, the fluorescence emission may change. This phenomenon is an useful tool to inquire the tertiary structure and the folding status of proteins.

Fluorescence spectra were recorded on a Cary Eclipse spectrofluorimeter by Varian, with the excitation wavelength at 280 nm and the emission spectrum in the range of 300 to 400 nm. The sample was pipetted in a quartz cuvette of 1 mL of volume, in a concentration between 0.01 and 0.2 mg/mL.

4.8.4 Chemical Unfolding

Unfolding of protein was measured with both CD and fluorescence techniques with the sample in different concentrations of denaturing agent (GdmCl or urea).

Sample preparation

The samples were prepared with the use of a pipetting robot, Microlab STAR Hamilton, hosted at the Robotein platform of the Center for Protein Engineering (CIP) of the

University of Liege. Three solutions were requested: the protein sample (minimal concentration = 2 mg/mL), the standard buffer of the protein and the denaturing buffer (GdmCl 6M or urea 8M). The three solutions were pipetted in a 48-wells plate with 2 mL of capacity each. The final protein concentration was 0.1 mg/mL and the concentration of the denaturing agent increased from 0 M to the maximum.

Chemical unfolding followed by CD

For each sample, the CD signal was measured on a JASCO J-810 spectropolarimeter at a fixed wavelength of 222 nm for 60 s. Protein samples were pipetted in quartz cuvettes of 1 mm of path length (maximum volume = 300 μL) and the recording was done at room temperature.

The average of the CD signal at 222 nm was plotted versus the concentration of the denaturant to obtain the unfolding curve of the protein.

Chemical unfolding followed by fluorescence

For each sample, the fluorescence spectrum was recorded with a Cary Eclipse spectrofluorimeter by Varian in the same conditions as described in Section 4.8.3, page 211. The intensity of the signal at a fixed-wavelength was plotted versus the concentration of the denaturing agent to obtain the unfolding curve of the protein. The fixed-wavelength was chosen as the wavelength at which the difference between the unfolded and folded protein spectra is the biggest.

4.8.5 Thermal Unfolding

Temperature-mediated protein unfolding was measured with both CD and fluorescence techniques. The samples were in the native buffer at a final concentration of 0.1 mg/mL. A drop of mineral oil was added in the cuvette to avoid evaporation of the protein sample. A thermo-coupler sensor (Testo 926 by Testo) was inserted in the protein solution to record the actual temperature of the sample.

Thermal unfolding followed by CD

A full CD spectrum was recorded at 25°C before the thermal unfolding experiment as described in Section 4.8.2, page 210. For the thermal unfolding, the CD signal was measured on a JASCO J-810 spectropolarimeter at a fixed wavelength of 222 nm. The initial and final temperatures were 25°C and 95°C respectively. Temperature increased at a rate of 0.5°C per minute. At the end of the thermal unfolding, at 95°C, a full CD

spectrum was recorded, as described in Section 4.8.2, page 210. The sample was then cooled down to 25°C and another full CD spectrum was recorded.

The CD signal at 222 nm was plotted versus the temperature to obtain the unfolding curve of the protein. The 3 full spectra were analyzed as described in Section 4.8.2, page 210.

Thermal unfolding followed by fluorescence

A fluorescence spectrum was recorded before unfolding with a Cary Eclipse spectrofluorimeter by Varian in the same conditions described in Section 4.8.3, page 211. Then the fluorescence was measured at a fixed-wavelength from 25°C to 95°C, with a speed of 0.5°C per minute and from 95°C to 25°C, at the same speed. At the end of the experiment a full fluorescence spectrum (described in Section 4.8.3, page 211) was recorded.

4.8.6 Near UV-CD

Contrary to the far-UV CD that measures secondary structure percentages in the protein sample, near-UV CD give information on the tertiary structure of the protein. The difference is due to the different targets of the 2 techniques, the first works on the peptide bonds, while the latter on the aromatic amino acids (tryptophan, tyrosine and phenylalanine).

Near-UV spectra were recorded on a JASCO J-810 spectropolarimeter with a wavelength range between 250 and 340 nm. Protein samples were pipetted in quartz cuvettes of 10 mm of path length (volume $\geq$ 1.7 mL) at a concentration between 0.5 and 2.5 mg/mL. The recording was done at room temperature and the spectrum of the buffer (blanc) was subtracted to the protein spectrum.

4.9 Crystallization

Protein crystallization was tested with the sitting drop vapor diffusion method. Water in the protein solution evaporates in order to equilibrate with solution in the reservoir, that is more concentrated. This vapor diffusion leads to a slow concentration of the protein sample, that eventually will form crystals.

480 different combinations of buffers, salts, pHs and detergents were tested with the use of the Mosquito robot by TTP Labtech, hosted at the Biological Macromolecule Crystallography lab at the Center of Protein Engineering (CIP) of the University of Liege.

iQ-plates from TTP Labtech contains 96-wells, and each of them is divided into 4 sections: a reservoir that contains 40 μL of buffer solutions and 3 spots of 0.6 μL for the protein drops. Crystallization kits for the 480 different conditions are reported in Table 4.17.

Once the plates were filled by the robot, they were hermetically closed with a film in order to create a closed system. The preparation and the storage of the crystallization plates were done at a constant temperature of 20°C. Plates were visually inspected at 3, 7 and 30 days after the preparation and every following months.

Kit name	**Producer**	**Conditions**
Crystal Screen	Hampton Research	48
Crystal Screen 2	Hampton Research	48
Index	Hampton Research	96
Wizard Classic	Emerald	48
Wizard Classic 2	Emerald	48
Salt	CIP	96
SPE	CIP	96

Table 4.17: Crystallization kits

Chapter 5

Conclusions and perspectives

Protein *de novo* design is a recent and challenging research area. In this work we presented the design of a new generation of Octarellins, artificial proteins modelled on the TIM-barrel fold. More than 8000 different sequences have been created with the software Rosetta and Modeller. They have been tested and ranked according to different parameters: the Rosetta energy, the amino acid composition, the prediction of secondary structures and molecular dynamic simulations. 10 among them have been chosen for experimental validation, and 5 mutants have been created meanwhile. During the experimental validation, we faced many issues that were not predictable during the design, as non-expression of the protein (OctaVII_01), N-terminal truncations (OctaVII_02, OctaVII_03 and OctaVII_07), and high insolubility of inclusion bodies (OctaVII_06 and OctaVII_08).

OctaVII_01 is not expressed in *E. coli* BL21(DE3) cells, and we showed that cells carrying this plasmid grow slower compared to cells with the OctaVII_02 gene (see Figure 2.57). These results suggest that the metabolism of the cell is altered by the presence of OctaVII_01 construct. In order to obtain the expression of OctaVII_01, it is possible to try different *E. coli* strains or alternative host (i.e. yeast). However, the problem might be bound to the DNA or mRNA sequences (i.e. sites recognized by DNA-binding or RNA-binding repressors). In this case, a re-optimization of the gene sequence might solve the problem.

This solution, the re-optimization of the gene sequences, might be applied also to the truncated proteins OctaVII_02, OctaVII_03 and OctaVII_07. We demonstrated for OctaVII_02 that the truncation is not due to endogenous proteolysis, since its mutant OctaVII_02 Y57Q also resulted truncated (see Figure 2.63). A second hypothesis to ex-

plain the N-terminal truncations of the three variants is the presence of internal ribosome binding sites (RBS) in the gene sequences. Re-optimizing the sequences to avoid internal RBS might be the solution to recover the full length of the proteins.

The third unexpected issue previously mentioned is the high insolubility of the inclusion bodies of OctaVII_06 and OctaVII_08. Indeed, these inclusion bodies were not solubilized in 8 M urea and just partially solubilized in 6 M GdmCl. This result is extremely uncommon and it is not reported in the literature. In order to fully solubilize the proteins it is possible to increase the incubation time of the inclusion bodies in the denaturing buffer to several days. Moreover, preliminary treatments with detergents such SDS and Tween may help in the solubilization of the inclusion bodies.

The high insolubility of the inclusion bodies of OctaVII_06 and OctaVII_08 is exceptional. On the contrary, protein insolubility upon expression and inclusion bodies formation are common problems in the field of *de novo* design and in the Octarellin history (see details in Chapter 1.3, page 19). In the present work, all the expressed proteins were produced in inclusion bodies, with only 4 partially present in the soluble fraction. We suggested, following experiments on OctaVII_05 (see Figure 2.90), that the presence of OctaVII_02, OctaVII_04 and OcatVII_05 in the soluble fraction was due to disulfide formation with endogenous proteins. This covalent interaction prevented the segregation of part of the protein in inclusion bodies. When free cysteines of OctaVII_04 and OctaVII_05 were substituted by other residues, the proteins were produced only in inclusion bodies. The fourth protein partially found in the soluble fraction is OctaVII_09 WS, which has no cysteines. This protein is a 14-residue mutant of OctaVII_09, which was produced only in inclusion bodies. The 14 mutations were rationally designed with the use of a recent version of Rosetta. The same method might be applied to OctaVII_04 and OctaVII_05 in order to shift part of their production to the soluble fraction of the cell.

Protein insolubility, however, is not only bound to protein over-expression in *E.coli*. With the exception of OctaVII_09 and OctaVII_09 WS, all the artificial proteins considered for further characterization were poorly soluble after refolding and prone to aggregation and precipitation, even at concentrations lower than 0.5 mg/mL. Despite the low solubility, these proteins presented defined secondary structures according to the far-UV CD analysis. The content of α-helices in the experimental proteins is lower than expected from the computational models. However, the content of secondary structures of most proteins corresponds to the one of the natural TIM-barrel of *Thermotoga maritima*, with about 36% of α-helices, 18% of β-strands and 19% of turns. Fluorescence measurements

confirmed the presence of compact tertiary structures. Since the λ_{max} of the refolded proteins ranged between 333 and 339 nm, we could exclude the presence in solution of molten globule states.

Among the 15 artificial proteins designed and produced in this work, OctaVII_09 and its mutant OctaVII_09 WS are the most promising. Only the first one was further characterized by chemical and thermal unfolding and by near-UV CD. It has a non-cooperative folding, however it is well folded and extremely thermostable. In order to confirm that the 14 mutations between OctaVII_09 and OctaVII_09 WS do not affect the secondary structure content and the overall folding, a full biophysical characterization has to be done on OctaVII_09 WS. Both proteins were soluble enough to be tested for crystallization, however we do not have yet information about the quality of the crystals that we obtained so far. As for Octarellin V.1 described in Chapter 1.3.5, page 24, crystallization helpers may help to stabilize the protein and to promote crystallization. Alternatively, structural analysis of both proteins may also be observed by cryogenic electron microscopy (cryo-EM). This is an emerging technique that has made huge progressions in terms of resolution (near-atomic), minimal size requirements (64 kDa) and applicability to challenging biological systems [142]. This technique is currently used in complement to NMR and X-ray crystallography. Its main advantage is that it does not require crystallization or labeling of the sample. Currently, many different biological structures obtained by cryo-EM are reported in the Protein Data Bank, including challenging biological molecules: viruses, membrane proteins, receptors, amyloid fibrils and large cellular machinery such as spliceosome and 26S proteasome [143]. Since OctaVII_09 and OctaVII_09 WS have been shown to form aggregates with a molecular mass higher than 70 kDa (see Section 2.4.14, page 155), they could potentially be interesting candidates for cryo-EM experiments. Indeed, our results obtained by circular dichroism and fluorescence on OctaVII_09 (see Section 2.4.13, page 144), confirmed the presence of secondary structures and a compact tertiary structure. This suggests that the aggregates are not composed by unordered structures and are therefore interesting to analyze. The cryo-EM technique can be extremely helpful not only to solve the structure of OctaVII_09, but also to analyze the aggregates that it forms in solution and elucidate their dimension, shape and heterogenity. The resolution of the structure remains the only method to judge the designs, and efforts have to be made in this direction.

To conclude, the design of artificial proteins is extremely challenging, and the results are not always the one that are expected. Improvements have to be done on both the computational side and the experimental one in order to reduce the “unexpected” results.

Bibliography

[1] Goraj, K. et al. "Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modelled on the α/β-barrel proteins". *Protein engineering* 3 (1990), pp. 259–266.

[2] Beauregard, M. et al. "Spectroscopic investigation of structure in octarellin (a de novo protein designed to adopt the α/β-barred packing)". *Protein engineering* 4 (1991), pp. 745–749.

[3] Houbrechts, A. et al. "Second-generation octarellins: two new de novo (β/α) 8 polypeptides designed for investigating the influence of β-residue packing on the α/β-barrel structure stability". *Protein engineering* 8 (1995), pp. 249–259.

[4] Offredi, F. et al. "De novo backbone and sequence design of an idealized α/β-barrel protein: evidence of stable tertiary structure". *Journal of molecular biology* 325 (2003), pp. 163–174.

[5] Figueroa, M. et al. "Octarellin VI: Using Rosetta to Design a Putative Artificial (β/α)8 Protein". *PLoS ONE* 8 (2013), e71858.

[6] Figueroa, M. et al. "The unexpected structure of the designed protein Octarellin V.1 forms a challenge for protein structure prediction tools". *Journal of Structural Biology* 195 (2016), pp. 19–30.

[7] Kuhlman, B. "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy". *Science* 302 (2003), pp. 1364–1368.

[8] Röthlisberger, D. et al. "Kemp elimination catalysts by computational enzyme design". *Nature* 453 (2008), pp. 190–195.

[9] Dahiyat, B. I. and Mayo, S. L. "De novo protein design: fully automated sequence selection". *Science* 278 (1997), pp. 82–87.

[10] Gutte, B. "A synthetic 70-amino acid residue analog of ribonuclease S-protein with enzymic activity." *Journal of Biological Chemistry* 250 (1975), pp. 889–904.

[11] Gutte, B. et al. "Design, synthesis and characterisation of a 34-residue polypeptide that interacts with nucleic acids". *Nature* 281 (1979), pp. 650–655.

[12] Moser, R. et al. "An artificial crystalline DDT-binding polypeptide". *FEBS Letters* 157 (1983), pp. 247–251.

[13] Pabo, C. "Molecular technology. Designing proteins and peptides". *Nature* 301 (1983), p. 200.

[14] Bowie, J. U. et al. "A method to identify protein sequences that fold into a known three-dimensional structure". *Science (New York, N.Y.)* 253 (1991), pp. 164–170.

[15] Jeschek, M. et al. "Directed evolution of artificial metalloenzymes for in vivo metathesis". *Nature* 537 (2016), pp. 661–665.

[16] Murphy, G. S. et al. "De Novo Proteins with Life-Sustaining Functions Are Structurally Dynamic". *Journal of Molecular Biology* 428 (2016), pp. 399–411.

[17] Regan, L. and DeGrado, W. F. "Characterization of a helical protein designed from first principles". *Science* 241 (1988), pp. 976–978.

[18] Hecht, M. H. et al. "De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence". *Science (New York, N.Y.)* 249 (1990), pp. 884–891.

[19] Yan, Y. and Erickson, B. W. "Engineering of betabellin 14D: Disulfide-induced folding of a β-sheet protein". *Protein Science* 3 (1994), pp. 1069–1073.

[20] Germann, H. P. and Heidemann, E. "A synthetic model of collagen: an experimental investigation of the triple-helix stability". *Biopolymers* 27 (1988), pp. 157–163.

[21] Urry, D. W. "Free energy (chemomechanical) transduction in elastomeric polypeptides by chemical potential modulation of an inverse temperature transition". *International Journal of Quantum Chemistry* 34 (1988), pp. 235–245.

[22] Talbot, J. A. and Hodges, R. S. "Tropomyosin: a model protein for studying coiled-coil and .alpha.-helix stabilization". *Accounts of Chemical Research* 15 (1982), pp. 224–230.

[23] Lear, J. D. et al. "Synthetic amphiphilic peptide models for protein ion channels". *Science* 240 (1988), pp. 1177–1181.

[24] Needleman, S. B. and Wunsch, C. D. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48 (1970), pp. 443–453.

[25] Smith, T. F. and Waterman, M. S. "Identification of common molecular subsequences". *Journal of Molecular Biology* 147 (1981), pp. 195–197.

[26] Bacon, D. J. and Anderson, W. F. "Multiple sequence alignment". *Journal of Molecular Biology* 191 (1986), pp. 153–161.

[27] Blundell, T. et al. "Knowledge-based protein modelling and design". *European Journal of Biochemistry* 172 (1988), pp. 513–520.

[28] Taylor, W. R. and Orengo, C. A. "Protein structure alignment". *Journal of Molecular Biology* 208 (1989), pp. 1–22.

[29] Brenner, S. E. and Berry, A. "A quantitative methodology for the de novo design of proteins". *Protein Science* 3 (1994), pp. 1871–1882.

[30] Dantas, G. et al. "A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins". *Journal of Molecular Biology* 332 (2003), pp. 449–460.

[31] Johansson, K. E. et al. "Computational Redesign of Thioredoxin Is Hypersensitive toward Minor Conformational Changes in the Backbone Template". *Journal of Molecular Biology* 428 (2016), pp. 4361–4377.

[32] Martina, C. E. et al. "HoLaMa: A Klenow sub-fragment lacking the 3'-5' exonuclease domain". *Archives of Biochemistry and Biophysics* 575 (2015), pp. 46–53.

[33] Jiang, L. et al. "De Novo Computational Design of Retro-Aldol Enzymes". *Science* 319 (2008), pp. 1387–1391.

[34] Korkegian, A. "Computational Thermostabilization of an Enzyme". *Science* 308 (2005), pp. 857–860.

[35] Balaram, P. "De novo design: backbone conformational constraints in nucleating helices andβ-hairpins". *The Journal of Peptide Research* 54 (1999), pp. 195–199.

[36] Kobayashi, N. et al. "Self-Assembling Supramolecular Nanostructures Constructed from de Novo Extender Protein Nanobuilding Blocks". *ACS Synthetic Biology* 7 (2018), pp. 1381–1394.

[37] Woolfson, D. N. et al. "De novo protein design: how do we expand into the universe of possible protein structures?" *Current Opinion in Structural Biology* 33 (2015), pp. 16–26.

[38] Liu, Y. and Kuhlman, B. "RosettaDesign server for protein design". *Nucleic Acids Research* 34 (2006), W235–W238.

[39] Gerlt, J. A. "New wine from old barrels". *Nature Structural & Molecular Biology* 7 (2000), pp. 171–173.

[40] Hegyi, H. and Gerstein, M. "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome 11Edited by G. von Heijne". *Journal of Molecular Biology* 288 (1999), pp. 147–164.

[41] Banner, D. W. et al. "Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 angstrom resolution using amino acid sequence data". *Nature* 255 (1975), pp. 609–614.

[42] Höcker, B. "Directed evolution of (betaalpha)(8)-barrel enzymes". *Biomolecular Engineering* 22 (2005), pp. 31–38.

[43] Wierenga, R. K. "The TIM-barrel fold: a versatile framework for efficient enzymes". *FEBS letters* 492 (2001), pp. 193–198.

[44] Cuff, A. L. et al. "The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies". *Nucleic Acids Research* 37 (2009), pp. D310–D314.

[45] McLachlan, A. D. "Gene duplications in the structural evolution of chymotrypsin". *Journal of Molecular Biology* 128 (1979), pp. 49–79.

[46] Murzin, A. G. et al. "Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis". *Journal of Molecular Biology* 236 (1994), pp. 1369–1381.

[47] Murzin, A. G. et al. "Principles determining the structure of β-sheet barrels in proteins II. The observed structures". *Journal of Molecular Biology* 236 (1994), pp. 1382–1400.

[48] Liu, W.-m. "Shear numbers of protein β-barrels: definition refinements and statistics11Edited by J. M. Thornton". *Journal of Molecular Biology* 275 (1998), pp. 541–545.

[49] Nagano, N. et al. "Barrel structures in proteins: automatic identification and classification including a sequence analysis of TIM barrels." *Protein Science : A Publication of the Protein Society* 8 (1999), pp. 2072–2084.

[50] Lesk, A. M. et al. "Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet". *Proteins* 5 (1989), pp. 139–148.

[51] Mancia, F. and Evans, P. R. "Conformational changes on substrate binding to methylmalonyl CoA mutase and new insights into the free radical mechanism". *Structure* 6 (1998), pp. 711–720.

[52] Heinz, D. W. et al. "Structural and mechanistic comparison of prokaryotic and eukaryotic phosphoinositide-specific phospholipases C11Edited by K. Nagai". *Journal of Molecular Biology* 275 (1998), pp. 635–650.

[53] Rouvinen, J. et al. "Three-dimensional structure of cellobiohydrolase II from Trichoderma reesei". *Science* 249 (1990), pp. 380–386.

[54] Spezio, M. et al. "Crystal structure of the catalytic domain of a thermophilic endocellulase". *Biochemistry* 32 (1993), pp. 9906–9916.

[55] Reardon, D. and Farber, G. K. "The structure and evolution of alpha/beta barrel proteins." *The FASEB Journal* 9 (1995), pp. 497–503.

[56] Urfer, R. and Kirschner, K. "The importance of surface loops for stabilizing an eightfold beta alpha barrel protein." *Protein science : a publication of the Protein Society* 1 (1992), pp. 31–45.

[57] Nagano, N. et al. "One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions". *Journal of Molecular Biology* 321 (2002), pp. 741–765.

[58] Hennig, M. et al. "A TIM barrel protein without enzymatic activity? Crystal-structure of narbonin at 1.8 Å resolution". *FEBS Letters* 306 (1992), pp. 80–84.

[59] Hennig, M. et al. "Crystal structure of narbonin at 1.8 A resolution". *Acta Crystallographica. Section D, Biological Crystallography* 51 (1995), pp. 177–189.

[60] Kesari, P. et al. "Structural and functional evolution of chitinase-like proteins from plants". *PROTEOMICS* 15 (2015), pp. 1693–1705.

[61] Priestle, J. P. et al. "Three-dimensional structure of the bifunctional enzyme N-(5'-phosphoribosyl)anthranilate isomerase-indole-3-glycerol-phosphate synthase from Escherichia coli." *Proceedings of the National Academy of Sciences of the United States of America* 84 (1987), pp. 5690–5694.

[62] Miller, B. G. et al. "Anatomy of a proficient enzyme: The structure of orotidine 5-monophosphate decarboxylase in the presence and absence of a potential transition state analog". *Proceedings of the National Academy of Sciences of the United States of America* 97 (2000), pp. 2011–2016.

[63] Knowles, J. R. "Enzyme catalysis: not different, just better". *Nature* 350 (1991), pp. 121–124.

[64] Sterner, R. and Höcker, B. "Catalytic Versatility, Stability, and Evolution of the ($\beta\alpha$)8-Barrel Enzyme Fold". *Chemical Reviews* 105 (2005), pp. 4038–4055.

[65] Farber, G. K. and Petsko, G. A. "The evolution of α/β barrel enzymes". *Trends in Biochemical Sciences* 15 (1990), pp. 228–234.

[66] Copley, R. R. and Bork, P. "Homology among $(\beta\alpha)8$ barrels: implications for the evolution of metabolic pathways11Edited by G. Von Heijne". *Journal of Molecular Biology* 303 (2000), pp. 627–641.

[67] Fani, R. et al. "The evolution of the histidine biosynthetic genes in prokaryotes: A common ancestor for the hisA and hisF genes". *Journal of Molecular Evolution* 38 (1994), pp. 489–495.

[68] Lang, D. et al. "Structural Evidence for Evolution of the β/α Barrel Scaffold by Gene Duplication and Fusion". *Science* 289 (2000), pp. 1546–1550.

[69] Miles, E. W. and Davies, D. R. "On the Ancestry of Barrels". *Science* 289 (2000), pp. 1490–1490.

[70] Höcker, B. et al. "Dissection of a $(\beta\alpha)_8$-barrel enzyme into two folded halves". *Nature Structural & Molecular Biology* 8 (2001), pp. 32–36.

[71] Höcker, B. et al. "Mimicking enzyme evolution by generating new $(\beta\alpha)8$-barrels from $(\beta\alpha)4$-half-barrels". *Proceedings of the National Academy of Sciences of the United States of America* 101 (2004), pp. 16448–16453.

[72] Luger, K. et al. "Correct folding of circularly permuted variants of a beta alpha barrel enzyme in vivo". *Science* 243 (1989), pp. 206–210.

[73] Eder, J. and Kirschner, K. "Stable substructures of eightfold beta alpha-barrel proteins: fragment complementation of phosphoribosylanthranilate isomerase". *Biochemistry* 31 (1992), pp. 3617–3625.

[74] Soberón, X. et al. "In vivo fragment complementation of a (beta/alpha)(8) barrel protein: generation of variability by recombination". *FEBS letters* 560 (2004), pp. 167–172.

[75] Bertolaet, B. L. and Knowles, J. R. "Complementation of Fragments of Triosephosphate Isomerase Defined by Exon Boundaries". *Biochemistry* 34 (1995), pp. 5736–5743.

[76] Forsyth, W. R. et al. "Topology and Sequence in the Folding of a TIM Barrel Protein: Global Analysis Highlights Partitioning between Transient Off-pathway and Stable On-pathway Folding Intermediates in the Complex Folding Mechanism of a $(\beta\alpha)8$ Barrel of Unknown Function from B. subtilis". *Journal of Molecular Biology* 372 (2007), pp. 236–253.

[77] Carstensen, L. et al. "Folding Mechanism of an Extremely Thermostable ($\beta\alpha$)8-Barrel Enzyme: A High Kinetic Barrier Protects the Protein from Denaturation". *Biochemistry* 51 (2012), pp. 3420–3432.

[78] Rudolph, R. et al. "Reversible unfolding and refolding behavior of a monomeric aldolase from Staphylococcus aureus". *Protein Science: A Publication of the Protein Society* 1 (1992), pp. 654–666.

[79] Moosavi-Movahedi, A. A. et al. "Thermodynamic and kinetic studies of competitive inhibition of adenosine deaminase by ring opened analogues of adenine nucleoside". *International Journal of Biological Macromolecules* 15 (1993), pp. 125–129.

[80] Pyrpassopoulos, S. et al. "Equilibrium heat-induced denaturation of chitinase 40 from Streptomyces thermoviolaceus". *Proteins* 64 (2006), pp. 513–523.

[81] Pan, H. et al. "Equilibrium and kinetic folding of rabbit muscle triosephosphate isomerase by hydrogen exchange mass spectrometry". *Journal of Molecular Biology* 336 (2004), pp. 1251–1263.

[82] Alvarez, M. et al. "Triose-phosphate Isomerase (TIM) of the Psychrophilic BacteriumVibrio marinus KINETIC AND STRUCTURAL PROPERTIES". *Journal of Biological Chemistry* 273 (1998), pp. 2199–2206.

[83] Beaucamp, N. et al. "Dissection of the gene of the bifunctional PGK-TIM fusion protein from the hyperthermophilic bacterium Thermotoga maritima: design and characterization of the separate triosephosphate isomerase". *Protein Science: A Publication of the Protein Society* 6 (1997), pp. 2159–2165.

[84] Mainfroid, V. et al. "Stabilization of human triosephosphate isomerase by improvement of the stability of individual alpha-helices in dimeric as well as monomeric forms of the protein". *Biochemistry* 35 (1996), pp. 4110–4117.

[85] Chánez-Cárdenas, M. E. et al. "Unfolding of triosephosphate isomerase from Trypanosoma brucei: identification of intermediates and insight into the denaturation pathway using tryptophan mutants". *Archives of Biochemistry and Biophysics* 399 (2002), pp. 117–129.

[86] Chánez-Cárdenas, M. E. et al. "Reversible equilibrium unfolding of triosephosphate isomerase from Trypanosoma cruzi in guanidinium hydrochloride involves stable dimeric and monomeric intermediates". *Biochemistry* 44 (2005), pp. 10883–10892.

[87] Zárate-Pérez, F. et al. "The Folding Pathway of Triosephosphate Isomerase". *Progress in Molecular Biology and Translational Science*. Ed. by P. Michael Conn. Vol. 84. Molecular Biology of Protein Folding, Part B. Academic Press, 2008, pp. 251–267.

[88] Silverman, J. A. and Harbury, P. B. "The equilibrium unfolding pathway of a (beta/alpha)8 barrel". *Journal of Molecular Biology* 324 (2002), pp. 1031–1040.

[89] Fujiwara, K. et al. "A systematic survey of in vivo obligate chaperonin-dependent substrates". *The EMBO journal* 29 (2010), pp. 1552–1564.

[90] Kerner, M. J. et al. "Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli". *Cell* 122 (2005), pp. 209–220.

[91] Georgescauld, F. et al. "GroEL/ES Chaperonin Modulates the Mechanism and Accelerates the Rate of TIM-Barrel Domain Folding". *Cell* 157 (2014), pp. 922–934.

[92] Williams, J. C. et al. "Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power". *Protein Engineering, Design and Selection* 12 (1999), pp. 243–250.

[93] Bharadwaj, A. et al. "The Critical Role of Partially Exposed N-Terminal Valine Residue in Stabilizing GH10 Xylanase from Bacillus sp.NG-27 under Poly-Extreme Conditions". *PLOS ONE* 3 (2008), e3063.

[94] Mahanta, P. et al. "Structural insights into N-terminal to C-terminal interactions and implications for thermostability of a $(\beta/\alpha)8$ -triosephosphate isomerase barrel enzyme". *The FEBS journal* 282 (2015), pp. 3543–3555.

[95] Maes, D. et al. "The crystal structure of triosephosphate isomerase (TIM) from Thermotoga maritima: A comparative thermostability structural analysis of ten different TIM structures". *Proteins: Structure, Function, and Bioinformatics* 37 (1999), pp. 441–453.

[96] Mavridis, I. M. et al. "Structure of 2-keto-3-deoxy-6-phosphogluconate aldolase at 2 . 8 A resolution". *Journal of Molecular Biology* 162 (1982), pp. 419–444.

[97] Carrell, H. L. et al. "X-ray crystal structure of D-xylose isomerase at 4-A resolution". *The Journal of Biological Chemistry* 259 (1984), pp. 3230–3236.

[98] Wierenga, R. K. et al. "Structure determination of the glycosomal triosephosphate isomerase from Trypanosoma brucei brucei at 2.4 A resolution". *Journal of Molecular Biology* 198 (1987), pp. 109–121.

[99] Stuart, D. I. et al. "Crystal structure of cat muscle pyruvate kinase at a resolution of 2.6 A". *Journal of Molecular Biology* 134 (1979), pp. 109–142.

[100] Matsuura, Y. et al. "Structure and possible catalytic residues of Taka-amylase A". *Journal of Biochemistry* 95 (1984), pp. 697–702.

[101] Lindqvist, Y. and Brändén, C. I. "Structure of glycolate oxidase from spinach". *Proceedings of the National Academy of Sciences of the United States of America* 82 (1985), pp. 6855–6859.

[102] Henrick, K. et al. "Structures of D-xylose isomerase from Arthrobacter strain B3728 containing the inhibitors xylitol and D-sorbitol at 2.5 A and 2.3 A resolution, respectively". *Journal of Molecular Biology* 208 (1989), pp. 129–157.

[103] Wintjens, R. et al. "Typical interaction patterns in alphabeta and betaalpha turn motifs". *Protein Engineering* 11 (1998), pp. 505–522.

[104] Street, A. G. and Mayo, S. L. "Pairwise calculation of protein solvent-accessible surface areas". *Folding & Design* 3 (1998), pp. 253–258.

[105] Connolly, M. L. "Solvent-accessible surfaces of proteins and nucleic acids". *Science (New York, N.Y.)* 221 (1983), pp. 709–713.

[106] Dunbrack, R. L. and Karplus, M. "Backbone-dependent rotamer library for proteins. Application to side-chain prediction". *Journal of Molecular Biology* 230 (1993), pp. 543–574.

[107] De Maeyer, M. et al. "All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination". *Folding & Design* 2 (1997), pp. 53–66.

[108] Dahiyat, B. I. et al. "De novo protein design: towards fully automated sequence selection". *Journal of Molecular Biology* 273 (1997), pp. 789–796.

[109] Das, R. and Baker, D. "Macromolecular modeling with rosetta". *Annual Review of Biochemistry* 77 (2008), pp. 363–382.

[110] Berendsen, H. J. C. et al. "GROMACS: A message-passing parallel molecular dynamics implementation". *Computer Physics Communications* 91 (1995), pp. 43–56.

[111] Waldo, G. S. et al. "Rapid protein-folding assay using green fluorescent protein". *Nature biotechnology* 17 (1999), pp. 691–695.

[112] Urvoas, A. et al. "Design, production and molecular structure of a new family of artificial alpha-helicoidal repeat proteins (αRep) based on thermostable HEAT-like repeats". *Journal of Molecular Biology* 404 (2010), pp. 307–327.

[113] Guellouz, A. et al. "Selection of specific protein binders for pre-defined targets from an optimized library of artificial helicoidal repeat proteins (alphaRep)". *PloS One* 8 (2013), e71512.

[114] Korotkov, K. V. et al. "Crystal structure of the N-terminal domain of the secretin GspD from ETEC determined with the assistance of a nanobody". *Structure (London, England: 1993)* 17 (2009), pp. 255–265.

[115] Domanska, K. et al. "Atomic structure of a nanobody-trapped domain-swapped dimer of an amyloidogenic beta2-microglobulin variant". *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011), pp. 1314–1319.

[116] Nagarajan, D. et al. "Design of symmetric TIM barrel proteins from first principles". *BMC Biochemistry* 16 (2015).

[117] Huang, P.-S. et al. "De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy". *Nature Chemical Biology* advance online publication (2016), pp. 29–34.

[118] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.8". 2015.

[119] Krissinel, E. and Henrick, K. "Multiple Alignment of Protein Structures in Three Dimensions". *Computational Life Sciences*. Ed. by M. R. Berthold et al. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2005, pp. 67–78.

[120] Berman, H. M. et al. "The Protein Data Bank". *Nucleic Acids Research* 28 (2000), pp. 235–242.

[121] Wang, G. and Dunbrack, R. L. "PISCES: a protein sequence culling server". *Bioinformatics* 19 (2003), pp. 1589–1591.

[122] Tsutsumi, M. and Otaki, J. M. "Parallel and Antiparallel β-Strands Differ in Amino Acid Composition and Availability of Short Constituent Sequences". *Journal of Chemical Information and Modeling* 51 (2011), pp. 1457–1464.

[123] Fiser, A. et al. "Modeling of loops in protein structures". *Protein Science* 9 (2000), pp. 1753–1773.

[124] Magnan, C. N. and Baldi, P. "SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity". *Bioinformatics* 30 (2014), pp. 2592–2597.

[125] Drozdetskiy, A. et al. "JPred4: a protein secondary structure prediction server". *Nucleic Acids Research* 43 (2015), W389–W394.

[126] Showalter, S. A. and Brüschweiler, R. "Validation of Molecular Dynamics Simulations of Biomolecules Using NMR Spin Relaxation as Benchmarks: Application to the AMBER99SB Force Field". *Journal of Chemical Theory and Computation* 3 (2007), pp. 961–975.

[127] Jorgensen, W. L. et al. "Comparison of simple potential functions for simulating liquid water". *The Journal of Chemical Physics* 79 (1983), pp. 926–935.

[128] Lange, O. F. et al. "Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data". *Biophysical Journal* 99 (2010), pp. 647–655.

[129] Wilkins, M. R. et al. "Protein identification and analysis tools in the ExPASy server". *Methods in Molecular Biology (Clifton, N.J.)* 112 (1999), pp. 531–552.

[130] Song, J. et al. "PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites". *PloS One* 7 (2012), e50300.

[131] Prinz, W. A. et al. "The Role of the Thioredoxin and Glutaredoxin Pathways in Reducing Protein Disulfide Bonds in the Escherichia coliCytoplasm". *Journal of Biological Chemistry* 272 (1997), pp. 15661–15667.

[132] Ribitsch, D. et al. "C-terminal truncation of a metagenome-derived detergent protease for effective expression in E. coli". *Journal of Biotechnology* 150 (2010), pp. 408–416.

[133] Rosano, G. L. and Ceccarelli, E. A. "Recombinant protein expression in Escherichia coli: advances and challenges". *Frontiers in Microbiology* 5 (2014).

[134] Kofoid, E. C. and Parkinson, J. S. "Tandem translation starts in the cheA locus of Escherichia coli." *Journal of Bacteriology* 173 (1991), pp. 2116–2119.

[135] Sachadyn, P. et al. "A cryptic ribosome binding site, false signals in reporter systems and avoidance of protein translation chaos". *Journal of Biotechnology* 143 (2009), pp. 169–172.

[136] Whitaker, W. R. et al. "Avoidance of Truncated Proteins from Unintended Ribosome Binding Sites within Heterologous Protein Coding Sequences". *ACS Synthetic Biology* 4 (2015), pp. 249–257.

[137] Tanaka, Y. et al. "How Oligomerization Contributes to the Thermostability of an Archaeon Protein PROTEIN l-ISOASPARTYL-O-METHYLTRANSFERASE FROM SULFOLOBUS TOKODAII". *Journal of Biological Chemistry* 279 (2004), pp. 32957–32967.

[138] Kabsch, W. and Sander, C. "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers* 22 (1983), pp. 2577–2637.

[139] Hornik, K. *R FAQ*. 2017.

[140] Guex, N. and Peitsch, M. C. "SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling". *Electrophoresis* 18 (1997), pp. 2714–2723.

[141] Dong, X. et al. "PlasMapper: a web server for drawing and auto-annotating plasmid maps". *Nucleic Acids Research* 32 (2004), W660–664.

[142] Khoshouei, M. et al. "Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate". *Nature Communications* 8 (2017), p. 16099.

[143] Cheng, Y. "Single-particle cryo-EM—How did it get here and where will it go". *Science* 361 (2018), pp. 876–880.

[144] Altschul, S. F. et al. "Basic local alignment search tool". *Journal of Molecular Biology* 215 (1990), pp. 403–410.

[145] "Comparative Protein Modelling by Satisfaction of Spatial Restraints". *Journal of Molecular Biology* 234 (1993), pp. 779–815.

[146] Dong, X. et al. "PlasMapper: a web server for drawing and auto-annotating plasmid maps". *Nucleic Acids Research* 32 (2004), W660–W664.

[147] Lovell, S. C. et al. "Structure validation by Cα geometry: ϕ,ψ and Cβ deviation". *Proteins: Structure, Function, and Bioinformatics* 50 (2003), pp. 437–450.

Chapter 6

Annexes

6.1 Annex 1, List of software and programs

1. **BLAST** [144]

   The Basic Local Alignment Search Tool (BLAST) is a web-service created in 1990 by Lipman group at the National Institute of Health in Bethesda (USA) to align and compare biological sequences. Although is possible to align nucleotide sequences, in this work the main use of BLAST is for protein sequences alignment (Protein BLAST). The amino acid sequence of interest can be aligned against a database (i.e. not redundant protein sequences), or against a personal list of sequences. The algorithm of the program calculates the total score of the alignment, the query coverage and the sequence identity for all the paired sequences. The link to the web-service is:

   www.blast.ncbi.nlm.nih.gov/Blast.cgi

2. **DATE, DAtabase for Tim-barrel Enzymes**

   The web-site was created by S. Kumar Singh and M. Madan Babu at the MRC Laboratory of Molecular Biology in Cambridge (UK). It is intended to be a support for all the researchers that are working with TIM-barrel proteins, in order to get quick and comprehensive information about the protein of interest. The database includes 85 TIM-barrel enzymes, that are analyzed for composition and residue conformation (Ramachandran plot). Information such as sequence, length, oligomeric status, function and metabolic pathways are included in the description of the enzymes. The web-site last update was in July 2001, and the information are limited to the protein structures that were available at that moment. It is a good starting point to work with TIM-barrel structures but, up to date, it is obsolete. The database

can be found at:

www.mrc-lmb.cam.ac.uk/genomes/date/

3. **DeepView, Swiss-Pdb Viewer** [140]

   The program was developed in 1994 by Nicolas Guex at the SIB Swiss Institute of Bioinformatics, Biozentrum in Basel (CH) and represents an user-friendly application for the visualization and modeling of protein structures. The software is free and the web-site have detailed user guides and tutorials for training. One of the advantages of this pdb viewer over the other ones is that it allows to calculate the distances by atoms in just few clicks. For this reason, it was used in this project. The program can be downloaded at the address:

   www.expasy.org/spdbv/

4. **DSSP, Dictionary of Protein Secondary Structure** [138]

   The program was designed in 1983 by Wolfgang Kabsch and Chris Sander at the Biophysics department of the Max Planck Institute of Medical Research in Heidelberg (DE). DSSP calculates the H-bond energy between all atoms of pdb files and assign to each residue its class of secondary structures. It can differentiate between α-helices (H), residue in isolated β-bridge (B), β-strands (E), 3-helix (G), 5-helix (I), hydrogen bonded turn (T) and bend (S). It also calculate geometrical features and solvent exposure of proteins. The program can be downloaded at:

   www.swift.cmbi.umcn.nl/gv/dssp/index.html.

5. **GROMACS** [110]

   The program for molecular dynamic simulation was developed in 1995 by the Berendsen's group at the department of Biophysical Chemistry of the University of Groningen (NL) for the analysis of biochemical molecules. Its high performance and accuracy make it the best tool overtime for dynamic simulation and it is used today worldwide, in complex biological system as proteins, lipids and nucleic acids, and in non-biological systems as polymers. Simulations can be done in both full-atom or coarse-grained mode, with implicit or explicit solvent and with different force-fields (AMBER, CHARMM, GROMOS and OPLS). The program can be found at:

   www.gromacs.org

6. **JPred4** [125]

   The Jpred server was developed the first time in 1998 by the Barton's group and Jpred4 is its more recent version released in 2015. The server does prediction of

secondary structures, of solvent accessibility and of coiled coil regions. For the secondary structure prediction the accuracy is more than 82%, calculated in a blind test. It can work with single sequences, multiple sequences (batch mode) or with multiple alignments as input files. Evolutionary information can be included in the prediction for a higher accuracy. Limitations of the server are the length of the sequences (maximum 800 residues) and the number of uploads (max 200 fasta sequences). The web-service is at:

www.compbio.dundee.ac.uk/jpred4/index.html

7. **Modeller** [145]

   Modeller is a program developed in 1993 by A. Sali and T.L. Blundell at the Crystallography Birkbeck College in London (UK). In its first version Modeller was developed to perform comparative protein modeling, in which a given sequence is threaded on related structures in order to obtain a 3D model (also called homologous modeling). Nowadays Modeller is implemented with new packages and can perform de novo modeling of loops [123], multiple alignment of sequences/structures, clustering and comparison of protein structures. The documentation and the user manual of the program are well organized and it can be used also from people that are not expert in structural biology. Modeller can be found at:

   www.salilab.org/modeller/documentation.html.

8. **PDBeFOLD** [119]

   The web-service for structural alignment was developed in 2003 by E. Krissinel and K. Henrick at the European Bioinformatic Institute in Cambridge (UK). It can perform 3D structure comparison between two structures (pairwise) or more (multiple), and it can look for protein similarity with the whole RCBS-PDB archive. The web-service is at:

   www.ebi.ac.uk/msd-srv/ssm/.

9. **PeptideCutter** [129]

   PeptideCutter is one of the multiple programs that are present on the ExPASy Bioinformatics Resource Portal (**Ex**pert **P**rotein **A**nalysis **Sy**stem), developed in 2011 by the Swiss Institute of Bioinformatic (SIB). The web-service is a predictor of cleavage sites in a protein sequence, that may be due to proteases activity or to chemical reactions. Different options of refinement are available, such as the selection of the enzymes or of the chemicals from their database. The web-service can be found at:

www.expasy.org/peptide_cutter/.

10. **PISCES** [121]

    The web-service was developed by the Dunbrack's group in 2003 at the Institute for Cancer Research in Philadelphia (USA) in order to cull protein sequences according to their structure quality and sequence identity. The culling can be done on the whole RCBS-PDB archive or on a personal list of proteins. It accepts different identifiers, such as FASTA, GenBank and SwissProt and the BLAST output. The web-service can be found at:

    www.dunbrack.fccc.edu/Guoli/PISCESOptionPage.php.

11. **PlasMapper** [146]

    PlasMapper is a web-service developed in 2004 by Wishart group at the University of Alberta in Edmonton, Canada. The program generates the graphical output of a plasmid starting from its DNA sequence. Multiple option of annotation are possible and it automatically identifies common sequences as promoters, terminators, reporter genes, replication origins, multi-cloning sites, marker genes and so on. It is also extremely versatile because it is possible to add personalized annotations defining the first and the last nucleotide of the sequence of interest and the interface is easy to use. The program can be found at:

    www.wishart.biology.ualberta.ca/PlasMapper/.

12. **PROSPER** [130]

    PROSPER is a web-service developed in 2012 by Pike group at the Monash University in Melbourne, Australia. It predicts the cleavage sites mediated by proteases on a protein sequence, as PeptideCutter, but it is integrated with advanced features, including a machine learning approach. The amino acid sequence of the target protein is analyzed for secondary structure predictions, solvent accessibility and native disorder predictions, in order to find the part of the proteins that are more accessible to a protease cleavage. The results are more precise and specific than PeptideCutter. On the other hand its database of proteases is limited to 24 enzymes and the program is less versatile than PeptideCutter because it does not include multiple options of search among enzymes. The web-service can be found at:

    https://prosper.erc.monash.edu.au/home.html.

13. **ProtParam** [129]

ProtParam is one of the multiple programs that are present on the ExPASy Bioinformatics Resource Portal (**Ex**pert **P**rotein **A**nalysis **Sy**stem), developed in 2011 by the Swiss Institute of Bioinformatic (SIB). This web-service calculates different physic-chemical properties starting from an amino acid sequence, such as: **a-** molecular weight (MW), **b-** extinction coefficient, **c-** theoretical isoelectric point (pI), **d-** amino acid content, **e-** atomic composition, **f-** protein estimated half-life, **g-** instability index, **h-** aliphatic index and **i-** average hydropathicity. The web-service is found at:

https://web.expasy.org/protparam/.

14. **Pymol** [118]

    Pymol is a software for structures visualization developed in 2000 by the company DeLano Scientific LLC (today it is commercialized b Schrodinger, Inc.). It allows the visualization and partial modeling of biomolecules such as proteins, DNA and small molecules. Different functionalities are available in the program, such as the structural alignment of multiple proteins. Its main features are the high quality of the graphical outputs, and the possibility to create video of the molecule of interest. All the picture of protein structures presented in this work are made with Pymol. The program is available at:

    www.pymol.org/2/.

15. **R-language** [139]

    R is a programming language for statistical analysis and graphics developed at Bell Laboratories (now maintained by Lucent Technologies). Despite it is extremely well-suited for statistical analysis (linear and non-linear models, clustering, basic statistic, and so on), R is mainly used in this project for its graphical features: all the scatter plots, histogram plots and bar plots that are presented in this work are made with R. The program is available at:

    www.r-project.org/.

16. **RAMPAGE** [147]

    RAMPAGE is a web-service developed in 2003 by Richardson group at the University of Cambridge in Cambridge, UK. The program calculates the dihedral angles for any structure that is uploaded on the web-site, and creates its Ramachandran plot. Despite the fact that multiple programs are available for the same task, the graphical output of RAMPAGE is better and nicer. The most populated area (i.e.

the α-helix or the β-strand regions) are highlighted in blue, while the less common are in orange. All the Ramachandran plots presented in this work are created with RAMPAGE, that is available at the address:

www.mordred.bioc.cam.ac.uk/ rapper/rampage.php.

17. **RCBS-PDB** [120] The RCBS Protein Data Bank is a database for 3D structures of proteins that was at first announced in 1971 as a collaboration between the University of Cambridge, UK, and the Brookhaven National Laboratory, USA. The objective of the database is to allow world-wide researchers to share protein structural data obtained through X-ray crystallography, NMR and, nowadays, cryo-electron microscopy. Up to date, it contains more than 45000 distinct protein sequences, and in average, 1000 new structures are deposited each year. The web-site is available at:

    www.rcsb.org/.

18. **Rosetta** [38]

    Rosetta was developed in the '90ies by David Baker and it is nowadays the best software for macromolecular modeling and analysis of protein structures. It is nowadays developed thanks to the collaboration between 49 Universities and Research Institutes world-wide and present numerous tools for protein modeling and design. Among the more interesting there are: **a-** RosettaAbinitio, for the prediction of a 3D structure starting from an amino acid sequence; **b-** RosettaDesign to find low free energy sequences for a given target backbone (used in this work); **c-** RosettaDock for the prediction of protein-protein interactions, **d-** RosettaLigand for small molecule-protein docking; **e-** RosettaEnzDes for the design of enzymes and **f-** RosettaMembrane for modeling of membrane proteins. The software is available at:

    https://www.rosettacommons.org/.

19. **SSpro** [124]

    SSpro is one of the computational tools that are available in the SCRATCH suite, developed in 2002 by Baldi's group at the University of California, USA. SSpro is a predictor of secondary structures as JPred4, but its recent implementations improved its accuracy in the prediction up to 93% against 75% of JPred4. This high accuracy is due to two main things: the use of evolutionary information (i.e. homologous sequences), and the use of machine learning approaches. It is definitely the best program nowadays for secondary structure prediction of natural proteins,

anyway in the context of this work with artificial sequences its performances drop to the same level of JPred4. The program is available at:

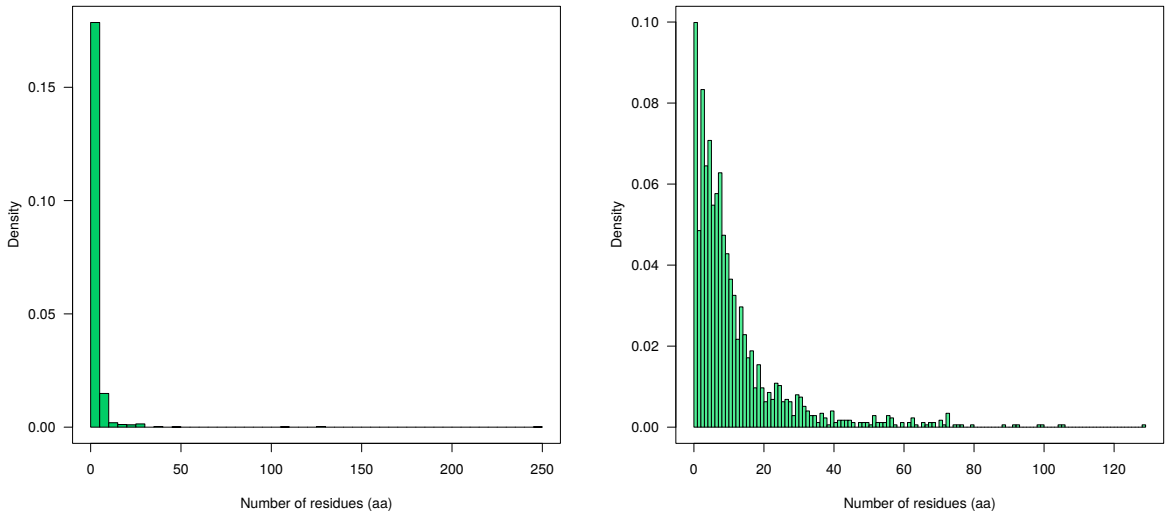
www.scratch.proteomics.ics.uci.edu/.

6.2 Annex 2, List of 219 natural TIM-barrels

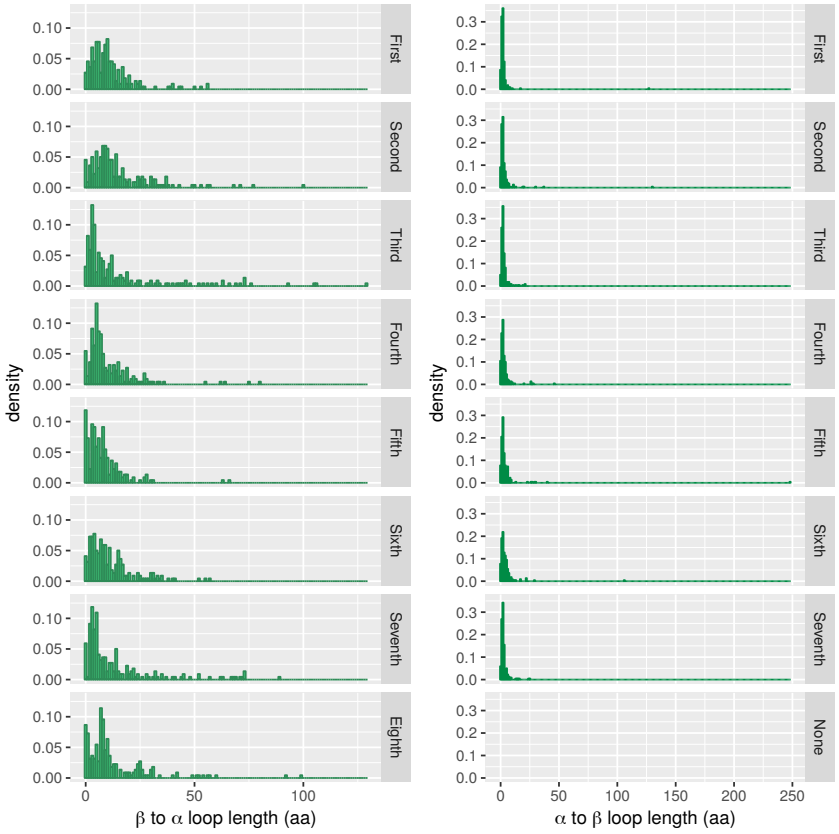
IDs	Length (AA)	Resolution (Å)	IDs	Length (AA)	Resolution (Å)	IDs	Length (AA)	Resolution (Å)
1OYCA	400	2	2NZLA	392	1.35	1LQAA	346	1.6
1KBLA	873	1.94	3E96A	316	1.8	1D3GA	367	1.6
2G50A	530	1.65	1WCGA	464	1.1	2VRQA	496	2
1WDDA	477	1.35	1VYRA	364	0.9	1LUCB	324	1.5
1GVFA	286	1.45	2QF7A	1165	2	1P7TA	731	1.95
1Z41A	338	1.3	3DAQA	292	1.45	1FOBA	334	1.8
1ZFJA	491	1.9	3ERPA	353	1.55	1KWGA	645	1.6
1X1ZA	252	1.45	2QULA	290	1.79	3DZ1A	313	1.87
1VQTA	213	2	1TQJA	230	1.6	1TVNA	293	1.41
1O1ZA	234	1.6	3EBVA	302	1.5	1G5AA	628	1.4
1US0A	316	0.66	1VD6A	224	1.3	2CZDA	208	1.6
1GTEA	1025	1.65	1WDDS	128	1.35	2HMCA	344	1.9
1EZWA	349	1.65	1MXGA	435	1.6	1WZAA	488	1.6
2ZBTA	297	1.65	2OO6A	409	1.8	1UC4A	554	1.8
2UY2A	294	1.6	1VC4A	254	1.8	1O94C	264	2
1AJ2A	282	2	1XKYA	301	1.94	1I1WA	303	0.89
1JI1A	637	1.6	3CNYA	301	1.85	2CYGA	312	1.45
3C8NA	356	1.9	1JNDA	420	1.3	1WZLA	585	2
1TWDA	256	1.7	3CBWA	353	1.27	1GOIA	499	1.45
2C0HA	353	1.6	2F2HA	773	1.95	1NTHA	458	1.55
2QW5A	335	1.78	2NT0A	497	1.79	2G0WA	296	1.7
1SR9A	644	2	2NQ5A	755	1.9	2NX9A	464	1.7
2I7GA	376	1.73	1W5QA	337	1.4	1VEMA	516	1.85
1EYEA	280	1.7	1VZWA	244	1.8	2OX4A	403	1.8
1M04A	512	1.95	2V2HA	242	1.18	1QO2A	241	1.85
2A4AA	281	1.84	1VCVA	226	2	2VEFA	314	1.8
3E49A	311	1.75	1YKWA	435	2	2H9AA	445	1.9
1P4CA	380	1.35	1HL2A	297	1.8	1NSJA	205	2
2NWRA	267	1.5	1QW9A	502	1.2	2GJLA	328	2
2D3NA	485	1.9	3EAUA	327	1.82	3BMVA	683	1.6
1ZJAA	557	1.6	5RUBA	490	1.7	1GCYA	527	1.6
1UWSA	489	1.95	1W3IA	293	1.7	1TA3A	274	1.7
1REQA	727	2	2OEMA	413	1.7	3CZGA	644	1.8
1XX1A	285	1.75	2C6QA	351	1.7	1KKOA	413	1.33
1BQCA	302	1.5	1H1NA	305	1.12	1ITXA	419	1.1
3CO4A	312	1.92	1TQXA	227	2	2ZOXA	469	1.9
2YW3A	207	1.67	2QAPA	391	1.59	1EOKA	290	1.8
1VHCA	224	1.89	2I57A	438	1.97	1NVMB	312	1.7
3CM4A	349	1.85	1A53A	247	2	1M5WA	243	1.96
2EPOA	627	1.56	3E2VA	401	1.5	3B9OA	440	1.9
3CIWA	348	1.35	2VM8A	501	1.9	2R8WA	332	1.8
2HK0A	309	2	1SGJA	284	1.84	2A0NA	265	1.64
1TZZA	392	1.86	1CNVA	299	1.65	1LUCA	355	1.5
2QJJA	402	1.8	1GVEA	327	1.38	3B4UA	294	1.2
1N7KA	234	2	2ZUVA	759	1.85	3CWNA	337	1.4
1U5HA	273	1.65	1VF8A	377	1.31	1VHNA	318	1.59
1UB3A	220	1.4	3F4WA	211	1.65	3BPWA	342	1.7
1I60A	278	1.6	1KFWA	435	1.74	2CHOA	716	1.85
1P1XA	260	0.99	2ISWA	323	1.75	1WA3A	205	1.9
3B40A	417	2	1YDYA	356	1.7	1YNPA	317	1.25
1WBHA	214	1.55	1DOSA	358	1.67	1P0KA	349	1.9

IDs	**Length** (AA)	**Resolution** (Å)	**IDs**	**Length** (AA)	**Resolution** (Å)	**IDs**	**Length** (AA)	**Resolution** (Å)
2DSKA	311	1.5	3DHUA	449	2	1US3A	530	1.85
3BOFA	566	1.7	1UJPA	271	1.34	1NOFA	383	1.42
1ZP4A	304	1.85	2RFGA	297	1.5	1L6WA	220	1.93
2CKSA	306	1.6	3BLEA	337	2	1W8SA	263	1.85
1HT6A	405	1.5	1GKPA	458	1.29	1WDPA	495	1.27
1XI3A	215	1.7	2PZ0A	252	1.91	1X38A	602	1.7
1F74A	293	1.6	3CHVA	284	1.45	1VR6A	350	1.92
1ZZMA	259	1.8	1AVAC	181	1.9	2BHUA	602	1.1
1SVDM	110	1.8	1O94A	729	2	1OHLA	342	1.6
1GQIA	708	1.48	1W9PA	433	1.7	3C6CA	316	1.72
1CB7B	483	2	3CUZA	532	1.04	2DDXA	333	0.86
2DVTA	327	1.7	3BMXA	642	1.4	1PIIA	452	2
1OF8A	370	1.5	1V93A	296	1.9	1CB7A	137	2
1ZGDA	312	1.7	1Q6OA	216	1.2	1YX1A	264	1.8
3CH0A	272	1.5	3CLMA	352	1.14	1PXGA	382	1.7
1O94D	320	2	2FHFA	1083	1.65	2VCCA	891	2
1UC4B	224	1.8	1UC4G	173	1.8	1QTWA	285	1.02
1UG6A	431	0.99	2GUYA	478	1.59	1D8WA	426	1.6
1REQB	637	2	2NLIA	368	1.59	2HS8A	402	1.9
1NVMA	345	1.7	1ONWA	390	1.65	3CMGA	667	1.9
2V3GA	283	1.2	1O5KA	306	1.8	2E6FA	314	1.26
2GDQA	382	1.8	1E6QM	501	1.35	2BG5A	324	1.82
7A3HA	303	0.95	3BC9A	599	1.35	1NARA	290	1.8
1H4PA	408	1.75	3CU2A	237	1.91	1VKFA	188	1.65
2TPSA	227	1.25	1EDQA	540	1.55	1RHCA	330	1.8

6.3 Annex 3, Loops length distribution



(a) (b)



(c)

Figure 6.1: Loops length distribution

Length distributions for N-term (**a**) and C-term loops (**b**), and their corresponding plot for individual position in the TIM-barrel fold (**c** and **d**).

6.4 Annex 4, DNA sequences of the OctaVIIs

>OctaVII_01
ATGGAGTTCCACATTTTCATTTTCGGTACCACCTGCAACCTGGATCAGTACTTCATCGAGGCATGGAAGATCCTGATGGAGGCAAAG
GACGCTCATCTGGGTCTGGGTATCCAAGTCGAGGATCAGGTAATCCGTTACCTGTTCAAGAAGTGGCACAAACTGGCCCTGGAATTC
GAACTGCGTGGTTGGATCAGCATCTTCGTCTACACCACGGGTGATGCAGATGCACTGTTCCGTGAATTCCTGGCATTCTGGCTGAAA
GTGGATCAGCGTTGTGGTGCTATCGCTCTGGGTGGTGGCACTGGTGATCTGTATAACGCTGTTAAAAAACACCTGGAAGACGCGAAA
CGCAGCAAAGCTGTGCACCACGCTCTGTGCGTAATGCTGCCTCCAGGTCCGATTAATGACCTGTTTATCCTGCTGATGATCCTGTGG
GAACTGTTTCGTAACGGTGGCGGCGCGATTTGGATTGGTGTTCAGTCTGGTGGCATCAAAGAAATGCTGGAACTGTGGATTCGTATC
ATCAAAAAAGGCAGCGAATCCCTGCTGGGCGTTGCGGTTGGCGGCAACTCTGGCGACTTTGACAAAGCCTGGGAAATTATGCTGGAA
ATCCTGACTAAAGACTCTCATGCCAACTATGCGGTGGGCATTATCATCTCCAACGGCCGCGCGAAAGACAAAACCAAAGCGTGGACC
CTGCGCTTTCTGAAAGAACAGAACTCCCtcgag

>OctaVII_02
ATGGGCAAGGTAATGGTCGTGCTGTTCGGCAAGACTAAGTTCGCAGAGAAGCGTTTCAAGGACGCTATGCAGATCATCAACGACTGC
GACGCAGACGGTCTGGCTGTAATGGTTGCAATCTTCGACACTACTGGTCTGGAACTGTTCAAAAAAGCGGCAGAGCTGGCACGTGAT
TACAGCTGTGGTGGTATGGGTCTGGCAATGTATGGTACCCAAACGGATGCTAAAAAAGTCGTGGCTGAGATCATCAAACAGCTGCAG
AACATCGACCACGACGATGTGGTTTGTATCGTGACCGGTGCTACCGATACCGCTCTGAAAATCCAGGAAATTGCTCGTGAGATGCTG
GAAAAAGCTGACATCCGTGGTGGCGGCCTGGGTATTACCGAACAGTCTGGTCCGCTGGAAAAATATGCGCGTCTGGCGATGGCGAAT
GCGAAAAAATTTACCTACGCCAACTTTCTGTTTGTAATCGTTATCTCTTCCGGCGATAAAGAAGACATCCTGCTGAAACTGCTGGAA
GAATGGAAAAAATCCGGCGTTGCGGGCGGTGGTGTTGGCATTTACGGCGATGATACGGCGATTGTTGAACATTTCAAAAAAATCGTT
AAAATCATCGCGAAACTGAAATGCACCAACGGCATCATTATGATTATGGACAGCCGCGGCGATTTCCTGGAACTGCTGCGCATTTTC
GCCGAAATTGCGGAAAAAGCCCAGCGCGCCCtcgag

>OctaVII_03
ATGGgcATCCTGTTTATTGCTTTCTCTTGTGAAACCACCGATAACGAAAAAGCGTTTGAACTGGCGGTGAAACTGGTTCTGGATGAG
CAGATGGAACACATCGGTATTCAGGTTGGTGGCCCGGGCGGTCCGCTGGAAGAGGCAGCTGCCAAATTCATCAAGAAAATGCAGCTG
GCTAAAACGTCCGGTCAGGGTTTTATCGTTAACTTCACCAGCCGTGACGGTAACGACTGGTTCGAAAAGGCGCGTAAACTGCATCAG
AAAAGCTGCCACGACAACATGGTGTTCCTGATCTCCGTGACCCACACCGAAGCGCTGGACCTGCTGGAAGCCTGGCTGCGCAAGCTG
CAAAAAGATAAATCATCCTGGATGGGCGTGCTGTTCAACCACGGCAACCGTAACGTTGAAAAAGCCTATCAGATTGCAGCCGAAATC
TTCAAAAAAGTGTTGTCTCTGTACTGCGTCTGGGCGATTCTGATGACCCAGGGTGACATCCGTGATTTAGCGGCAAAATGGGCAGAA
CAGGCGGCAAACATCAAAATCTGGGGCTGCGAAATTTACCTGTACAACACCAACGGCGATCTGGAAGCTATCGCGCGCGAGCTGGCA
AAAATTGCGAAAAAATACACCGGTACCTTCTGCGGCTTCGGTGTGGTGGGCGCAGGTGAAGACCTGTATAACCTGAATGCACGTCTG
ATTAAAGCGGCGAAAGAAGAAATGCTGAACctcgag

>OctaVII_04
ATGGgcACCATCTTTGTTGGTCTGCAGGGTCAGGAAACCGGTGCGGACGATAAATTCAAACGCATTGTTGAAATTGTGCGTGCGCTG
AAAAGCGGTGAATGTGGCGTTGCCGTTACCCTGGGCACCGGGGACACCGAAGCGCAGCTGGAAAAGTGGATCCGTATCGCGCAGAAA
TCCGAATGCCGTAGTGAATGTATAGGTGTAGGTGGTAGTGAAGGTGACGCTGAAGCTAGCTGGCGTAAAGCGGCGGAACTGCACAAC
AAATGCGATAACAGCGACTCCATGCTCTACAGCATTGCCTCGGGTTCTAACAAGCAGGAAATGTTCGATCGCCACCTGAAAGCAGCT
GAAGAACACTCTAAAACGCTGATCGCCTTCTTCTTTGAACACGATGATACCCGGGCTGACGACAAATGGCTGGAATTCTTCCTTAAA
CTGCTGAACTCTTCTGGCGGTGTTATCTTCACCGGCATCGTTGCGAGCCGTGGTGACGTTAAAAACGCCCTGCATAAATGGTTGGAA
ATCGCGATGAAACAGAAACAGGGCGGCTGGGGCGTAGGTATCAACGTTAGCGGCGATCCGGTAGAAGAGTGGTGGAAATTGATCCTT
AAATTCATCAAAAAATACTGCGGCGAACAGTGCGCCATTTTCATTGTTGGCACGGGCAGCAAAATGGAAAAACTTCTGGAGAAATTT

GCGAAAGAACTTGAAAAACTGTTGCAGGCGCtcgag

>OctaVII_04 NoCys
ATGGgcACCATCTTTGTTGGTCTGCAGGGTCAGGAAACCGGTGCGGACGATAAATTCAAACGCATTGTTGAAATTGTGCGTGCGCTG
AAAAGCGGTGAAACCGGCGTTGCCGTTACCCTGGGCACCGGGGACACCGAAGCGCAGCTGGAAAAGTGGATCCGTATCGCGCAGAAA
TCCGAAATCCGTAGTGAAGTGATAGGTGTAGGTGGTAGTGAAGGTGACGCTGAAGCTAGCTGGCGTAAAGCGGCGGAACTGCACAAC
AAACTGGATAACAGCGACTCCATGCTCTACAGCATTGCCTCGGGTTCTAACAAGCAGGAAATGTTCGATCGCCACCTGAAAGCAGCT
GAAGAACACTCTAAAACGCTGATCGCCTTCTTCTTTGAACACGATGATACCCGGGCTGACGACAAATGGCTGGAATTCTTCCTTAAA
CTGCTGAACTCTTCTGGCGGTGTTATCTTCACCGGCATCGTTGCGAGCCGTGGTGACGTTAAAAACGCCCTGCATAAATGGTTGGAA
ATCGCGATGAAACAGAAACAGGGCGGCTGGGGCGTAGGTATCAACGTTAGCGGCGATCCGGTAGAAGAGTGGTGGAAATTGATCCTT
AAATTCATCAAAAAATACGCGGGCGAACAGATGGCCATTTTCATTGTTGGCACGGGCAGCAAAATGGAAAAACTTCTGGAGAAATTT
GCGAAAGAACTTGAAAAACTGTTGCAGGCGCtcgag

>OctaVII_04 WS
ATGGgcACCATCTTTGTTGGTGTGAACAGCCAGGAAACCGGTGCGGACGATAAATTCAAACGCATTGTTGAAATTGTGCGTGCGCTG
AAAAGCGGTGAAACCGGCGTTGCCGTTACCCTGGGCACCGGGGACACCGAAGCGCAGCTGGAAAAGTGGATCCGTATCGCGCAGAAA
TCCGAAATCCGTAGTGAAGTGATAGGTGTAGGTGGTAGTGAAGGTGACGCTGAAGCTAGCTGGCGTAAAGCGGCGGAACTGCACAAC
GAAATTGAAGGCAGCGACTCCATGATTTTTAACATTGCCTCGGGTTCTAACAAGCAGGAAATGTTCGATCGCCACCTGAAAGCAGCT
GAAGAACACTCTAAAAGCGTGGTGGCCTTCTTCTTTGAACACGATGATACCCGGGCTGACGACAAATGGCTGGAATTCTTCCTTAAA
CTGCTGAACTCTTCTGGCGGTGTTATCTTCGTGAGCTTTGTTGCGAGCCGTGGTGACGTTAAAAACGCCCTGCATAAATGGTTGGAA
ATCGCGATGAAACAGAAACAGGGCGGCTTTAGCATTGGTATCAACGTTAGCGGCGATCCGGTAGAAGAGTGGTGGAAATTGATCCTT
AAATTCATCAAAAAATACGCGGGCGAACAGATGGCCATTTTCATTGTTGGCACGGGCAGCAAAATTCGTGAACTTCTGGAGAAATTT
GCGAAAGAACTTGAAAAACTGTTGCAGGCGCtcgag

>OctaVII_05
ATGGAAGGTGGCATCGGTTTTTCCGGTACTGGCACCGCGAACGAAAAAGAATGGGAAAAAGCGCGTGAAGCGGTGCGTAAAATCGAT
CACGAAGAACTGTTCCTGATTTTCATCGGCTGTACCACCGCTGAACGTGATGAATTCAAAAAATTCGCTGAGAAAGCCTATAAAGCA
GATATCGCCAGTTTTATCCTGGCAGTGGGCGGTCACGGTACCGAACGGAAAAACTACATCGAAATCGCACTGCAGATCTACCTGAAC
CTGTCTGTTGCTTCTAACGGCTGGATGGTTGTCGGTAGCCCGGACGGCTTTCTGGACGATTTTAAATGGGCGGTGAAACGCAGCATC
GAGTCTGACAGCAAACACCTGGGCCTGTGCCTGGAAGGTCCAAACGGTGACGTTGAAAAAGCGATCCGCGAAATGCTGAAGATGTGG
CAGAAAGCTTCCGACGGTGAACTGATGCTAGGTTTCGTGGCTACCAGCGGCAACACCCTGGAAATTCTGAAAATTGCTCTGGAGTAC
TTCGCCAAACAGACCAACCACCGTGCGTGCATCTTCGTTAAAATGAGTTACGGCGATATCGATAACATGGCGGCGATCATCGCGAAA
CTGATCAACATCGCGGATCTGGGCCACCGTGCGGAAGCGTACGTTGGTTCTGGCGAATACCAGGAAGAACTGCTGAAAGAATGGATC
CGTCGTCTGAAAGCGAACCTGCTGAAACtcgag

>OctaVII_05_4
ATGGAAGGTGGCATCGGTTTTTCCGGTACTGGCACCGCGAACGAAAAAGAATGGGAAAAAGCGCGTGAAGCGGTGCGTAAAATCGAT
CACGAAGAACTGTTCCTGATTTTCATCGGCACCACCACCGCTGAACGTGATGAATTCAAAAAATTCGCTGAGAAAGCCTATAAAGCA
GATATCGCCAGTTTTATCCTGGCAGTGGGCGGTCACGGTACCGAACGGAAAAACTACATCGAAATCGCACTGCAGATCTACCTGAAC
CTGTCTGTTGCTTCTAACGGCTGGATGGTTGTCGGTAGCCCGGACGGCTTTCTGGACGATTTTAAATGGGCGGTGAAACGCAGCATC
GAGTCTGACAGCAAACACCTGGGCCTGGCGCTGGAAGGTCCAAACGGTGACGTTGAAAAAGCGATCCGCGAAATGCTGAAGATGTGG
CAGAAAGCTTCCGACGGTGAACTGATGCTAGGTTTCGTGGCTACCAGCGGCAACACCCTGGAAATTCTGAAAATTGCTCTGGAGTAC
TTCGCCAAACAGACCAACCACCGTGCGGTGATCTTCGTTAAAATGAGTTACGGCGATATCGATAACATGGCGGCGATCATCGCGAAA
CTGATCAACATCGCGGATCTGGGCCACCGTGCGGAAGCGTACGTTGGTTCTGGCGAATACCAGGAAGAACTGCTGAAAGAATGGATC

CGTCGTCTGAAAGCGAACCTGCTGAAACtcgag

>OctaVII_06
ATGGGAACTGTTGTCGTGTTGACATACGGACACACTTCTGATTTTTGGAAAGAAATGGAAAAACACCTGCAAGAATTACAAAAGGCG
GGAGACGCAGCCTTGGAATTCGGATTTATCATTCACTCGGGGAACTTGTCCGAGGATTTATGGTGGTTCGTGTATCTGGCTAAAAAG
TATGTTACTCGTTCTGTGGCACTGTTTTTTGCAGGCACCGGCACAAAATGGGAAAAGGAGTTTCGCACTGCCCTGAAGATTTTAGAG
ATGATTGGTACTACTGGGTCGGCTTTCGGTTTCATCTCAGGAAATACGGTTACGGATGAGTGGATGCGCAAAGCCCACGCTGAGTTT
TTGAAAATGCGCGAGGGCAAGATTCACATTGGTATGGAAGGCAATAAGGGCGACGAGGTGGAGCTTTTTAAACGCGCTTTGGCCGAA
TGGCTTAACGCAGGAAAACGCGCGAACATCTTGTTCGTTGCACGTCACAAAACGGAAGAACTTAAAAAAGCCGAGGAATTCATCAAG
ATGGCTTTGAAACAGCAGGCCATTAGCATCGCTTTAGCCTTGAATGAAGATACTGGTGATGCCTTAAAAGTGTGGGCCGAGATCTTG
AAGTTGTTAAAAAAGACTAAGGACGGTGAATTTCATGCACTGGTCATTGGCACGGGGACCACGGCCAAGAAATTATTAGAGATTATG
CGTAAGATGGCAATCAAAATGGAGCTGGGGctcgag

>OctaVII_07
ATGGGGTACTTCACTATCGGCCATATTCGCTCGACGGGGGCACAGGACAAGTACTTCGCAGTAGCCCTTGAACTGATTTTGAAAAGT
ACAGGGCGCGATGGTGCAATTATTATTATCGGCGCGGAGACGAAGGAGCTGAAACTGGCTGAGGAGTGGATGAAACGTGCACTGAAG
GCCGAGACCCGTATCACGGGTCTGGCGATCGGGGGAGACACCACCAACATCGATCAAGTGTTCCTGGAGTTATGGAAAATTTGGCTT
AAAATCACCTCGACTCTGTCGTTTTTTATGATTTTTGCCTCCGGGGGCGATTTTAAGGCGTTGTTGCACAAGTGGCTTCGCCTTTTG
GAAAAGTGGACTGATGTAGACTTCGGTACAGGCGTAGTTCTGACCGACACAAAAGAATCTGCGTTATTCGAGGAGTGGTTAAAGGAG
TTAGAGAAGTTTCAGGGCAAAACAGGATTAGTGATTATTATCGCCGGTGATGGTAATACTCGTGATGCACTGGAAGAATGGCTTCGT
AAGGCTATCAAAGCGTCGACCGGTCACCTTGGCGTAGGCATTGCTAGCAGTGGGAAGAATGCGCGTGACTACACCCAAGAAGCGATC
AAACTTTTACGCAACACTCAATCAAACGATGGTGCCCTGACTGTTTCAGGAAGTGACGACCGTGCACGTGACTGGTTGGAAATCGCA
ATCCGCGAAGCGGCCAAAGAGGCTTTGAATctcgag

>OctaVII_08
ATGGGGAAGGCCGCAGTCGGGCTTGCAGGTCAAACCAAGTTGGCCGATGAAATTTTTAAGCGCATCGTGCAGTTACTTTTAGAGGCA
GACACGGGTGGTTTGGGCCTGGTAGTCGCCGGGAACTCTAAATCTTTGCGTGACCGTTTTTATGAGTGGGCACGCCAGGCGGCTGAG
TTTAAGAGCGGGGTGATCAATATTGGGGTGGATGGCGATTCAGGAGATATGGAAGCCAAATTTAAGGAAGCTTTCGCCATTTATTTG
AAGTTTCTTTATGGCGCACTTAATTGGTTCTTGTCCGATTTCACTAAAAAACAACTGGAATTATTGCGTAAGTTTATGGAGTTGGCG
ATGCGTGCCGATTCTACAGGGATGGCTGTTTCCTTCGGTAAAGGAACAGGTGACTCTGATCAAATTGCCAAAGAAATTGCCAAAATT
TGGATCGATTATACTACACAGTATGGTGGACTGGGGGTAACCTACACGGATGAACATTTTCTGGAGTTATTGGAAAAATACTTACGC
ATTTATGCCAAAACCAGCACTGCTGAGCTTATGATCATCATTAAAACTACGGACTCGGACAATTCCGCGATTTTCGAGACAGCCATT
CGTATTCTTTTGAAAATTGTTGGACCTGACCTGGAGTTGATCGTGATCGGGAGTGGGGATAACATGACTAAGATCATTGAAAAAATT
TTGCGCATTGCCGCAAAGGCAGAGAAGACTctcgag

>OctaVII_09
ATGGAGATTGGTGTAGCTTTGGTTGGGGGTACTACTGCTATGGACAAACTTTTCGATGAATTGCTGAAGATCCTGCAGAAGTTGCAG
ACGGATTCAAGCATGTTGGGGGTGGTAGGGTTCGACAGTCGTGACCGCGCCTTATTCCAGGAATGGTTGAAAAAGGCGAAGAAATCT
GATAGTGCCCTTTTTGCCTTTGGCCACGGGGGTAGCAACGGGGACGAAGAGAAAAACTATAAGGAGGCACAAAAGGACTTCTTGAAA
ATCGACGCAGCTTATATCAGCAACGTCACGGGTTCAGCGAAGGGAGATACCGAAAAAATGTTCAAGGAGTTCGTGAAAATTGCTTTG
AAAGCCGAGGACCGTGGGGCTGGATTTTTGTTTGCATCGGGCTCTGGCGACTTACATCGCGATTGGCTTGAGGCTATGAAGGAAGCG
TTGAAACAGACCGGACAGGCGATCCGTTTTTACATCTCTATCACTAATACCCAGTTGCAGGAGTTACTTGATAAGTGGTTCAAGGAA
TCCGCTAAAGTAACGGACGGTCACTCAGGAGTAGCAATTCTGGGCTATAAAGGAAATCAGGACAAGTTATTTGAGGAGTTATTACAG
AAAATTAAGAAGTATGCAGCCGGAGACGGCAGCCTGATGATCGCTATCGGAGGCACCGGATTTCGCGACCGTACGGCAGAACACCTT

AAACGTATGAAGAAGGAAGATGAAAGCctcgag

>OctaVII_09 WS
ATGGAGATTGGTGTAGCTTTGGTTGGGGGTACTACTGATATGGACGGCCTTTTCGATGAATTGCTGAAGATCCTGCAGAAGTTGCAG
ACGGATTCAAGCATGTTGGGGGTGGTAGGGTTCGACAGTCGTGACCGCGCCTTATTCCAGGAATGGTTGAAAAAGGCGAAGAAATCT
GATAGTGCCCTTTTTGCCTTTGGCCACGGGGGTAGCAGCGGGGACGAAGAGAAAAACTTTAAGGAGGCACAAAAGGACTTCTTGAAA
ATCGACGCAGCTTATATCCTGAACGTCGTGGGTGCAGCGAAGGGAGATACCGAAACACTGTTCAAGGAGTTCGTGAAAATTATGTTG
AAAGCCGAGGACCGTGGGGCTGGATTTTTGTTTGCAAACGGCTCTGGCGACTTACATCGCGATTGGCTTGAGGCTATGAAGGAAGCG
TTGAAACAGCAGGGACAGGCGATCCGTTTTTACATCGCTATCACTAATACCCAGTTGCAGGAGTTACTTGATAAGTGGTTCAAGGAA
TCCGCTAAAGTAACGGACGGTCACTCAGGAGTAGCAATTCTGGGCTATAAAGGAAATCAGGACAAGTTATTTGAGGAGTTATTACAG
AAAATTAAGAAGTATGCAGCCGGAGACGGCAGCCTGATGATCGCTATCGGAGGCCCGGGATTTCGCGACCGTACGGCAGAACACCTT
AAACGTATGAAGAAGGAAGATGAAAGCctcgag

>OctaVII_10
ATGGGTACGTTTGGCATCATTCTGGCCGGAACCAGTACCGATCGCGAAAAGGCAGCCAAAATCATCATCGAACTTGTCGAATGGGCC
AACAGTGGCGCATGGCTGTTGGGCTTTGCCGAGGGGGGAACTGAGGCTCGTGACAAATTGTTGGAGTGGGCTAAAGAAGCCTTGAAA
ACTACGGCAGAGGGATTCATGATGGGCTTTGATGGTGGTAGTACGCGCACTTTAGACGAGTTTGAGGAGTTTCTGAAGGTCGCAGAA
TGGACCAAGACCTCCACTATCCAGGTGCTTGTAATCCTTCGTCGTGGCGAATCAGATAAGTTCGCTAAGGAAGCACTTCGTCGTTTA
GAGGAATCGCGCAGTGTAGGCCAGGGTGCCGGGATGATGGAAACGGGCCCTGGTGCGGATGAAATCTTTAAGAAGATCCTGTATTAT
GCCGAAGTGTTTACCGGAGACTATTTTTCGCTTGCAATCTTTGCTTTTAGCGGAACGGATGCCAAACAGGGTGAGAAATGGGCACGT
TTAGCTGCAAAGAGCACAGAAGGGGTTATCGTAATTGGAATTGAACTGCATTCGGGACGCTTACGTGATGGATTCCACAAATTGCGT
GCCATCTGGGAAAAGTTGGAATCGCGTACTGGCGGATTAGTCATCATCGGGGGCAGGTGATGATCAGAAAGCGGAAGCCTTGGAGTTC
CTTAAAGAGATGGAAAAGCATCTGAAAAAActcgag

6.5 Annex 5, Command lines and scripts

6.5.1 Secondary structures assignment

Secondary structure assignment is performed with the program DSSP. the pdb file is the only input requested.

Command line:

```
mkdssp INPUT.pdb -o OUTPUT.txt
```

6.5.2 Energy minimization of the natural TIM-barrels

Energy minimization of the natural TIM-barrels is performed with the package Relax of the Rosetta software (version 2015.19.57819_bundle). All the input pdb structures are listed in the file listpdb.

Command line:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/relax.linuxgcc
release -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/databas
e/ -l listpdb -in:file:fullatom
```

6.5.3 Parametric design

The parametric design of the artificial TIM-barrel backbones is performed with the package BundleGridSampler of the Rosetta software (version 2015.19.57819_bundle). Three input files are requested: FASTA, OPTIONS and PARAMETERS.

Command line:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase/ -in:file:fasta sequence.fasta @flags -parser:protocol basic.xml
```

FASTA: sequence.fasta

```
AAAAA
```

OPTIONS: @flag

```
-nstruct 1
-mute all
-unmute protocols.helical_bundle
```

```
-inout:dump_connect_info
-inout:connect_info_cutoff 0.0
-score:weights talaris2014
-chemical:exclude_patches LowerDNA UpperDNA Cterm_amidation SpecialRotamer VirtualBB Sho
veBB VirtualNTerm VirtualDNAPhosphate CTermConnect sc_orbitals pro_hydroxylated_case1 pr
o_hydroxylated_case2 ser_phosphorylated thr_phosphorylated tyr_phosphorylated tyr_sulfat
ed lys_dimethylated lys_monomethylated lys_trimethylated lys_acetylated glu_carboxylated
cys_acetylated tyr_diiodinated N_acetylated C_methylamidated MethylatedProteinCterm
```

PARAMETERS: basic.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<tala weights="talaris2014.wts" />
</SCOREFXNS>
<TASKOPERATIONS>
</TASKOPERATIONS>
<FILTERS>
</FILTERS>
<MOVERS>
<BundleGridSampler name=Octa7 scorefxn="tala" set_bondlengths=true set_bondangles=true
set_dihedrals=true r0=17.4 dump_pdbs=false pdb_prefix="out">
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=0.000000000 delta_omega1=1.57
crick_params_file=beta_strand helix_length=9 delta_t=1 />
<Helix omega0=0.04 delta_omega0=5.890486230 delta_omega1=3.14 helix_length=21/>
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=5.497787140 delta_omega1=4.31
crick_params_file=beta_strand helix_length=9 delta_t=0.25/>
<Helix omega0=0.04 delta_omega0=5.105088060 delta_omega1=3.14 helix_length=21/>
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=4.712388980 delta_omega1=1.57
crick_params_file=beta_strand helix_length=9 delta_t=1 />
<Helix omega0=0.04 delta_omega0=4.319689900 delta_omega1=3.14 helix_length=21/>
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=3.926990820 delta_omega1=4.31
crick_params_file=beta_strand helix_length=9 delta_t=0.25/>
<Helix omega0=0.04 delta_omega0=3.534291740 delta_omega1=3.14 helix_length=21/>
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=3.141592650 delta_omega1=1.57
crick_params_file=beta_strand helix_length=9 delta_t=1 />
<Helix omega0=0.04 delta_omega0=2.748893570 delta_omega1=3.14 helix_length=21/>
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=2.356194490 delta_omega1=4.31
crick_params_file=beta_strand helix_length=9 delta_t=0.25/>
<Helix omega0=0.04 delta_omega0=1.963495410 delta_omega1=3.14 helix_length=21/>
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=1.570796330 delta_omega1=1.57
crick_params_file=beta_strand helix_length=9 delta_t=1 />
<Helix omega0=0.04 delta_omega0=1.178097250 delta_omega1=3.14 helix_length=21/>
<Helix r0=7.5 invert=1 omega0=0.32 delta_omega0=0.785398163 delta_omega1=4.31
```

```
crick_params_file=beta_strand helix_length=9 delta_t=0.25/>
<Helix omega0=0.04 delta_omega0=0.392699082 delta_omega1=3.14 helix_length=21/>
</BundleGridSampler>
</MOVERS>
<APPLY_TO_POSE>
</APPLY_TO_POSE>
<PROTOCOLS>
<Add mover=Octa7 />
</PROTOCOLS>
</ROSETTASCRIPTS>
```

6.5.4 Alanine substitution

The alanine substitution is performed in 4 steps with the Design package of Rosetta software (version 2015.19.57819_bundle). Three input files are requested: OPTIONS, PARAMETERS and RESFILE.

COMMAND LINE step 1:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag1.txt -s INPUT.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrach
i_cutoff 0 -nstruct 200
```

OPTIONS step 1: @flag1.txt

```
-options
-user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
-protocol design1.xml
```

PARAMETERS step 1: design1.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<s weights=talaris2013_cart />
</SCOREFXNS>
<TASKOPERATIONS>
<InitializeFromCommandline name=ifcl/>
<ReadResfile name="rrf" filename="design1.resfile" />
</TASKOPERATIONS>
```

```
<MOVERS>
<SavePoseMover name=init_struct reference_name=init_struct/>
<FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
<PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
</MOVERS>
<PROTOCOLS>
<Add mover_name=design/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE step 1: design1.resfile

```
NATAA
START

2 A PIKAA VIL
3 A APOLAR NOTAA GC
4 A PIKAA VIL
5 A APOLAR NOTAA GC
6 A PIKAA VIL
7 A APOLAR NOTAA GC
8 A PIKAA VIL
13 A APOLAR NOTAA GC
17 A APOLAR NOTAA GC
20 A APOLAR NOTAA GC
21 A APOLAR NOTAA GC
24 A APOLAR NOTAA GC
32 A PIKAA VIL
33 A APOLAR NOTAA GC
34 A PIKAA VIL
35 A APOLAR NOTAA GC
36 A PIKAA VIL
37 A APOLAR NOTAA GC
38 A PIKAA VIL
43 A APOLAR NOTAA GC
47 A APOLAR NOTAA GC
50 A APOLAR NOTAA GC
51 A APOLAR NOTAA GC
54 A APOLAR NOTAA GC
62 A PIKAA VIL
63 A APOLAR NOTAA GC
64 A PIKAA VIL
65 A APOLAR NOTAA GC
66 A PIKAA VIL
67 A APOLAR NOTAA GC
68 A PIKAA VIL
73 A APOLAR NOTAA GC
77 A APOLAR NOTAA GC
80 A APOLAR NOTAA GC
81 A APOLAR NOTAA GC
84 A APOLAR NOTAA GC
92 A PIKAA VIL
93 A APOLAR NOTAA GC
```

```
94 A PIKAA VIL
95 A APOLAR NOTAA GC
96 A PIKAA VIL
97 A APOLAR NOTAA GC
98 A PIKAA VIL
103 A APOLAR NOTAA GC
107 A APOLAR NOTAA GC
110 A APOLAR NOTAA GC
111 A APOLAR NOTAA GC
114 A APOLAR NOTAA GC
122 A PIKAA VIL
123 A APOLAR NOTAA GC
124 A PIKAA VIL
125 A APOLAR NOTAA GC
126 A PIKAA VIL
127 A APOLAR NOTAA GC
128 A PIKAA VIL
133 A APOLAR NOTAA GC
137 A APOLAR NOTAA GC
140 A APOLAR NOTAA GC
141 A APOLAR NOTAA GC
144 A APOLAR NOTAA GC
152 A PIKAA VIL
153 A APOLAR NOTAA GC
154 A PIKAA VIL
155 A APOLAR NOTAA GC
156 A PIKAA VIL
157 A APOLAR NOTAA GC
158 A PIKAA VIL
163 A APOLAR NOTAA GC
167 A APOLAR NOTAA GC
170 A APOLAR NOTAA GC
171 A APOLAR NOTAA GC
174 A APOLAR NOTAA GC
182 A PIKAA VIL
183 A APOLAR NOTAA GC
184 A PIKAA VIL
185 A APOLAR NOTAA GC
186 A PIKAA VIL
187 A APOLAR NOTAA GC
188 A PIKAA VIL
193 A APOLAR NOTAA GC
197 A APOLAR NOTAA GC
200 A APOLAR NOTAA GC
201 A APOLAR NOTAA GC
204 A APOLAR NOTAA GC
212 A PIKAA VIL
213 A APOLAR NOTAA GC
214 A PIKAA VIL
215 A APOLAR NOTAA GC
216 A PIKAA VIL
217 A APOLAR NOTAA GC
218 A PIKAA VIL
223 A APOLAR NOTAA GC
227 A APOLAR NOTAA GC
```

```
230 A APOLAR NOTAA GC
231 A APOLAR NOTAA GC
234 A APOLAR NOTAA GC
```

COMMAND LINE step 2:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag2.txt -s INPUTs.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrac
hi_cutoff 0 -nstruct 50
```

OPTIONS step 2: @flag2.txt

```
-options
-user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
-protocol design2.xml
```

PARAMETERS step 2: design2.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<s weights=talaris2013_cart />
</SCOREFXNS>
<TASKOPERATIONS>
<InitializeFromCommandline name=ifcl/>
<ReadResfile name="rrf" filename="design2.resfile" />
</TASKOPERATIONS>
<MOVERS>
<SavePoseMover name=init_struct reference_name=init_struct/>
<FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
<PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
</MOVERS>
<PROTOCOLS>
<Add mover_name=design/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE step 2: design2.resfile

```
NATAA
START
```

```
14 A NOTAA CPGA
16 A NOTAA CPGA
23 A NOTAA CPGA
25 A NOTAA CPGA
27 A NOTAA CPGA
44 A NOTAA CPGA
46 A NOTAA CPGA
53 A NOTAA CPGA
55 A NOTAA CPGA
57 A NOTAA CPGA
74 A NOTAA CPGA
76 A NOTAA CPGA
83 A NOTAA CPGA
85 A NOTAA CPGA
87 A NOTAA CPGA
104 A NOTAA CPGA
106 A NOTAA CPGA
113 A NOTAA CPGA
115 A NOTAA CPGA
117 A NOTAA CPGA
134 A NOTAA CPGA
136 A NOTAA CPGA
143 A NOTAA CPGA
145 A NOTAA CPGA
147 A NOTAA CPGA
164 A NOTAA CPGA
166 A NOTAA CPGA
173 A NOTAA CPGA
175 A NOTAA CPGA
177 A NOTAA CPGA
194 A NOTAA CPGA
196 A NOTAA CPGA
203 A NOTAA CPGA
205 A NOTAA CPGA
207 A NOTAA CPGA
224 A NOTAA CPGA
226 A NOTAA CPGA
233 A NOTAA CPGA
235 A NOTAA CPGA
237 A NOTAA CPGA
```

COMMAND LINE step 3:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag3.txt -s INPUTs.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrac
hi_cutoff 0 -nstruct 50
```

OPTIONS step 3: @flag3.txt

```
-options
-user
```

```
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
-protocol design3.xml
```

PARAMETERS step 3: design3.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<s weights=talaris2013_cart />
</SCOREFXNS>
<TASKOPERATIONS>
<InitializeFromCommandline name=ifcl/>
<ReadResfile name="rrf" filename="design3.resfile" />
</TASKOPERATIONS>
<MOVERS>
<SavePoseMover name=init_struct reference_name=init_struct/>
<FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
<PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
</MOVERS>
<PROTOCOLS>
<Add mover_name=design/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE step 3: design3.resfile

```
NATAA
START

15 A POLAR NOTAA GC
18 A POLAR NOTAA GC
19 A POLAR NOTAA GC
22 A POLAR NOTAA GC
26 A POLAR NOTAA GC
45 A POLAR NOTAA GC
48 A POLAR NOTAA GC
49 A POLAR NOTAA GC
52 A POLAR NOTAA GC
56 A POLAR NOTAA GC
75 A POLAR NOTAA GC
78 A POLAR NOTAA GC
79 A POLAR NOTAA GC
82 A POLAR NOTAA GC
86 A POLAR NOTAA GC
105 A POLAR NOTAA GC
108 A POLAR NOTAA GC
109 A POLAR NOTAA GC
```

```
112 A POLAR NOTAA GC
116 A POLAR NOTAA GC
135 A POLAR NOTAA GC
138 A POLAR NOTAA GC
139 A POLAR NOTAA GC
142 A POLAR NOTAA GC
146 A POLAR NOTAA GC
165 A POLAR NOTAA GC
168 A POLAR NOTAA GC
169 A POLAR NOTAA GC
172 A POLAR NOTAA GC
176 A POLAR NOTAA GC
195 A POLAR NOTAA GC
198 A POLAR NOTAA GC
199 A POLAR NOTAA GC
202 A POLAR NOTAA GC
206 A POLAR NOTAA GC
225 A POLAR NOTAA GC
228 A POLAR NOTAA GC
229 A POLAR NOTAA GC
232 A POLAR NOTAA GC
236 A POLAR NOTAA GC
```

COMMAND LINE step 4:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag4.txt -s INPUTs.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrac
hi_cutoff 0 -nstruct 50
```

OPTIONS step 4: @flag4.txt

```
-options
-user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
-protocol design4.xml
```

PARAMETERS step 4: design4.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<s weights=talaris2013_cart />
</SCOREFXNS>
<TASKOPERATIONS>
<InitializeFromCommandline name=ifcl/>
<ReadResfile name="rrf" filename="design4.resfile" />
```

```
</TASKOPERATIONS>
<MOVERS>
<SavePoseMover name=init_struct reference_name=init_struct/>
<FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
<PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
</MOVERS>
<PROTOCOLS>
<Add mover_name=design/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE step 4: design4.resfile

```
NATAA
START

1 A NOTAA GCWYKF
9 A NOTAA GCWYKF
10 A NOTAA GCWYKF
11 A NOTAA GCWYKF
12 A NOTAA GCWYKF
28 A NOTAA GCWYKF
29 A NOTAA GCWYKF
30 A NOTAA GCWYKF
31 A NOTAA GCWYKF
39 A NOTAA GCWYKF
40 A NOTAA GCWYKF
41 A NOTAA GCWYKF
42 A NOTAA GCWYKF
58 A NOTAA GCWYKF
59 A NOTAA GCWYKF
60 A NOTAA GCWYKF
61 A NOTAA GCWYKF
69 A NOTAA GCWYKF
70 A NOTAA GCWYKF
71 A NOTAA GCWYKF
72 A NOTAA GCWYKF
88 A NOTAA GCWYKF
89 A NOTAA GCWYKF
90 A NOTAA GCWYKF
91 A NOTAA GCWYKF
99 A NOTAA GCWYKF
100 A NOTAA GCWYKF
101 A NOTAA GCWYKF
102 A NOTAA GCWYKF
118 A NOTAA GCWYKF
119 A NOTAA GCWYKF
120 A NOTAA GCWYKF
121 A NOTAA GCWYKF
129 A NOTAA GCWYKF
130 A NOTAA GCWYKF
131 A NOTAA GCWYKF
132 A NOTAA GCWYKF
```

```
148 A NOTAA GCWYKF
149 A NOTAA GCWYKF
150 A NOTAA GCWYKF
151 A NOTAA GCWYKF
159 A NOTAA GCWYKF
160 A NOTAA GCWYKF
161 A NOTAA GCWYKF
162 A NOTAA GCWYKF
178 A NOTAA GCWYKF
179 A NOTAA GCWYKF
180 A NOTAA GCWYKF
181 A NOTAA GCWYKF
189 A NOTAA GCWYKF
190 A NOTAA GCWYKF
191 A NOTAA GCWYKF
192 A NOTAA GCWYKF
208 A NOTAA GCWYKF
209 A NOTAA GCWYKF
210 A NOTAA GCWYKF
211 A NOTAA GCWYKF
219 A NOTAA GCWYKF
220 A NOTAA GCWYKF
221 A NOTAA GCWYKF
222 A NOTAA GCWYKF
238 A NOTAA GCWYKF
239 A NOTAA GCWYKF
240 A NOTAA GCWYKF
```

6.5.5 Loop closure

Loop closure is performed with the package Loopmodel of Modeller version 9.15. The INSTRUCTION file is requested as input together with the pdb file.

Command line:

```
python inputname.py
```

INSTRUCTION: inputname.py

```
from modeller import *
from modeller.automodel import *
log.verbose()
env = environ()
env.io.atom_files_directory = ['.', '../atom_files']

class MyLoop(loopmodel):
      def select_loop_atoms(self):
           return selection(self.residue_range('8:A', '12:A'),
                            self.residue_range('27:A', '31:A'),
                            self.residue_range('38:A', '42:A'),
```

```
                    self.residue_range('57:A', '61:A'),
                    self.residue_range('68:A', '72:A'),
                    self.residue_range('87:A', '91:A'),
                    self.residue_range('98:A', '102:A'),
                    self.residue_range('117:A', '121:A'),
                    self.residue_range('128:A', '132:A'),
                    self.residue_range('147:A', '151:A'),
                    self.residue_range('158:A', '162:A'),
                    self.residue_range('177:A', '181:A'),
                    self.residue_range('188:A', '192:A'),
                    self.residue_range('207:A', '211:A'),
                    self.residue_range('218:A', '222:A'))

m = MyLoop(env, inimodel='Model_Name.pdb', sequence='1', loop_assess_methods=assess.DOPE)
m.loop.starting_model= 1
m.loop.ending_model  = 1
m.loop.md_level = refine.very_fast
m.make()
```

6.5.6 Energy minimization of the backbone structures

Energy minimization of the natural TIM-barrels is performed with the package Relax of the Rosetta software (version 2015.19.57819_bundle). All the input pdb structures are listed in the file listpdb.

Command line:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/relax.linuxgcc
release -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/databas
e/ -l listpdb -in:file:fullatom -relax:constrain_relax_to_start_coords
```

6.5.7 Sequence design

The alanine substitution is performed in 4 cycles with the packages Design and Relax of Rosetta software (version 2015.19.57819_bundle). Three input files are requested: OPTIONS, PARAMETERS and RESFILE.

COMMAND LINE cycle 1:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag1.txt -s INPUT.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrach
i_cutoff 0 -nstruct 100
```

OPTIONS cycle 1: @flag1.txt

```
-options
 -user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
 -protocol design1.xml
```

PARAMETERS cycle 1: design1.xml

```
<ROSETTASCRIPTS>
       <SCOREFXNS>
             <s weights=talaris2013_cart />
       </SCOREFXNS>
       <TASKOPERATIONS>
             <InitializeFromCommandline name=ifcl/>
             <ReadResfile name="rrf" filename="design1.resfile" />
       </TASKOPERATIONS>
       <MOVERS>
             <SavePoseMover name=init_struct reference_name=init_struct/>
             <FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
             <PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
      </MOVERS>
      <PROTOCOLS>
      <Add mover_name=design/>
      </PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE cycle 1: design1.resfile

```
NATAA
START

2 A APOLAR
3 A APOLAR
4 A APOLAR
5 A APOLAR
6 A APOLAR
7 A APOLAR
8 A APOLAR
13 A APOLAR
17 A APOLAR
20 A APOLAR
21 A APOLAR
24 A APOLAR
32 A APOLAR
```

```
33 A APOLAR
34 A APOLAR
35 A APOLAR
36 A APOLAR
37 A APOLAR
38 A APOLAR
43 A APOLAR
47 A APOLAR
50 A APOLAR
51 A APOLAR
54 A APOLAR
62 A APOLAR
63 A APOLAR
64 A APOLAR
65 A APOLAR
66 A APOLAR
67 A APOLAR
68 A APOLAR
73 A APOLAR
77 A APOLAR
80 A APOLAR
81 A APOLAR
84 A APOLAR
92 A APOLAR
93 A APOLAR
94 A APOLAR
95 A APOLAR
96 A APOLAR
97 A APOLAR
98 A APOLAR
103 A APOLAR
107 A APOLAR
110 A APOLAR
111 A APOLAR
114 A APOLAR
122 A APOLAR
123 A APOLAR
124 A APOLAR
125 A APOLAR
126 A APOLAR
127 A APOLAR
128 A APOLAR
133 A APOLAR
137 A APOLAR
140 A APOLAR
141 A APOLAR
144 A APOLAR
152 A APOLAR
153 A APOLAR
154 A APOLAR
155 A APOLAR
156 A APOLAR
157 A APOLAR
158 A APOLAR
163 A APOLAR
```

```
167 A APOLAR
170 A APOLAR
171 A APOLAR
174 A APOLAR
182 A APOLAR
183 A APOLAR
184 A APOLAR
185 A APOLAR
186 A APOLAR
187 A APOLAR
188 A APOLAR
193 A APOLAR
197 A APOLAR
200 A APOLAR
201 A APOLAR
204 A APOLAR
212 A APOLAR
213 A APOLAR
214 A APOLAR
215 A APOLAR
216 A APOLAR
217 A APOLAR
218 A APOLAR
223 A APOLAR
227 A APOLAR
230 A APOLAR
231 A APOLAR
234 A APOLAR
```

COMMAND LINE cycles 2-4:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag2.txt -s INPUT.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrach
i_cutoff 0 -nstruct 10
```

OPTIONS cycles 2-4: @flag2.txt

```
-options
-user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
-protocol design2.xml
```

PARAMETERS cycles 2-4: design2.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<s weights=talaris2013_cart />
```

```
</SCOREFXNS>
<TASKOPERATIONS>
<InitializeFromCommandline name=ifcl/>
<ReadResfile name="rrf" filename="design2.resfile" />
</TASKOPERATIONS>
<MOVERS>
<SavePoseMover name=init_struct reference_name=init_struct/>
<FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
<PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
</MOVERS>
<PROTOCOLS>
<Add mover_name=design/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE cycles 2-4: design2.resfile

```
NATAA
START

1 A ALLAA
2 A APOLAR
3 A APOLAR
4 A APOLAR
5 A APOLAR
6 A APOLAR
7 A APOLAR
8 A APOLAR
9 A ALLAA
10 A ALLAA
11 A ALLAA
12 A ALLAA
13 A APOLAR
14 A ALLAA
15 A POLAR
16 A ALLAA
17 A APOLAR
18 A POLAR
19 A POLAR
20 A APOLAR
21 A APOLAR
22 A POLAR
23 A ALLAA
24 A APOLAR
25 A ALLAA
26 A POLAR
27 A ALLAA
28 A APOLAR
29 A ALLAA
30 A ALLAA
31 A ALLAA
32 A APOLAR
```

```
33 A APOLAR
34 A APOLAR
35 A APOLAR
36 A APOLAR
37 A APOLAR
38 A APOLAR
39 A ALLAA
40 A ALLAA
41 A ALLAA
42 A ALLAA
43 A APOLAR
44 A ALLAA
45 A POLAR
46 A ALLAA
47 A APOLAR
48 A POLAR
49 A POLAR
50 A APOLAR
51 A APOLAR
52 A POLAR
53 A ALLAA
54 A APOLAR
55 A ALLAA
56 A POLAR
57 A ALLAA
58 A ALLAA
59 A ALLAA
60 A ALLAA
61 A ALLAA
62 A APOLAR
63 A APOLAR
64 A APOLAR
65 A APOLAR
66 A APOLAR
67 A APOLAR
68 A APOLAR
69 A ALLAA
70 A ALLAA
71 A ALLAA
72 A ALLAA
73 A APOLAR
74 A ALLAA
75 A POLAR
76 A ALLAA
77 A APOLAR
78 A POLAR
79 A POLAR
80 A APOLAR
81 A APOLAR
82 A POLAR
83 A ALLAA
84 A APOLAR
85 A ALLAA
86 A POLAR
87 A ALLAA
```

```
88 A ALLAA
89 A ALLAA
90 A ALLAA
91 A ALLAA
92 A APOLAR
93 A APOLAR
94 A APOLAR
95 A APOLAR
96 A APOLAR
97 A APOLAR
98 A APOLAR
99 A ALLAA
100 A ALLAA
101 A ALLAA
102 A ALLAA
103 A APOLAR
104 A ALLAA
105 A POLAR
106 A ALLAA
107 A APOLAR
108 A POLAR
109 A POLAR
110 A APOLAR
111 A APOLAR
112 A POLAR
113 A ALLAA
114 A APOLAR
115 A ALLAA
116 A POLAR
117 A ALLAA
118 A ALLAA
119 A ALLAA
120 A ALLAA
121 A ALLAA
122 A APOLAR
123 A APOLAR
124 A APOLAR
125 A APOLAR
126 A APOLAR
127 A APOLAR
128 A APOLAR
129 A ALLAA
130 A ALLAA
131 A ALLAA
132 A ALLAA
133 A APOLAR
134 A ALLAA
135 A POLAR
136 A ALLAA
137 A APOLAR
138 A POLAR
139 A POLAR
140 A APOLAR
141 A APOLAR
142 A POLAR
```

```
143 A ALLAA
144 A APOLAR
145 A ALLAA
146 A POLAR
147 A ALLAA
148 A ALLAA
149 A ALLAA
150 A ALLAA
151 A ALLAA
152 A APOLAR
153 A APOLAR
154 A APOLAR
155 A APOLAR
156 A APOLAR
157 A APOLAR
158 A APOLAR
159 A ALLAA
160 A ALLAA
161 A ALLAA
162 A ALLAA
163 A APOLAR
164 A ALLAA
165 A POLAR
166 A ALLAA
167 A APOLAR
168 A POLAR
169 A POLAR
170 A APOLAR
171 A APOLAR
172 A POLAR
173 A ALLAA
174 A APOLAR
175 A ALLAA
176 A POLAR
177 A ALLAA
178 A ALLAA
179 A ALLAA
180 A ALLAA
181 A ALLAA
182 A APOLAR
183 A APOLAR
184 A APOLAR
185 A APOLAR
186 A APOLAR
187 A APOLAR
188 A APOLAR
189 A ALLAA
190 A ALLAA
191 A ALLAA
192 A ALLAA
193 A APOLAR
194 A ALLAA
195 A POLAR
196 A ALLAA
197 A APOLAR
```

```
198 A POLAR
199 A POLAR
200 A APOLAR
201 A APOLAR
202 A POLAR
203 A ALLAA
204 A APOLAR
205 A ALLAA
206 A POLAR
207 A ALLAA
208 A ALLAA
209 A ALLAA
210 A ALLAA
211 A ALLAA
212 A APOLAR
213 A APOLAR
214 A APOLAR
215 A APOLAR
216 A APOLAR
217 A APOLAR
218 A APOLAR
219 A ALLAA
220 A POLAR
221 A ALLAA
222 A ALLAA
223 A APOLAR
224 A ALLAA
225 A POLAR
226 A ALLAA
227 A APOLAR
228 A ALLAA
229 A POLAR
230 A APOLAR
231 A APOLAR
232 A POLAR
233 A ALLAA
234 A APOLAR
235 A ALLAA
236 A POLAR
237 A ALLAA
238 A ALLAA
239 A ALLAA
240 A ALLAA
```

COMMAND LINE cycles 5-7:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag5.txt -s INPUT.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrach
i_cutoff 0 -nstruct 10
```

OPTIONS cycles 5-7: @flag5.txt

```
-options
```

```
-user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
-protocol design5.xml
```

PARAMETERS cycles 5-7: design5.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<s weights=talaris2013_cart />
</SCOREFXNS>
<TASKOPERATIONS>
<InitializeFromCommandline name=ifcl/>
<ReadResfile name="rrf" filename="design5.resfile" />
</TASKOPERATIONS>
<MOVERS>
<SavePoseMover name=init_struct reference_name=init_struct/>
<FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
<PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
</MOVERS>
<PROTOCOLS>
<Add mover_name=design/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE cycles 5-7: design5.resfile

```
NATAA
START

1 A ALLAA
2 A APOLAR
3 A APOLAR
4 A APOLAR
5 A APOLAR
6 A APOLAR
7 A APOLAR
8 A APOLAR
9 A ALLAA
10 A ALLAA
11 A ALLAA
12 A ALLAA
13 A APOLAR
14 A ALLAA
15 A POLAR
16 A ALLAA
17 A APOLAR
```

```
18 A POLAR
19 A POLAR
20 A APOLAR
21 A APOLAR
22 A POLAR
23 A ALLAA
24 A APOLAR
25 A ALLAA
26 A POLAR
27 A ALLAA
28 A APOLAR
29 A ALLAA
30 A ALLAA
31 A ALLAA
32 A APOLAR
33 A APOLAR
34 A APOLAR
35 A APOLAR
36 A APOLAR
37 A APOLAR
38 A APOLAR
39 A ALLAA
40 A ALLAA
41 A ALLAA
42 A ALLAA
43 A APOLAR
44 A ALLAA
45 A POLAR
46 A ALLAA
47 A APOLAR
48 A POLAR
49 A POLAR
50 A APOLAR
51 A APOLAR
52 A POLAR
53 A ALLAA
54 A APOLAR
55 A ALLAA
56 A POLAR
57 A ALLAA
58 A ALLAA
59 A ALLAA
60 A ALLAA
61 A ALLAA
62 A APOLAR
63 A APOLAR
64 A APOLAR
65 A APOLAR
66 A APOLAR
67 A APOLAR
68 A APOLAR
69 A ALLAA
70 A ALLAA
71 A ALLAA
72 A ALLAA
```

```
73 A APOLAR
74 A ALLAA
75 A POLAR
76 A ALLAA
77 A APOLAR
78 A POLAR
79 A POLAR
80 A APOLAR
81 A APOLAR
82 A POLAR
83 A ALLAA
84 A APOLAR
85 A ALLAA
86 A POLAR
87 A ALLAA
88 A ALLAA
89 A ALLAA
90 A ALLAA
91 A ALLAA
92 A APOLAR
93 A APOLAR
94 A APOLAR
95 A APOLAR
96 A APOLAR
97 A APOLAR
98 A APOLAR
99 A ALLAA
100 A ALLAA
101 A ALLAA
102 A ALLAA
103 A APOLAR
104 A ALLAA
105 A POLAR
106 A ALLAA
107 A APOLAR
108 A POLAR
109 A POLAR
110 A APOLAR
111 A APOLAR
112 A POLAR
113 A ALLAA
114 A APOLAR
115 A ALLAA
116 A POLAR
117 A ALLAA
118 A ALLAA
119 A ALLAA
120 A ALLAA
121 A ALLAA
122 A APOLAR
123 A APOLAR
124 A APOLAR
125 A APOLAR
126 A APOLAR
127 A APOLAR
```

```
128 A APOLAR
129 A ALLAA
130 A ALLAA
131 A ALLAA
132 A ALLAA
133 A APOLAR
134 A ALLAA
135 A POLAR
136 A ALLAA
137 A APOLAR
138 A POLAR
139 A POLAR
140 A APOLAR
141 A APOLAR
142 A POLAR
143 A ALLAA
144 A APOLAR
145 A ALLAA
146 A POLAR
147 A ALLAA
148 A ALLAA
149 A ALLAA
150 A ALLAA
151 A ALLAA
152 A APOLAR
153 A APOLAR
154 A APOLAR
155 A APOLAR
156 A APOLAR
157 A APOLAR
158 A APOLAR
159 A ALLAA
160 A ALLAA
161 A ALLAA
162 A ALLAA
163 A APOLAR
164 A ALLAA
165 A POLAR
166 A ALLAA
167 A APOLAR
168 A POLAR
169 A POLAR
170 A APOLAR
171 A APOLAR
172 A POLAR
173 A ALLAA
174 A APOLAR
175 A ALLAA
176 A POLAR
177 A ALLAA
178 A ALLAA
179 A ALLAA
180 A ALLAA
181 A ALLAA
182 A APOLAR
```

```
183 A APOLAR
184 A APOLAR
185 A APOLAR
186 A APOLAR
187 A APOLAR
188 A APOLAR
189 A ALLAA
190 A ALLAA
191 A ALLAA
192 A ALLAA
193 A APOLAR
194 A ALLAA
195 A POLAR
196 A ALLAA
197 A APOLAR
198 A POLAR
199 A POLAR
200 A APOLAR
201 A APOLAR
202 A POLAR
203 A ALLAA
204 A APOLAR
205 A ALLAA
206 A POLAR
207 A ALLAA
208 A ALLAA
209 A ALLAA
210 A ALLAA
211 A ALLAA
212 A APOLAR
213 A APOLAR
214 A APOLAR
215 A APOLAR
216 A APOLAR
217 A APOLAR
218 A APOLAR
219 A ALLAA
220 A POLAR
221 A ALLAA
222 A ALLAA
223 A APOLAR
224 A ALLAA
225 A POLAR
226 A ALLAA
227 A APOLAR
228 A ALLAA
229 A POLAR
230 A APOLAR
231 A APOLAR
232 A POLAR
233 A ALLAA
234 A APOLAR
235 A ALLAA
236 A POLAR
237 A ALLAA
```

```
238 A ALLAA
239 A ALLAA
240 A ALLAA
```

COMMAND LINE cycles 8-10:

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag8.txt -s INPUT.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrach
i_cutoff 0 -nstruct 50
```

OPTIONS cycles 8-10: @flag8.txt

```
-options
-user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
-protocol design8.xml
```

PARAMETERS cycles 8-10: design8.xml

```
<ROSETTASCRIPTS>
<SCOREFXNS>
<s weights=talaris2013_cart />
</SCOREFXNS>
<TASKOPERATIONS>
<InitializeFromCommandline name=ifcl/>
<ReadResfile name="rrf" filename="design8.resfile" />
</TASKOPERATIONS>
<MOVERS>
<SavePoseMover name=init_struct reference_name=init_struct/>
<FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
<PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
</MOVERS>
<PROTOCOLS>
<Add mover_name=design/>
</PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE cycles 8-10: design8.resfile

```
NATAA
START
```

```
1 A ALLAA
2 A ALLAA
3 A ALLAA
4 A ALLAA
5 A ALLAA
6 A ALLAA
7 A ALLAA
8 A ALLAA
9 A ALLAA
10 A ALLAA
11 A ALLAA
12 A ALLAA
13 A ALLAA
14 A ALLAA
15 A ALLAA
16 A ALLAA
17 A ALLAA
18 A ALLAA
19 A ALLAA
20 A ALLAA
21 A ALLAA
22 A ALLAA
23 A ALLAA
24 A ALLAA
25 A ALLAA
26 A ALLAA
27 A ALLAA
28 A ALLAA
29 A ALLAA
30 A ALLAA
31 A ALLAA
32 A ALLAA
33 A ALLAA
34 A ALLAA
35 A ALLAA
36 A ALLAA
37 A ALLAA
38 A ALLAA
39 A ALLAA
40 A ALLAA
41 A ALLAA
42 A ALLAA
43 A ALLAA
44 A ALLAA
45 A ALLAA
46 A ALLAA
47 A ALLAA
48 A ALLAA
49 A ALLAA
50 A ALLAA
51 A ALLAA
52 A ALLAA
53 A ALLAA
54 A ALLAA
55 A ALLAA
```

```
56 A ALLAA
57 A ALLAA
58 A ALLAA
59 A ALLAA
60 A ALLAA
61 A ALLAA
62 A ALLAA
63 A ALLAA
64 A ALLAA
65 A ALLAA
66 A ALLAA
67 A ALLAA
68 A ALLAA
69 A ALLAA
70 A ALLAA
71 A ALLAA
72 A ALLAA
73 A ALLAA
74 A ALLAA
75 A ALLAA
76 A ALLAA
77 A ALLAA
78 A ALLAA
79 A ALLAA
80 A ALLAA
81 A ALLAA
82 A ALLAA
83 A ALLAA
84 A ALLAA
85 A ALLAA
86 A ALLAA
87 A ALLAA
88 A ALLAA
89 A ALLAA
90 A ALLAA
91 A ALLAA
92 A ALLAA
93 A ALLAA
94 A ALLAA
95 A ALLAA
96 A ALLAA
97 A ALLAA
98 A ALLAA
99 A ALLAA
100 A ALLAA
101 A ALLAA
102 A ALLAA
103 A ALLAA
104 A ALLAA
105 A ALLAA
106 A ALLAA
107 A ALLAA
108 A ALLAA
109 A ALLAA
110 A ALLAA
```

```
111 A ALLAA
112 A ALLAA
113 A ALLAA
114 A ALLAA
115 A ALLAA
116 A ALLAA
117 A ALLAA
118 A ALLAA
119 A ALLAA
120 A ALLAA
121 A ALLAA
122 A ALLAA
123 A ALLAA
124 A ALLAA
125 A ALLAA
126 A ALLAA
127 A ALLAA
128 A ALLAA
129 A ALLAA
130 A ALLAA
131 A ALLAA
132 A ALLAA
133 A ALLAA
134 A ALLAA
135 A ALLAA
136 A ALLAA
137 A ALLAA
138 A ALLAA
139 A ALLAA
140 A ALLAA
141 A ALLAA
142 A ALLAA
143 A ALLAA
144 A ALLAA
145 A ALLAA
146 A ALLAA
147 A ALLAA
148 A ALLAA
149 A ALLAA
150 A ALLAA
151 A ALLAA
152 A ALLAA
153 A ALLAA
154 A ALLAA
155 A ALLAA
156 A ALLAA
157 A ALLAA
158 A ALLAA
159 A ALLAA
160 A ALLAA
161 A ALLAA
162 A ALLAA
163 A ALLAA
164 A ALLAA
165 A ALLAA
```

```
166 A ALLAA
167 A ALLAA
168 A ALLAA
169 A ALLAA
170 A ALLAA
171 A ALLAA
172 A ALLAA
173 A ALLAA
174 A ALLAA
175 A ALLAA
176 A ALLAA
177 A ALLAA
178 A ALLAA
179 A ALLAA
180 A ALLAA
181 A ALLAA
182 A ALLAA
183 A ALLAA
184 A ALLAA
185 A ALLAA
186 A ALLAA
187 A ALLAA
188 A ALLAA
189 A ALLAA
190 A ALLAA
191 A ALLAA
192 A ALLAA
193 A ALLAA
194 A ALLAA
195 A ALLAA
196 A ALLAA
197 A ALLAA
198 A ALLAA
199 A ALLAA
200 A ALLAA
201 A ALLAA
202 A ALLAA
203 A ALLAA
204 A ALLAA
205 A ALLAA
206 A ALLAA
207 A ALLAA
208 A ALLAA
209 A ALLAA
210 A ALLAA
211 A ALLAA
212 A ALLAA
213 A ALLAA
214 A ALLAA
215 A ALLAA
216 A ALLAA
217 A ALLAA
218 A ALLAA
219 A ALLAA
220 A ALLAA
```

```
221 A ALLAA
222 A ALLAA
223 A ALLAA
224 A ALLAA
225 A ALLAA
226 A ALLAA
227 A ALLAA
228 A ALLAA
229 A ALLAA
230 A ALLAA
231 A ALLAA
232 A ALLAA
233 A ALLAA
234 A ALLAA
235 A ALLAA
236 A ALLAA
237 A ALLAA
238 A ALLAA
239 A ALLAA
240 A ALLAA
```

6.5.8 Molecular dynamics

The molecular dynamic is performed in 7 steps with the GROMACS software (version 5.0.7). The input file is the protein structure (pdb). PARAMETER files are reported when necessary.

COMMAND LINE step 1, Generate the topology:

```
pdb2gmx -v -f Protein.pdb -o Protein.gro -p Protein.top -ignh -water tip3p -ff amber99sb
```

COMMAND LINE step 2, Add periodic boundaries:

```
editconf -f Protein.gro -o 3_PBC.gro -bt dodecahedron -d 1.0
```

COMMAND LINE step 3, Add water molecules:

```
genbox -cp 3_PBC.gro -cs spc216.gro -p Protein.top -o 4_Water.gro
```

COMMAND LINE step 4, Add ions:

```
grompp -v -f minim.mdp -c 4_Water.gro -p Protein.top -o 5_Setup.trp
```

COMMAND LINE step 5, Energy minimization:

```
grompp -v -f minim.mdp  -c 5_Ions.gro -p Protein.top -o 6_Setup.tpr
mdrun -v -deffnm 6_Setup -c 6_EM.gro
```

PARAMETERS step 5: minim.mdp

```
title         = Energy Minimization
cpp           = /lib/cpp
define        = -DFLEXIBLE
integrator    = steep
emtol         = 1000.0
emstep        = 0.01
nsteps        = 50000
nstlist       = 20
cutoff-scheme = Verlet
ns_type       = grid
coulombtype   = PME
rcoulomb      = 1.0
rvdw          = 1.0
pbc           = xyz
```

COMMAND LINE step 6, Temperature and pressure coupling:

```
grompp -v -f TempCoupling.mdp -c 6_EM.gro -p Protein.top -o 7_Setup.tpr
mdrun -v -deffnm 7_Setup
grompp -v -f PressCoupling.mdp -c 7_Setup.gro -p Protein.top -o 8_Setup.tpr
mdrun -v -deffnm 8_Setup
```

PARAMETERS step 6: TempCoupling.mdp

```
title                    = NVT simulation
define                   = -DPOSRES
; RUN CONTROL PARAMETERS
integrator               = md
dt                       = 0.002
nsteps                   = 50000
; OUTPUT CONTROL OPTIONS
nstxout                  = 500
nstvout                  = 500
nstlog                   = 500
nstenergy                = 500
; NEIGHBORSEARCHING PARAMETERS
cutoff-scheme            = Verlet
nstlist                  = 10
ns_type                  = grid
rcoulomb                 = 1.0
rvdw                     = 1.0
; OPTIONS FOR BONDS
continuation             = no
```

```
constraint_algorithm     = lincs
constraints              = all-bonds
lincs_iter               = 1
lincs_order              = 4
; OPTIONS FOR ELECTROSTATICS AND VDW
coulombtype              = PME
pme_order                = 4
fourierspacing           = 0.16
; DISPERSION CORRECTION
DispCorr                 = EnerPres
; VELOCITY GENERATION
gen_vel                  = yes
gen_temp                 = 300
gen_seed                 = -1
; TEMPERATURE COUPLING
tcoupl                   = V-rescale
tc-grps                  = Protein Non-Protein
tau_t                    = 0.1      0.1
ref_t                    = 300      300
; PRESSURE COUPLING
pcoupl                   = no
```

PARAMETERS step 6: PressCoupling.mdp

```
title                    = NPT equilibration
define                   = -DPOSRES
; RUN CONTROL PARAMETERS
integrator               = md
nsteps                   = 50000
dt                       = 0.002
; OUTPUT CONTROL OPTIONS
nstxout                  = 500
nstvout                  = 500
nstenergy                = 500
nstlog                   = 500
; NEIGHBORSEARCHING PARAMETERS
cutoff-scheme            = Verlet
ns_type                  = grid
nstlist                  = 10
rcoulomb                 = 1.0
rvdw                     = 1.0
; OPTION FOR BONDS
continuation             = yes
constraint_algorithm     = lincs
```

```
constraints             = all-bonds
lincs_iter              = 1
lincs_order             = 4
; OPTIONS FOR ELECTROSTATICS AND VDW
coulombtype              = PME
pme_order                = 4
fourierspacing           = 0.16
; DISPERSION CORRECTION
DispCorr                 = EnerPres
; VELOCITY GENERATION
gen_vel                  = no
; TEMPERATURE COUPLING
tcoupl                   = V-rescale
tc-grps                  = Protein Non-Protein
tau_t                    = 0.1   0.1
ref_t                    = 300    300
; PRESSURE COUPLING
pcoupl                   = Parrinello-Rahman
pcoupltype               = isotropic
tau_p                    = 2.0
ref_p                    = 1.0
compressibility          = 4.5e-5
refcoord_scaling         = com
; PBC
pbc                      = xyz
```

COMMAND LINE step 7, Molecular dynamic simulation:

```
grompp -v -f MD.mdp -c 8_Setup.gro -p Protein.top -o topol.tpr
mdrun -v
```

PARAMETERS step 7: MD.mdp

```
title = Production Simulation
; RUN CONTROL PARAMETERS
integrator           = md
nsteps               = 25000000
dt                   = 0.002
; OUTPUT CONTROL OPTIONS
nstxout              = 5000
nstvout              = 5000
nstenergy            = 5000
nstlog               = 5000
nstxout-compressed   = 5000
```

```
; nstxout-compressed replaces nstxtcout
compressed-x-grps    = System
; Bond parameters
continuation         = yes
constraint_algorithm = lincs
constraints          = all-bonds
lincs_iter           = 1
lincs_order          = 4
; Neighborsearching
cutoff-scheme        = Verlet
ns_type              = grid
nstlist              = 10
rcoulomb             = 1.0
rvdw                 = 1.0
; Electrostatics
coulombtype      = PME
pme_order        = 4
fourierspacing   = 0.16
; Temperature coupling is on
tcoupl           = V-rescale
tc-grps          = Protein Non-Protein
tau_t            = 0.1   0.1
ref_t            = 300    300
; Pressure coupling is on
pcoupl           = Parrinello-Rahman
pcoupltype       = isotropic
tau_p            = 2.0
ref_p            = 1.0
compressibility  = 4.5e-5
; Periodic boundary conditions
pbc              = xyz
; Dispersion correction
DispCorr         = EnerPres
; Velocity generation
gen_vel          = no
```

6.5.9 Cystein removals with Rosetta

The alanine substitution is performed in 4 cycles with the packages Design and Relax of Rosetta software (version 2015.19.57819_bundle). Three input files are requested: OPTIONS, PARAMETERS and RESFILE.

COMMAND LINE

```
/path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/source/bin/rosetta_scripts.
linuxgccrelease -database /path-to-rosetta/rosetta_bin_linux_2015.19.57819_bundle/main/d
atabase @flag1.txt -s INPUT.pdb -ex1 -ex2 -ex1aro -ex2aro -linmem_ig 10 -packing:extrach
i_cutoff 0 -nstruct 1
```

OPTIONS: @flag1.txt

```
-options
 -user
-ignore_unrecognized_res
-add_orbitals
-relax:dualspace
-parser
 -protocol design1.xml
```

PARAMETERS: design1.xml

```
<ROSETTASCRIPTS>
       <SCOREFXNS>
             <s weights=talaris2013_cart />
       </SCOREFXNS>
       <TASKOPERATIONS>
             <InitializeFromCommandline name=ifcl/>
             <ReadResfile name="rrf" filename="design.resfile" />
       </TASKOPERATIONS>
       <MOVERS>
             <SavePoseMover name=init_struct reference_name=init_struct/>
             <FastRelax name=fast_relax scorefxn=s task_operations=ifcl/>
             <PackRotamersMover name=design scorefxn=s task_operations=ifcl,rrf />
      </MOVERS>
      <PROTOCOLS>
      <Add mover_name=design/>
      </PROTOCOLS>
</ROSETTASCRIPTS>
```

RESFILE OctaVII_06: design.resfile

```
NATAA
START

92 A PIKAA S
98 A PIKAA S
```

RESFILE OctaVII_07: design.resfile

```
NATAA
START

89 A PIKAA S
92 A PIKAA S
178 A PIKAA S
209 A PIKAA S
217 A PIKAA S
219 A PIKAA S
```

RESFILE OctaVII_08: design.resfile

```
NATAA
START

42 A PIKAA S
59 A PIKAA S
97 A PIKAA S
119 A PIKAA S
125 A PIKAA S
191 A PIKAA S
```

RESFILE OctaVII_09: design.resfile

```
NATAA
START

32 A PIKAA S
93 A PIKAA S
98 A PIKAA S
128 A PIKAA S
157 A PIKAA S
```

RESFILE OctaVII_10: design.resfile

```
NATAA
START

29 A PIKAA S
70 A PIKAA S
90 A PIKAA S
103 A PIKAA S
153 A PIKAA S
209 A PIKAA S
```

6.6 HMM physical-chemical features

List of the 17 features selected for the training of the HMM software. The first 16 are obtained from Aaindex (www.genome.jp/aaindex), while the last one is obtained from DynaMine analysis.

1. Hydrophobicity coefficient in RP-HPLC
2. Amino acid abundance
3. Normalized positional residue frequency at helix termini C
4. Unfolding Gibbs energy in water
5. Information measure for extended without H-bond
6. Information measure for N-terminal turn
7. Zimm-Bragg parameter sigma x 1.0E4
8. Normalized relative frequency of coil
9. Normalized positional residue frequency at helix termini
10. Average relative fractional occurrence in AL(i)
11. Relative preference value at C
12. STERIMOL length of the side chain
13. Alpha-helix indices for beta-proteins
14. Weights for coil at the window position of 6
15. Relative preference value at C1
16. Electron-ion interaction potential
17. DynaMine predicted backbone dynamics

OctarellinVII

A new generation of *de novo* designed $(\beta/\alpha)_8$ - barrel proteins

De novo protein design is a growing field in protein chemistry, aiming at the production of artificial proteins. On a purely fundamental basis, the design of proteins from scratch allows testing the accuracy of the current protein knowledge and, possibly, to improve it. A deep knowledge of the sequence, structure, function relationships in proteins is necessary to design new proteins with specific functions. This facet of *de novo* protein design has numerous applications in biotechnology and biomedicine. On the other hand, in the context of a post-genomic era, advanced computational methods for protein analysis, modelling and design are needed to decode the massive amount of genomic data.

There is a long tradition at the University of Liège in the design of artificial $(\beta/\alpha)_8$-barrel proteins, called Octarellins. This fold, also known as TIM-barrel, is widespread in nature, particularly in enzymes, and represents an interesting target for therapeutic or biological applications. Several generations of Octarellins were designed with the help of very different approaches. Lessons from these previous works has served as a rational basis for this study, which consists in the design of a new generation of artificial TIM-barrels, termed OctaVII. This thesis is divided in four sections that are shortly described hereafter:

The first section describes a pool of natural TIM-barrels, which structural features were analyzed in order to extract useful information for the following steps of design and validation.

The second section is dedicated to the design of OctaVII models. Backbone structures were designed with the use of the modelling software Rosetta and Modeller. This led to the selection of 28 backbones structures, which were used for the design of sequences, using Rosetta. Finally, more than 8000 artificial sequences were designed.

The third section includes the *in silico* validation of the design. Information obtained from natural TIM-barrels was used to screen the 8000 artificial sequences and to select 10 of them for experimental characterization. Various structural features were tested, including hydrogen bond content and amino acid composition, and both secondary structure predictions and molecular dynamic simulations were performed.

The fourth section is dedicated to the experimental validation of the design through protein expression, purification and biophysical characterization. In addition to the ten original sequences that were designed in this work, five additional variants were tested for their possibly improved properties (collaborations at the Vrije Universiteit Brussel, Belgium, and at the Vanderbilt University, USA).

This thesis contributes to the development of the *de novo* design of proteins as an emerging methodology for both a better understanding of proteins and the design of new functional proteins with applications in biomedicine and nanotechnology.