Communauté Française de Belgique
Académie Universitaire Wallonie-Europe
Faculté Universitaire des Sciences Agronomiques de Gembloux

# DÉVELOPPEMENT D'UNE MÉTHODE DE PRÉDICTION DES SITES D'INTERACTION PROTÉINES-MOLÉCULES À PARTIR DE LA STRUCTURE PRIMAIRE

Promoteurs :
**Professeur R. Brasseur**
**Docteur B. Charloteaux**

Dissertation originale présentée par :
**N. Delsaux**

en vue de l'obtention du garde de docteur
en sciences agronomiques et ingénierie biologique

2008

**Nicolas Delsaux (2008).** Développement d'une méthode de prédiction des sites d'interaction protéines-molécules à partir de la structure primaire (thèse de doctorat). Gembloux, Belgique, Faculté Universitaires des Sciences Agronomiques, 204p., 20 tabl., 81 fig.

**Résumé :** Ce travail a pour but d'améliorer nos connaissances des interfaces et d'aider les scientifiques à caractériser au mieux les protéines de fonction et de structure encore inconnues. Pour cela, nous avons construit des banques de données de structures tridimensionnelles de complexes et leurs interfaces ont été analysées au niveau atomique. L'analyse détaillée des interfaces a permis de confirmer le rôle important des acides aminés aromatiques et de l'arginine ainsi que de montrer quels couples de résidus sont significativement favorisés dans celles-ci. De plus, l'importance du volume des résidus voisins des sites d'interaction et de la conformation des acides nucléiques a pu être montrée. Les principales variables corrélées aux interfaces sont : trois propensions à être en interaction, le type de résidu et sa position dans la séquence, les prédictions d'accessibilité et de structures secondaires, la prédiction en 'Receptor Binding Domain', et la présence de certains motifs protéiques. Finalement, une méthode de prédiction des sites d'interaction été mise au point. Cette méthode est l'une des seules à n'utiliser que des informations directement accessible à partir de la séquence et donne des résultats très encourageants. La spécificité obtenue est en effet suffisante pour améliorer les résultats expérimentaux obtenus par mutagénèse dirigée.

**Nicolas Delsaux (2008).** Development of a prediction method of protein-molecule interaction sites from primary structure (doctoral thesis, in French). Gembloux, Belgium, Agricultural University, 204p., 20 tabl., 81 fig.

**Summary:** The objective of this work is to improve the knowledge about interfaces and to assist the characterization of proteins with unknown function and structure. To this aim, we constructed databases of three-dimensional structures of complexes and their interfaces were analyzed at the atomic scale. The in-depth analysis of interfaces confirms the preponderance of aromatic amino acids and of arginine. Residue pairs that are significantly favored at the interfaces were also identified. Moreover, the importance of interaction sites neighboring residue volumes and of the nucleic acid's conformation has been highlighted. The main parameters correlated to interaction sites are: three interaction propensities, residue's type and its location within the sequence, predictions of accessibility and of secondary structures, prediction to be a Receptor Binding Domain, and the presence of protein patterns. Finally, a prediction method of interaction sites has been developed. This method is one of the few which only use information directly accessible from protein sequence and gives promising results. The achieved specificity is indeed sufficient to improve experimental results of site-directed mutagenesis.

# Remerciements

A l'occasion de la présentation de ce travail, il m'est agréable de pouvoir remercier toutes les personnes qui ont contribué de près ou de loin à l'élaboration de cette thèse *:*

- Je tiens à remercier le professeur Robert Brasseur, PhD, Directeur de Recherche au FNRS et Directeur du Centre de Biophysique Moléculaire Numérique (CBMN), pour son soutien, ses conseils avisés et, évidemment, pour m'avoir permis d'effectuer cette thèse de doctorat au sein du CBMN.

- Je tiens également à remercier Benoit Charloteaux, PhD, et Annick Thomas, PhD, pour les nombreuses discussions et le travail accompli ensemble tout au long de ces quatre années.

- Il m'est aussi agréable de remercier May Morris, PhD, du Centre de Recherche de Biochimie Macromoléculaire de Montpellier ainsi que les professeurs Daniel Portetelle, Richard Kettmann et Bernard Wathelet de la Faculté Universitaire des Sciences Agronomiques de Gembloux pour avoir accepté de faire partie de mon jury.

- Merci à Marie-Hélène Van Eyck pour son soutien lors de cette thèse et les corrections apportées au manuscrit mais aussi pour m'avoir permis d'acquérir une expérience professionnelle au sein de Biosiris.

- Merci aussi à tous les autres membres du CBMN, les doctorants, les post-doctorants et les mémorants, les informaticiens, le personnel scientifique, les stagiaires et autres personnes de passage pour les 'pause time', les temps de midi à la sauce CS et les nombreuses discussions.

- Un tout grand merci à mes parents et à ma famille, merci à tous mes amis pour leur soutien et les moments de détente apportés lors de cette thèse et surtout en dehors du cadre professionnel. Et finalement, merci à Emeline que j'embrasse de tout mon cœur et à Simon qui a apporté un peu de piment supplémentaire à cette fin de thèse.

# Liste des abréviations

3D : tridimensionnel
Å : Angstrœm
ADN : Acide DésoxyriboNucléique
ARN : Acide RiboNucléique
ARNr : ARN ribosomique
ARNt : ARN de transfert
ARNm : ARN messager
ARNsn : ARN small nuclear
ASA : Accessible Surface Area
BK : Backbone
BRET : Bioluminescence Resonance Energy Transfert
CBMN : Centre de Biophysique Moléculaire Numérique
CDR : Complementarity Determining Region
CRF : Conditional Random Field
dsRBD : double stranded RNA Binding Domain
FN : False Negative
FP : False Positive
FRET : Fluorescence Resonance Energy Transfert
HLH : Helix-Loop-Helix
HTH : Helix-Turn-Helix
lien H : lien Hydrogène
LR : Linear Regression
MCC : Matthews Correlation Coefficient
NB : Naive Bayes
NN : Neural Network
P : Propension
PDB : Protein Data Bank
RBD Receptor Binding Domain
RRM : RNA Recognition Motif
RMN : Résonnance Magnétique Nucléaire
Rx : Rayons x
SC : Side Chain
SVM : Support Vector Machine
TN : True Negative
TP : True Positive

A : adénine
G : guanine
C : cytosine
T : thymine
U : uracile

A : Ala : alanine
C : Cys : cystéine
D : Asp : acide aspartique
E : Glu : acide glutamique
F : Phe : phénylalanine
G : Gly : glycine
H : His : histidine
I : Ile : isoleucine
K : Lys : lysine
L : Leu : leucine
M : Met : méthionine
N : Asn : asparagine
P : Pro : proline
Q : Gln : glutamine
R : Arg : arginine
S : Ser : sérine
T : Thr : thréonine
V : Val : valine
W : Trp : tryptophane
Y : Tyr : tyrosine

# Tables des matières

# I. INTRODUCTION

## *I.1. Avant-propos*

Les protéines sont des macromolécules impliquées dans un grand nombre des activités de la cellule et les types de protéines présents dans les systèmes biologiques sont donc nombreux : enzymes, protéines de structure, hormones, facteurs de croissance, activateurs de gènes, récepteurs et transporteurs membranaires, anticorps, toxines…[1] Il n'est dès lors pas surprenant de constater que les premières recherches biochimiques et les recherches actuelles soient largement tournées vers ces molécules. Les objectifs de ces études sont variés et vont de la compréhension des phénomènes fondamentaux du repliement protéique ('folding') à la création de nouveaux médicaments en passant par une étape primordiale de compréhension des cascades d'interactions cellulaires.

Une classification grossière des protéines permet de les séparer en trois grandes catégories : les protéines fibreuses, les protéines membranaires et les protéines solubles. Les protéines fibreuses sont principalement impliquées comme matériaux de construction extracellulaires (collagène, élastine, kératine…) et possèdent une structure allongée résistante aux forces de traction et de cisaillement. Les protéines membranaires se situent à l'interface entre l'intérieur de la cellule et le milieu extérieur et composent en moyenne 50% de la masse des membranes naturelles.[2] Ces protéines agissent comme transporteurs, catalyseurs, transducteurs de signal ou soutiennent la structure membranaire. Finalement, les protéines solubles : celles-ci ont une structure compacte qui leur a conféré leur qualificatif de globulaires. Certaines de ces protéines peuvent occuper jusqu'à 35% du volume d'une solution saturée et possèdent de très nombreuses fonctions nécessaires à la vie intracellulaire.

Ces dernières années, l'étude des interactions entre protéines a pris une importance de plus en plus grande notamment grâce à l'accumulation de structures protéiques mais aussi grâce au séquençage de génomes et à la construction d'interactomes d'organismes de référence. Le Centre de Biophysique Moléculaire Numérique (CBMN) est un centre de recherche offrant de nombreux outils bioinformatiques puissants et, dans ce travail, ils nous seront d'une grande utilité pour poser un pas de plus vers la compréhension et la prédiction des sites d'interaction.

## *I.2. Rappel sur les Protéines*

### I.2.1. Structure primaire

Les protéines sont des polymères qui peuvent incorporer 20 monomères différents. Ces monomères sont les 20 acides aminés naturels (des α-amino acides) dont la structure générale est présentée ci-dessous (Figure I-1). La proline est un α-imino acide dont la chaîne latérale est liée à l'atome d'azote. Excepté le cas de la glycine dont la chaîne latérale est constituée d'un seul hydrogène, chaque acide aminé contient un carbone asymétrique, le carbone alpha ($C_\alpha$) qui est presque toujours sous la forme isomèrique L. On retrouve la forme D dans des toxines, des surfactants excrétés par des bactéries, des constituants des parois bactériennes…

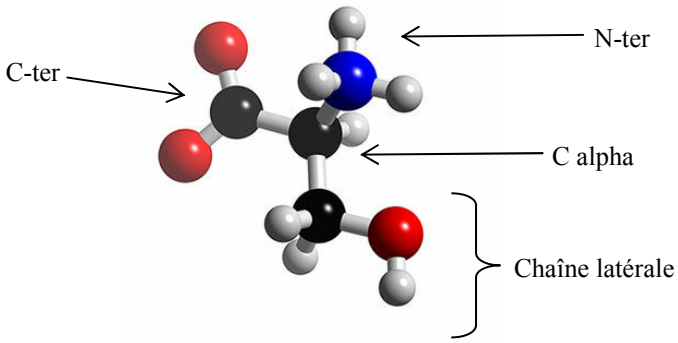



**Figure I-1 : Modèle tridimensionnel de la sérine. Les atomes d'hydrogène sont représentés en blanc, l'azote en bleu, l'oxygène en rouge et les carbones en noir.**

Les protéines naturelles sont linéaires et ne peuvent être différenciées que selon la séquence des acides aminés qui les constituent et la longueur de celle-ci. A l'heure actuelle, on admet que la séquence détermine la conformation de la protéine et donc, en principe, ses fonctions biologiques (paradigme structure-fonction). Pour former ce polymère linéaire, les acides aminés sont liés entre eux par condensation et forment un lien peptidique entre un groupe α-carboxyle et un groupe α-amine (Figure I-2). Dans la majorité des cas, cette liaison est en position *trans* ce qui permet de limiter les encombrements stériques en éloignant les chaînes latérales les unes des autres. Le lien peptidique est polaire et est considéré comme plan et rigide (plan amide). Cette planarité est due au caractère partiellement (40%) doublement lié de la liaison C-N mais l'angle de la liaison n'est pas totalement fixe et peut varier faiblement (dans presque toutes les protéines) ou plus fortement (protéines de petite taille et peptides cycliques).[3]

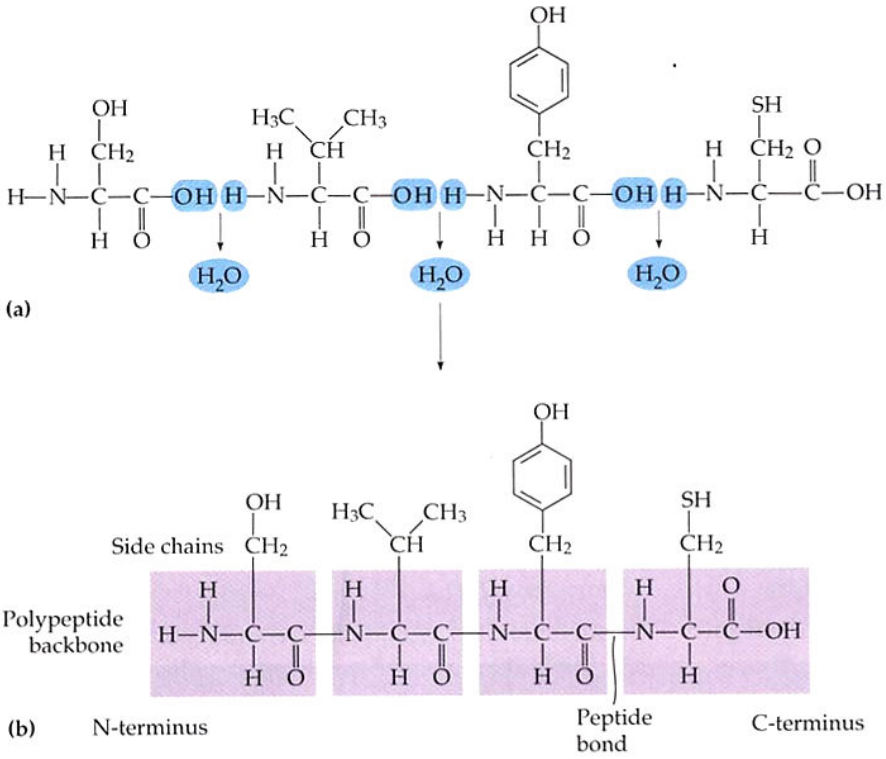



**Figure I-2 : Formation d'un lien peptidique entre quatre acides aminés.**

## I.2.2. Structure secondaire

### Configuration du squelette peptidique

Le plan presque rigide de la liaison polypeptidique limite fortement la rotation des atomes des fonctions acides et amines. Les seules liaisons dont l'orientation reste libre sont celles qui entourent chacun des carbones asymétriques. La chaîne polypeptidique possède donc deux degrés de liberté en rotation (voir Figure I-3) :

- La liberté de rotation autour de la liaison $C_\alpha$ – N-H caractérisée par l'angle $\Phi$.
- La liberté de rotation autour de la liaison $C_\alpha$ – C=O caractérisée par l'angle $\psi$.

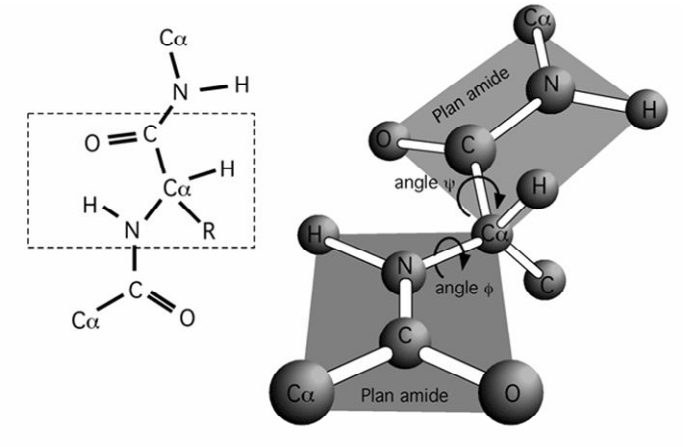



**Figure I-3 : Représentation du plan de la liaison peptidique et des angles de rotation Φ et ψ. Le symbole R représente la chaîne latérale de l'acide aminé.**

Ces angles sont les deux principaux paramètres permettant de définir les différents types de structures secondaires. Les structures secondaires retrouvées habituellement dans les protéines sont présentées dans les points suivants.

### Structures en hélice

L'hélice est l'une des deux grandes catégories de structures secondaires rencontrées dans les protéines. Les hélices sont caractérisées par le nombre d'acides aminés par tour d'hélice (n), par le pas de l'hélice c'est-à-dire la translation (en Å) par tour d'hélice (p), par la présence de liens H et par le couple Φ/ψ. La Figure I-4 représente le type d'hélice le plus couramment rencontré qu'est l'hélice α de pas droit (n = 3,6 résidus ; p = 5,41 Å ; lien H entre l'oxygène du résidu n et l'azote du résidu n + 4 ; Φ = -57° et ψ = -47°).

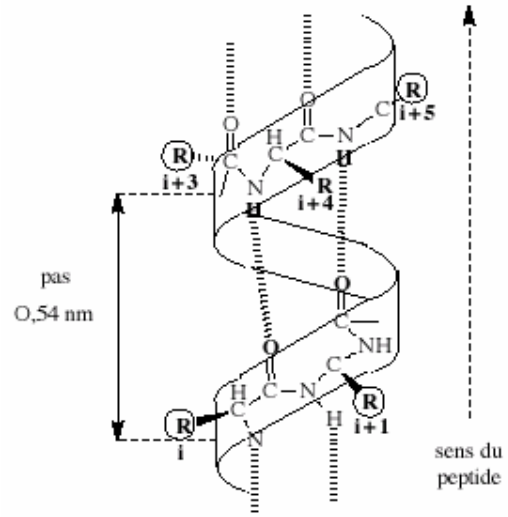



**Figure I-4 : Représentation d'une hélice α de pas droit.**

D'autres types d'hélices existent naturellement dans les protéines bien qu'elles soient moins fréquentes : les hélices $3^{10}$, les hélices π et les hélices ω. Ces trois types d'hélices diffèrent principalement par l'écart entre les résidus liés par pont hydrogène. Les hélices $3^{10}$ représentent une forme plus étirée de l'hélice α tandis que les hélices π et ω sont des formes plus compressées.

**Structures β**

Les feuillets plissés β sont, après les hélices α, les structures les plus couramment rencontrées dans la nature. Les feuillets sont constitués de brins β qui sont reliés entre eux par des liens H entre le groupement C=O d'un premier brin et le groupement N-H d'un brin adjacent. Si les brins β sont orientés dans le même sens, on parle de feuillets β parallèles sinon, on parle de feuillets β antiparallèles (Figure I-5).

Remarque : les feuillets β sont des structures secondaires particulières dans le sens où elles comprennent plus d'un segment de séquence. En effet, les différents brins β d'un feuillet sont séparés par d'autres structures (turns en général).

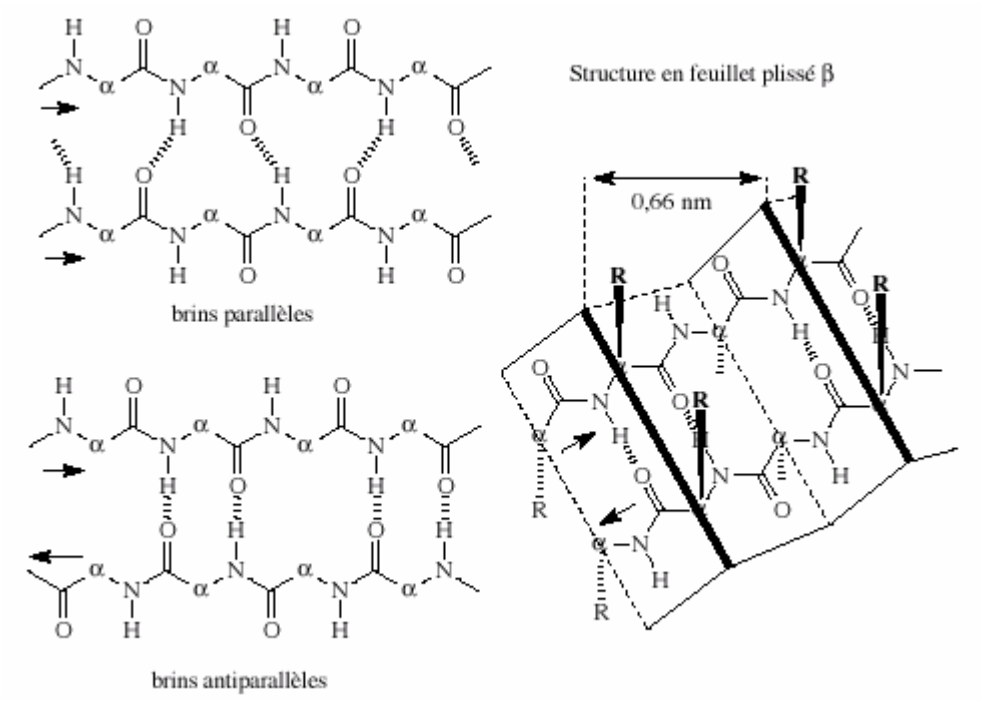



**Figure I-5 : Brins β parallèles et brins β antiparallèles.**

**Turns et structures non-régulières (random coils)**

Turns et random coils sont des motifs particuliers différents des deux principales structures secondaires que sont les hélices α et les structures β.

Dans les random coils, les boucles correspondent à des structures non-répétitives peu ordonnées, qui comprennent de 4 à 20 résidus, qui relient des hélices α entre elles ou avec des feuillets β et qui sont généralement exposées au solvant en présentant leurs chaînes latérales chargées ou polaires. Les structures des 'bras terminaux' (résidus C ou N – terminaux) et les hélices gauches font également partie des random coils.

Les turns sont de petites structures secondaires d'au plus 6 résidus qui permettent, le plus souvent, de relier deux brins β d'un feuillet en conduisant à une rotation de 180° de la chaîne polypeptidique (Figure I-6). Différents types de turns existent : les α à 5 résidus, les β à 4 résidus, les γ à 3 résidus, les δ à 2 résidus et les π à 6 résidus).[4]

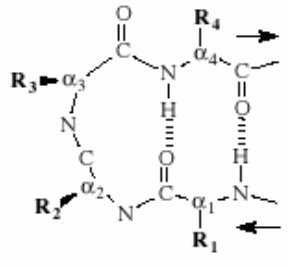



**Figure I-6 *:* Turn β à 4 résidus.**

**Structures polyproline**

Les prolines sont des acides aminés qui ne conviennent pas pour des structures secondaires types (hélices α et structures β). Dès lors, lorsque plusieurs prolines se retrouvent l'une à la suite de l'autre dans la séquence, elles vont former une structure particulière : l'hélice de polyproline représentée à la Figure I-7. Les prolines de l'hélice peuvent se trouver en forme *trans* ou *cis* et passer d'une forme à l'autre selon les conditions du milieu.[2]

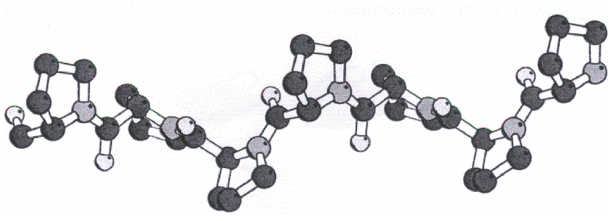


**Figure I-7 : Hélice de polyproline.**

### I.2.3. Structure tertiaire

La structure tertiaire est formée par l'agencement des structures secondaires entre elles pour aboutir à la formation de domaines. Dans un domaine, les structures secondaires vont se replier sur elles-mêmes pour former un ensemble dont la géométrie spatiale est presque stable et autonome. Cela va permettre à des acides aminés qui étaient séquentiellement éloignés de se retrouver les uns à côté des autres et leur donner la possibilité d'effectuer des interactions diverses (liens H, ponts disulfures, interactions hydrophobes…).

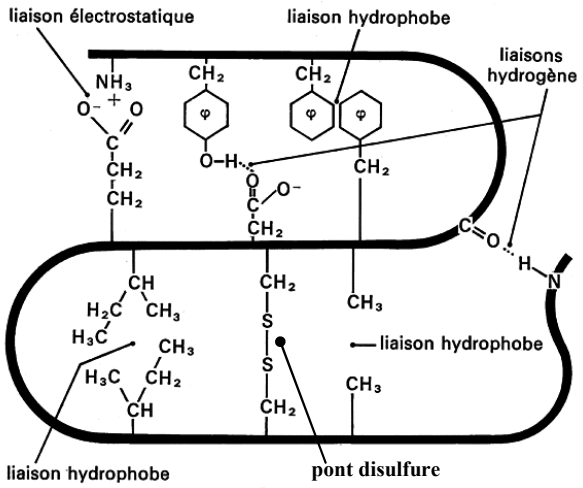



**Figure I-8 : Interactions entre chaînes latérales impliquées dans la structure tertiaire des protéines.**

### I.2.4. Structure quaternaire

La structure quaternaire résulte de l'agencement de plusieurs domaines entre eux et/ou de plusieurs protéines entre elles. La structure quaternaire correspond donc aux interactions entre protéines et ce type d'association sera décrit en détail au paragraphe I.4.3.

## *I.3. Rappel sur les Nucléotides*

### I.3.1. Structure primaire

Les acides nucléiques sont composés de 5 bases différentes appartenant à deux classes : A (adénine) et G (guanine) font partie de la famille des purines alors que C (cytosine), T (thymine) et U (uracile) font partie de la famille des pyrimidines.

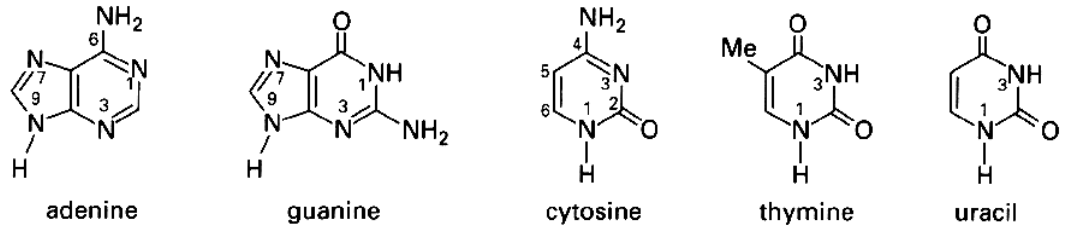



**Figure I-9 : Structure des 5 bases nucléiques.**[5]

Ces bases sont liées par un atome d'azote (l'azote 9 pour les purines et l'azote 1 pour les pyrimidines) au carbone 1 d'un sucre (ribose ou désoxyribose) pour former un nucléoside (Figure I-10). Les différents nucléosides sont, pour l'ADN, la désoxyadénosine, la désoxyguanosine, la désoxythymidine et la désoxycytidine, et pour l'ARN, l'adénosine, la guanosine, l'uridine et la cytidine.

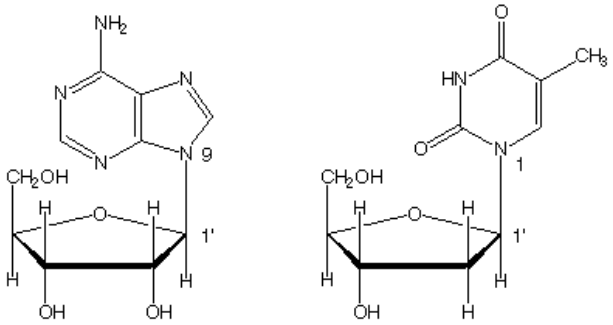


**Figure I-10 : Structure de l'adénosine (à gauche) et de la désoxythymidine (à droite).**

Enfin, un phosphate vient se fixer au sucre par estérification au carbone 5 du sucre comme représenté sur la Figure I-11. Les nucléotides sont nommés par le nom du nucléoside correspondant suivi du mot monophosphate.

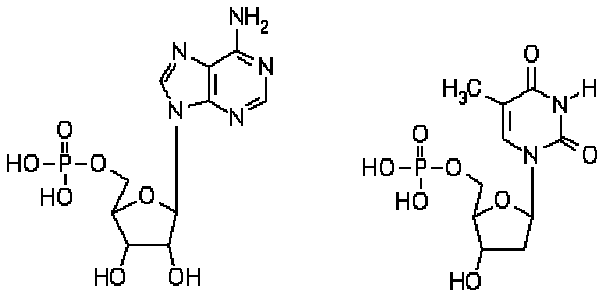


**Figure I-11 : Exemples de structures de nucléotides (adénosine monophosphate, à gauche et désoxythymidine monophosphate, à droite).**

Ces nucléotides peuvent polymériser par formation de liaisons esters entre le phosphate d'un nucléotide et le carbone 3 du sucre du nucléotide suivant. Ils forment ainsi un acide nucléique (cf. Figure I-12) : acide désoxyribonucléique (ADN) ou acide ribonucléique (ARN).

Finalement, deux brins complémentaires peuvent s'apparier par liaisons hydrogènes pour former l'hélice bicaténaire (Figure I-12). Les bases sont tournées vers l'intérieur de l'hélice afin de pouvoir se lier par liaisons hydrogènes avec la base complémentaire, le squelette sucre-phosphate étant dirigé vers l'extérieur. Ce squelette sucre-phosphate est également appelé le 'backbone' de l'ADN.

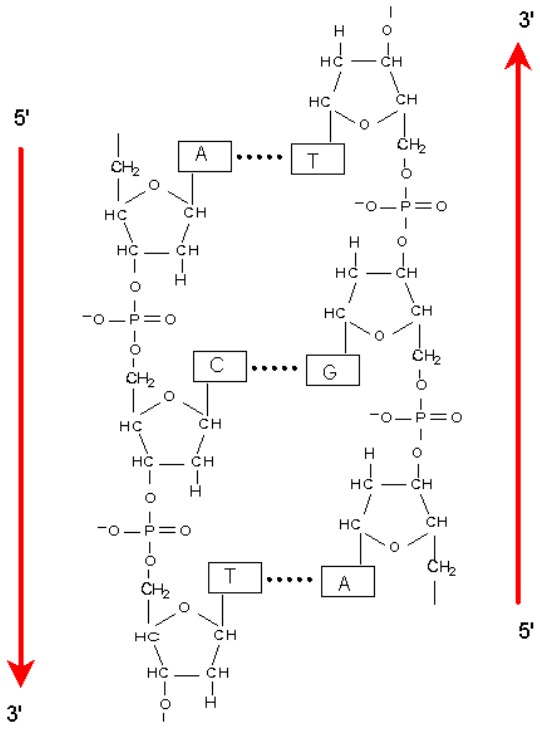



**Figure I-12 : Appariement des bases et formation d'une double hélice d'ADN.**

## I.3.2. Structure secondaire

### Acide désoxyribonucléique (ADN)

Il y a un peu plus de 50 ans, Watson & Crick[6] mettaient en évidence la structure de l'ADN. Cette structure est de type bicaténaire hélicoïdale, c'est-à-dire que l'ADN est formé de 2 brins d'acides nucléiques complémentaires et anti-parallèles. Ces deux brins sont maintenus en contact par des liaisons hydrogènes et des interactions hydrophobes. L'appariement de bases nucléotidiques appartenant à des brins opposés se fait de manière très stricte et précise grâce à des liens H. L'adénine (A) s'apparie toujours avec la thymine (T) (ou l'uracile « U » dans le cas de l'ARN) par deux liens H et la guanine (G) s'apparie toujours avec la cytosine (C) par trois liens H. Cette règle d'appariement n'est cependant pas toujours respectée et certains appariements se font en dehors de cette règle (Figure I-13).

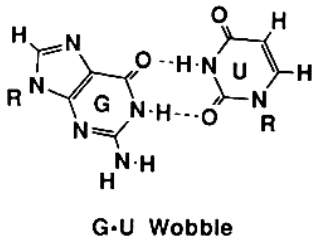



**Figure I-13 : Exemple d'appariement non-conventionnel ; le wobble entre une guanine et une uracile.**

Les interactions hydrophobes et de van der Waals contribuent de façon significative à la stabilité globale de l'hélice. En principe, les brins peuvent former une hélice soit de pas droit, soit de pas gauche. Cependant, la structure du squelette sucre-phosphate est plus compatible avec la première configuration. C'est donc la configuration en pas droit que l'on retrouvera préférentiellement dans la nature. Il existe trois types principaux de double hélice d'ADN : les hélices de type B, de type A et de type Z (Figure I-14). Le type d'hélice adopté par deux brins d'ADN dépend principalement de la séquence nucléotidique mais aussi de la concentration saline du milieu.

La configuration la plus courante est celle de type B dans laquelle un tour d'hélice compte environ 10 paires de bases, les paires de bases étant perpendiculaires au plan de l'axe de l'hélice. L'hélice forme un tour tous les 3,4Å. La forme A est relativement fréquente mais beaucoup moins que la B. Cette forme A compte 11 paires de bases par tour d'hélice, chaque paire de base est inclinée selon un angle de 19° par rapport au plan perpendiculaire à l'axe de l'hélice. Cette forme, moins étirée que la forme B, effectue un tour tous les 2,6Å. Il existe une troisième forme beaucoup plus rare appelée Z pour 'ZigZag'. Contrairement aux deux

premières formes, celle-ci est de pas gauche. Elle comporte 12 paires de bases par tour et avance de 3,7Å à chaque tour. Cette forme est très majoritairement composée d'une alternance de C et de G.[7] Finalement, l'ADN peut former des triples hélices droites avec deux brins complémentaires antiparallèles et un brin parallèle au premier qui s'insère dans le petit sillon d'une hélice B étirée.

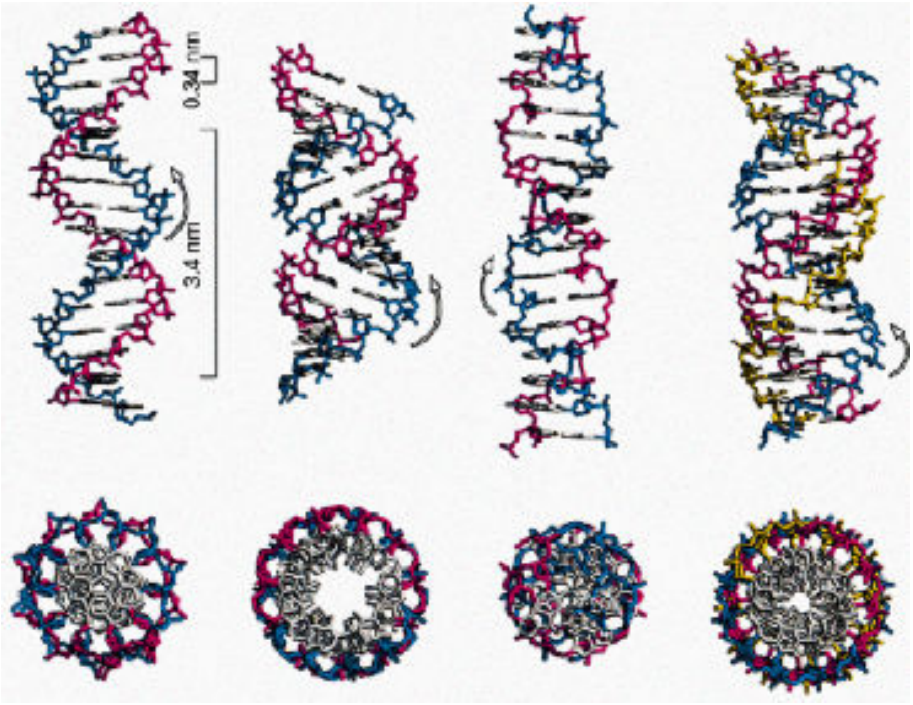



**Figure I-14 : Les trois types principaux de double hélice d'ADN. De gauche à droite, le type B, A et Z. La quatrième structure représente une triple hélice d'ADN, la troisième hélice est représentée en jaune.**

Pour l'ADN de type B, les brins torsadés forment entre eux deux sillons hélicoïdaux de largeur différente : le grand et le petit sillon. Par l'intermédiaire de ceux-ci, une partie de chaque base est accessible de l'extérieur par d'autres types de molécules. Dans la forme A, le petit sillon est écrasé et par conséquent inaccessible. La forme Z quant à elle, possède deux sillons de tailles équivalentes.

La définition de ces trois types d'hélices permet d'avoir une idée globale de la structure de l'ADN. Pour une analyse plus détaillée de la conformation des nucléotides, il existe un grand nombre de paramètres définis sur base d'un couple de bases complémentaires ou sur base de deux couples successifs ('two-base pair step'). Ces paramètres dépendent principalement de l'angle et de la distance par rapport au plan de la paire de bases et par rapport à l'axe de l'hélice. On définit de cette manière des mouvements de rotation et de translation des couples de paires de bases[8,9] (voir aussi le site Internet de la IUPAC -

International Union of Pure and Applied Chemistry : http://www.imb-jena.de/~csc/NANA.html).

Divers programmes permettant de déterminer le type de conformation de l'ADN et un calcul précis de ces différents paramètres ont été créés. Ce sont par exemple les programmes : ADAPT[10] et 3DNA.[11]

Remarque : L'ADN peut dans certains cas se retrouver sous une configuration en double hélice circulaire (virus, plasmide, bactérie).

**Acide ribonucléique (ARN)**

On distingue plusieurs types d'ARN au sein d'une cellule : ARNr (ribosomique), ARNt (de transfert), ARNm (messager), ARNsn ('small nuclear'), miRNA (micro ARN) et siRNA ('small interfering').

L'ARN se retrouve principalement sous forme de simple brin mais ces simples brins peuvent, sous l'influence des forces d'empilement, prendre la forme d'une simple hélice droite irrégulière. Des doubles hélices peuvent exister entre deux brins d'ARN, dans des hybrides ADN-ARN et par association de deux segments distants d'un simple brin d'ARN. Les doubles hélices d'ARN suivent les mêmes règles d'appariement entre bases nucléiques que les doubles hélices d'ADN (de type Watson et Crick). Mais, d'une manière plus prononcée que pour l'ADN, contiennent régulièrement des appariements non-conventionnels.[12]

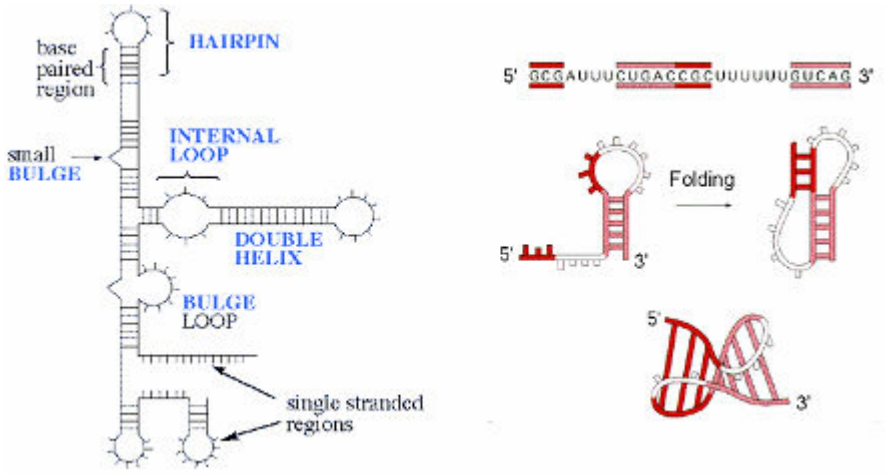



**Figure I-15 : Structure secondaire de l'ARN.**

L'ARN se retrouve classiquement sous forme de tiges et de boucles ou « épingle à cheveux » ('hairpin'), de boucles internes ('internal loop') et de renflements ('bulge') en plus

des zones en simple brin et en double hélice (Figure I-15 à gauche). Finalement, les pseudo-nœuds (‘pseudo-knots’) qui correspondent plutôt à une structure tertiaire, sont aussi régulièrement retrouvés (Figure I-15 à droite). Les ARN de transfert quant à eux possèdent tous une structure particulière bien déterminée (Figure I-16).

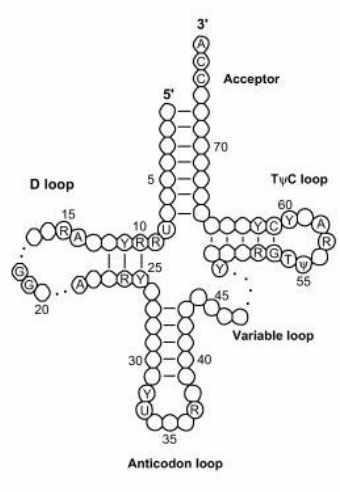



**Figure I-16 : Structure d’un ARN de transfert.**

### I.3.3. Interactions non-covalentes

L'interaction la plus importante, énergétiquement et structurellement, entre deux atomes (ou deux molécules) est la répulsion qui peut éventuellement exister lorsqu'ils s'approchent l'un de l'autre à très courte distance. On parle alors de répulsion stérique. Si on considère chaque atome comme une sphère, le volume 'impénétrable' dépend du rayon de van der Waals de cet atome (cf. tableau en Annexe 1).

Tous les atomes (même les atomes non-chargés) s'attirent les uns les autres sous l'effet d'interactions mutuelles dues à des phénomènes de polarisation induite. Ces attractions omniprésentes sont faibles, agissent sur de courtes distances (3-4 Å) et sont connues comme interactions de van der Waals (voir Figure I-17). Ce type d'interaction existe entre deux dipôles permanents, un dipôle permanent et un dipôle induit ou entre deux dipôles induits.

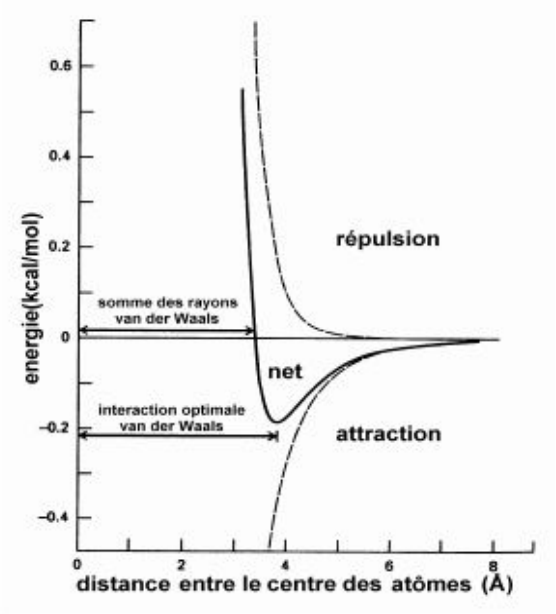



**Figure I-17 : Evolution de la force d'interaction de van der Waals en fonction de la distance.**

Les forces d'interaction non-covalentes les plus fondamentales sont les forces électrostatiques. L'attraction entre des atomes de charges opposées est décrite par la loi de Coulomb prenant en compte la charge de ces atomes, la distance les séparant et la constante diélectrique du milieu (ε). Ce type d'interaction est efficace sur des distances relativement longues. La distance optimale d'interaction est de 2,8 Å et dépend de l'état de charge des

acides aminés ioniques et donc du pH du milieu. Les interactions entre des groupements de charge opposée sont connues, dans les protéines, sous le nom de ponts salins.

Loi de Coulomb :

$$F = \frac{1}{4\pi\varepsilon} \times \frac{q_1 x q_2}{d^2}$$

Avec $q_i$ = charge de l'atome i

d = distance entre les atomes 1 et 2

ε = constante diélectrique

Finalement, les liens hydrogène (ou liens H) apparaissent quand deux atomes électronégatifs entrent en compétition pour le même atome d'hydrogène. L'hydrogène est lié de façon covalente à un des deux atomes (le donneur) mais interagit favorablement avec l'autre (l'accepteur). Il existe une interaction électrostatique entre les charges partielles négatives des atomes électronégatifs et la charge partielle positive de l'hydrogène. Dans les systèmes biologiques, les donneurs sont préférentiellement des groupements N-H et O-H (provenant de l'eau par exemple ; voir Figure I-18) et, moins fréquemment, des groupements S-H et C-H.[13,14] Les accepteurs les plus courants sont : O=, −O−, −N=, et, moins fréquemment, $-S^-$, −S− et les électrons π des cycles aromatiques. Cela implique que la presque totalité des groupements d'une protéine peuvent être impliqués dans un lien H.

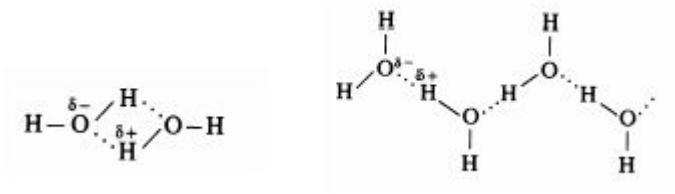


**Figure I-18 : Exemples de liens H impliquant des molécules d'eau.**

### I.3.4. Effet hydrophobe

Les molécules non-polaires ou hydrophobes, sont de très mauvaises partenaires en vue d'une interaction polaire avec le solvant. Cette absence relative d'interaction entre des molécules apolaires et l'eau provoque des interactions entre les groupes apolaires eux-mêmes. La tendance qu'ont les résidus apolaires à se regrouper entre eux afin d'exclure les molécules d'eau est connue sous le nom d'effet hydrophobe et est le facteur principal stabilisant les protéines. Ces interactions diffèrent des liaisons classiques et apparaissent essentiellement comme d'origine entropique.

L'effet hydrophobe est rarement pris en compte ou décrit dans la littérature[15,16] et, en général, quand on parle d'interaction hydrophobe, on parle d'interaction de London (1930). Les interactions de London représentent en fait des contacts hydrophobes dus à des interactions de van der Waals de type dipôle induit-dipôle induit.

## *I.4. Importance des protéines et de leurs Interactions*

### I.4.1. Pourquoi porter autant d'intérêt aux interactions avec les protéines ?

L'ambitieux objectif de l'étude des protéines est d'élucider la structure, les interactions et les fonctions de toutes les protéines d'une cellule et/ou d'un organisme entier.[17] Les objectifs sont multiples : meilleure connaissance des réseaux et processus cellulaires, meilleure compréhension des maladies, production de nouveaux médicaments…

Au cours des dix dernières années, plus de 550 génomes bactériens (voir la revue de Binnewies *et al.*)[18] et plus de 70 génomes eucaryotiques dont ceux de *Saccharomyces cerevisiae,*[19-21] *Caenorhabditis elegans*[22] et *Homo sapiens*[23-25] ont été entièrement séquencés (Figure I-19). Il est clair que, même si les gènes compris dans ces génomes ne sont pas clairement définis, nous avons maintenant accès à des séquences codant pour des dizaines de milliers de protéines. Dès lors, une question fondamentale se pose : quelles sont les fonctions biologiques de ces protéines ?

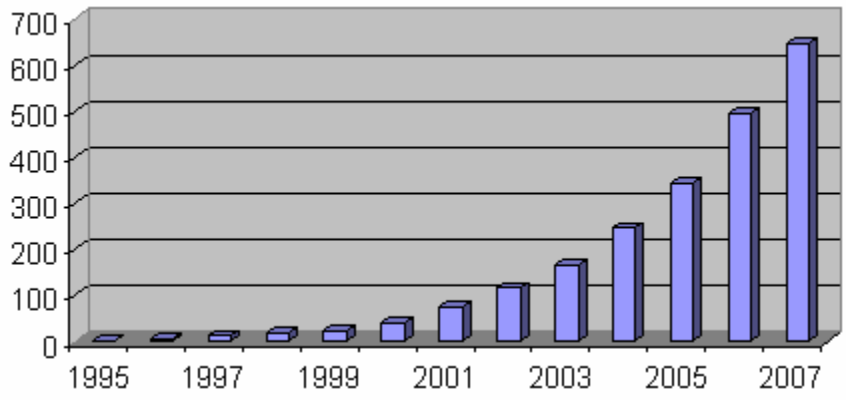



**Figure I-19 : Evolution du nombre de génomes complètement séquencés (données de septembre 2007 issues de la 'GOLD'[26]).**

Pour exprimer leur fonction biologique, une grande partie des protéines a besoin d'interagir avec un partenaire car peu de protéines ont une fonction sous forme de monomère. Les sites d'interaction jouent donc des rôles essentiels en biologie. Ceux-ci concernent notamment la régulation des réseaux métaboliques, la reconnaissance immune, la réplication de l'ADN, la progression à travers le cycle cellulaire, la transduction de signaux, la synthèse

protéique, la régulation des différentes enzymes et hormones, et encore la correction du 'misfolding' par des protéines chaperonnes.

La compréhension des interactions entre protéines et la prédiction des sites impliqués ouvrent donc de nombreuses voies dont deux principales qui dépendent l'une de l'autre : la compréhension de la biologie cellulaire et la production de nouveaux médicaments.

D'un point de vue fondamental, la connaissance des différents réseaux moléculaires d'interaction présente un intérêt scientifique énorme. Si on connaissait la fonction et les différents partenaires d'interaction d'une protéine, les nombreuses cascades cellulaires qui permettent aux organismes de vivre et de se développer pourraient être expliquées et certaines interrogations pourraient être levées.

Un corollaire à la connaissance des mécanismes d'interaction biologique est la compréhension des mécanismes suivis par certaines maladies. En effet, une partie des maladies existantes sont dues à une interaction anormale entre protéines, à la perte d'une interaction donnée ou à l'agrégation de plusieurs protéines.[27] De plus, la détection des molécules impliquées dans une maladie nouvelle serait plus rapidement réalisée. D'un point de vue pratique, la production de médicaments serait facilitée par une identification plus aisée des cibles pharmaceutiques potentielles. Une fois les cibles pharmaceutiques détectées avec précision grâce aux nouvelles connaissances, la production de peptides imitant le partenaire de l'interaction permettrait par exemple de bloquer les effets néfastes du système impliqué. Si, actuellement, il est encore impossible de prédire avec certitude la position dans la séquence des sites d'interaction, il est possible de s'aider des données et techniques déjà mises au point. Golemis *et al.*[28] proposent dans leur article une démarche en cinq étapes permettant de puiser le maximum des informations actuelles pour générer des médicaments utiles :

1. Identifier les cibles potentielles en se basant sur l'évidence des fonctions biologiques dans des conditions cliniques semblables à celles qui seront adressées et en se basant sur l'évidence que les fonctions de la cible peuvent être manipulées de manière productive.
2. À partir des données existantes, trouver un réseau physique et fonctionnel d'interactions pour la cible.
3. Déterminer la bibliographie des mutations des interactions protéine-protéine impliquant la cible.
4. Développer une stratégie discriminante pour identifier les inhibitions spécifiques/non-spécifiques de la protéine cible, des protéines en interaction et du réseau de contrôle.

5. Etre prêt à exploiter des découvertes très intéressantes…

L'auteur soulève ici un des problèmes majeurs dans la compréhension des phénomènes d'interaction : la difficulté de retirer les données essentielles de la mine d'informations disponibles.

Finalement, si les mécanismes d'interaction sont compris et connus avec précision, les répercussions sur les tentatives de prédiction du 'folding' protéique seront non-négligeables. En effet, les forces qui sont importantes dans les processus de 'folding' (hydrophobicité, liens H, ponts salins, interactions de van der Waals, etc.) sont précisément celles responsables des interactions protéines-protéines.

### I.4.2. Evolution des connaissances

C'est dans les années 20, avec les travaux de Svedberg,[29,30] et presque par hasard, que l'on établit pour la première fois l'existence d'associations entre protéines. Celui-ci allait d'ailleurs se voir attribuer le prix Nobel de chimie en 1926. Par la suite, le monde scientifique allait très vite se rendre compte de l'omniprésence de ce type d'interactions.

Les études sur les interactions protéines-protéines peuvent être abordées selon diverses voies, néanmoins, deux approches principales peuvent être différenciées[31]:

- La première voie analyse les interactions au niveau moléculaire et/ou atomique dans le but de comprendre les phénomènes physiques expliquant pourquoi une protéine va interagir spécifiquement avec une (ou plusieurs) autre(s) alors qu'elle est en présence de nombreux partenaires potentiels, d'expliquer les dysfonctionnements dans les interactions et de développer de nouveaux agents pharmaceutiques
- La deuxième voie consiste en une étude de l'interactome c'est-à-dire l'étude de l'ensemble des interactions dans un organisme ou dans un compartiment cellulaire donné dans le but principal de comprendre le fonctionnement du système cellulaire dans sa globalité ('systems biology' voir Cusick *et al.*[32] pour une revue sur le sujet).

**La 'voie atomique'**

Avant de tenter de répondre à la question « Quelle protéine interagit avec quelle protéine ? » et, ainsi, d'améliorer nos connaissances de l'interactome, les scientifiques ont étudié en détail les résidus et les atomes impliqués dans les interfaces biologiques. L'étude de ces résidus et de ces atomes est d'une grande importance pour permettre la compréhension des mécanismes mis en jeu, pour arriver à moduler les interactions (augmentation de la spécificité, blocage de l'interaction) mais aussi pour arriver à prédire quelle partie d'une séquence protéique est impliquée dans l'interaction. Il est essentiel, avant d'arriver à prédire correctement la position de ces sites d'interaction (cf. paragraphe I.5), d'effectuer une étude détaillée des caractéristiques des résidus à l'interface.

Les interactions entre protéines ont été les premières à pouvoir être analysées avec précision et sont décrites au point I.4.3. Les interactions entre protéines et acides nucléiques quant à elles, ont souffert plus longtemps du nombre limité de structures 3D disponibles. Les caractéristiques principales de ces interactions sont décrites au point I.4.4.

N.B. : Une partie des résultats sur la localisation des sites d'interaction obtenus expérimentalement ont été regroupés dans des bases de données accessibles sur Internet. On peut notamment citer : AANT,[33] ProNIT,[34,35] ProtTherm,[35,36] BindingDB,[37] BID,[38] NTDB,[39] NPIDB,[40] ASEdb.[41] Sur ces sites, on retrouve des informations sur les macromolécules mises en jeu, des liens vers d'autres sites caractérisant les protéines étudiées et les données thermodynamiques mesurées. Les différentes conditions expérimentales sont également généralement mentionnées avec éventuellement un lien vers l'article de référence.

**La 'voie de l'interactome'**

L'étude des différentes cascades d'interactions entre protéines observées dans tout système biologique a attiré et attire encore actuellement de nombreux scientifiques. La compréhension de ces phénomènes a des répercussions dans de nombreux domaines (voir paragraphe précédent, I.4.1). Dès lors, il n'est pas surprenant de voir les efforts effectués pour décrire au mieux possible l'interactome.

Les méthodes expérimentales utilisées sont diverses, on peut citer la technique du double hybride[42] qui permet une « observation directe de l'interaction »[43,44] entre deux protéines connues ou une protéine inconnue (proie) et une protéine connue (appât), les techniques d'immuno-précipitation qui permettent d'isoler les complexes protéiques,

l'altération un par un des gènes d'un génome suivie de l'observation du phénotype, les techniques de transfert d'énergie de fluorescence (FRET pour 'Fluorescence Resonance Energy Transfert') et de bioluminescence (BRET pour 'Bioluminescence Resonance Energy Transfert'),[45] etc. Au fil des années, ces expérimentations ont permis de comprendre de nombreux mécanismes biologiques et les données fournies ont été en partie regroupées dans des banques de données d'interactions protéiques (DIP,[46,47] BIND, MIPS,[48,49] MINT,[50] HPRD,[51] IntAct,[52,53] String,[54] BioGRID[55]…). Ces banques de données permettent de trier les interactions selon le type d'organisme étudié et la ou les méthodes utilisées tout en apportant des informations sur les protéines concernées et sur l'effet de l'association/dissociation du complexe. Les données ainsi rassemblées doivent toutefois être utilisées avec prudence. En effet, les expériences amènent des erreurs comme des faux positifs (interactions non-significatives), des faux négatifs ('folding' incorrect, localisation sub-cellulaire inadéquate, manque de modifications post-traductionnelles spécifiques), des imprécisions dues à la méthode et variant selon le type de protéine étudiée… Pour discerner les interactions effectives des erreurs expérimentales, certaines méthodes sont proposées comme l'utilisation de structures tri-dimensionnelles observées aux rayons-X et l'établissement d'un degré de certitude en fonction du nombre de méthodes différentes ayant mis en évidence l'interaction.[56,57]

Uetz *et al.*[58] en 2000 et Ito *et al.*[59] en 2001 ont présenté le premier interactome d'un organisme de référence : la levure *Saccharomyces cerevisiae*. Cet interactome comprend 1870 protéines impliquées dans 2240 interactions physiques directes provenant principalement d'analyses systématiques par la technique du double hybride (Figure I-20). Le réseau d'interactions décrit a permis de montrer[60] que plus une protéine est connectée à d'autres dans la cellule, plus celle-ci est importante pour la survie de la levure. Il existe donc des protéines se trouvant à des 'carrefours d'interactions' qui sont essentielles au fonctionnement cellulaire. Ce type de protéine ne représente qu'une faible proportion des protéines de l'interactome, ce qui pourrait expliquer la robustesse de la levure face aux mutations. Plus tard, d'autres interactomes ont été présentés[61] : la drosophile *Drosophilia malanogaster*[62] en 2003, le ver *Caenorhabditis elegans*[63] en 2004, l'humain *Homo sapiens*[64] en 2005, *Plasmodium falciparum*[65] en 2006, etc. Il est important de se souvenir que ces interactomes ne sont pas complets. Dans certains cas, un nombre très réduit de protéines, toutes impliquées dans le même mécanisme, ont été étudiées et mises en relation.

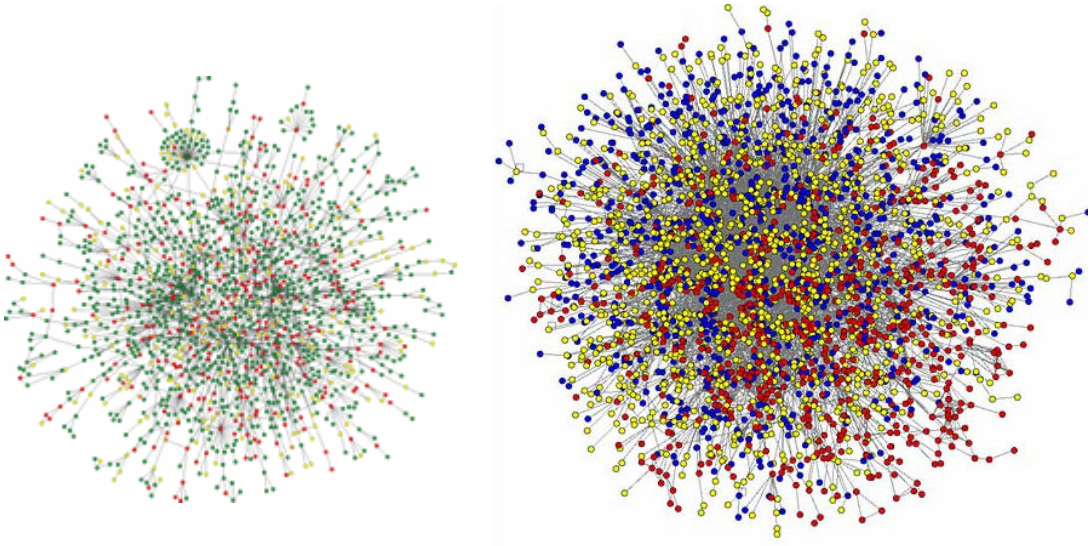


**Figure I-20 : Interactome de la levure *S. cerevisiae* (à gauche) et de *C. elegans* (à droite). Chaque rond représente une protéine et chaque ligne, une interaction.**

Les méthodes expérimentales citées précédemment ont le principal défaut d'être longues à mettre en place. Des méthodes informatiques dédiées à la prédiction des partenaires ont été développées pour palier à ce défaut. Une première catégorie de méthodes utilisent les événements de fusion de gènes.[66] Le postulat de départ se base sur le fait que la fusion de gènes provient d'une pression de sélection pour associer des gènes qui sont co-régulés et peuvent interagir. En d'autres termes, certaines protéines d'une espèce donnée sont constituées de domaines qui, dans d'autres espèces, correspondent à des protéines individualisées et complètes. En détectant les zones ayant subi une fusion de gènes dans la séquence, il est dès lors possible de prédire une interaction. D'autres méthodes se basent sur l'observation de 'mutations corrélées'.[67] En effet, si des changements apparaissent dans une zone en interaction de la séquence d'une protéine lors du processus d'évolution, on peut postuler que la séquence de la protéine partenaire subira certaines mutations permettant de garder une interaction la plus spécifique possible, on parle de mutations corrélées. Les analyses d'homologie de séquence, de structures secondaires sont également des méthodes largement utilisées pour prédire les zones d'interaction.[68-70] Finalement, certains interactomes sont construits par recoupement des informations contenues dans les bases de données protéiques (SPIDer[71] ou PIPE[72] p.ex.) ou par analyse (automatique ou non) de la littérature scientifique ('literature-curated interactome' - cf. Roberts[73] pour une revue sur le sujet).

### I.4.3. Interactions protéines-protéines

#### Introduction

En 1975, Chotia et Janin[74] furent dans les premiers à proposer un modèle d'interaction entre protéines : « L'hydrophobicité est le principal facteur stabilisant les associations protéines-protéines, alors que la complémentarité joue un rôle sélectif en décidant quelles protéines peuvent s'associer ». Et, bien que de nombreux progrès aient été réalisés dans le domaine, ce modèle de base est encore approuvé par de nombreux auteurs.

Par la suite, les travaux effectués permirent une amélioration des connaissances mais souffrirent longtemps du faible nombre de structures 3D connues (Figure I-21) et du peu de types d'interactions représentées : principalement des complexes enzyme(protéase)-inhibiteur et antigène-anticorps. Pour exemple, l'article de Chotia et Janin[74] se base sur seulement 3 dimères alors que les travaux actuels utilisent fréquemment plus de 1000 structures[75] et même jusque 1981 complexes pour Aytuna *et al.*[76] Actuellement, des efforts supplémentaires sont effectués en vue de collecter des structures de complexes protéiques par le biais de projets de 'structural genomics' (cf. Grabowski *et al.*[77] pour une revue récente sur le sujet). Ces projets permettent de récolter à haut débit de telles structures mais avec peu ou pas d'informations sur la fonction de celles-ci.

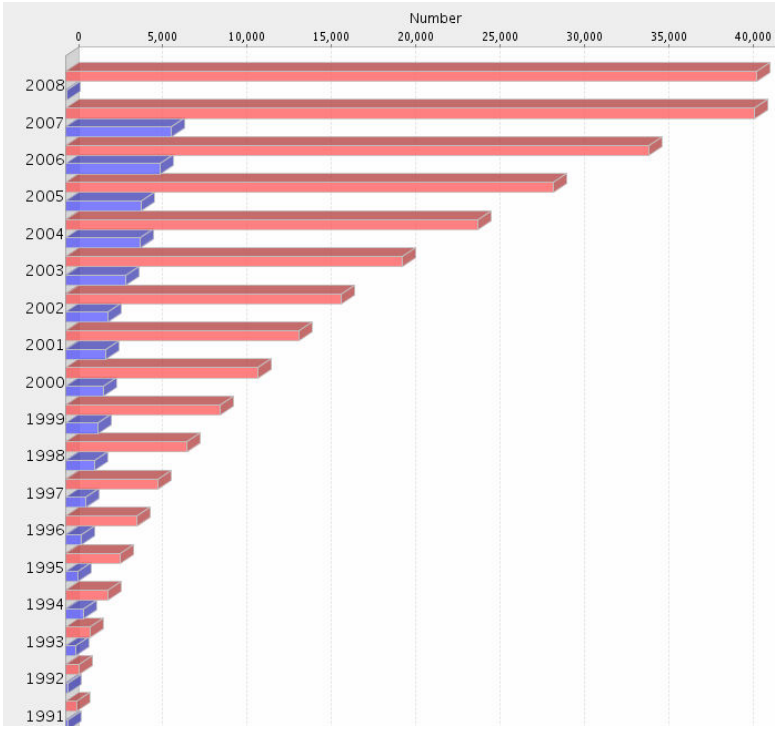



**Figure I-21 : Evolution du nombre de structures 3D Rx contenues dans la Protein Data Bank. En bleu, les nouvelles structures déposées dans l'année considérée et en rouge le nombre total de structures. Données de janvier 2008.**

En 1995, Jones et Thornton[78] marquent une nouvelle étape dans les études sur les interactions protéine-protéine en réalisant un travail très complet sur une banque de données de 32 dimères. La composition en acides aminés et en atomes, la forme, les structures secondaires, les liens H, les ponts salins, la segmentation, la complémentarité des interfaces furent analysés. Il ressort de cette étude que, bien que l'interface soit plus hydrophobe que le reste de la surface, certains acides aminés chargés et polaires (arginine, méthionine) montrent une certaine affinité pour l'interface. D'autres travaux ont permis de confirmer que les acides aminés hydrophiles interviennent de manière spécifique dans les interactions protéiques. Larsen *et al.*[79] ont réalisé une étude plus 'visuelle' d'une banque de 136 homodimères et ont montré que seulement un tiers des complexes possédaient un cœur hydrophobe alors que les autres deux tiers possèdent des interfaces où l'on retrouve des zones hydrophobes et des zones hydrophiles (avec des liens H, présence de molécules d'eau) entremêlées.

**Le rôle du solvant**

Le rôle de l'eau qui, ne l'oublions pas, reste le solvant biologique par excellence, fut lui aussi de plus en plus pris en compte. Janin[80] montre que, malgré un échange très rapide des molécules d'eau entre l'interface et le milieu aqueux (moins d'une milliseconde pour l'eau au centre de l'interface), les interactions avec les molécules d'eau sont aussi importantes pour la stabilité et la spécificité de la liaison que les liens H directs entre protéines. En 2005, le groupe de Janin,[81] propose de diviser les interfaces en deux types : les interfaces 'mouillées' et les interfaces 'sèches' où les molécules d'eau sont situées tout autour de l'interface. Ce sont ces dernières qui correspondraient plutôt aux interactions spécifiques. De même, les fragments de protéines qui ne sont pas suffisamment hydratés à l'état monomérique seraient des sites préférentiels pour les interactions intermoléculaires[82] et la conservation des molécules d'eau dans les interfaces à récemment été étudiée.[83]

Toutefois, il reste un problème majeur à l'étude de ce type d'interaction eau-protéine : pour pouvoir situer avec précision la position d'une molécule d'eau, il est nécessaire de posséder des structures cristallines aux rayons-X hautement résolues (de l'ordre de 2Å ou moins) ce qui ne représente pas la majorité des structures connues.

**Les changements de conformation**

En 2000, Sundberg et Mariuzza[84] ont publié une revue sur les changements de conformation subis par une protéine lors de son association à une autre protéine. Ces changements sont étudiés par de nombreux auteurs[78,85-92] et connus depuis longtemps (ex. :

allostérie enzymatique) mais les avis divergent quant à leur importance. Sundberg et Mariuzza[84] ont montré que certains atomes bougent d'une distance allant jusque 3Å. Heifetz et Eisenstein[93] ont pour leur part analysé les différences de flexibilité des chaînes latérales des vingt acides aminés naturels et ont montré que les chaînes latérales de la lysine et de l'arginine sont les plus flexibles. Ces résultats pourraient permettre d'expliquer certains comportement des protéines comme leur faculté d'interagir avec plusieurs partenaires, la reconnaissance du système immunitaire (mouvement des régions CDR pour faire tourner les domaines $V_H$ et $V_L$). Les mouvements effectués par les protéines lors de l'association pourraient aussi permettre d'amener un plus grand nombre de résidus hydrophobes à l'interface en vue de la stabiliser.[94] Gunasekaran et Nussinov[95] ont essayé de différencier les interfaces rigides de celles subissant des changements de conformation. Il apparaît que le tryptophane, les interactions entre résidus hydrophobes, entre résidus aromatiques, et entre résidus polaires et aromatiques sont associées à des grands changements de conformation.

Néanmoins, pour étudier les changements de conformation lors de la liaison, il faut posséder les structures 3D à la fois des monomères et du complexe (cf. 'the Database of Macromolecular Movements'),[90] ce qui n'est pas toujours le cas. De plus, il faut remarquer que les protéines en solution présentent naturellement de faibles déformations de leur structure.[89] Ces deux derniers points montrent que ce problème est encore loin d'être résolu mais le développement des techniques par Résonance Magnétique Nucléaire (RMN), qui permettent l'observation de ces changements de conformation en solution, devrait permettre une compréhension plus poussée du phénomène.

**Classification des interfaces**

Une des tactiques employée pour permettre de simplifier le problème des interactions entre protéines est de séparer les différents types de complexes. Les complexes peuvent être différenciés selon la taille de l'interface, leur caractère obligatoire[96] (protéines fonctionnelles uniquement sous-forme complexée, on parle de 'folding' deux-états,[97] c'est le cas des oligomères et des homodimères) ou leur caractère non-obligatoire[96] (protéines actives à l'état monomérique, on parle de 'folding' trois-états,[97] comme pour les complexes enzyme-inhibiteur).[27,98] Mais la distinction la plus simple consiste à séparer les complexes selon les différentes familles biochimiques. C'était le cas de la plupart des articles où l'on étudie uniquement ou distinctement des homodimères, des oligomères, des complexes enzyme-inhibiteur ou protéase-inhibiteur (hétérodimères en général), des complexes antigène-anticorps, des complexes intervenant dans le cycle cellulaire, la transduction de signal… En

effectuant cette séparation somme toute assez simple, on arrive à un deuxième niveau de différenciation. En effet, Jones et Thornton[78] ont décrit d'un point de vue cinétique trois types de complexes. Les interactions antigène-anticorps sont l'équivalent moléculaire d'une première rencontre avec une constante de liaison de l'ordre de $10^{-9}$ $mol^{-1}$ (bien que lors de la réponse immune subséquente, des mutations somatiques puissent augmenter la spécificité et la force de l'interaction). Pour certains complexes dimériques, cette constante peut avoir des valeurs de l'ordre de $10^{-16}$ $mol^{-1}$ et on est donc en présence d'interactions aussi fortes qu'à l'intérieur d'un monomère, il faut dénaturer le dimère pour casser la liaison. Les complexes enzyme-inhibiteur montrent un caractère intermédiaire avec des constantes de liaison de l'ordre de $10^{-7}$ $mol^{-1}$ à $10^{-13}$ $mol^{-1}$.

En 2003, Ofran et Rost[99] ont analysés les interfaces protéiques en divisant celle-ci en 6 types distincts : interaction intramoléculaires, entre domaines, intermoléculaires pour les complexes homomériques permanents, homomériques transitoires, hétéromériques permanents et hétéromériques transitoires. L'originalité de ce travail réside dans la mise au point d'une méthode permettant de classer efficacement de grandes bases de données en deux types d'interfaces (transitoires et permanentes), types connus pour être impliqués dans la diversité des interactions protéine-protéine.[96,100]

Remarque : bien que les complexes obligatoires correspondent, dans la plupart des cas, à des complexes permanents et les complexes non-obligatoires à des complexes transitoires, cette correspondance n'est pas toujours vraie et il est important de différencier ces deux classifications.[96]

**État des lieux**

Le concept de structuration de l'interface le plus utilisé, bien qu'il représente une simplification de la réalité, est celui du 'O-ring'. Conte *et al.*[87] y divisent l'interface en trois zones (Figure I-22) : les zones A et C restent partiellement accessibles au solvant tandis que la zone B ne l'est plus après interaction, les zones B et C contiennent des atomes qui établissent des contacts de type van der Waals avec la protéine partenaire et les atomes de la zone A deviennent moins accessibles lors de la liaison mais n'établissent pas de contact. La zone C peut aussi être considérée comme une zone accessible au solvant sauf si on inclut dans la structure les molécules d'eau ayant cristallisé.[101] Il faut noter que les résidus de la zone B ne représentent pas nécessairement un cœur hydrophobe comme décrit par exemple par Young *et al*.[102]

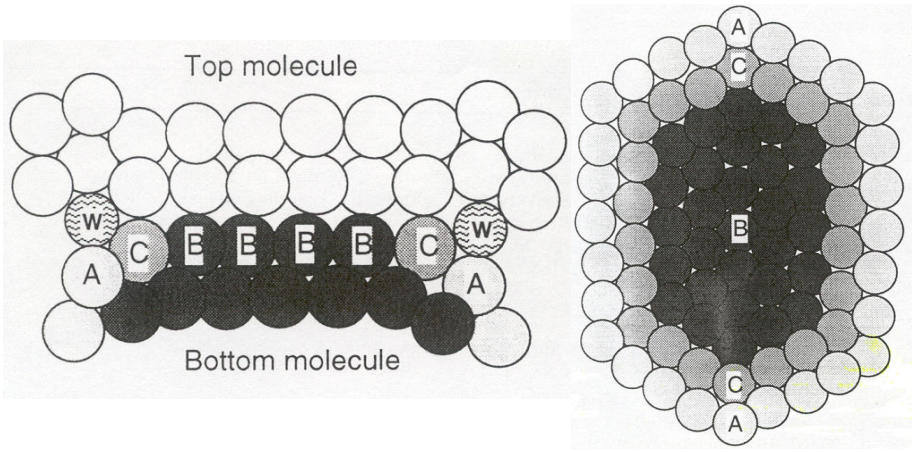



**Figure I-22 : Différentes classes d'atomes à l'interface par Conte *et al.*[87] La figure de gauche représente une vue de côté et la figure de droite une vue du dessus de l'interface.**

De leur côté, Bogan et Thorn[103] ont utilisé une banque de mutants alanine pour étudier les hot-spots (résidus contribuant grandement à l'énergie de liaison - ddG > 2 kcal/mol) au niveau de l'interface. Les mutants alanine sont issus d'une séquence protéique de départ dans laquelle on remplace systématiquement chaque résidu par une alanine,[104] cette méthode a été utilisée pour la première fois lors de l'étude de l'hormone de croissance et de sa protéine de liaison.[105] Ces mutations permettent une étude individuelle du rôle énergétique des chaînes latérales des acides aminés bien que certaines limitations sont à prendre en compte.[106] Les hot-spots sont enrichis en tryptophane, tyrosine et arginine, et sont entourés de résidus moins importants énergétiquement. Ce deuxième type de résidus jouerait un rôle important en protégeant les hot-spots du solvant, les molécules d'eau perturberaient moins facilement les résidus du centre de l'interface et augmenteraient la stabilité de l'interaction en ralentissant la dissociation. Il a également été montré que les hot-spots correspondent généralement à des résidus conservés dans les structures protéiques.[107-110]

Ces deux théories (hot spots et O-ring) peuvent donc être combinées d'une manière assez élégante : les résidus ayant un impact énergétique élevé lors de l'interaction, les hot-spots se situeraient dans la zone complètement isolée du solvant (zone B du O-ring) et seraient protégés par des acides aminés moins importants énergétiquement mais ayant un rôle majeur à jouer en empêchant le passage des molécules d'eau (résidus des zones A et C du O-ring).[111] Une revue sur le sujet a récemment été réalisée par Moreira *et al.*[112]

Finalement, la correspondance entre résidus en interaction et résidus conservés dans la séquence est, elle, beaucoup plus controversée. En effet, il n'est pas certain que les acides aminés en interaction soient plus conservés que les autres résidus de la surface protéique.[113] Même si certaines études démontrent que la corrélation n'existe pas pour les interactions

protéine-protéine mais bien pour les interactions protéine-ligand,[114] il semblerait en fait que ce soit le nombre de résidus particulièrement bien conservés qui importe et pas une conservation moyenne.[115]

L'étude de la conservation des acides aminés en interaction a permis de mettre en évidence un autre type de résidu appelés 'anchor residues'.[116] Ces acides aminés se trouvent sous une conformation figée dès l'état monomérique et servent de point d'ancrage pour la protéine partenaire de l'interaction. Finalement, au sein des interfaces 'sèches', les acides aminés qui échangent le moins rapidement les molécules d'eau (résidus 'secs') seraient les plus conservés suite à la pression évolutive.[117]

En résumé, l'effet hydrophobe, dont l'importance dans les phénomènes de 'folding' n'est plus à prouver,[94] joue un rôle énergétique pour stabiliser l'interaction. Cette effet n'est pas seul à agir et une complémentarité de forme et de charge des interfaces permet d'obtenir une spécificité élevée entre les sites d'interaction.[31,78,118] Par ailleurs, les ponts salins auraient un rôle supplémentaire a jouer : les charges portées par les sous-unités protéiques permettraient une 'pré-orientation' de celles-ci,[119] ce qui expliquerait les vitesses élevées d'interaction observées. En effet, les ponts salins sont les plus efficaces à longue distance. Dans les interfaces, un nombre réduit de résidus, les hot-spots, auraient un impact énergétique majeur lors de l'interaction. Ces résidus sont isolés du solvant par un anneau ('ring') de résidus moins importants énergétiquement.

Donc, un concept actuel permettant d'expliquer le mécanisme d'interaction se diviserait en deux étapes : la première impliquerait les charges de surface et conduirait à la spécificité de l'interaction, la deuxième impliquerait les résidus hydrophobes de l'interface et stabiliserait le complexe au niveau énergétique. De plus, les résidus hot-spots seraient les acteurs clés lors de ces deux étapes.

Finalement, il ne faut pas perdre de vue que ces différents concepts ne sont pas absolus et comportent d'importantes simplifications. Chaque site fonctionnel est unique et une étude statistique ne permet pas de mettre en évidence certains détails caractéristiques d'une interaction donnée. Mais, bien que chaque site d'interaction soit unique, Kini et Evans[120] sont les premiers à proposer cinq règles générales et fondamentales qui caractérisent les sites d'interaction :

- Certaines interactions impliquent plus d'un site séquentiel d'interaction par molécule (et une seule molécule peut interagir avec plusieurs partenaires).

- Le nombre d'acides aminés impliqués dans une zone d'interaction varie souvent de 3 à 6.
- Les résidus en interaction ne sont pas toujours liés par un lien peptidique, ils sont rapprochés par la structure tertiaire de la protéine.
- Les chaînes latérales sont largement impliquées dans les interactions tandis que la chaîne principale l'est dans une moindre mesure.
- L'affinité des sites d'interaction est due à des facteurs stériques, électrostatiques et hydrophobes.

Res & Lichtarge[121] ont réalisé une très bonne revue sur le sujet en 2003. Et en 2005, une description plus globale des interfaces est tentée par Keskin *et al.*[122] et sera en partie reprise plus tard par Reichmann *et al.*[123] :

- L'association entre protéine a lieu de manière coopérative : la stabilité d'un complexe est plus que la somme des contributions individuelles des acides aminés impliqués.
- Les sites d'interaction sont constitués d'une ou d'un petit nombre de régions bien compactes et indépendantes. Ces régions contiennent les hot-spots qui interagissent avec d'autres hot-spots sur la protéine partenaire et ne correspondent pas à des segments séquentiels.
- Les interfaces possèdent des patchs hydrophobes qui contiennent les résidus clés pré-orientés dans la structure monomérique.
- Les protéines désordonnées possèdent souvent de grandes interfaces intermoléculaires, la taille de celle-ci étant dictée par la fonction de la protéine.
- Bien qu'ayant des interfaces de structures similaires, les protéines desquelles proviennent ces interfaces peuvent être différentes.

**Remarque**

Dans tous les travaux présentés ci-dessus, les auteurs apportent une information plus ou moins complète sur les fréquences des acides aminés se trouvant à l'interface, les paires d'acides aminés en interaction et les structures secondaires impliquées. Ces résultats seront synthétisés en comparaison à ceux obtenus lors de ce travail dans le Chapitre IV : Discussion Générale.

## I.4.4. Interactions protéines-acides nucléiques

### Introduction

Les interactions entre molécules biologiques, spécialement celles impliquant les acides nucléiques, sont la base de la vie cellulaire. Les interactions entre protéines et acides nucléiques sont impliquées dans des rôles aussi fondamentaux que la réplication, la traduction et la réparation des acides nucléiques, la régulation des gènes et la liaison des virus. Les protéines liant l'ADN sont codées par 2-3% du génome chez les procaryotes et 6-7% chez les eucaryotes.[124]

Dans les paragraphes suivants, nous allons décrire les mécanismes directs (contacts atomiques intermoléculaires) ainsi que les mécanismes indirects (structure intramoléculaire des acides nucléiques) de reconnaissance. Ces deux types de reconnaissance sont importants pour la spécificité des interactions et peuvent agir seuls ou de concert,[125-127] et ce, dans des proportions variables.[125,128] Finalement, nous présenterons les divers motifs protéiques caractéristiques des interactions avec l'ADN et l'ARN.

### Reconnaissance directe

La reconnaissance d'une séquence nucléotidique spécifique par une protéine est, *a priori*, déterminée par les interactions atomiques entre les acides aminés et les nucléotides. Les premiers travaux ayant pour but de dégager les lois guidant cette reconnaissance ont débutés dans les années 70 avec les travaux de Seeman *et al*.[129] Dans ce travail, tous les atomes des bases nucléiques susceptibles de former des liens H ont été identifiés (Figure I-23). Ensuite, ces atomes ont été utilisés pour décrire tous les couples acide aminé-base possibles. Sur base de la seule structure tridimensionnelle disponible à cette époque (une structure d'ARN de transfert), Seeman et ses collaborateurs[129] ont conclu que la reconnaissance spécifique ne pouvait être expliquée par un lien H unique mais plutôt par des liens H multiples (liens H bifurqués ou bidentés).

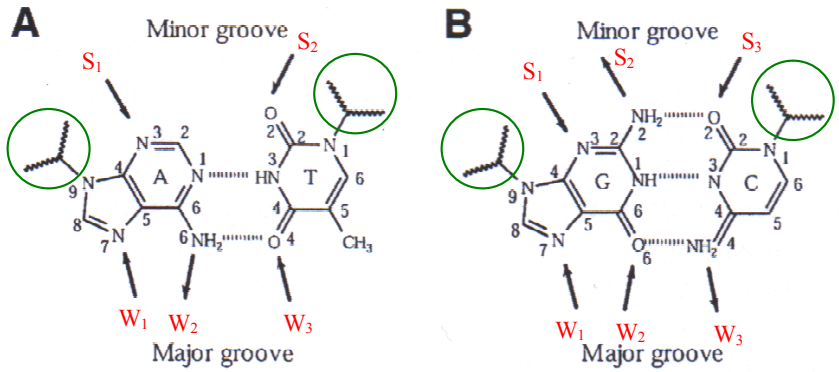



**Figure I-23: Position des sites potentiels pour un lien H des quatre bases de l'ADN. Ces sites sont notés $S_i$ (S = small) dans le sillon mineur et $W_i$ (W = wide) dans le sillon majeur. Le squelette nucléique est symbolisé par un cercle vert. Image adaptée de Luscombe *et al.*[130]**

Par la suite, la plupart des travaux se sont particulièrement intéressés aux liens H intervenant dans les complexes protéine-acide nucléique et ont montré que les liens H entre les protéines et les bases nucléiques étaient les plus significatifs pour la reconnaissance.[131,132] La plupart de ces liens H impliquent les chaînes latérales des acides aminés et sont décrits comme le type d'interaction le plus important pour la spécificité des interactions.[133] Les interactions entre les nucléotides et le squelette peptidique protéique semblent pour leur part être importantes pour la stabilisation et l'orientation du complexe.[134] Comme proposé par Seeman *et al.*,[129] la spécificité des liens H de la chaîne latérale protéique ne peut être expliquée par un mécanisme 'one-to-one' mais plutôt par des interactions de type 'one-to-many'.[130] De plus, la structure des différentes bases nucléotidiques associée aux liens H permet une reconnaissance spécifique de chacune de ces bases.[135] Avec l'augmentation du nombre de structures 3D disponibles et l'augmentation de la précision sur la position des atomes les composant (meilleure résolution), les études sur les liens H impliqués dans les interactions protéine-acide nucléique ont pu être approfondies. En plus des liens H classiques (O---H-N, O---H-O, N---H-O, N---H-N, O---H-S ou N---H-S), au moins deux autres types de liens H ont été décrits : les liens H non-classiques et les liens H modulés par l'eau. Les liens H non-classiques sont principalement du type O---H-C. Ces liens sont assez fréquents entre les groupements C-H du sillon majeur de l'ADN et les oxygène de la protéine.[13] Ce type de lien H correspond à 33% des liens H pour les interactions entre protéines et ARN.[136] De leur côté, les liens H modulés par l'eau permettent de faire interagir entre eux deux atomes accepteurs d'hydrogène de manière indirecte. Ils permettent notamment de minimiser l'impact défavorable des charges négatives des acides aspartique et glutamique lors de l'approche des charges négatives des phosphates nucléiques. Ce type de lien H est beaucoup plus temporaire

qu'un lien H classique à cause de la mobilité des molécules d'eau.[80] Les liens H modulés par l'eau représentent environ 15% de l'ensemble des interactions protéine-ADN[13,130,131] (cf. Jayaram et Jain[137] pour une revue sur le sujet) et semblent largement impliqués dans les contacts protéine-ARN.[133,136,138] Il faut remarquer que certains auteurs pensent que le rôle majeur de l'eau ne serait pas d'intervenir dans les interactions protéine-acide nucléique mais plutôt de stabiliser soit la protéine, soit l'acide nucléique.[139] N.B. : une revue très complète sur le rôle des liens H dans les complexes protéine-ADN a été publiée en 2007 par Coulocheri *et al*.[140]

Outre les liens H, les ponts salins entre les acides aminés chargés positivement et les atomes d'oxygène des groupements phosphate semblent être plutôt impliquées dans la stabilisation des complexes[136,141] et seraient impliquées dans la pré-orientation des chaînes polypeptidiques.[119] Alors que l'effet hydrophobe est largement pris en compte dans les interactions protéine-protéine, ce type d'interaction est moins souvent considéré dans les interactions avec les acides nucléiques. Néanmoins, lors des interactions avec les sucres au niveau du sillon mineur de l'ADN (et plus particulièrement dans les hélice de type A), les acides aminés les plus impliqués seraient de type hydrophobe.[142] Finalement, les interactions de van der Waals, qui représentent parfois plus de 75% des interactions avec l'ADN double brins et plus de 90% des interactions avec de l'ADN mono-caténaire,[143] joueraient également un rôle stabilisateur plutôt qu'un rôle dans la spécificité de la reconnaissance.[130]

Une grande partie des protéines interagissant avec l'ADN le font sous forme d'homodimères (voir ci-après, *Motifs protéiques d'interaction*). Mais, alors que l'on pourrait s'attendre à ce que les chaînes protéiques se fixent de manière symétrique, il semble que ce type d'interaction se fasse régulièrement de manière asymétrique. Les légères différences de reconnaissance entre les deux chaînes du complexe et l'acide nucléique pourraient avoir un rôle à jouer dans la spécificité de la reconnaissance.[144]

**Reconnaissance indirecte**

Après les liens H, le principal facteur influençant la spécificité des interactions réside dans la conformation des acides nucléiques, elle-même influencée par la séquence nucléotidique. Il est connu que l'ADN est une macromolécule très flexible.[145-148] De plus, certaines protéines liant l'ADN induisent une courbure de l'ADN modifiant de cette manière la reconnaissance par les acides aminés.[147,149,150] La capacité de l'ADN à se courber sous une pression extérieure est influencée par la succession de bases le long du polymère d'ADN.[151] A partir d'une banque de données de doubles hélices d'ADN de type Watson-Crick, il a été

montré[152,153] qu'une succession des bases de type pyrimidine-purine favorisait l'accessibilité de l'ADN. En effet, des successions de type CA/TG et TA/TA (pyrimidine-purine) permettent à la chaîne d'ADN d'adopter la conformation la plus flexible alors que des successions de type AA/TT, AT/AT et GA/TC conduisent à une structure plus rigide.[152] Cette conclusion a été confirmée plus tard par Olson *et al*.[153] sur base de complexes protéine-ADN. Dans ce dernier travail, les successions pyrimidine-purine sont décrites comme des charnières flexibles qui permettent à l'ADN d'ajuster sa structure à la surface protéique. Steffen *et al*.[154] ont eux étudié l'impact énergétique de la courbure de l'ADN pour identifier des sites potentiels de liaison des protéines. Le calcul de propension à la déformation au niveau de chaque paire de base permet d'arriver à des résultats encourageants bien que la simulation de la reconnaissance indirecte ne soit pas suffisante à elle seule.[154]

Comme décrit au point 2.2.1., l'ADN bicaténaire peut adopter trois formes principales : A, B et Z. Néanmoins, plusieurs travaux ont montré qu'il peut exister des formes intermédiaires à ces structures[155,156] et que des transitions continues entre les formes A et B sont possibles.[157-159] Ces transitions, dépendantes de la séquence nucléique et de l'activité de l'eau, correspondraient à un moyen supplémentaire de réguler la liaison de protéines à l'ADN. Varnai et ses collaborateurs[160] ont montré qu'au sein même de complexes contenant des structures d'ADN de type B, il existe des variations dans les angles du squelette sucre-phosphate (α et γ). Ces transitions ne seraient possibles que sous l'effet de l'interaction avec une protéine et la présence de couples α/γ inhabituels permettrait des ajustements structurels et diminuerait l'énergie du complexe.[160] De plus, les interactions avec l'ADN de type B seraient gouvernées par les acides aminés polaires alors que les acides aminés hydrophobes auraient plutôt tendance à interagir avec l'ADN de type A.[142]

**Motifs protéiques d'interaction**

L'analyse de la conformation des sites d'interaction protéine-acide nucléique a permis de décrire différentes familles de protéines liant l'**ADN**. La description présentée ci-dessous se base principalement sur les travaux de Pabo & Sauer,[134] Luscombe *et al*.,[124] et Garvie & Wolberger.[161] N.B. : cette classification a été remise en cause par Prabakaran *et al*. qui proposent une classification basée sur différents descripteurs structuraux.[162] Cette classification en 7 types serait plus adaptée à la description des interactions protéine-acide nucléique.

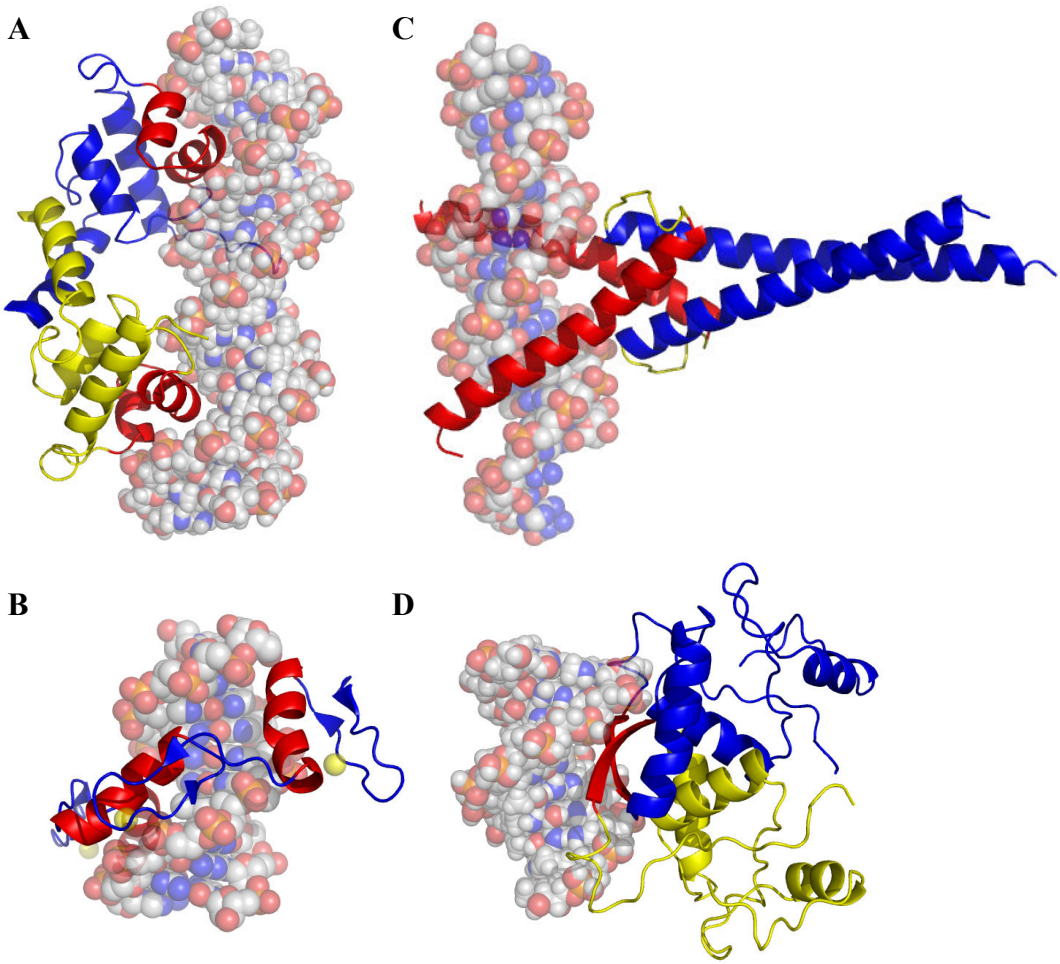



**Figure I-24 : Représentations 3D de complexes protéine-ADN (adaptées de Luscombe *et al.*[124]). L'ADN est représenté en vrai volume et couleurs. A : Motif HTH du complexe 1LMB[163] ('Cro and repressor family'). L'hélice de reconnaissance (en rouge) est insérée dans le sillon majeur. La protéine est sous forme de dimère, un des monomères est en bleu et l'autre en jaune. B : Motif ββα 'zinc finger' (1AAY).[164] Les brins β sont représentés en bleu et les hélices de reconnaissance en rouge. Les ions de zinc sont en vrai volume et en jaune. C : Motif HLH du complexe 1AM9[165] (SERBP). Les hélices intervenant dans la dimérisation sont en bleu, les hélices de reconnaissance à l'ADN en rouge et les boucles intermédiaires en jaune. D : Les répresseurs MetJ et Arc interagissent avec l'ADN (1CMA)[166] par l'intermédiaire de brins β (en rouge). Les deux protéines (en bleu et en jaune) fournissent chacune un brin qui forment un 'feuillet anti-parallèle' au niveau du sillon de l'ADN. Image générée par le logiciel PyMol.[167]**

Le premier motif mis en évidence correspond au motif Hélice-Coude-Hélice ('Helix-Turn-Helix' - HTH) et est utilisé par les régulateurs de transcription et les enzymes aussi bien chez les procaryotes que chez les eucaryotes. Les deux hélices du motif sont reliées entre elles par un turn β d'environ 4 résidus et, bien que la première hélice puisse interagir avec l'ADN, c'est la deuxième hélice (appelée 'recognition helix') qui interagit de manière spécifique en

s'insérant dans le sillon majeur. Le reste de la protéine est de composition et de structure fort variable.

Le deuxième groupe est composé de protéines liant le zinc c'est à dire que le motif est caractérisé par la coordination de un ou deux atomes de zinc par des résidus conservés : cystéines et histidines. Ce motif intervient principalement dans les facteurs de transcription eucaryotiques. Les motifs en « doigt de zinc » ('zinc finger') sont composés d'un court feuillet β antiparallèle à deux brins et d'une hélice α (Figure I-24 B) et constituent la classe la plus abondante de protéines liant l'ADN dans le génome humain.[24] La deuxième famille liant le zinc est la famille des récepteurs d'hormones. Ces protéines, après s'être liées à une hormone, se déplacent jusqu'au noyau pour transcrire la séquence cible. Le motif est constitué de deux hélices perpendiculaires et de deux boucles, chaque paire hélice-boucle liant un atome de zinc à l'aide de quatre cystéines.

Le troisième groupe reprend les « tirettes à leucine » ('leucine zipper') et les motifs Hélice-Boucle-Hélice ('Helix-Loop-Helix' - HLH). Ces deux motifs utilisent un système de dimérisation entre deux hélices qui interagissent avec le sillon majeur de l'ADN. Les tirettes à leucine utilisent une structure de type 'coiled coil' pour lier les deux hélices. Celles-ci sont composées d'une partie de type 'tirette' contenant une leucine (ou une valine ou une isoleucine) tous les 8 résidus. De leur côté, les motifs HLH dimérisent également par une structure 'coiled coil' mais sont plus flexibles suite à l'insertion d'une boucle entre la zone de liaison et la zone de dimérisation (Figure I-24 C). Le dimère est donc composé de quatre hélices, deux boucles et une structure 'coiled coil'.

En plus de ces principales familles utilisant des hélices α pour interagir avec le sillon majeur de l'ADN, certains motifs permettent de faire interagir une hélice α avec le sillon mineur d'un ADN, généralement distordu.

Certaines protéines interagissent à l'aide de brins β ou de feuillets β. Parmi celles-ci, la principale famille représentée est la famille des protéines de type TATA box qui constituent un élément essentiel du complexe multi-protéique d'initiation de la transcription.

Finalement, les histones interagissent de manière très étroite avec l'ADN au niveau du nucléosome. Néanmoins, elles ne possèdent pas de motif particulier d'interaction.

Le regroupement des motifs protéiques liant l'**ARN** s'est limité à une description au cas par cas et/ou à une comparaison aux interactions avec l'ADN[168] jusqu'en 2001, où une nouvelle classification a été proposée.[169] Trois classes principales peuvent être dégagées : les motifs de reconnaissance de l'ARN (RRM - 'RNA-Recognition Motif'), les domaines

d'homologie K (KH - 'K-Homology domain') et les domaines de liaison à l'ARN double brins (dsRBD - 'double stranded RNA Binding Domain').

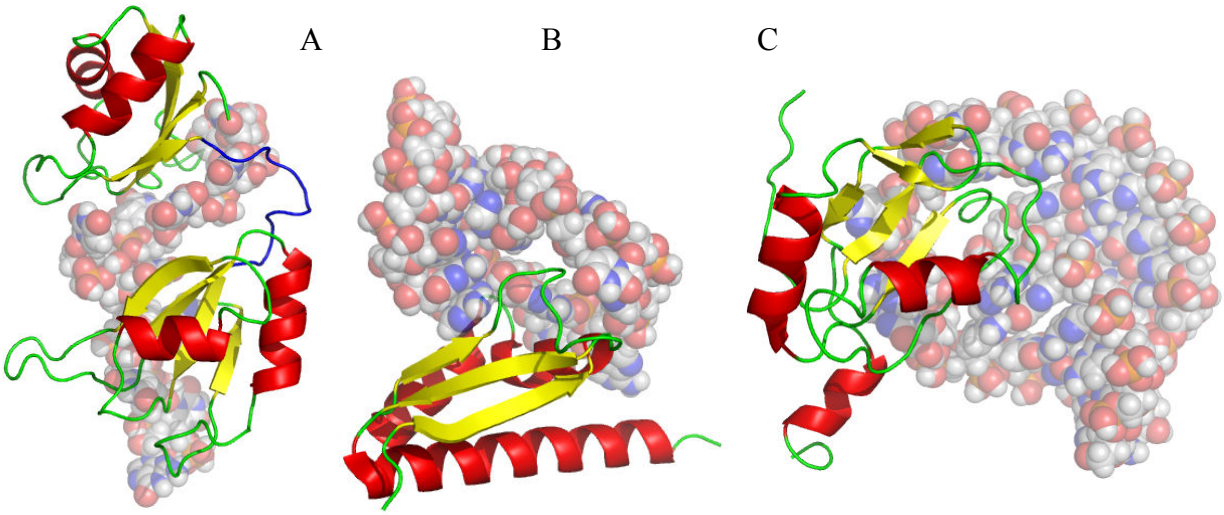



**Figure I-25 : Représentation 3D de complexes protéine-ARN (adaptées de Perez-Canadillas & Varani).[169] L'ARN est représentée en vrai volume et couleurs. Les protéines sont colorées selon leurs structures secondaires : hélices en rouge et brins en jaune. A : Motifs RRM (4 brins et 2 hélices) de la protéine Sxl de la drosophile (1B7F).[170] Le linker entre les deux motifs RRM est représenté en bleu. B : Motifs en domaine KH de la protéine Nova (1EC6)[171] interagissant avec une « épingle à cheveux » ('hairpin') d'ARN. Notez que deux boucles (en vert) interagissent également avec l'ARN). C : Motifs dsRBD de la protéine Staufen de la drosophile (1AUD)[172] interagissant avec une « épingle à cheveux » ('hairpin') d'ARN. De nouveau, les boucles sont impliquées dans les interactions avec l'ARN. Image générée par le logiciel PyMol.[167]**

Les RRM ou ribonucléoprotéines (RNP) sont les domaines les mieux caractérisés et les plus répandus. Ils sont constitués d'un feuillet β à 4 brins accolé à deux hélices α (Figure I-25 A). C'est la deuxième hélice qui joue le rôle le plus important dans la reconnaissance de l'ARN. En effet, les résidus la composant se trouvent dans une structure variable dans la protéine libre et prennent la configuration en hélice α seulement en interaction avec l'ARN.

Les domaines KH possèdent la même structure que les RRM et que les dsRBD avec en plus un feuillet β anti-parallèle recouvrant la surface des hélices α. La boucle reliant les deux hélices ainsi qu'une des boucles du feuillet β anti-parallèle semblent être impliquées dans la reconnaissance de l'ARN (Figure I-25 B). Elles interagissent notamment avec le squelette sucre-phosphate de l'ARN.

Le troisième motif le plus représenté correspond aux dsRBD. Ces motifs reconnaissent sélectivement les ARN double brins mais peuvent aussi se lier à de l'ADN double brins ou à des hybrides ADN-ARN. De plus, les dsRBD ne lient pas de séquences spécifiques d'ARN et

le fonctionnement biologique de ces motifs n'est pas encore décrit avec précision bien que deux boucles hautement conservées semblent avoir un rôle à jouer.

Finalement, il semblerait que certaines structures protéiques soient favorisées en interactions avec l'ADN et l'ARN. C'est notamment le cas des hélices gauches de type poly-proline.[173]

Remarque : certaines méthodes ont été mises au point pour tenter de détecter ces motifs sur des structures protéiques de fonction inconnue.[174,175]

**Etat des lieux**

En résumé, les interactions avec les acides nucléiques peuvent donc être divisées en trois étapes principales en partie confirmées par des études sur l'énergie des interactions protéine-acide nucléique[176] :

- Des interactions peu spécifiques agissent tout d'abord entre les acides aminés chargés positivement et les phosphates chargés négativement des acides nucléiques. Ces interactions permettraient une pré-orientation des protéines et la stabilisation des complexes.
- Les interactions spécifiques ont lieu entre les chaînes latérales des acides aminés et les bases nucléotidiques, elles sont principalement composées de liens H. Les motifs protéiques présentés dans la section précédente favorisent l'accès au sillons et donc aux bases nucléotidiques. Des liens H modulés par l'eau sont eux aussi très nombreux mais ne sont pas spécifiques.
- La courbure de l'ADN est influencée par ces interactions et permet de tendre vers une configuration de type A-ADN. Cette configuration autorise un accès plus facile au sillon mineur et, par voie de conséquence, augmente la possibilité de créer des interactions hydrophobes.

**Remarques**

- Dans tous les travaux présentés ci-dessus, les auteurs apportent une information plus ou moins complète sur les fréquences des acides aminés se trouvant à l'interface, les paires de résidus en interaction et les structures secondaires impliquées. Ces résultats seront résumés en comparaison à ceux obtenus lors de ce travail dans le Chapitre IV : Discussion Générale.

- Un certains nombre de méthodes permettant d'analyser la conformation des complexes protéine-acide nucléique sont disponibles sur Internet (3DNA,[11] NUCPLOT,[177] ENTANGLE[135]).

## *I.5. Méthodes de prédiction des sites d'interaction*

Avec l'intérêt porté aux interactions protéiques et le grand nombre de résultats engrangés, sont apparues différentes méthodes qui ont pour but de prédire la position des sites d'interaction. Les méthodes utilisées sont nombreuses et utilisent des alignements de séquences et autres arbres d'évolution, et/ou des informations provenant de la séquence et/ou de la structure tridimensionnelle (cf. Tableau I-1 et Tableau I-2). Il serait très fastidieux de décrire ici toutes les méthodes existantes et nous allons donc décrire ci-dessous un échantillon représentatif de ces méthodes.

| Référence | Année de publication | Méthode | SEQUENCE | HOMOLOGIE | Profil de séquence | Evolutionary tracing | STRUCTURE | Surface/Accessibilité | Structure secondaire | Alignement de structure |
|---|---|---|---|---|---|---|---|---|---|---|
| **Kini & Evans[120,178]** | **1996** | **-** | **OUI** | **NON** | | | **NON** | | | |
| **Gallet *et al.*[179] - RBD** | **2000** | **-** | **OUI** | **NON** | | | **NON** | | | |
| **Ofran & Rost[180]** | **2003** | **NN** | **OUI** | **NON** | | | **NON** | | | |
| Koike & Takagi[181] | 2003 | SVM | OUI | NON | | | OUI | x | | |
| Hoskins *et al.*[182] | 2006 | - | OUI | NON | | | OUI | x | x | |
| Murakami & Jones[183] | 2006 | - | OUI | NON | | | OUI | x | | |
| Kufareva *et al.*[184] | 2007 | LR | OUI | NON | | | OUI | x | | |
| Porollo & Meller[185] | 2007 | NN | OUI | NON | | | OUI | x | | |
| Koike & Takagi[186] - a | 2004 | SVM | OUI | OUI | x | | NON | | | |
| Res *et al.*[187] - a | 2005 | SVM | OUI | OUI | x | | NON | | | |
| Res *et al.*[187] - b | 2005 | SVM | OUI | OUI | | x | NON | | | |
| Chen & Zhou[188] - a | 2005 | NN | OUI | OUI | x | | NON | | | |
| Zhou & Shan[189] | 2001 | NN | OUI | OUI | x | | OUI | x | | |
| Fariselli *et al.*[31] | 2002 | NN | OUI | OUI | x | | OUI | x | | |
| Koike & Takagi[186] - b | 2004 | SVM | OUI | OUI | x | | OUI | x | | |
| Bradford & Westhead[190] | 2005 | SVM | OUI | OUI | x | | OUI | x | | |
| Bordner & Abagyan[191] | 2005 | SVM | OUI | OUI | | x | OUI | x | | |
| Chen & Zhou[188] - b | 2005 | NN | OUI | OUI | x | | OUI | x | | |
| Chung *et al.*[192] | 2006 | SVM | OUI | OUI | x | | OUI | x | | x |
| Wang *et al.*[193,194] | 2006 | SVM | OUI | OUI | x | x | OUI | x | | |
| Dong *et al.*[195] | 2007 | SVM | OUI | OUI | x | | OUI | x | | |
| Li *et al.*[196] | 2007 | CRF | OUI | OUI | x | | OUI | x | | |

**Tableau I-1 : Informations utilisées par les méthodes de prédiction des sites d'interaction entre protéines. Les méthodes n'utilisant que la séquence sont données en rouge. SVM = 'Support Vector Machine', NN = 'Neural Network', LR = 'Linear Regression' et CRF = 'Conditional Random Field'.**

| | Référence | Année de publication | Méthode | SEQUENCE | HOMOLOGIE | Profil de séquence | Evolutionary tracing | STRUCTURE | Surface/Accessibilité | Structure secondaire |
|---|---|---|---|---|---|---|---|---|---|---|
| ADN | Mandel-Gutfreund & Margalit[197] | 1998 | - | OUI | NON | | | NON | | |
| ADN | Ahmad *et al.*[198] - a | 2004 | NN | OUI | NON | | | NON | | |
| ARN | Terribilini *et al.*[199] | 2006 | NB | OUI | NON | | | NON | | |
| ADN | Kuznetsov *et al.*[200] - a | 2006 | SVM | OUI | NON | | | NON | | |
| ADN ARN | Wang & Brown[201] | 2006 | SVM | OUI | NON | | | NON | | |
| ADN | Yan *et al.*[202] - a | 2006 | NB | OUI | NON | | | NON | | |
| ADN | Ahmad *et al.*[198] - b | 2004 | NN | OUI | NON | | | OUI | x | x |
| ARN | Jeong *et al.*[203] | 2004 | NN | OUI | NON | | | OUI | x | |
| ADN | Bhardwaj & Lu[204] | 2007 | SVM | OUI | NON | | | OUI | x | x |
| ADN | Ahmad & Sarai[205] | 2005 | NN | OUI | OUI | x | | NON | | |
| ARN | Kim *et al.*[206] | 2006 | - | OUI | OUI | x | | NON | | |
| ADN | Yan *et al.*[202] - b | 2006 | NB | OUI | OUI | x | | NON | | |
| ADN | Ho *et al.*[207] | 2007 | SVM | OUI | OUI | | x | NON | | |
| ADN | Ofran *et al.*[208] | 2007 | SVM | OUI | OUI | x | | NON | | |
| ADN | Jones *et al.*[209] | 2003 | - | OUI | OUI | x | | OUI | x | |
| ADN | Kuznetsov *et al.*[200] - b | 2006 | SVM | NON | OUI | x | | OUI | x | x |
| ADN | Tjong & Zhou[210] | 2007 | NN | OUI | OUI | x | | OUI | x | |

**Tableau I-2 : Informations utilisées par les méthodes de prédiction des sites d'interaction entre protéines et acides nucléiques. Les méthodes n'utilisant que la séquence sont données en rouge. SVM = 'Support Vector Machine', NN = 'Neural Network' et NB = 'Naive Bayes classifier'.**

### I.5.1. Kini et Evans

La première méthode de prédiction des sites d'interaction date de 1996. Elle se base uniquement sur la séquence des protéines. En effet, l'étude de certaines suites d'acides aminés a permis à Kini et Evans[120,178] de mettre en évidence la présence de prolines dans les segments entourant les zones en interactions et à une distance un peu plus grande, de cystéines. La proline a 2,5 fois plus de chance de se trouver dans ce type de segment que dans le reste de la protéine. Elle semble agir comme un casseur de structures secondaires et permettrait au site d'interaction de garder sa conformation et son intégrité ce qui est indispensable à la spécificité de la liaison. Si, dans une séquence, on trouve de 3 à 7 résidus entourés de deux prolines, on peut considérer que ces résidus ont une grande chance d'être impliqués dans une interaction.

L'avantage principal de cette méthode est qu'elle semble être robuste c'est-à-dire applicable à de nombreux types de complexes. Néanmoins, peu de valeurs permettant

d'évaluer la qualité de cette méthode sont données dans l'article de référence et cette méthode se base sur des fragments réduits de la protéine (site d'interaction et quelques résidus avoisinants). Il est donc difficile de savoir quel est la qualité de la méthode et quelle serait l'influence du reste de la séquence sur la précision de la méthode.

### I.5.2. Receptor Binding Domains

En 2000, une méthode uniquement basée sur la séquence, elle aussi, a été présentée par Gallet *et al.*[179] Cette méthode utilise l'hydrophobicité des acides aminés pour construire des graphiques d'Eisenberg[211] et se base sur le concept expérimental des 'Receptor-Binding Domain' (RBD).[212] Le principe de cette méthode sera expliqué plus précisément au point II.5. Les RBD sont enrichis en acides aminés chargés (arginine et lysine) et polaires.

Les résultats obtenus sont, en moyenne, de 59% à 80% et sont encore supérieur pour les interactions avec l'ADN (~95%). Néanmoins, ces résultats représentent le pourcentage de fragments protéiques connus expérimentalement pour être impliqué dans les interactions dans lesquels un RBD est prédit. Dès lors, il n'est pas exclu que d'autres RBD soient prédits ailleurs dans la séquence ce qui diminuerait la spécificité obtenue par l'introduction d'un grand nombre de faux positifs. De plus, la méthode s'avère inefficace pour la détection de sites hydrophobes.

### I.5.3. Réseaux neuronaux

Avec les machines à vecteur de support (cf. paragraphe suivant), les réseaux neuronaux ('Neural Network' - NN) sont les systèmes d'intelligence artificielle les plus utilisés en biologie et pour la prédiction des sites d'interaction (cf. Tableau I-1). Globalement, un réseau de neurones est un graphe dont les sommets (neurones) ont une capacité à transformer un signal d'entrée en un signal de sortie. Les connexions sont pondérées et servent à transférer les signaux d'un (ou plusieurs) sommet(s) vers un (ou plusieurs) autre(s). On distingue les couches d'entrée et de sortie ainsi qu'un nombre variable de couches cachées (Figure I-26). Par ailleurs les connexions peuvent être équipées de fonction de *retard* et/ou de *retour en arrière* permettant ainsi une plus grande souplesse du modèle.

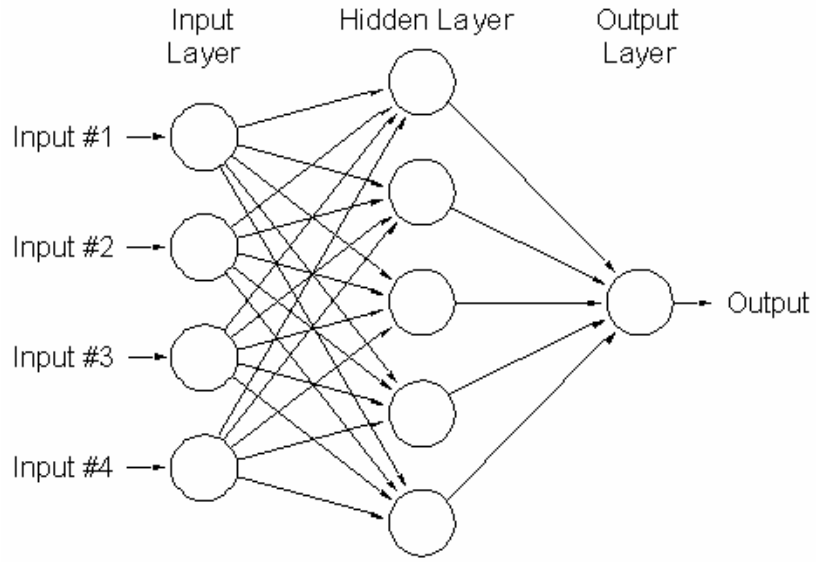



**Figure I-26: Représentation schématique d'un réseau neuronal : données d'entrée (input#), couches d'entrée (input layer), couche cachée (hidden layer) et couche de sortie (output layer) et donnée de sortie (output).**

Dans leur travail, Ofran & Rost[180] utilisent un réseau neuronal dont l'originalité est d'utiliser une fenêtre de 9 acides aminés comme entrée du réseau pour un total de 189 entrées. Les résultats obtenus sont encourageants dans l'optique de la mise au point d'une méthode de prédiction se basant sur la séquence uniquement avec des résultats s'écartant des résultats obtenus aléatoirement. Néanmoins, les résultats restent encore très faibles. En effet, le système permet difficilement d'optimiser à la fois la sensibilité (pourcentage de sites d'interaction détectés ; 30%) et la spécificité (pourcentage de prédictions correspondant réellement à un site d'interaction ; de 20 à 42%).

### I.5.4. Homologie de séquence et machines à vecteur de support

La méthode présentée dans ce paragraphe[186] utilise les deux principes suivants :

- Les « profils de séquence ». Afin de créer ces profils, il est nécessaire de trouver des protéines homologues. Pour cela, des banques de données de séquences (UniProt/SwissProt ou NCBI) vont être scannées (PSI-BLAST). Ensuite, les alignements construits entre les protéines homologues vont permettre d'établir des matrices de substitution qui informent sur les résidus qui sont les plus/moins variables c'est-à-dire les résidus qui sont les moins/plus conservés. On parle dès lors de « profil de séquence » ('sequence profile'). Remarque : dans certains cas,

des arbres d'évolutions sont construits afin de regrouper les homologues en catégories phylogénétiques ('evolutionnary tracing')(Tableau I-1 et Tableau I-2).

- Les machines à vecteurs de support ('Support Vector Machines' - SVM). Les SVM sont des méthodes de classification binaire par apprentissage supervisé, qui furent introduites en 1995 par Vapnik.[213] Elles sont basés sur l'utilisation de fonction dites à noyau ('kernel') qui permettent une séparation optimale des données. Les applications les plus courantes sont la classification de documents, la reconnaissance 3D (biométrie)... L'application à la prédiction des sites d'interaction des SVM est courante et, particulièrement, quand la structure des protéines est utilisée (cf. Tableau I-1 et Tableau I-2).

La méthode de Koike et Takagi[186] utilise les deux principes présentés ci-dessus de la manière suivante : les « profils de séquences » vont permettre de récolter des informations spécifiques aux interfaces et ensuite, ces informations seront introduites dans le système de machines à vecteur de support (SVM). Il est intéressant de voir que les meilleurs résultats sont obtenus quand le pourcentage de résidus en interaction est connu (et, dans une moindre mesure, prédit). La fraction de résidus en interaction est donc un paramètre influençant la qualité de la méthode de prédiction. De nouveau, bien que les résultats s'écartent de ce qui aurait été obtenu de manière aléatoire, la précision est seulement de l'ordre de 36%. Remarque : dans le même article, les auteurs ont créé plusieurs modèles dont certains utilisent des informations provenant de la structure tridimensionnelle. Les résultats repris ci-dessus correspondent aux résultats obtenus uniquement sur base de la séquence.

### I.5.5. Structure tridimensionnelle

Finalement, de nombreuses méthodes de prédiction utilisent les informations provenant de la structure tridimensionnelle des protéines (Tableau I-1 et Tableau I-2). Dans ce cas, des patchs d'interactions sont définis et permettent d'isoler les acides aminés voisins (dans l'espace) des résidus en interactions. Régulièrement, les acides aminés de surface sont extraits par calcul de la surface accessible (ASA) et parfois, on a recours à des alignements de structures ou à la reconnaissance des structures secondaires.

## *I.6. Objectifs de la Thèse de Doctorat*

Alors que nous sommes entrés à grands pas dans l'ère de la post-génomique, de la protéomique, et des méthodes à haut débit, un nombre de plus en plus conséquent de données biologiques sont disponibles pour les scientifiques du monde entier. Mais, en plus de la difficulté technique de rassembler toutes ces informations dans des banques facilement utilisables, la recherche d'informations pertinentes à partir de ces banques reste un problème difficile à résoudre. Dans le cadre des interactions protéiques, de plus en plus de recherches se penchent sur le problème de l'interactome c'est-à-dire de la détection des interactions physiques protéine-protéine de l'ensemble du protéome et donc, du ou des partenaires de l'interaction parmi de nombreux candidats possibles. Un deuxième niveau de connaissance peut ensuite être atteint par la détermination du site d'interaction proprement dit.

La thèse de doctorat ici présentée a pour but de contribuer à l'étude des interfaces protéiques en deux étapes : premièrement, par la compréhension des mécanismes mis en jeu au niveau atomique et deuxièmement, par la mise au point d'une méthode de prédiction des sites d'interaction sur base de la séquence uniquement.

Afin d'analyser les interfaces, des banques de données de complexes protéine-protéine et protéine-acide nucléique issus de la PDB[214] seront construites. Dans le cadre de ces analyses, la méthode Pex[14,215] sera optimisée et une méthodologie optimisée de sélection des acides aminés en interaction sera mise au point. Ceci nous permettra d'étudier avec précision la grande quantité d'informations contenue dans notre banque de structures 3D. Nous allons ensuite réaliser une étude statistique sur ces complexes en analysant en détail les propriétés des acides aminés en interaction afin de mettre en évidence les particularités de ceux-ci. Cette analyse statistique de structures de complexes protéiques devrait nous permettre de poser les bases quant aux propriétés fondamentales nécessaires à une interaction optimale.

A l'aide de ces résultats, une méthode de prédiction des sites d'interaction à partir de la séquence sera développée. La méthode sera mise au point sur base d'un modèle de régression logistique et la qualité de la méthode sera analysée et comparée aux résultats trouvés dans la littérature.

# II. MATÉRIEL ET MÉTHODES

## *II.1. Banques de Données*

### II.1.1. Sélection des complexes

Les différentes structures de complexes protéiques utilisées dans ce travail ont été extraites de la 'Protein Data Bank' (PDB).[214] La PDB est accessible gratuitement via un site Internet (http://www.rcsb.org/pdb/home/home.do) et permet d'accéder à la structure tridimensionnelle de nombreuses molécules, que ce soient des protéines, des acides nucléiques, des phospholipides ou des substrats, sous forme monomérique ou complexée. Lors des recherches effectuées sur la PDB, nous avons extrait les structures obtenues par rayons X répondant aux critères suivants :

- Structure contenant au moins deux chaînes protéiques et de résolution inférieure ou égale à 2.5Å.

OU

- Structure contenant au moins une chaîne protéique et une chaîne de nucléotides (ADN ou ARN) et résolution inférieure ou égale à 3Å.

Cette première sélection a conduit à l'extraction de 9131 complexes protéine-protéine, 564 complexes protéines-ADN et 128 complexes protéines-ARN.

Néanmoins, dans ces complexes, de nombreux homologues et certaines redondances (complexes identiques ayant des codes PDB différents) étaient présents. Afin d'obtenir des banques de données reprenant des complexes de séquences non-homologues, nous avons utilisé le programme PISCES.[216] Ce programme, accessible librement sur Internet (http://www.fccc.edu/research/labs/dunbrack/pisces), permet de détecter des séquences peptidiques ayant une homologie supérieure à un certain pourcentage en basant ses calculs sur des alignements locaux. Dans ce travail, nous avons exclu les chaînes ayant une homologie supérieure à 30%. Ce seuil a été choisi car au-delà de 30% d'homologie, il est communément admis que deux séquences auront des structures 3D similaires. Ensuite, les complexes provenant d'une co-cristallisation ont été mis en évidence par le serveur PQS[217] puis supprimés. Finalement, nous avons observé manuellement les différents complexes sélectionnés pour détecter les fichiers PDB contenant plusieurs unités biologiques et supprimer ces répétitions.

Ces différents critères (résumés dans la Figure II-1) nous ont permis d'arriver aux banques de données de référence qui contiennent 1.297 complexes protéine-protéine, 139

complexes protéine-ADN et 49 complexes protéine-ARN. Le code PDB de chacun des ces complexes est repris en Annexe 2.

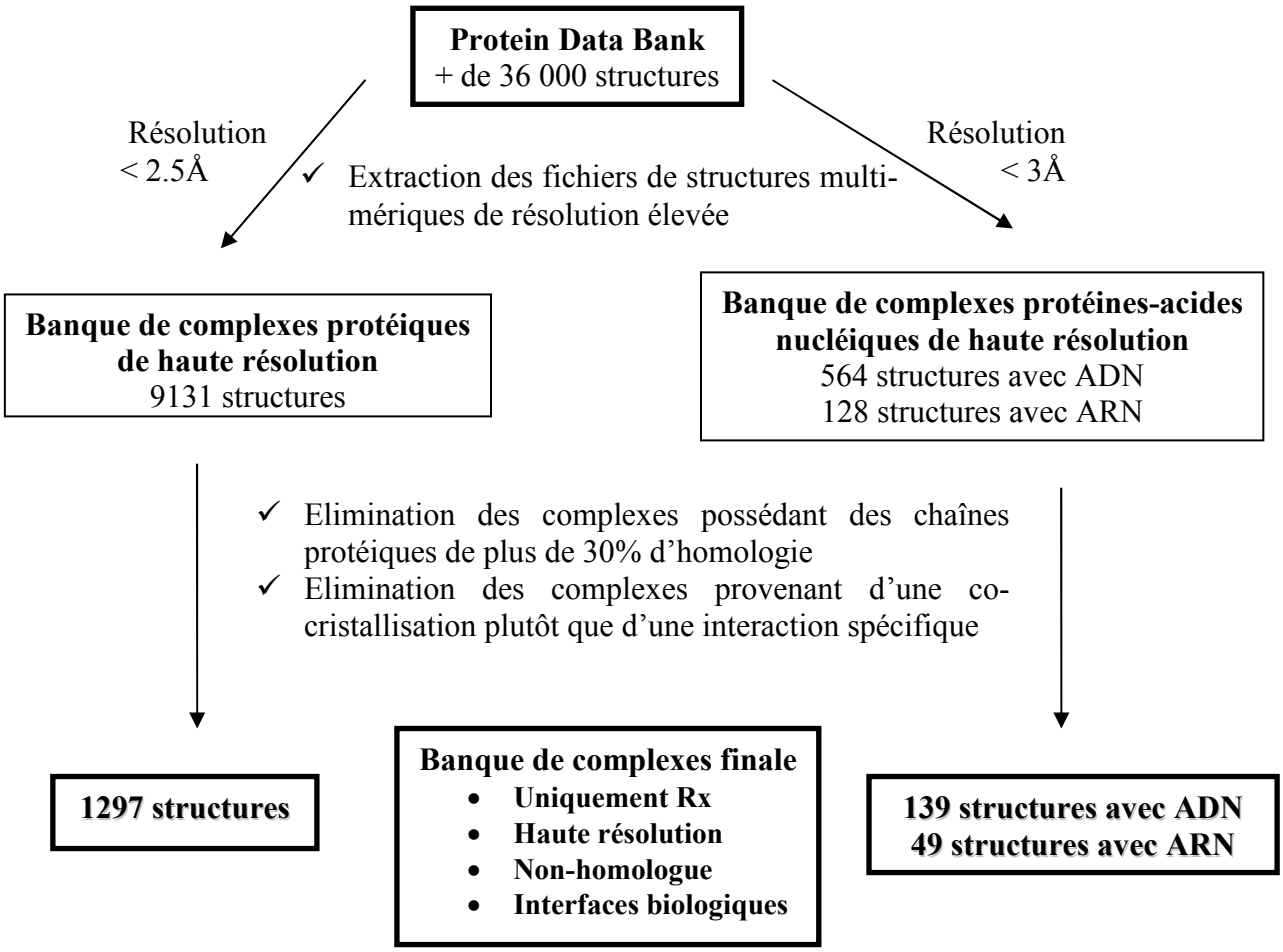



**Figure II-1 : Etapes de construction de la banque finale de complexes protéiques.**

### II.1.2. Préparation des complexes en vue de leur utilisation dans Z-Ultime

Les fichiers PDB sont des fichiers texte contenant des informations indispensables sur la protéine ou le complexe étudié (description biologique, chaînes impliquées, mode de cristallisation…), des remarques diverses et les coordonnées spatiales (x, y, z) de chacun des atomes constitutifs des molécules.

Ces fichiers ne peuvent pas être importés tels quels dans le programme Z-Ultime (cf. point II.2.1) et ils ont premièrement été 'nettoyés' : toutes les lignes non-nécessaires à la construction de la molécule ainsi que la plupart des hétéroatomes (principalement des résidus de solvant de cristallisation comme p.ex. l'eau, certains ions, l'ATP, le GDP…) ont été éliminés en prenant soin de ne pas enlever ceux qui font partie intégrante de la molécule. Certains hétéroatomes sont des acides aminés ou des bases modifiés, les nouveaux types

atomiques présents dans ces hétéroatomes ont été ajoutés à notre liste des types atomiques en vue de leur utilisation dans la définition des types d'interactions (cf. point II.3.1).

Les atomes d'hydrogènes qui ne sont pas présents dans les fichiers PDB ont été placés à l'aide du programme Reduce.[218] Ce sont ces fichiers qui seront utilisés par Z-Ultime.

### II.1.3. Classification des complexes protéine-protéine

Sur base du travail d'Ofran et Rost,[99] pour chaque couple de chaines protéiques en interaction, nous avons déterminé si l'interaction était transitoire ou permanente. En effet, si deux protéines possèdent des codes SwissProt/UniProt différents, on peut supposer qu'elles sont fonctionnelles par elles-mêmes (complexe non-obligatoire) et elles sont considérées comme ayant une interaction transitoire. En effet, des codes différents signifient que les protéines ont été découvertes et étudiées séparément. Inversement, si deux protéines possèdent le même code SwissProt/UniProt, elles ont été étudiées comme un seul élément biologique (complexe obligatoire) et peuvent être considérées comme interagissant de manière permanente. Nous avons aussi différencié les complexes faits de deux séquences identiques ('homomer') ou différentes ('hétéromer'). On parlera donc d'interfaces hétéromer-permanentes, hétéromer-transitoires et homomériques. Remarques :

- Environ 13% des interactions étudiées n'ont pas pu être classées par manque d'information SwissProt/UniProt.
- Étant donné la composition de nos complexes (jusque 10 chaines différentes), plusieurs types d'interfaces peuvent être trouvés dans le même complexe (PDB).
- Dans notre banque de données, les interfaces homomériques correspondent toujours à des complexes permanents.
- Comme signalé dans l'introduction (cf. point I.4.3 : Classification des interfaces), la correspondance entre complexe obligatoire/non-obligatoire et complexe permanent/transitoire n'est pas absolue mais a été considérée comme telle lors de notre classification automatique.

## *II.2. Programmes Utilisés*

### II.2.1. Programme Z-Ultime et fichiers Pex

Les outils Pex[14,215] ont été créés par le CBMN afin de faciliter la comparaison et le regroupement des informations contenues dans la PDB. L'utilisation du programme Z-Ultime permet de générer ce type de fichier dans lesquels les lignes représentent les différents acides aminés et les colonnes correspondent aux nombreux paramètres calculés. En général, les trois premières colonnes identifient les résidus de la protéine depuis l'extrémité N-terminale (première ligne) jusqu'à l'extrémité C-terminale (dernière ligne). Dans notre cas, un seul fichier comprendra les informations sur les différentes chaînes de tous les complexes de la banque. Un exemple d'une partie d'un fichier Pex est présenté dans le Tableau II-1. Le programme permet de calculer un grand nombre d'autres paramètres selon la demande de l'utilisateur. De plus, à partir des fichiers générés, il est possible d'extraire les lignes ou colonnes d'intérêt ce qui est très utile pour, par exemple, sélectionner les différents types d'acides aminés.

| Numéro du résidu | Code 1 lettre | Code 3 letres | Nom du PDB | Nom de la chaîne | Coordonnée (x) du carbone alpha | Coordonnée (y) du carbone alpha | Coordonnée (z) du carbone alpha | Structure secondaire | Distance d'interaction | Paire de résidus en interaction | Atome en interaction (i) | Atome en interaction (j) | Surface accessible au solvant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | M | MET | 1A0A.ent | A | 4,785 | -2,489 | 57,15 | --- | 4,591 | MET_&___C | HE_ | O2P | 177,4 |
| 2 | K | LYS | 1A0A.ent | A | 4,43 | -0,952 | 53,67 | C | 2,813 | LYS_&___T | HZ_ | H5M | 143,1 |
| 3 | R | ARG | 1A0A.ent | A | 8,138 | -1,715 | 53,03 | C | 2,133 | ARG_&___C | HH2 | H5_ | 182,5 |
| 9 | E | GLU | 1A0A.ent | A | 7,945 | 8,055 | 45,74 | H3 | 1,86 | GLU_&___C | OE1 | H5_ | 79,01 |
| 10 | Q | GLN | 1A0A.ent | A | 6,296 | 9,705 | 42,72 | H3 | 1,973 | GLN_&___C | HG_ | O1P | 96,18 |
| 50 | G | GLY | 1A3Q.ent | A | 6,245 | 58,61 | 14,27 | C | 4,785 | GLY_&___G | O__ | O1P | 35,5 |
| 52 | R | ARG | 1A3Q.ent | A | 6,704 | 64,36 | 11,11 | B | 1,924 | ARG_&___G | HH2 | O6_ | 89,78 |
| 54 | R | ARG | 1A3Q.ent | A | 7,017 | 70,08 | 8,594 | Ba | 2,104 | ARG_&___G | HH2 | N7_ | 52,05 |

**Tableau II-1 : Exemple simplifié de fichier Pex.**

Lors de ce travail, parmi les nombreux paramètres calculés (une vingtaine de colonnes), nous allons nous intéresser particulièrement aux colonnes représentant les résidus

considérés, les résidus en interaction, la distance d'interaction, les noms des atomes impliqués et les structures secondaires.

**Définition des structures secondaires**

Dans les fichiers PDB, aucune protéine ou complexe protéique ne possède de description complète de sa structure secondaire excepté dans l'entête du fichier. Dès lors, le programme Zpex va redéfinir l'ensemble des structures secondaires du cristal étudié en se basant sur la valeur des angles Φ et ψ ainsi que sur la présence de liens H sur le groupement N-H de la chaîne principale (distance maximale entre le donneur et l'accepteur de 3,5Å).

Les différents types de structures secondaires présentés dans l'introduction (point I.2.2) sont définis selon les paramètres suivants :

- Les hélices sont caractérisées par des angles Φ/ψ compris dans un cercle de 45° autour du couple Φ = -57° et ψ = -47°. Le type d'hélice est fonction du lien H et de l'écart séparant les deux résidus impliqués. L'hélice α possède un lien H entre le résidu n et n + 4, l'hélice $3^{10}$ entre le résidu n et n + 3, l'hélice π entre le résidu n et n + 5 et l'hélice ω entre le résidu n et n + 6.
- Les structures β sont divisées en brins β, feuillets β parallèles et feuillets β antiparallèles. Les structures β sont caractérisées par des angles Φ/ψ compris dans un cercle de 90° autour du couple Φ = -129° et ψ = 123°. Les brins β ne possèdent pas de lien H ou forment un lien H avec une structure non-β. Les feuillets β sont constitués de différents brins β reliés entre eux par des liens H et on utilise un vecteur reliant les carbones α entre eux pour déterminer le caractère parallèle ou anti-parallèle.
- Les turns sont définis selon la nomenclature de Srinivasan.[219] Les random coils (structures désordonnées) couvrent une large gamme d'angles Φ/ψ et comprennent notamment les hélices gauches, les hélices droites et les feuillets β avec un lien H entre le résidu n et n + i (i > 6) et n et n + j (j < 3) respectivement.

### II.2.2. Autres programmes

Les programmes YAGME et PyMOL[167] ont été utilisés pour permettre la visualisation en trois dimensions des complexes protéiques. Cette visualisation en trois dimensions a permis d'observer dans l'espace les protéines étudiées et de clarifier certaines hypothèses émises à partir de l'analyse des fichiers Pex. De plus, grâce à ce logiciel, certaines interactions spécifiques sont présentées en figure dans le Chapitre III : Résultats.

SAS Enterprise Guide est un outil statistique qui a été utilisé pour la construction de modèle par régression logistique. Ce programme a l'avantage de permettre la construction du modèle logistique et la sélection des variables d'intérêt ce qui n'est pas le cas de tous les logiciels de statistique (p.ex. Minitab).

## *II.3. Démarche Suivie*

### II.3.1. Définitions et sélections

**Types d'acides aminés**

Pour faciliter l'analyse des résultats, les différents acides aminés ont été regroupés en deux familles et cinq sous-familles. Cette classification se base sur l'échelle d'hydrophobicité consensus d'Eisenberg.[211] Parmi les résidus hydrophobes, nous distinguerons les résidus dits aliphatiques (Ala, Ile, Leu, Met et Val) des acides aminés aromatiques (Phe, Trp et Tyr) et des résidus dits particuliers (Cys, Gly et Pro). Pour les acides aminés hydrophiles, résidus chargés (Arg, Asp, Glu et Lys) et polaires (Asn, Gln, His, Ser et Thr) sont classés dans deux sous-familles distinctes.

**Types d'atomes**

De même, chaque atome a été classé selon sa position dans le résidu. Pour les acides aminés, les atomes du squelette peptidique ('backbone' - BK) et de la chaîne latérale ('side chain' - SC) ont été différenciés : N, CA, C, O, H et HA pour le squelette peptidique et tous les autres atomes pour la chaîne latérale. Du côté des acides nucléiques, trois groupes sont définis : le phosphate (P, O1P et O2P), le sucre (O2*, O3*, O4*, O5*, C1*, C2*, C3*, C4*, C5*, H1*, H2*, H3*, H4*, H5* et HO2) et la base (tous les autres atomes). La notation des atomes correspond à celle utilisée classiquement dans les fichiers de structure (PDB). Cette nomenclature est présentée dans les Annexes 3.1. et 3.2.

Ensuite, les atomes ont été regroupés selon leurs types : carbone hydrophobe (Cpho); carbone polaire (Cpolar); atome d'hydrogène lié à un carbone (Hc), à un azote (Hn), à un oxygène (Ho) ou à un soufre (Hs); atome d'hydrogène chargé (Hcharged), atome d'azote (N) ou atome d'azote chargé (Ncharged); atome d'oxygène (O) ou atome d'oxygène chargé négativement (Ocharged); atome de phosphore (P); atome de soufre (S); et hétéroatome (hetero) pour les atomes listés comme HETATM dans les fichiers PDB.

**Types d'interactions**

Dans l'ensemble des couples atome-atome extraits, cinq types d'interaction ont été différenciés. Premièrement, les ponts salins correspondent aux interactions entre atomes possédant une charge nette. Les atomes chargés négativement sont principalement les oxygènes des acides aspartiques et glutamiques et du groupement phosphate des nucléotides

tandis que les atomes positifs sont principalement les atomes des fonctions amines de la lysine et de l'arginine.

Deuxièmement, les ponts hydrogènes reprennent les liens H classiques : partage d'un hydrogène entre deux atomes électronégatifs : X-**H••Y** avec X et Y correspondant à des atomes d'azote, d'oxygène et parfois de soufre (O, Ocharged, N, Ncharged et S). Mais aussi des liens H où l'hydrogène est lié de manière covalente à un carbone (C-**H••O**). Les atomes d'hydrogène utilisés sont : Hn, Ho or Hcharged.

Le troisième type d'interaction reprend les interactions de type hydrophobe entre deux atomes peu électronégatifs et de différence d'électronégativité faible (Hc et Cpho).

Ensuite, les ponts disulfures sont exclusivement des interactions entre deux atomes de soufre (S) et les hétéro-interactions impliquent obligatoirement un hétéroatome (hetero).

Finalement, le dernier groupe contient les interactions de van der Waals. Dans notre cas, nous y avons classé toutes les interactions ne répondant pas aux critères des quatre premiers types d'interactions et elles seront aussi notées 'autres interactions de van der Waals'.

| | HC | Hcharged | HN | HO | HS | N | Ncharged | O | Ocharged | Cpho | Cpolar | S |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **HC** | pho | vdW | vdW | vdW | vdW | H bond | H bond | H bond | H bond | pho | vdW | vdW |
| **Hcharged** | vdW | vdW | vdW | vdW | vdW | H bond | H bond | H bond | salt bridge | vdW | vdW | vdW |
| **HN** | vdW | vdW | vdW | vdW | vdW | H bond | H bond | H bond | H bond | vdW | vdW | vdW |
| **HO** | vdW | vdW | vdW | vdW | vdW | H bond | H bond | H bond | H bond | vdW | vdW | vdW |
| **HS** | vdW | vdW | vdW | vdW | vdW | H bond | H bond | H bond | H bond | vdW | vdW | vdW |
| **N** | H bond | H bond | H bond | H bond | H bond | vdW | vdW | vdW | vdW | vdW | vdW | vdW |
| **Ncharged** | vdW | vdW | H bond | H bond | H bond | vdW | vdW | vdW | salt bridge | vdW | vdW | vdW |
| **O** | H bond | H bond | H bond | H bond | H bond | vdW | vdW | vdW | vdW | vdW | vdW | vdW |
| **Ocharged** | H bond | salt bridge | H bond | H bond | H bond | vdW | salt bridge | vdW | vdW | vdW | vdW | vdW |
| **Cpho** | pho | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | pho | vdW | vdW |
| **Cpolar** | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW |
| **S** | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | vdW | diS bridge |

**Tableau II-2: Définition du type d'interaction en fonction des types d'atomes en interactions (pho = contact hydrophobe, vdW = van der Waals, H bond = lien hydrogène, 'salt bridge' = pont salin et diS bridge = pont disulfure).**

## II.3.2. Résidus en interaction

### Extraction des résidus en interaction

Dans ce travail, le critère d'extraction des résidus en interaction est exclusivement lié à la distance entre deux atomes situés sur des chaines différentes.

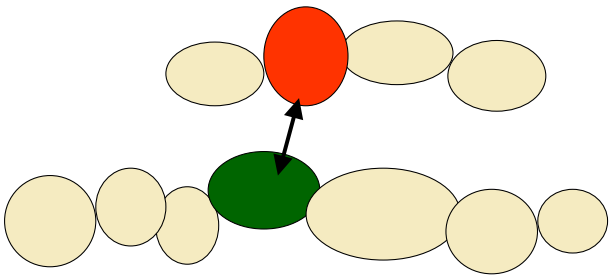


**Figure II-2 : Représentation schématique d'une interaction entre deux chaînes polypeptidiques. Les deux acides aminés en interaction sont représentés en vert et en rouge.**

Pour chaque atome de chaque résidu, le plus proche atome appartenant à une autre chaîne est détecté et la distance les séparant est calculée. Néanmoins, pour un résidu donné, un seul couple atome-atome est retenu de sorte qu'un acide aminé donné ne sera comptabilisé qu'une seule fois.

Ensuite, un cut-off de distance est appliqué et seuls les couples d'atomes les plus proches sont conservés. La distance minimale d'interaction est fixée à 1 Å car une distance inférieure est impossible physiquement pour cause de conflits stériques (voir point I.3.3). La distance maximale utilisée dans la littérature est variable : 3,5Å,[198,201,201,220] 4Å,[191,197] 4,5Å,[11,97,200,200,204,204,221] 5Å,[181,186,189,192,192,195,196,222] 6Å,[180,208,208,223] 7Å[118] (entre les carbones beta), 8Å,[224] 12Å[31,193] (entre les carbones alpha)… Certains auteurs n'emploient pas une distance cut-off fixe mais considèrent que deux résidus sont en interaction si la distance qui les sépare est inférieure à la somme de leurs rayons de van der Waals + 0,5Å[225,226] ou + 1Å.[227] Cette méthode est utilisée par d'autres[119] pour différencier les contacts forts des contacts distants en ajoutant à la somme des rayons de van der Waals des valeurs de 0,3Å et 3Å respectivement.

Les distances choisies dans ce travail sont définies au point III.1.2 - « Composition de la banque des résidus en interaction » pour les interactions avec les acides nucléiques et au point III.2.2 - « Extraction des résidus en interactions » pour les interactions protéine-protéine.

**Définition des zones de résidus en interaction**

Etant donné que les résidus en interaction ne forment pas nécessairement des fragments continus de séquence mais sont bien souvent dispersés dans celle-ci (mais proches dans la structure 3D), nous avons défini des zones d'acides aminés en interaction. La définition est la suivante : des résidus en interaction séparés d'au maximum trois résidus sont reliés dans une même zone et cette zone sera élargie de un résidu à chaque extrémité. Cette définition est illustrée ci-après :

- _ _ _ _ X _ _ _ X _ _ devient _ _ _ Z X Z Z Z X Z _
- _ _ X _ X X X _ X _ _ devient _ Z X Z X X X Z X Z _

Chaque tiret représente un acide aminé, les acides aminés en interaction sont notés par un X et les acides aminés dans une zone en interaction sont notés par un Z. Ces zones seront particulièrement utile pour la mise au point d'un modèle de régression logistique.

**Analyse et comparaison des résidus en interaction**

Lors de l'analyse des résultats, nous avons comparé les fréquences des résidus dans la banque de départ, dans la banque contenant les résidus en interaction et dans d'autres sous-banques.

La comparaison de ces différentes fréquences va être facilitée par le calcul de <u>propensions</u> :

$$P_i = \frac{I_i / \sum_i I_i}{T_i / \sum_i T_i}$$

où le numérateur correspond à la fréquence de l'acide aminé i dans la banque des résidus en interaction ($I_i$ = nombre d'acides aminés i dans la banque des résidus en interaction) et le dénominateur correspond à la fréquence du même acide aminé dans la banque totale ($T_i$ = nombre de l'acide aminé i dans la banque de départ). $\sum T_i$ = 792.015 pour les complexes protéine-protéine, 60.316 pour les complexes protéine-ADN et 28.152 pour les complexes protéine-ARN. $\sum I_i$ = 131.531 pour les interactions dans les complexes avec les protéines dont 104.782 pour les interactions protéine-protéine, 82.531 pour les interactions homomériques, 9.859 pour les interactions hétéromer-transitoires et 3.148 pour les interactions hétéromer-permanentes. $\sum I_i$ = 7.671 pour les interactions protéine-ADN et 3.367 pour les interactions protéine-ARN.

Par convention, un résidu ayant une propension supérieure à 1,2 (1,1) indique un résidu qui est (légèrement) favorisé dans les sites d'interaction. Pour les résidus ayant des valeurs de propension variant entre 0,8 et 1,1, on considèrera que la différence de fréquence est trop faible pour déterminer un comportement particulier. En dessous de 0,8, le résidu impliqué sera considéré comme étant défavorisé dans les sites d'interaction.

### II.3.3. Résidus de surface

Afin de bien déterminer si un résidu est en surface, il est important de définir deux notions: la surface accessible au solvant ('Accessible Surface Area', notée ASA) et la surface relative accessible (notée relativeASA).

Le concept de surface accessible au solvant est un concept largement employé par les scientifiques qui étudient les protéines au niveau atomique. La description de l'ASA remonte à 1971 avec les travaux et la création de l'algorithme de Lee et Richards.[228] L'ASA représente la surface sur laquelle le centre d'une molécule d'eau peut être déplacé tout en maintenant des contacts de van der Waals avec un atome et en ne pénétrant aucun autre atome. La Figure II-3 schématise ce concept : les cercles noirs représentent les acides aminés d'une protéine, le cercle contenant un W schématise une molécule d'eau et la bordure extérieure de la zone grise-ondulée représente la surface accessible au solvant. La précision du calcul de l'ASA dépend notamment du nombre de points sur lesquels on place le centre de la molécule d'eau. Dans ce travail, nous avons utilisé une méthode en 642 points[229] et le rayon de la molécule d'eau est fixé à 1,4Å.

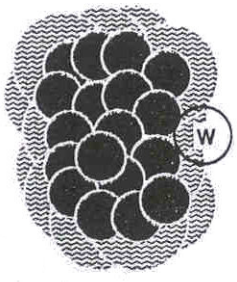



**Figure II-3 : Représentation schématique de la surface accessible au solvant (ASA).**

Ensuite, la surface relative accessible au solvant sera utilisée pour sélectionner les acides aminés de surface et internes. La relativeASA est le rapport entre l'ASA d'un résidu et son ASA totale. L'ASA totale d'un résidu est la surface accessible de ce résidu calculée par Lins *et al.*[230] en utilisant le programme Zpex et une fenêtre de trois acides aminés. Les

surfaces de référence sont celles déterminées expérimentalement sur des peptides de type Gly-X-Gly par Creighton en 1993.[2] Ce type de modèle permet d'éviter de surestimer la surface totale d'un résidu en prenant en compte le fait qu'un acide aminé est (presque) toujours lié à deux autres acides aminés par les liens peptidiques de sa chaîne principale. Dans le Tableau II-3, on peut voir que les valeurs utilisées sont proches des valeurs expérimentales.

| Code 1 lettre | Code 3 lettres | Surface selon Creighton | ASA totale | Code 1 lettre | Code 3 lettres | Surface selon Creighton | ASA totale |
|---|---|---|---|---|---|---|---|
| A | ALA | 113 | 111 | L | LEU | 180 | 179 |
| R | ARG | 241 | 250 | K | LYS | 211 | 212 |
| N | ASN | 158 | 166 | M | MET | 204 | 201 |
| D | ASP | 151 | 160 | F | PHE | 218 | 208 |
| C | CYS | 140 | 157 | P | PRO | 143 | 135 |
| Q | GLN | 189 | 194 | S | SER | 122 | 125 |
| E | GLU | 183 | 187 | T | THR | 146 | 144 |
| G | GLY | 85 | 86 | W | TRP | 259 | 249 |
| H | HIS | 194 | 191 | Y | TYR | 229 | 227 |
| I | ILE | 182 | 173 | V | VAL | 160 | 149 |

**Tableau II-3 : Valeurs (en Å$^2$) de surfaces accessibles totales expérimentales et théoriques.**

Dans la littérature, de nombreux cut-off différents de relativeASA ont été utilisés : 4%,[185] 5%,[76,78,98,185,187,209,231] 6%,[227] 10%,[115,181,186,189,189] 15%,[192,192,196,232] 16%,[31] 20%,[225] 40%[204]… Notre choix s'est porté sur une valeur de 10% (cf. point III.3.1). Cela signifie que tout résidu ayant une relativeASA supérieure à 10% sera considéré comme étant en surface et, inversement, une relativeASA inférieure à 10% décrit un résidu considéré comme interne à la protéine. La Figure II-4 représente les acides aminés sélectionnés comme étant en surface du complexe 1CSE.[233]

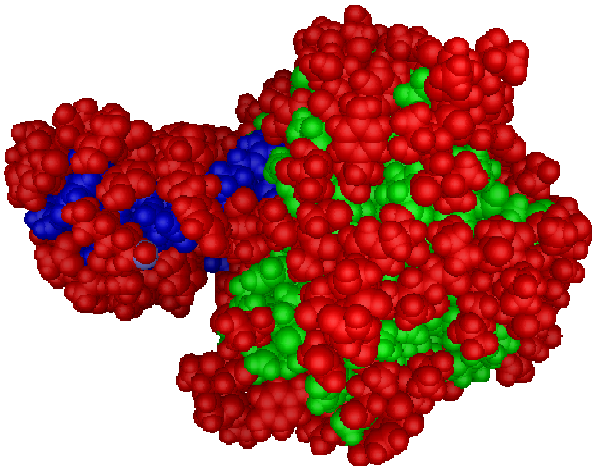


**Figure II-4 : Représentation en vrai volume d'un complexe (1CSE) protéase (en vert) – inhibiteur (en bleu). Les résidus sélectionnés comme étant en surface pour un cut-off de 10% sont représentés en rouge. Image générée par le logiciel YAGME.**

### II.3.4. Analyse des matrices d'interactions

L'analyse des couples atome-atome peut se réaliser à deux niveaux différents : au niveau des résidus ou au niveau des atomes. Dans le premier cas, l'ensemble des 210 couples acide aminé-acide aminé ou des 80 couples acide aminé-nucléotide est comptabilisé. Dans le second cas, pour un couple de résidus donné, toutes les paires atome-atome possibles sont analysées.

**Matrices d'interactions en valeurs absolues**

A partir des banques de résidus en interaction, trois matrices différentes sont construites : une pour les complexes entre protéines, une pour les complexes avec l'ADN et une pour les protéines liant l'ARN. Ces matrices permettent de détecter quelles sont les paires les plus nombreuses.

Ce point de vue (uniquement numérique), bien qu'important, n'est pas suffisant pour comprendre et discuter des différents couples présents. Pour mettre en évidence les couples qui sont favorisés au sein des interfaces biologiques (même s'ils sont peu fréquents), une analyse statistique est nécessaire et la méthode utilisée est développée au point suivant.

**Analyse statistique des matrices d'interactions**

Pour comparer les fréquences observées, il faut choisir une distribution de référence. Dans notre cas, les fréquences choisies sont celles des acides aminés et des nucléotides dans la banque totale. A partir de ces fréquences, il est possible de calculer une table de probabilités (P) dans laquelle la valeur de chaque cellule ($P_{ij}$) correspond au produit des fréquences de référence des deux résidus considérés. En multipliant ces valeurs par le nombre total de couples atome-atome observés, nous obtenons les fréquences attendues ou fréquences théoriques. D'un point de vue biologique, ces fréquences correspondent à celles qui auraient été observées si les interactions entre protéines ou entre protéines et acides nucléiques se déroulaient aléatoirement.

L'analyse statistique des matrices se base sur un test $\chi^2$ de Pearson.[234] Cette méthode a notamment déjà été utilisée dans deux études sur des complexes protéiques : Mandel-Gutfreund *et al.*,[132] et Treger & Westhof.[136] Dans un premier temps, la valeur du $\chi^2$ observé pour toute la matrice va être calculée selon la formule développée ci-après.

$$\chi^2_{obs} = \left[ \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}^2}{nP_{ij}} \right] - n$$

où $n_{ij}$ = nombre de couples entre les résidus i et j

n = nombre total de couples observés

$P_{ij}$ = probabilité du couple ij

$\chi^2_{obs}$ = khi carré observé

La réalisation du test d'indépendance se fait par la comparaison du $\chi^2$ observé au $\chi^2$ théorique. Pour les complexes entre protéines, le $\chi^2$ théorique ($\chi^2_{1-\alpha}$) vaut 433,96 pour 361 degrés de liberté [k = (p-1)(q-1) = (20-1)(20-1) = 361] et $\alpha$ = 0,005. Pour les complexes entre protéines et acides nucléiques, le $\chi^2$ théorique ($\chi^2_{1-\alpha}$) vaut 88,24 pour 57 degrés de liberté [k = (p-1)(q-1) = (20-1)(4-1) = 57] et $\alpha$ = 0,005. L'hypothèse d'indépendance est rejetée si $\chi^2_{obs} \geq \chi^2_{1-\alpha}$. Dans notre cas, il semble logique que l'hypothèse d'indépendance soit rejetée car dans le cas contraire, les conclusions seraient de dire que les interactions entre protéines et entre protéines et acides nucléiques se déroulent de manière aléatoire.

Dès lors, nous pouvons comparer les valeurs de $\chi^2$ observé pour chaque couple au $\chi^2$ théorique à un seul degré de liberté (k=1, $\alpha$ = 0,005 et $\chi^2$ = 7,88) en utilisant la formule suivante :

$$\chi^2_{obs} = \frac{(n_{ij} - nP_{ij})^2}{nP_{ij}}$$

où $n_{ij}$ = n nombre de couples entre les résidus i et j

n = nombre total de couples observés

$P_{ij}$ = probabilité du couple ij

$\chi^2_{obs}$ = khi carré observé

La comparaison des valeurs de $\chi^2$ va nous permettre de déterminer quels couples ont des fréquences significativement différentes des fréquences attendues 'aléatoirement'. De plus, en regardant si la fréquence observée de ces cellules est supérieure ou inférieure à la fréquence attendue, nous pourrons déterminer si un couple est favorisé ou défavorisé.

En calculant les différences entre les $\chi^2$ théoriques et les $\chi^2$ observés, nous pouvons classer les différents couples : plus la différence est grande, plus le couple est (dé)favorisé.

Remarque : les valeurs attendues étant toutes supérieures à 5, le test $\chi^2$ de Pearson est applicable sans restriction.

## *II.4. Structure de l'ADN*

Pour étudier l'influence de la structure de l'ADN sur la composition des sites protéiques d'interaction, une méthode de définition du type d'hélice en fonction des angles de torsion des nucléotides a été mise au point. Ensuite, les complexes protéine-ADN à haute résolution et comprenant une double hélice ont été analysés séparément selon leur classement en type A, B ou Z.

### II.4.1. Définition des angles de torsion des trois types principaux de doubles hélices

Comme présenté dans l'introduction (point I.3.2), les nucléotides peuvent être décrits par divers paramètres. Une majeure partie de ces paramètres sont définis par rapport à l'axe du brin nucléotidique (roll, twist…) et font intervenir plusieurs nucléotides successifs. D'un autre côté, certains paramètres peuvent être calculés uniquement sur base des propriétés d'un seul nucléotide (angles de torsion notamment).

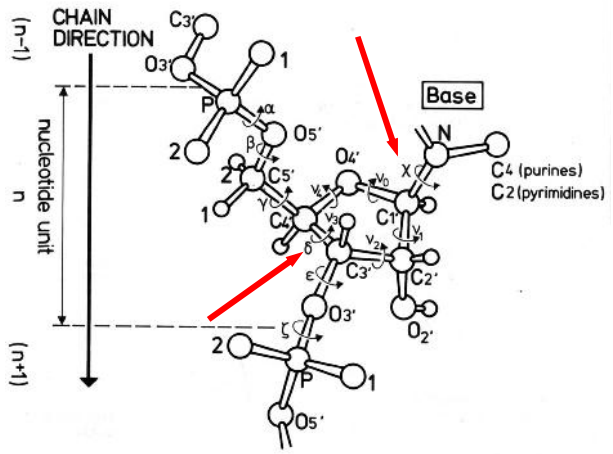



**Figure II-5 : Représentation des angles de torsion des nucléotides selon les normes de la 'International Union of Pure and Applied Chemistry' (IUPAC). Les angles χ et δ sont mis en évidence par des flèches rouges.**

Lu *et al.*[235] ont montré que l'utilisation de deux angles de torsion permet, d'une manière similaire aux diagrammes de Ramachandran pour les acides aminés, de définir des zones correspondant à tel ou tel type de double hélice d'ADN (Figure II-6). Ces deux angles sont : les angles de torsion δ (entre atomes du désoxyribose de l'ADN - atomes C5*-C4*-

C3*-O3*) et χ (liaison entre le désoxyribose et la base du nucléotide - atomes O4*-C1*-N1-C2 pour les pyrimidines et O4*-C1*-N9-C4 pour les purines) (cf. Figure II-5).

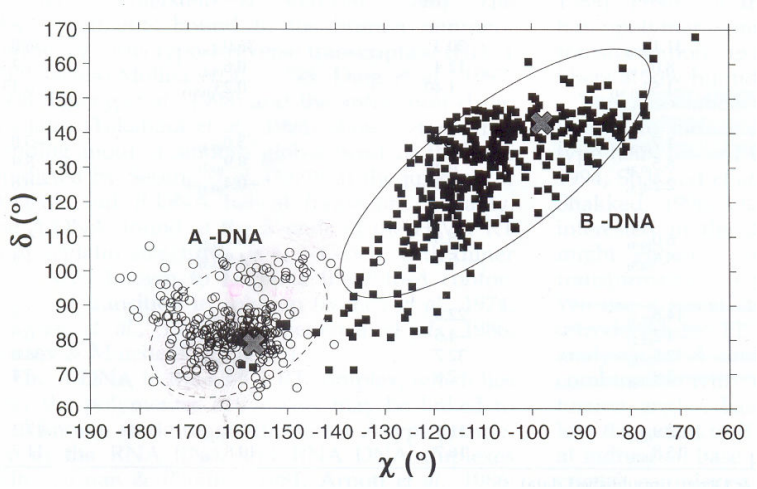



**Figure II-6 : Délimitation des angles δ et χ pour les ADN de type A et B selon Lu *et al.*[235]**

Sur base des structures utilisées dans le travail de Lu *et al.*[235] (24 structures de A-ADN et 22 structures de B-ADN) et d'une banque de 15 structures de Z-ADN issues de la PDB construite par nos soins, nous avons redéfini les valeurs limites de ces angles de torsion. Sur la Figure II-7, les valeurs d'angle des nucléotides de notre banque de données sont reprises ainsi que les valeurs de trois structures obtenues par modélisation informatique. Les rectangles de couleur correspondent aux zones qui seront utilisées dans la suite de ce travail. Pour l'A-ADN (en bleu), δ varie de 60 à 110 degrés et χ de 150 à -140 degrés. Le B-ADN (en rose) possède des valeurs de δ variant de 70 à 180 degrés et des valeurs de χ de -140 à -60 degrés. Pour les structures en Z-ADN, deux zones distinctes sont définies (en rouge). La première comprend des angles δ variant de 120 à 170 degrés et χ de -170 à -140 degrés. La seconde, des angles δ compris entre 40 et 180 degrés et χ entre 40 et 100 degrés. Les nucléotides n'appartenant pas à l'une de ces zones ont été classés à part dans une famille nommée « --- ». Les codes PDB de toutes les structures utilisées pour définir les valeurs limites des angles de torsion sont soulignés dans l'Annexe 2.

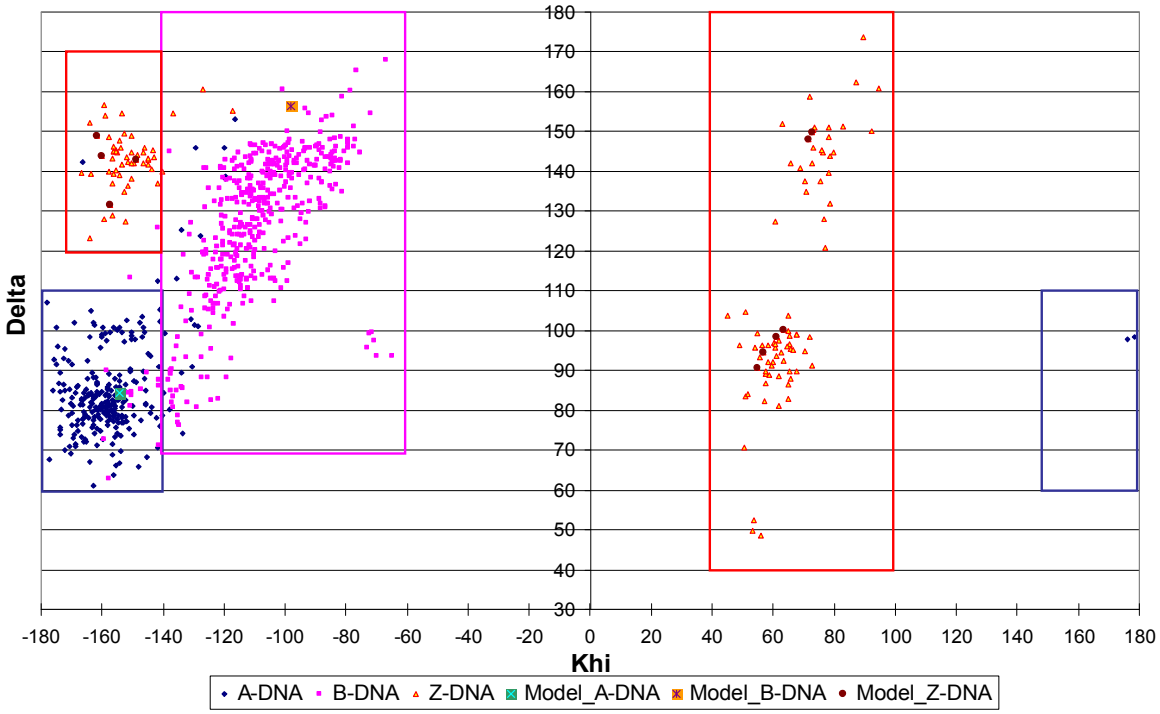



**Figure II-7 : Couples χ-δ des nucléotides et zones de délimitation des types d'ADN. Les nucléotides d'A-ADN sont représentés par des losanges bleus, le B-ADN par des carrés roses et le Z-ADN par des triangles orange. Les zones correspondantes sont délimitées par des rectangles de couleurs. Les grands carrés ainsi que les ronds bruns correspondent aux angles de structures modélisées.**

### II.4.2. Création d'une sous-banque de complexes comprenant des doubles hélices

De la banque de données des complexes protéines-ADN, les structures ayant une résolution égale ou inférieure à 2Å ont été extraites. Cette étape permet de sélectionner les complexes ayant une résolution suffisamment élevée pour étudier la structure de l'ADN avec précision. Ensuite, les complexes possédant une double hélice nucléotidique ont été extraits. Au final, 43 complexes satisfont ces critères et les angles δ et χ des nucléotides de ces doubles hélices ont été calculés. Par la suite, les acides aminés interagissant avec de l'A-ADN, du B-ADN ou du Z-ADN ont été classés dans trois sous-banques différentes.

## *II.5. Les Receptors Binding Domains*

Le concept de Receptor Binding Domain est utilisé depuis de nombreuses années pour caractériser les sites de liaison d'une chaîne polypeptidique et a permis d'aider à la progression d'études expérimentales. Il a été décrit pour la première fois en 1986 par DeLoof *et al*.[212] qui ont pu prédire la présence de deux sites d'interaction sur l'apolipoprotéine E. Le premier site était déjà connu expérimentalement tandis que la fonction du second ne le fut que beaucoup plus tard. Plus récemment, Marrec *et al.*[236] ont étudié la protéine de liaison à la pénicilline d'*Escherichia coli* en s'aidant des prédictions fournies par les RBD. Finalement, Gallet *et al.*[179] ont comparé les résultats obtenus par les prédictions RBD de deux banques d'interactions : la DIP et une banque fournie par Kini et Evans.[237] Les résultats obtenus furent très satisfaisants et les différents paramètres de la méthode (fenêtre, angle de rotation) furent optimisés. La prédiction de sites d'interaction à l'aide des RBD se base sur l'utilisation de graphiques d'Eisenberg, la construction de ces graphiques est expliquée dans le détail au point suivant.

### II.5.1. Graphiques d'Eisenberg

C'est en 1982 qu'Eisenberg[211] proposa pour la première fois un graphique dans lequel les acides aminés sont caractérisés par leur hydrophobicité moyenne <H> en fonction de leur moment hydrophobe moyen <μH>. L'échelle d'hydrophobicité utilisée est l'échelle consensus d'Eisenberg dont les valeurs pour chaque acide aminé sont reprises dans le Tableau II-4. Dans ce tableau est aussi présentée l'échelle de Kyte & Doolittle, une autre échelle couramment utilisée. Plus la valeur du résidu est positive, plus ce résidu est hydrophobe.

| Code 1 lettre | Code 3 lettres | Echelle consensus d'Eisenberg | Kyte & Doolittle | Code 1 lettre | Code 3 lettres | Echelle consensus d'Eisenberg | Kyte & Doolittle |
|---|---|---|---|---|---|---|---|
| A | ALA | 0,62 | 1,8 | L | LEU | 1,1 | 3,8 |
| R | ARG | -2,5 | -4,5 | K | LYS | -1,5 | -3,9 |
| N | ASN | -0,78 | -3,5 | M | MET | 0,64 | 1,9 |
| D | ASP | -0,9 | -3,5 | F | PHE | 1,2 | 2,8 |
| C | CYS | 0,29 | 2,5 | P | PRO | 0,12 | -1,6 |
| Q | GLN | -0,85 | -3,5 | S | SER | -0,18 | -0,8 |
| E | GLU | -0,74 | -3,5 | T | THR | -0,05 | -0,7 |
| G | GLY | 0,48 | -0,4 | W | TRP | 0,81 | -0,9 |
| H | HIS | -0,4 | -3,2 | Y | TYR | 0,26 | -1,3 |
| I | ILE | 1,4 | 4,5 | V | VAL | 1,1 | 4,2 |

**Tableau II-4 : Echelle d'hydrophobicité selon Eisenberg et Kyte & Doolittle.**

La valeur d'hydrophobicité moyenne <H> est la moyenne (scalaire) des valeurs consensus d'hydrophobicité des acides aminés de la fenêtre considérée. La valeur calculée est attribuée au résidu central de la fenêtre (Figure II-8). Au départ, la taille de la fenêtre proposée par Eisenberg était de 11 acides aminés mais toute valeur impaire peut convenir.

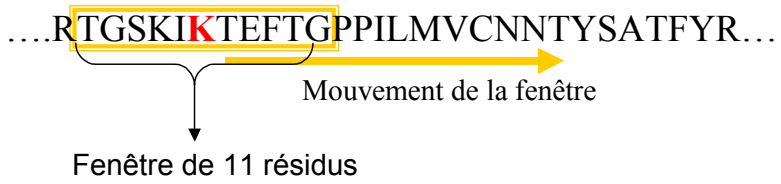



**Figure II-8 : Représentation d'une fenêtre utilisée pour le calcul de divers paramètres. Le cadre jaune représente la fenêtre et l'acide aminé central (en rouge) se voit attribuer la valeur calculée. La fenêtre se déplace d'un acide aminé à chaque étape.**

Le moment hydrophobe moyen <μH> représente la moyenne, sur la fenêtre, de la somme vectorielle des vecteurs hydrophobes. En plus de la valeur de l'échelle consensus, intervient dans le calcul du moment hydrophobe moyen un angle de rotation appliqué entre chaque vecteur. Cet angle peut se voir attribuer plusieurs valeurs correspondant à un type de structure secondaire : 100° pour l'hélice α, 170° pour le feuillet β, 85° pour le turn β. La méthode de référence prévoyait quant à elle un angle de 100°.

Ci-dessous sont reprises les formules de calcul de l'hydrophobicité moyenne et du moment hydrophobe moyen :

$$\langle H_i \rangle = \frac{1}{N} \sum_{n=1}^{N} h_n$$

$$\langle \mu_{Hi} \rangle = \frac{1}{N} \left[ \left( \sum_{n=1}^{N} h_n \sin(\delta n) \right)^2 + \left( \sum_{n=1}^{N} h_n \cos(\delta n) \right)^2 \right]^{1/2}$$

Avec : $h_n$ = hydrophobicité de l'acide aminé n selon l'échelle consensus d'Eisenberg,

N = nombre de résidus dans la fenêtre de calcul,

$<\mu H_i>$ = moment hydrophobe moyen de l'acide aminé central (i),

$<H_i>$ = hydrophobicité moyenne de l'acide aminé central (i),

δ = angle de giration entre deux résidus successifs.

Dans leur travail, Gallet *et al.*[179] ont pu montrer qu'en vue de la prédiction des sites d'interaction, une fenêtre de 5 acides aminés et un angle de rotation de 100° conduisent aux pourcentages de prédiction les plus élevés. Toutefois, la variation de l'angle de rotation, bien qu'affectant peu le taux de prédiction, conduit à la sélection d'acides aminés différents en

décalant les zones sélectionnées. Dans ce travail, l'angle est fixé à 100° ce qui correspond à une hélice α droite.

Après avoir calculé l'hydrophobicité moyenne et le moment hydrophobe moyen, on peut construire un graphique reprenant par exemple tous les acides aminés d'une protéine (Figure II-9).

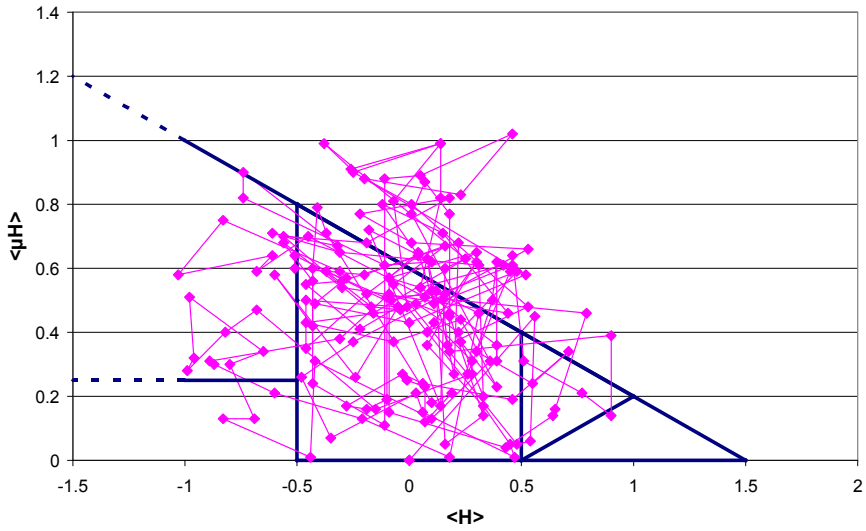



**Figure II-9 : Graphique d'Eisenberg représentant les acides aminés de l'hormone de croissance humaine (1HWG).[238] Les différentes zones sont décrites dans le paragraphe suivant.**

### II.5.2. Zone RBD

Sur le graphique Eisenberg, différentes zones ont été déterminées (Figure II-10). Ces différentes zones contiennent des acides aminés dont les caractéristiques d'hydrophobicité/hydrophilicité suggèrent qu'ils appartiennent à une partie donnée de la protéine. La zone S pour les résidus de surface, M pour les résidus membranaires et G pour les résidus globulaires. Par après, une zone T pour les résidus trans-membranaires et une zone RBD furent rajoutées. Les limites de la zone RBD correspondent à une hydrophobicité moyenne (<H>) inférieure à -0,5 et à un moment hydrophobe moyen <μH> compris entre 0.25 et la droite <μH> = -0.4<H> + 0.6.

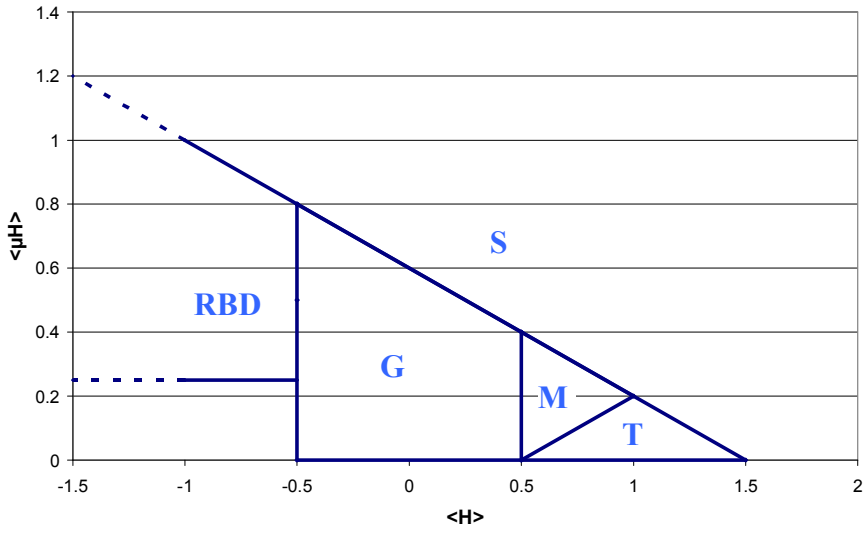



**Figure II-10 *:* Graphique d'Eisenberg et délimitation des différentes zones.**

## *II.6. Serveurs Internet de Prédiction se basant sur la Séquence*

Lors de la création du modèle de prédiction des sites d'interaction, certaines méthodes de prédiction ont été utilisées en plus des informations récoltées et extrapolées des banques de complexes. De cette manière, nous avons utilisé le maximum d'informations récoltables à partir de la séquence. Les serveurs utilisés sont brièvement décrits ci-dessous.

### II.6.1. Prédiction de structure secondaire

Les prédictions de structures secondaires ont été réalisées à l'aide de deux méthodes : PsiPred (http://bioinf.cs.ucl.ac.uk/psipred/) et NPSA (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html).

La méthode PsiPred[239] est une méthode simple et fiable basée sur deux réseaux neuronaux placés en 'feed-forward' et qui utilisent des données de similarité issues de PSI-BLAST (Position Specific Iterated - Basic Local Alignment Search Tool). A l'aide d'une méthode de validation croisée, les performances de PsiPred sont évaluées à 78% (Q3 moyen).

La méthode NPSA[240] donne une prédiction consensus des structures secondaires. Elle utilise donc différentes méthodes comme GOR, PHD, PREDATOR, etc. et constitue un bon complément de la méthode PsiPred.

### II.6.2. Prédiction d'accessibilité

Deux serveurs ont été testés : PredAcc (http://bioserv.rpbs.jussieu.fr/RPBS/cgi-bin/Ressource.cgi?chzn_lg=fr&chzn_rsrc=PredAcc) et NetASA (http://www.netasa.org/).

Les prédictions données par PredAcc[241] se basent sur une fonction logistique dont les données entrantes principales sont la longueur de la séquence, la distance séparant le résidu considéré des extrémités C-ter et N-ter de la protéine ainsi que différentes fréquences de résidus dans les banques de données utilisées dans un autre travail.[242] Le taux de prédictions correctes varie de 70.7% à 85.7%. Ce serveur permet de faire varier le seuil de surface accessible relative de 0% à 55% par pas de 5%.

NetASA[243] quant à lui, est basé sur un réseau neuronal qui utilise une fenêtre de 17 acides aminés et prend donc en compte les 8 résidus avant et après le résidu considéré. Une précision de 63% à 88% est obtenue sur l'échantillon test (185 structures).

Les résultats de ces deux serveurs ont été comparés aux résultats obtenus par calcul de la surface accessible relative (relativeASA, cf. II.3.3) par le programme Pex sur base des structures tridimensionnelles. Les pourcentages de prédictions correctes sur notre banque de données sont de 72% et 70% pour PredAcc et NetASA, respectivement. Ceci se trouve dans la gamme de résultats donnés dans les articles de référence.[241,243] En divisant les résultats selon le type d'acide aminé, on peut voir que les serveurs sont beaucoup plus performants pour la prédiction des résidus en surface (78%) que pour la prédiction des résidus internes (60% et 55%). Ci-dessous (Figure II-11), la comparaison des résidus en surface et des prédictions est donnée dans le cas de la serine carboxypeptidase.

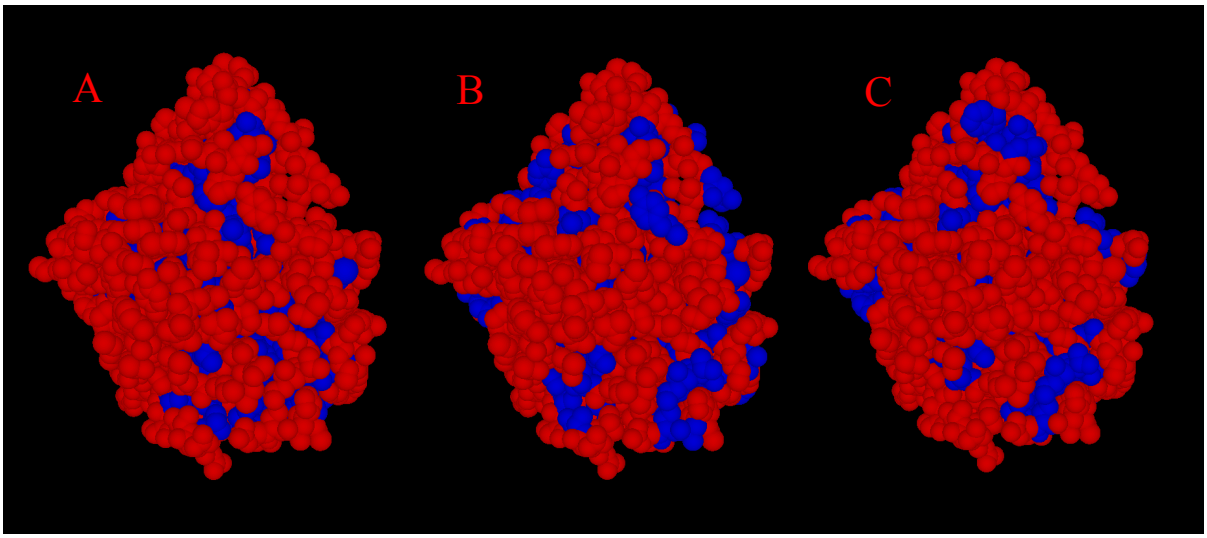



**Figure II-11 : Représentation en mode vrais volumes d'une enzyme (Serine carboxypeptidase - code PDB = 1BCS).[244] Les résidus considérés comme étant en surface sont représentés en rouge. A : surface accessible relative calculée sur base de la structure 3D (valeurs Pex) ; B : surface accessible relative prédite par NetASA ; C : surface accessible relative prédite par PredAcc. Images générées par le logiciel YAGME.**

### II.6.3. Détection de motifs ou domaines protéiques

Une recherche sur internet a été réalisée afin de trouver des serveurs permettant de détecter des motifs/domaines à partir d'une séquence. Les deux principaux serveurs trouvés sont : MOTIFsearch (http://motif.genome.jp/) et InterProScan (http://www.ebi.ac.uk/InterProScan/).

MOTIFsearch effectue des recherches dans les bases de données suivantes :

- PROSITE Pattern
- PROSITE Profile
- BLOCKS

- ProDom
- PRINTS
- Pfam
- Pfam fragment search
- Recherche de motif définis par l'utilisateur

InterProScan utilise 13 bases de données différentes :

- ProDom (BastProDom)
- PRINTS (FPrintScan)
- PIR (HMMPIR)
- Pfam (HMMPfam)
- SMART (HMMSmart)
- TigrFam (HMMTigr)
- PROSITE Profile (ProfileScan)
- PROSITE (ScanRegExp)
- Superfamily
- SignalP (SignalPHMM)
- Transmembrane helix (TMHMM)
- Panther (HMMPanther)
- Gene3D

On voit donc que le serveur InterProScan est plus complet que MOTIFsearch et que la plupart des banques étudiées se recoupent. Seuls BLOCKS et 'Pfam fragment search' ne sont pas repris dans InterProScan. De plus, le software InterProScan est téléchargeable et peut donc être utilisé en interne. Cette utilisation en interne présente deux avantages importants : possibilité de lancer un grand nombre de séquences en une seule fois et la possibilité de choisir un format de sortie plus adapté à une analyse statistique ultérieure.

### II.6.4. Prédiction du désordre des protéines

Une protéine (ou un domaine) désordonnée est une protéine (domaine) qui peut se retrouver sous plusieurs conformations. Ce désordre peut-être relié à une variation de structure secondaire ou de structure tertiaire. L'équilibre entre les différentes formes sera influencé, par exemple, par des variations de composition du milieu, l'addition de substrat….

La méthode utilisée dans ce travail a été présentée par Linding *et al.*[245] GloPlot (http://globplot.embl.de/) se base sur les tendances des résidus à induire du désordre et leurs propensions à se trouver soit dans des structures de type random coil ou dans des structures secondaires régulières (hélices α et structures β).

# III. RÉSULTATS

# *III.1. Interactions Protéines - Acides Nucléiques*

## III.1.1. Caractéristiques générales

### Complexes protéine-ADN

Cette banque de données comporte 62.715 résidus répartis dans 139 complexes ce qui fait une moyenne de 451 résidus par complexe. Le plus grand (1PV4)[246] en compte 2524 et le plus petit (1J75)[247] en compte 74. Le nombre de chaînes protéiques par complexe varie de 1 à 4, de 1 à 8 pour les chaînes nucléotidiques et la résolution varie de 1,25Å à 3Å.

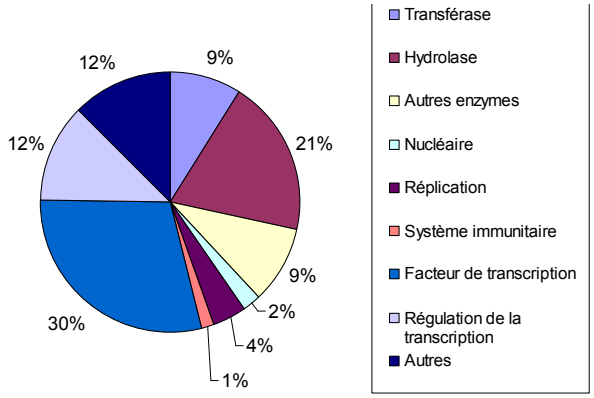



**Figure III-1 : Différents types de complexes protéine-ADN.**

Les fonctions biologiques des protéines impliquées sont diverses avec une part importante de complexes intervenant dans la transcription (42%) et d'enzymes (39%). Dans ces enzymes, les hydrolases (21%) et les transférases (9%) sont les mieux représentées alors que l'on retrouve en moindre quantité des endonucléases, hélicases, isomérases, ligases, polymérases, recombinases et transposases.

### Complexes protéine-ARN

L'ensemble des complexes entre protéines et ARN est composé de 27.297 résidus répartis dans 49 complexes (557 résidus en moyenne par complexe). 124 résidus pour le plus petit (2A8V)[248] et 1874 pour le plus grand (1GTF).[249] Le nombre de chaînes protéiques par complexe varie de 1 à 6, de 1 à 4 pour les chaînes nucléotidiques et la résolution varie de 1,8Å à 3Å.

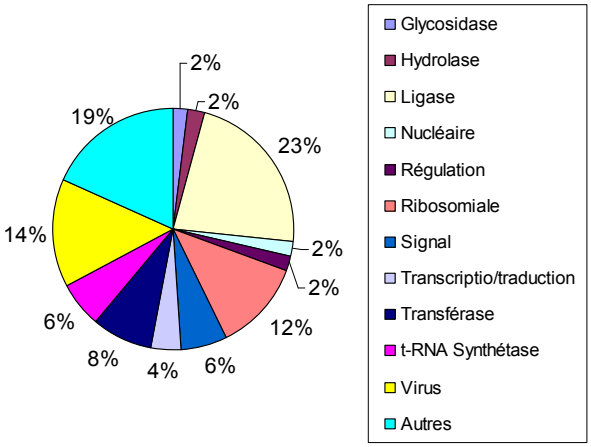


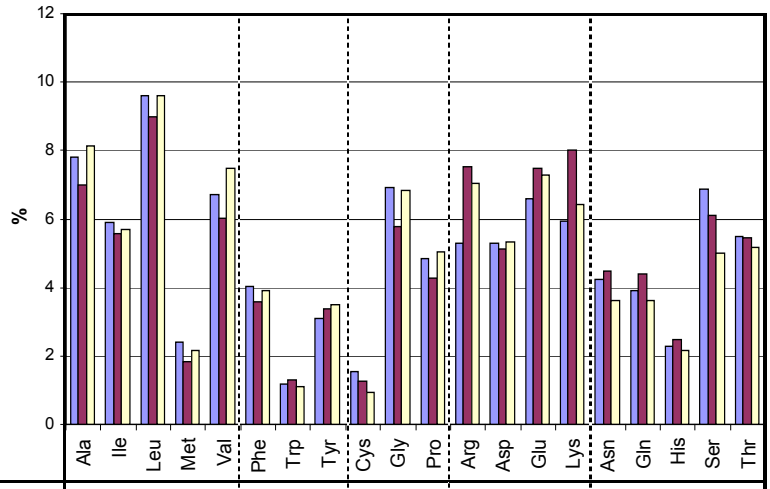

**Figure III-2 : Différents types de complexes protéine-ARN.**

Les complexes avec l'ARN sont principalement composés d'enzymes (41%) et spécialement de ligases (23%). Les protéines présentes au niveau du ribosome (12%) ainsi que les protéines constitutives des virus à ARN (14%) sont également fortement représentées. En moindre proportion, on trouve des complexes impliqués dans les processus de régulation, de transduction du signal, de transcription et de traduction.

## III.1.2. Composition des banques et sous-banques de données

### Analyse de la banque totale

Avant d'analyser en détail les caractéristiques des sites d'interaction, les banques totales (ADN et ARN) ont été comparées à la distribution des résidus dans la SwissProt/UniProt. Celle-ci a été considérée comme une référence pour la composition en acides aminés des protéines (Figure III-3).




| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| Swiss-Prot | 32,5 | 8 | 13,5 | 23 | 23 |
| DNA | 29,5 | 8 | 11,5 | 28 | 23 |
| RNA | 33 | 8,5 | 13 | 26 | 19,5 |

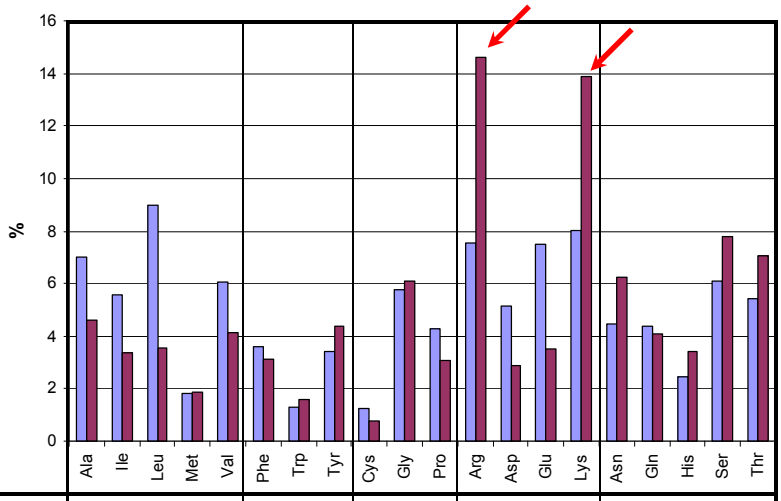
**Figure III-3 : Comparaison des fréquences des acides aminés dans la banque de référence (SwissProt/UniProt - ) et dans les deux banques de départ (ADN - et ARN - ). Les fréquences par familles sont reprises dans le tableau en bas du graphique.**

Si on regarde la Figure III-3, on peut voir que, globalement, les distributions sont fort proches. Néanmoins, en comparant le rapport en résidus hydrophobes/hydrophiles des trois banques, on remarque que la SwissProt/UniProt est légèrement enrichie en résidus hydrophobes avec 54% (Figure III-3, résidus aliphatiques, aromatiques et particuliers), alors que les deux banques de complexes sont plus hydrophiles. Pour les complexes protéines-ADN, 51% des résidus sont hydrophiles. Cette différence est principalement due à une proportion élevée en résidus positivement chargés avec 7,5% pour l'arginine et 8% pour la lysine. En comparaison, la SwissProt/UniProt contient seulement 5,5% d'arginine et 6% de lysine. La diminution du pourcentage de résidus hydrophobes est surtout observée pour les résidus aliphatiques et particuliers (41% pour les complexes avec l'ADN contre 46% dans la SwissProt/UniProt) alors que les acides aminés aromatiques sont distribués de manière équivalente (8%). Dans les complexes protéines-ARN, le rapport hydrophile/hydrophobe est semblable à celui de la SwissProt/UniProt (45,5/54,5 pour 46/54). Cependant, dans la famille des résidus hydrophiles, la balance entre résidus chargés et polaires penche du côté des résidus chargés avec 26% pour seulement 19,5% de résidus polaires (23% et 23% dans la banque de référence).

**Composition de la banque des résidus en interaction**

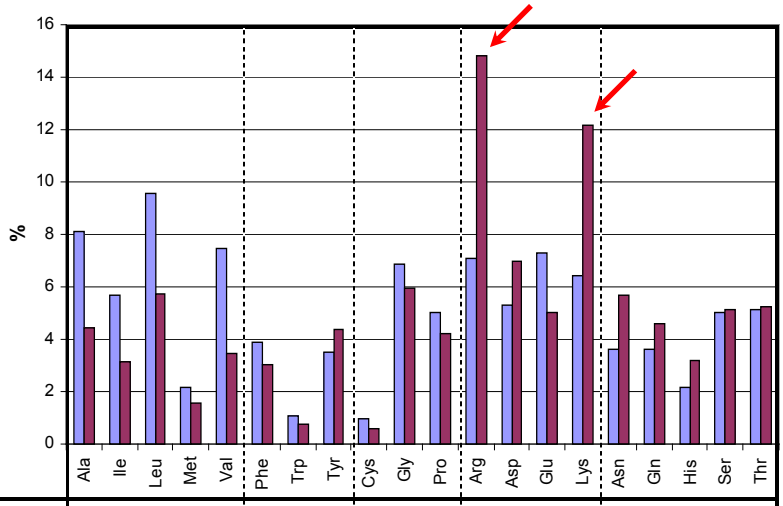
Les banques de résidus en interaction contiennent uniquement les résidus qui entrent en contact avec un autre résidu (cf. point II.3.2). La distance maximale d'interaction choisie étant de 5Å. L'analyse des résidus en interaction est détaillée ci-dessous.

Bien que les banques totales de complexes ADN/ARN soient déjà enrichies en résidus hydrophiles, et particulièrement en résidus chargés positivement, les banques de résidus en interaction le sont d'une manière encore plus marquée (Figure III-4 pour l'ADN et Figure III-5 pour l'ARN) : plus de 60% des acides aminés en interaction sont hydrophiles, 35% des sites d'interaction protéine-ADN et 39% des sites d'interaction avec l'ARN sont composés d'acides aminés chargés. L'enrichissement en résidus chargés positivement est d'autant plus important que les résidus chargés négativement sont eux, sous-représentés.




| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| ADN total | 29,5 | 8 | 11,5 | 28 | 23 |
| ADN en interaction | 17,5 | 9 | 10 | 35 | 28,5 |

**Figure III-4 : Comparaison des fréquences des acides aminés dans la banque des complexes protéine-ADN ( ) et dans la banque des acides aminés en interaction avec l'ADN ( ). Les fréquences par familles sont reprises dans le tableau en bas du graphique.**




| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| ARN total | 33 | 8,5 | 13 | 26 | 19,5 |
| ARN en interaction | 18,5 | 8 | 10,5 | 39 | 24 |

**Figure III-5 : Comparaison des fréquences des acides aminés dans la banque des complexes protéine-ARN ( ) et dans la banque des acides aminés en interaction avec l'ARN ( ). Les fréquences par familles sont reprises dans le tableau en bas du graphique.**

Pour analyser plus facilement l'importance de certains résidus, nous avons calculé les propensions ($P_i$) des acides aminés à interagir avec les nucléotides (cf. point II.3.2). Dans les complexes protéines-ADN (Figure III-6, en bleu), l'arginine et la lysine ont les valeurs de propension les plus élevées : 1,9 et 1,7 respectivement, alors que l'asparagine et l'histidine (1,4), la sérine, la thréonine et la tyrosine (1,3) et le tryptophane (1,2) sont également favorisés mais dans une moindre mesure. La leucine (0,4), l'acide glutamique (0,5), l'isoleucine, la cystéine et l'acide aspartique (0,6) et l'alanine, la valine et la proline (0,7) sont défavorisés. Les acides aminés en interaction avec l'ARN ont un comportement similaire et présentent les valeurs de propension favorables suivantes (Figure III-6, en rouge) : arginine (2,1), lysine (1,9), asparagine (1,6), histidine (1,5), glutamine et acide aspartique (1,3) et tyrosine (1,2). Les résidus défavorisés sont l'alanine et la valine (0,5), l'isoleucine, la leucine et la cystéine (0,6), la méthionine, le tryptophane et l'acide glutamique (0,7) et la phénylalanine (0,8).

Les principales différences entre les complexes avec l'ADN et l'ARN sont observées pour l'acide aspartique et le tryptophane (Figure III-6, flèches noires). L'acide aspartique a une propension de 1,3 et représente 7% des résidus en interaction dans la banque d'interaction avec l'ARN alors qu'en comparaison, avec l'ADN, il possède une propension de 0,6 et une fréquence de 3%. Le tryptophane a une propension de 0,7 quand il interagit avec l'ARN contre une propension de 1,2 avec l'ADN.

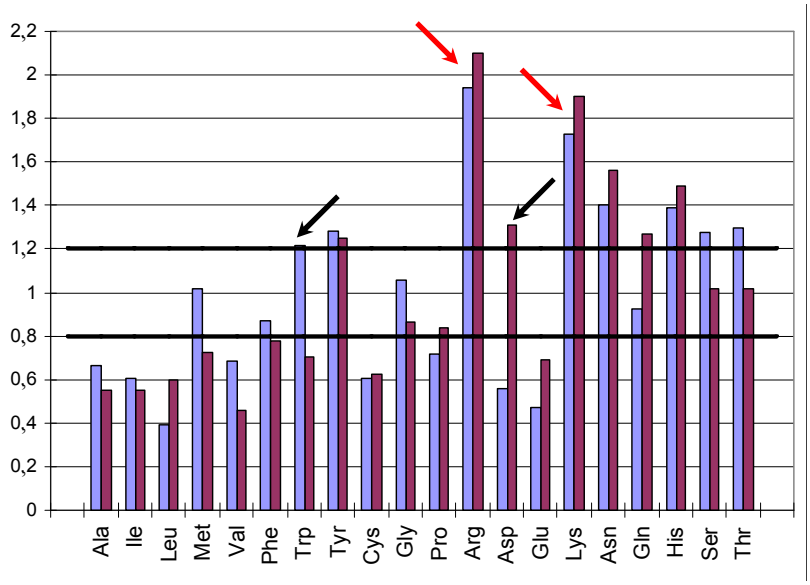



**Figure III-6 : Propensions des acides aminés à interagir avec les nucléotides. Les valeurs pour les complexes protéines-ADN sont représentées en bleu ( ) et, en rouge ( ), sont représentées les valeurs pour les interactions avec l'ARN.**

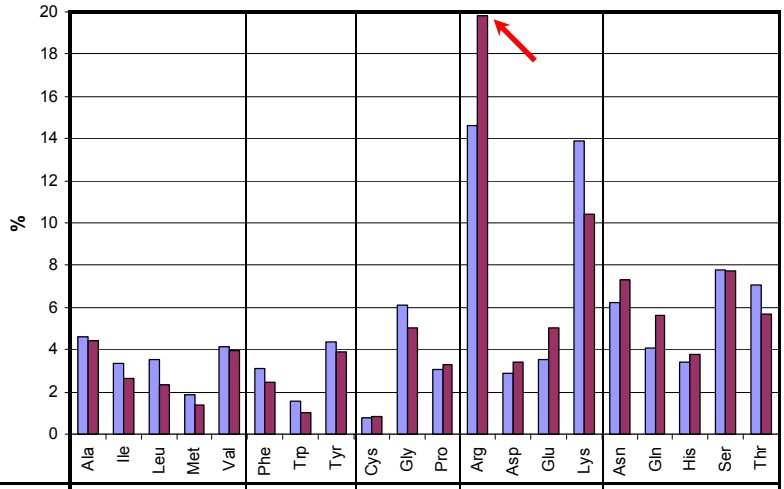
Ces premiers résultats confirment l'importance des deux acides aminés chargés positivement (arginine et lysine) lors des interactions avec des nucléotides. Ils mettent aussi en évidence le rôle des acides aminés polaires et certaines différences entre les interactions protéine-ADN et protéine-ARN.

### Interaction avec les bases nucléotidiques

Pour déterminer si les acides aminés chargés positivement interagissent essentiellement avec les groupes phosphates chargés négativement des acides nucléiques, nous avons extrait les interactions des protéines avec les bases de l'ADN/ARN. La Figure III-7 et la Figure III-8 montrent que les résidus chargés positivement sont largement impliqués dans les interactions avec les bases nucléiques.

L'arginine est largement favorisée avec une propension à interagir avec les bases d'ADN de 2,6 (Figure III-9, en bleu). Le résultat de la lysine est plus surprenant : elle est moins favorisée (1,3) à interagir avec les bases qu'avec l'ensemble de l'ADN (1,7). Malgré cela, la lysine est encore largement représentée avec 10,5% (Figure III-7). Outre les résidus chargés positivement, les acides aminés polaires sont aussi favorisés au contact des bases d'ADN et correspondent à 30% de ces interactions. L'asparagine a une propension à interagir avec les bases de 1,6, l'histidine de 1,5 et la glutamine et la sérine de 1,3. Les résidus hydrophobes sont défavorisés. Seule la tyrosine qui est classée comme résidu hydrophobe

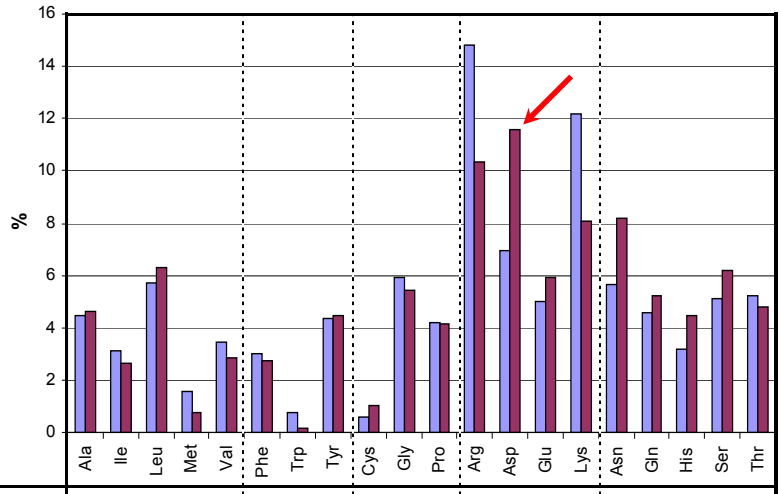
mais possède les caractéristiques d'un acide aminé aromatique polaire, a tendance à interagir avec les bases (1,2).




| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| ADN en interaction | 17,5 | 9 | 10 | 35 | 28,5 |
| ADN en interaction avec les bases | 15 | 7,5 | 9 | 38,5 | 30 |

**Figure III-7 : Comparaison des fréquences des acides aminés dans la banque des acides aminés en interaction avec l'ADN ( ) et dans la banque des résidus en interaction avec les bases de l'ADN ( ). Les fréquences par familles sont reprises dans le tableau en bas du graphique.**

Pour les interactions avec les bases de l'ARN, les acides aminés impliqués sont relativement différents (Figure III-8 et Figure III-9, en rouge). L'arginine (1,5) et la lysine (1,3) sont toujours favorisés mais l'asparagine (2,2), l'histidine (2,1) et, étonnamment, l'acide aspartique (2,2) possèdent les propensions les plus élevées (Figure III-9, flèches rouges). De plus, l'acide aspartique est l'acide aminé le plus fréquent au contact des bases de l'ARN (11,5%) (Figure III-8, flèche rouge). De nouveau, les résidus classés comme hydrophobes sont défavorisés (excepté la tyrosine - 1,2).




| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| ARN en interaction | 18,5 | 8 | 10,5 | 39 | 24 |
| ARN en interaction avec les bases | 17 | 7,5 | 10,5 | 36 | 29 |

**Figure III-8 : Comparaison des fréquences des acides aminés dans la banque des acides aminés en interaction avec l'ARN ( ) et dans la banque des résidus en interaction avec les bases de l'ARN ( ). Les fréquences par familles sont reprises dans le tableau en bas du graphique.**

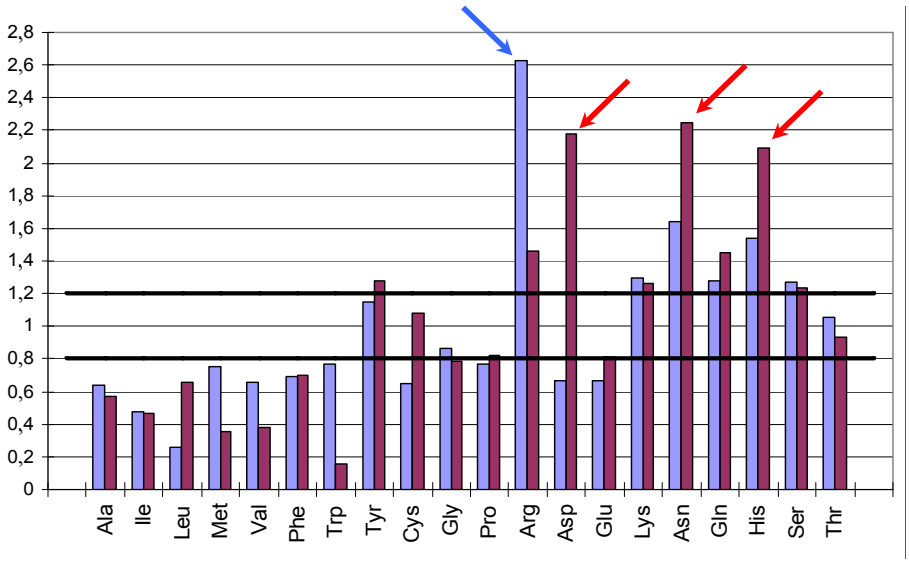



**Figure III-9 : Propensions des acides aminés à interagir avec les bases des nucléotides. Les valeurs pour les complexes protéines-ADN sont représentées en bleu ( ) et, en rouge ( ), sont représentées les valeurs pour les interactions avec l'ARN.**

### III.1.3. Distribution des atomes nucléotidiques impliqués dans les interactions

Bien que les liens H avec les bases nucléotidiques soient connus pour être importants dans la spécificité des interactions entre protéines et acides nucléiques, une protéine peut aussi entrer en contact avec les autres parties des nucléotides. Comme expliqué au point II.3.1 et comme on peut le voir dans la Figure III-10, nous avons regroupé les atomes nucléotidiques en trois familles distinctes : les atomes appartenant au phosphate, au sucre ou à la base. Contrairement aux résultats obtenus pour les acides aminés, les résultats montrent, ici, de claires différences entre les complexes avec l'ADN et ceux avec l'ARN (Figure III-10, graphique de gauche et de droite, respectivement).

Dans les complexes protéine-ADN, en moyenne, 47% des interactions impliquent les atomes du groupement phosphate alors que les interactions avec les bases en représentent seulement 24% (avec des valeurs limites de 18% pour l'adénine et de 27% pour la cytosine). La situation est très différente pour les complexes protéines-ARN. En moyenne, seulement 22% des contacts avec les protéines font intervenir les atomes du phosphate. A l'opposé, les interactions impliquant les atomes des bases et des sucres sont plus nombreuses avec 35% et 43%, respectivement.

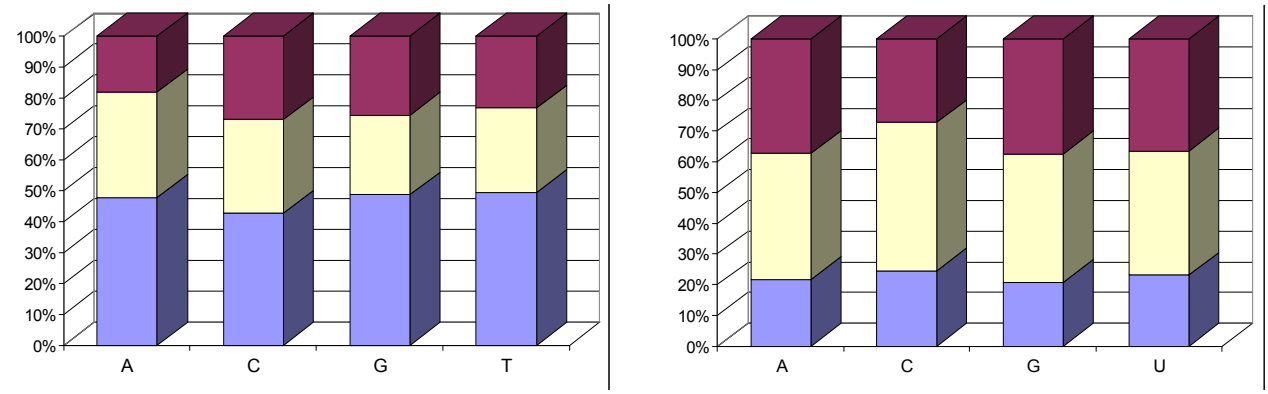



**Figure III-10 : Distribution des interactions selon la partie du nucléotide impliquée. Les types d'atomes sont colorés comme suit : la base = ; le sucre = et le phosphate = . A gauche, interactions avec l'ADN et à droite, interactions avec l'ARN.**

Ces divergences résultent probablement des conformations de l'ADN et de l'ARN. En effet, 118 des 139 complexes avec l'ADN de notre banque de données contiennent de l'ADN sous forme de double hélice alors que seulement 15 des 49 complexes avec l'ARN contiennent plus d'une chaîne d'ARN. Dès lors, les bases nucléotidiques de l'ADN sont moins accessibles que celles de l'ARN. Ceci peut expliquer que les bases de l'ARN soient

impliquées dans plus d'un tiers des contacts avec les protéines. Cette plus grande accessibilité des bases de l'ARN conduit également à une réduction des interactions avec les phosphates et les atomes du ribose deviennent les plus impliqués dans la banque des résidus en interaction avec l'ARN. Ce résultat suggère que le ribose, qui est rarement pris en compte, pourrait avoir un rôle important dans les interactions protéine-ARN.

Si on décompose les interactions protéine-acide nucléique en fonction du type d'atome protéique impliqué, on obtient la Figure III-11. Que ce soit pour les complexes avec l'ADN ou avec l'ARN, un peu plus de 70% des interactions font intervenir les chaînes latérales des acides aminés. Pour les interactions avec l'ADN (Figure III-11 graphique de gauche), les interactions entre les chaînes latérales et les phosphates sont les plus nombreuses (29%). Viennent ensuite les interactions entre chaînes latérales et le désoxyribose (22%) et seulement en troisième position les interactions chaînes latérales-bases (19%). Les interactions avec le squelette peptidique des protéines se font majoritairement avec le groupement phosphate (17%). Du côté des interactions avec l'ARN (Figure III-11, graphique de droite), viennent en premier lieu les interactions chaîne latérale-sucre et chaîne latérale-base avec 33% et 21% respectivement alors que les interactions chaîne latérale-phosphate ne représentent que 17%. Les interactions avec le squelette peptidique suivent le schéma suivant : base (13%) > sucre (11%) > phosphate (5%).

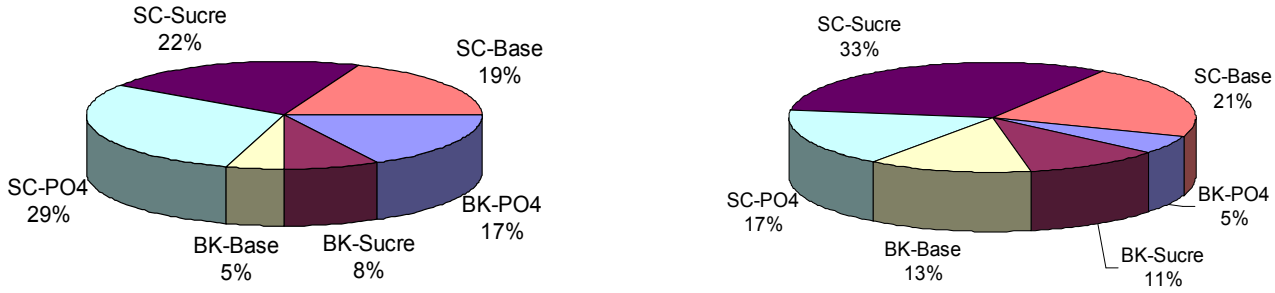



**Figure III-11 : Distribution des interactions dans les interfaces avec l'ADN (à gauche) et avec l'ARN (à droite). Les interactions sont différenciées selon le type d'atome impliqué (BK= squelette peptidique ou 'backbone' ; SC = chaîne latérale) : Backbone-phosphate ( ) ; Backbone-sucre ( ) ; Backbone-base ( ) ; Chaîne latérale-phosphate ( ) ; Chaîne latérale-sucre ( ) ; Chaîne latérale-base ( ).**

### III.1.4. Distribution des types d'interactions

La divergence dans les résultats obtenus pour l'ADN et l'ARN en termes de types d'atomes nous a conduit à étudier le type d'interaction impliqué. La Figure III-12 reprend les fréquences des quatre types d'interactions décrits au point II.3.1 avec, en bleu, les résultats pour l'ADN et, en rouge, ceux pour l'ARN.

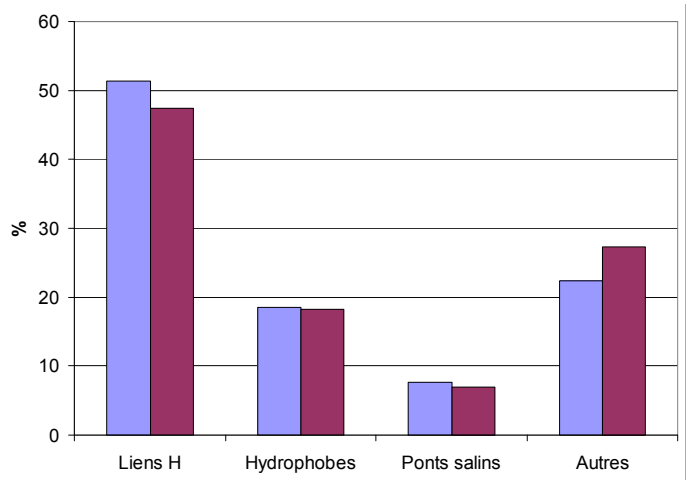



**Figure III-12 : Comparaison des distributions des types d'interactions dans les complexes protéines-ADN ( ) et protéines-ARN ( ).**

Globalement, les deux distributions sont assez proches avec une majorité de liens H (51% et 47%) et environ 8% et 7% de ponts salins. Le nombre similaire de ponts salins dans les complexes avec l'ADN et l'ARN est assez surprenant si on se souvient que les interactions avec les phosphates de l'ARN sont deux fois moins nombreuses que celles avec les phosphates de l'ADN (22% et 47%, respectivement). Or, les atomes du groupement phosphate sont les seuls atomes nucléotidiques à pouvoir former un pont salin avec les protéines.

En fait, dans les interactions avec les phosphates de l'ADN (47%), seulement 16% sont des ponts salins alors que 31% des interactions avec les phosphates d'ARN (22%) en sont. Cette différence conduit à un nombre semblable de ponts salins.

L'analyse de la distribution des liens H en fonction du type d'atome nucléotidique montre une nette différence entre les complexes avec l'ADN et ceux avec l'ARN (Figure III-13). Dans les interactions protéines-ADN, 62% des liens H font intervenir les atomes du groupement phosphate alors que ces atomes ne sont impliqués que dans 27% des liens H avec l'ARN. Il semble donc que, bien que les protéines liant l'ADN interagissent deux fois plus avec les phosphates que celles liant l'ARN, ce sont les liens H et non pas les ponts salins qui sont responsables de ce nombre élevé de contacts avec les phosphates de l'ADN. Cela

correspond en quelque sorte à une saturation des charges négatives des phosphates : les groupements phosphates ne sont plus capables de réaliser des ponts salins supplémentaires et les nouvelles interactions se font par l'intermédiaire de liens H.

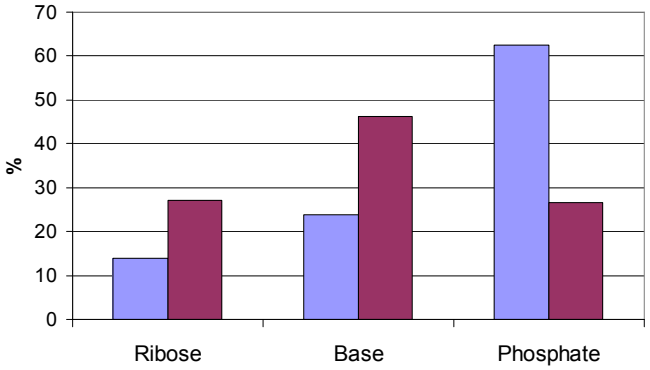



**Figure III-13 : Comparaison des distributions des liens H en fonction du type d'atome dans les complexes protéines-ADN ( ) et protéines-ARN ( ).**

Les interactions hydrophobes correspondent à 19% des interactions entre les acides nucléiques et les protéines de nos banques (Figure III-12). Les couples d'atomes hydrophobes sont composés d'atomes d'hydrogène du sucre et d'atomes d'hydrogène des chaînes latérales aliphatiques ou aromatiques. Les interactions hydrophobes sont le principal type d'interaction observé pour les sucres. Approximativement, 63% des interactions protéine-désoxyribose et 42% des interactions protéine-ribose sont des contacts hydrophobes. Ce type d'interaction est rarement pris en compte mais pourrait avoir un rôle stabilisant dans les interactions nucléotides-protéines.

Finalement, 22% des interactions protéine-ADN et 27% des interactions protéine-ARN sont décrites comme 'autres van der Waals'. Le pourcentage plus élevé d'interactions de van der Waals dans les complexes protéine-ARN peut être expliqué par la présence d'un groupe hydroxyle sur le cycle du ribose. Concrètement, ce groupe est impliqué dans 38% des contacts de van der Waals ce qui correspond à 11% de toutes les interactions protéine-ARN. Du point de vue des nucléotides, les interactions de van der Waals font principalement intervenir le squelette sucre-phosphate des acides nucléiques (62% pour l'ADN et 60% pour l'ARN). Du point de vue des acides aminés, bien que les atomes des chaînes latérales soient encore fréquemment impliqués, l'implication du squelette peptidique protéique est plus grande dans les complexes protéine-ADN (43%) que dans les complexes protéine-ARN (27%).

### III.1.5. Matrices d'interactions

Les matrices d'interactions reprenant les 80 couples possibles sont présentées dans le Tableau III-1. Pour les analyser, la fréquence observée des couples acide aminé-nucléotide a été comparée à celle attendue si les interactions se déroulaient aléatoirement. Une analyse statistique de l'ensemble de la table a montré que la distribution des fréquences des couples était différente d'une distribution aléatoire. En effet, le $\chi^2$ statistique (total) est largement plus grand que le $\chi^2$ théorique (2101 > 88,25 pour les complexes protéines-ADN et 837 > 88,25 pour les complexes protéines-ARN) et permet de refuser l'hypothèse d'indépendance des acides aminés et des nucléotides (cf. point II.3.4). Ce résultat n'est guère surprenant car on sait que les interactions n'interviennent pas de manière aléatoire mais cette étape était nécessaire avant de passer à une analyse plus détaillée des différents couples.

| ADN | A | C | G | T |
|---|---|---|---|---|
| Ala | ↓ | ↓ | ↓ | ↓ |
| Ile | ↓ | ↓ | ↓ | ↓ |
| Leu | ↓ | ↓ | ↓ | ↓ |
| Met | -- | -- | -- | -- |
| Val | ↓ | ↓ | ↓ | -- |
| Phe | ↓ | -- | -- | -- |
| Trp | -- | **53 ↑** | -- | -- |
| Tyr | -- | -- | **102 ↑** | **95 ↑** |
| Cys | -- | -- | -- | ↓ |
| Gly | -- | -- | -- | -- |
| Pro | -- | -- | -- | -- |
| Arg | **201 ↑** | **300 ↑** | **359 ↑** | **262 ↑** |
| Asp | ↓ | -- | ↓ | ↓ |
| Glu | ↓ | ↓ | ↓ | ↓ |
| Lys | **249 ↑** | **204 ↑** | **297 ↑** | **314 ↑** |
| Asn | -- | **134 ↑** | **109 ↑** | **137 ↑** |
| Gln | -- | -- | -- | -- |
| His | -- | -- | **87 ↑** | **76 ↑** |
| Ser | -- | -- | **143 ↑** | **174 ↑** |
| Thr | -- | **146 ↑** | **132 ↑** | **154 ↑** |

| ARN | A | C | G | U |
|---|---|---|---|---|
| Ala | ↓ | ↓ | ↓ | ↓ |
| Ile | ↓ | ↓ | -- | -- |
| Leu | ↓ | ↓ | -- | -- |
| Met | -- | -- | -- | -- |
| Val | ↓ | ↓ | -- | ↓ |
| Phe | -- | -- | -- | ↓ |
| Trp | -- | -- | -- | -- |
| Tyr | -- | -- | -- | **61 ↑** |
| Cys | -- | -- | ↓ | -- |
| Gly | -- | -- | **79 ↑** | ↓ |
| Pro | -- | -- | -- | -- |
| Arg | **90 ↑** | **153 ↑** | **141 ↑** | **116 ↑** |
| Asp | -- | -- | **89 ↑** | -- |
| Glu | ↓ | -- | -- | ↓ |
| Lys | **107 ↑** | **132 ↑** | **104 ↑** | -- |
| Asn | -- | -- | -- | **75 ↑** |
| Gln | -- | -- | -- | -- |
| His | -- | -- | **40 ↑** | **40 ↑** |
| Ser | -- | -- | -- | -- |
| Thr | -- | -- | -- | -- |

**Tableau III-1 : Matrices d'interactions. Les couples favorisés dans les sites d'interaction sont donnés en gras avec une flèche vers le haut. Les couples défavorisés sont signalés par une flèche vers le bas et les couples indifféremment trouvés dans les sites d'interaction ou ailleurs dans la protéine sont signalés par un ' -- '. A gauche, interactions avec l'ADN et à droite, interactions avec l'ARN.**

Un test $\chi^2$ de Pearson à un degré de liberté nous a dès lors permis de détecter quels couples contribuaient le plus à cette différence par rapport à une distribution aléatoire. Cette étape permet, en plus de l'observation directe c'est-à-dire de l'observation des couples les

plus/moins représentés, de détecter lesquels sont importants pour la spécificité de la reconnaissance.

Globalement, les couples entre un nucléotide et un acide aminé aliphatique ou chargé négativement sont défavorisés alors que les couples avec un acide aminé chargé positivement ou, dans une moindre mesure, avec un résidu polaire sont favorisés. Pour les protéines interagissant avec l'ADN, l'alanine, l'isoleucine, la leucine, la valine, l'acide glutamique et l'acide aspartique sont défavorisés (flèches vers le bas, Tableau III-1, table de gauche) alors que les acides aminés chargés positivement sont favorisés. L'asparagine et la thréonine sont aussi favorisées excepté quand elles interagissent avec l'adénosine alors que l'histidine et la sérine sont favorisées quand ils interagissent avec la guanosine et la thymidine. Les résultats obtenus pour les acides aminés aromatiques montrent que la phénylalanine adopte un comportement similaire aux résidus aliphatiques (le couple Phe-A est défavorisé) alors que le tryptophane et la tyrosine adoptent un comportement semblable aux résidus polaires (Trp-C, Tyr-G et Tyr-T sont favorisés).

Dans la matrice des interactions avec l'ARN (Tableau III-1, table de droite), les couples avec les acides aminés chargés positivement sont à nouveau favorisés. Parmi les couples avec les acides aminés polaires, seulement Asn-U, His-G et His-U sont favorisés et, globalement, les couples avec l'alanine, la leucine, l'isoleucine et la valine sont défavorisés. Le couple Phe-U est défavorisé alors que le couple Tyr-U est favorisé ce qui confirme le comportement des résidus aromatiques observé dans la matrice des interactions avec l'ADN. Les acides aminés chargés négativement donnent quant à eux des résultats plus surprenants : seulement Glu-A et Glu-U sont défavorisés alors que l'acide aspartique est favorisé quand il interagit avec la guanosine. Le fait que les charges négatives des acides aspartique et glutamique soient moins contraignantes dans les complexes protéines-ARN semble être dû à la structure des acides nucléiques d'ARN (tout comme c'est le cas pour la distribution des interactions en fonction du type d'atome nucléotidique ; cf. point III.1.3). En effet, la structure principalement en simple brin de l'ARN doit permettre à une protéine d'entrer en contact avec l'ARN sans avoir besoin de passer la 'barrière' de phosphates chargés négativement (cf. paragraphe : « Couples les plus significativement favorisés dans les complexes protéines-ARN »).

### III.1.6. Couples significativement favorisés

Les couples significatifs mis en évidence par le test $\chi^2$ ont été classés selon la différence entre le $\chi^2$ observé et le $\chi^2$ théorique pour des interactions aléatoires. Ce classement n'a été effectué que pour les résidus favorisés (fréquence supérieure à celle attendue). Ensuite, nous avons construit des matrices d'interaction au niveau atomique (cf. point II.3.4) pour les résidus les plus favorisés.

**Couples les plus significativement favorisés dans les complexes protéines-ADN**

Dans les complexes avec l'ADN, les couples favorisés sont les suivants : Arg-G >> Arg-C > Lys-G > Lys-T >> Arg-T >> Lys-A >> His-G > Trp-C > Asn-C > Tyr-G > Asn-T > Arg-A > Ser-T > Thr-T > Thr-C > Lys-C > His-T > Thr-G > Tyr-T > Asn-G > Ser-G. Les caractéristiques des couples les plus favorisés vont être détaillées ci-après.

***Arg-G***

Avec 359 interactions sur les 7671 de la banque de données protéine-ADN, Arg-G est le couple le plus fréquent et, de loin, le plus favorisé. 47% des couples Arg-G correspondent à des interactions avec des atomes de la base de la guanosine (Figure III-15). C'est deux fois plus que les contacts avec les bases dans la distribution globale des fréquences d'interaction (Figure III-11). De ces 47%, 87% (145 couples) sont des liens H entre les atomes d'hydrogène de la fonction amine de l'arginine et les atomes accepteurs d'hydrogène de la guanine (N7 et O6 cf. Figure I-23). Sur la Figure III-14, on peut voir une arginine (atome HH1) d'une hélice α interagir avec une guanine (atome O6). Cette guanine est appariée avec une cytosine au cœur de la double hélice d'ADN (Figure III-14). Les atomes d'hydrogène de la fonction amine de l'arginine sont en fait impliqués dans 80% de l'ensemble des couples Arg-G. Les ponts salins entre les atomes chargés positivement de l'arginine et les charges négatives des oxygènes du phosphate de G comptent pour 28% des couples Arg-G. Les atomes du squelette peptidique de l'arginine sont rarement impliqués dans des contacts directs avec l'ADN et ceux-ci ont lieu presque exclusivement avec les oxygènes du phosphate (4,5%).

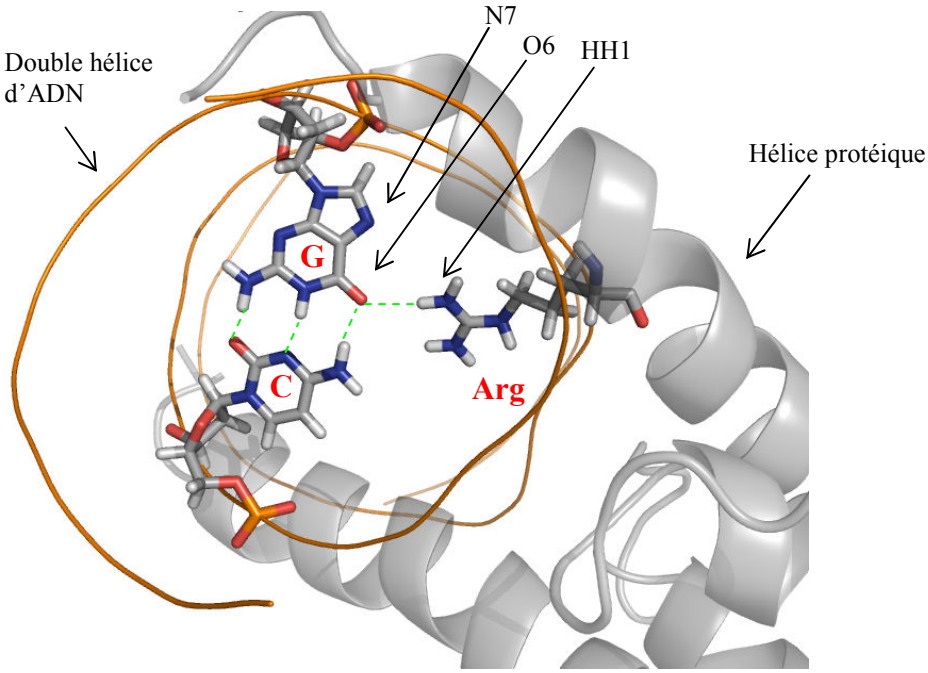



**Figure III-14 : Représentation d'un lien H entre une arginine (atome HH1) et une guanine (atome O6) du complexe à ADN, 1A0A[250] (facteur de transcription de type hélice-boucle-hélice). La guanine est appariée à une cytosine dans la double hélice (lignes orangées) et une arginine interagit avec l'oxygène (O6) de cette même guanine. L'arginine fait partie d'une hélice ('ruban' gris) qui est insérée dans le sillon. Imagé générée par le logiciel PyMol.[167]**

### *Arg-C*

Dans le cas du couple Arg-C, alors que les ponts salins sont encore très nombreux (29%), les liens H entre la chaîne latérale de l'arginine et la base de la cytidine sont moins nombreux avec seulement 22 contacts parmi les 300 (7,5%). Cette faible quantité de liens H peut être expliquée par le fait que la cytidine possède seulement un site accepteur d'hydrogène sur le sillon mineur et seulement un atome donneur sur le sillon majeur (Figure I-23). En compensation, les 'autres' interactions de van der Waals (cf. point II.3.1) entre l'arginine et les atomes du sucre et de la base de la cytidine correspondent à un tiers des 300 contacts. Elles sont presque exclusivement constituées de contacts entre les atomes chargés positivement de l'arginine et les atomes d'hydrogène de la cytosine. Finalement, les interactions de la chaîne latérale de l'arginine avec le ribose de la cytidine constituent 30.8% des interactions (Figure III-15), ce qui est largement supérieur à la moyenne trouvée dans l'ensemble des interactions protéine-ADN (22%, Figure III-11). Plus précisément, 28% des interactions avec le désoxyribose impliquent des contacts hydrophobes avec les hydrogènes aliphatiques de la chaîne latérale de l'arginine.

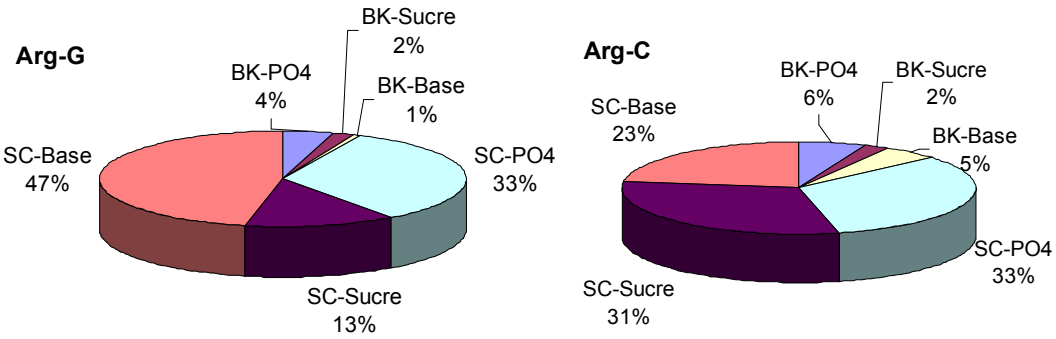



**Figure III-15 : Distribution des interactions dans les couples Arg-G et Arg-C. Les interactions sont différenciées selon le type d'atome impliqué (BK= squelette peptidique ou 'backbone' ; SC = chaîne latérale) : Backbone-phosphate ( ) ; Backbone-sucre ( ) ; Backbone-base ( ) ; Chaîne latérale-phosphate ( ) ; Chaîne latérale-sucre ( ) ; Chaîne latérale-base ( ).**

### *Arg-T*

Le couple Arg-T est le troisième couple le plus favorisé dans les interactions protéine-ADN. Plus de 50% de ces contacts impliquent les atomes du groupement phosphate (Figure III-16) alors que les liens H avec la base de thymidine sont moins fréquents que dans les couples Arg-G (13% au lieu de 41%). Comme pour les couples Arg-C, la faible quantité de liens H peut être expliquée par le fait que la thymine possède seulement deux sites accepteurs : un sur le sillon mineur et un sur le sillon majeur.

### *Lys-G*

De la même manière que pour les couples Arg-G, les contacts entre la lysine et la guanosine sont principalement soit des liens H entre les hydrogènes de la fonction amine et les deux atomes accepteurs d'hydrogène de la base de la guanosine (N7 et O6 ; 18%), soit des ponts salins (30%). Les atomes du squelette peptidique de la lysine sont impliqués dans seulement 15% des couples et 60% de ceux-ci sont des liens H entre l'hydrogène du lien peptidique et les oxygènes du phosphate.

### *Lys-T*

61% des interactions Lys-T font intervenir les phosphates (47% en moyenne dans la banque de données des résidus en interactions avec l'ADN (cf. Figure III-11) : 57% sont des ponts salins et 40% sont des liens H. Parmi les contacts hydrophobes (17%), 67,5% impliquent les atomes d'hydrogène du désoxyribose et 32,5% impliquent le groupe méthyle de la thymine. Comme on peut le voir dans la Figure III-16, la distribution des interactions de

la lysine et de l'arginine avec la thymidine sont similaires. On remarque néanmoins que les interactions entre le squelette peptidique et les groupements phosphates sont plus fréquents chez la lysine que l'arginine (13% au lieu de 5%).

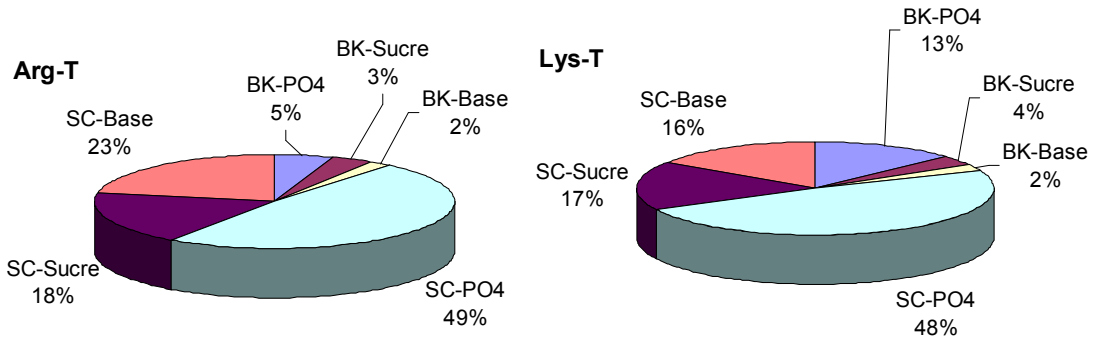



**Figure III-16 : Distribution des interactions dans les couples Arg-T et Lys-T. Les interactions sont différenciées selon le type d'atome impliqué (BK= squelette peptidique ou 'backbone' ; SC = chaîne latérale) : Backbone-phosphate ( ) ; Backbone-sucre ( ) ; Backbone-base ( ) ; Chaîne latérale-phosphate ( ) ; Chaîne latérale-sucre ( ) ; Chaîne latérale-base ( ).**

### *Lys-A*

Les principaux types d'interaction pour les couples Lys-A sont les ponts salins et les liens H avec 34% et 30,5% respectivement. Les liens H ont lieu principalement avec les atomes d'oxygène du phosphate (23% de l'ensemble des interactions). 14% des interactions sont des contacts hydrophobes entre les atomes d'hydrogène de la chaîne latérale de la lysine et les atomes d'hydrogène du sucre. D'une manière assez inattendue, les liens H spécifiques avec les atomes accepteurs d'hydrogène de l'adénine (N3 et N7 cf. Figure I-23) ne représentent que 4 interactions parmi les 248 recensées.

### *His-G*

Le couple His-G est le septième couple le plus favorisé et implique pour la première fois un acide aminé polaire. 86% des contacts sont des liens H (75/87) qui lient principalement les atomes d'hydrogène du cycle de l'histidine aux atomes accepteurs de la guanine (40%) ou aux atomes d'oxygène du phosphate (34% - cf. Figure III-17).

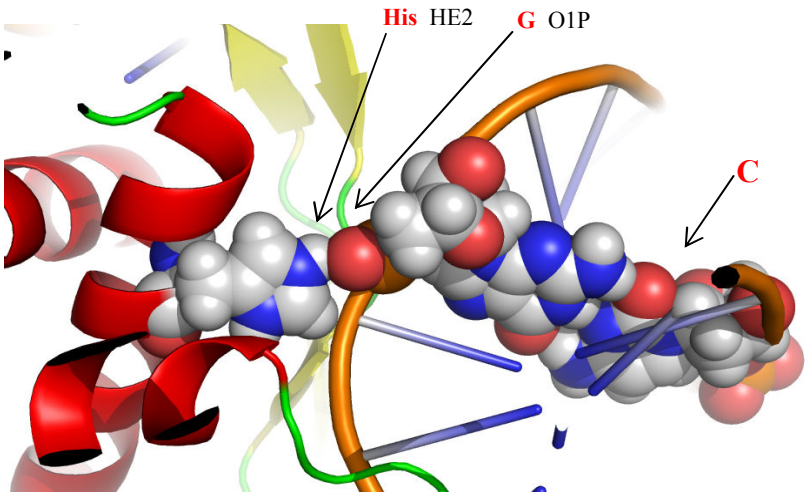



**Figure III-17 : Représentation d'un lien H entre une histidine (atome HE2) et une guanosine (atome O1P) au cœur du complexe 1CEZ.[251] La guanosine est appariée à une cytosine dans la double hélice d'ADN. La protéine est colorée en fonction de la structure secondaire : hélice en rouge, structures β en jaune et random coil en vert. Image générée par le logiciel PyMol.[167]**

### *Trp-C*

Le tryptophane correspond seulement à 1,5% des acides aminés en interaction mais près de la moitié interagit avec la cytidine (53 des 120 couples). 51% des interactions Trp-C sont des liens H avec les oxygènes du phosphate et 26% sont des contacts hydrophobes entre les hydrogènes de la chaîne latérale et les hydrogènes du désoxyribose (cf. Figure III-18).

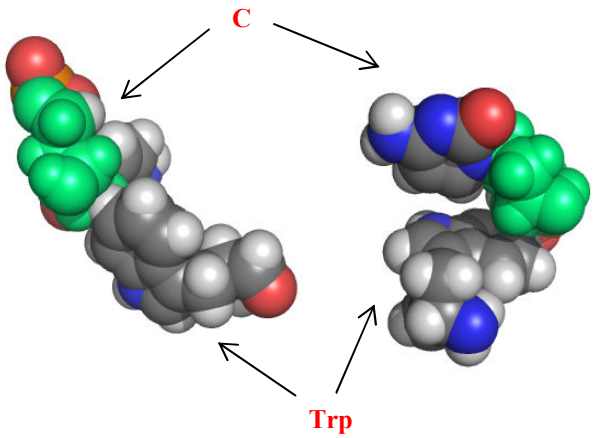



**Figure III-18 : Représentation de deux contacts hydrophobes entre la chaîne latérale d'un tryptophane et le désoxyribose d'une cytosine (complexe 1EYG).[252] Les atomes du désoxyribose sont mis en évidence en vert. Image générée par le logiciel PyMol.[167]**

### *Asn-C*

L'asparagine interagit préférentiellement avec la cytidine. Le principal type d'interaction est le lien H (82 interactions sur 134). Les atomes du squelette peptidique sont impliqués dans 34% de ces contacts. Les hydrogènes du squelette peptidique interagissent avec les oxygènes du phosphate et l'oxygène du lien peptidique interagit avec le groupement N-H en position 4 de la cytosine (Figure I-23). Les oxygènes du groupement phosphate forment également des liens H avec les atomes de la chaîne latérale des acides aminés (34% des liens H Asn-C). 38% des 37 interactions de van der Waals font intervenir les atomes d'hydrogène de la fonction amine et les atomes d'hydrogène du désoxyribose.

Les couples *Asn-T* qui sont également favorisés, montrent un comportement similaire.

### *Tyr-G*

Des quatre couples possibles avec la tyrosine, seul le couple avec la guanosine est favorisé. Bien que la tyrosine ait été classée dans ce travail comme un résidu hydrophobe, la présence d'un groupe hydroxyle sur son cycle phényle lui permet de réaliser des liens H avec des atomes accepteurs et notamment avec les groupements phosphate (cf. Figure III-19). Les liens H avec l'hydrogène du groupement hydroxyle correspondent à 30% de tous les contacts et à 55% de l'ensemble des liens H. D'un autre côté, les contacts hydrophobes entre les hydrogènes aromatiques et les hydrogènes du sucre correspondent seulement à 7% des interactions. Il apparait donc que la nature polaire de la tyrosine est prédominante quand elle interagit avec l'ADN. Finalement, les atomes accepteurs d'hydrogène de la guanine sont rarement impliqués (dans 5 cas seulement).

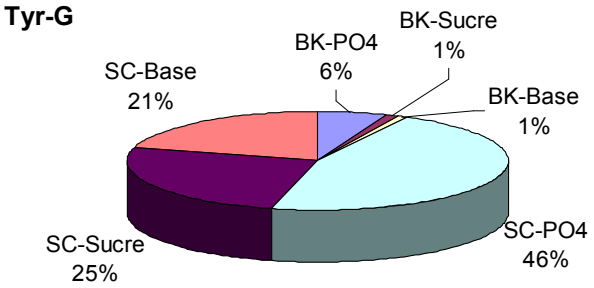



**Figure III-19 : Distribution des interactions entre la tyrosine et la guanosine. Les interactions sont différenciées selon le type d'atome impliqué (BK= squelette peptidique ou 'backbone' ; SC = chaîne latérale) : Backbone-phosphate ( ) ; Backbone-sucre ( ) ; Backbone-base ( ) ; Chaîne latérale-phosphate ( ) ; Chaîne latérale-sucre ( ) ; Chaîne latérale-base ( ).**

**Couples les plus significativement favorisés dans les complexes protéines-ARN**

Pour les complexes protéines-ARN (revoir le tableau Tableau III-1), le classement des couples favorisés est le suivant: Arg-C > Arg-G >> Lys-C > Asp-G >> Arg-U > Tyr-U > Asn-U > Lys-G > Gly-G > His-G > Lys-A > His-U > Arg-A. Comme dans le cas des couples impliquant l'ADN, seuls les couples les plus favorisés vont être détaillés.

### *Arg-C*

De tous les couples détectés dans les complexes protéine-ARN, Arg-C est le plus favorisé avec 153 interactions sur les 3367. Les résultats montrent que 83% des interactions se font avec la chaîne latérale de l'arginine (cf. Figure III-21) et que 64% des contacts impliquent les hydrogènes de la fonction amine. 21% des interactions sont de type pont salin entre les oxygènes du groupement phosphate et les hydrogènes de la fonction amine (Figure III-20), 15% sont des liens H impliquant le site accepteur de la base (O3 et N2 cf. Figure I-23) et 9% sont des contacts hydrophobes entre les hydrogènes aliphatiques de la chaîne latérale et les hydrogènes du ribose. Les hydrogènes du groupe hydroxyle du sucre de l'ARN sont impliqués dans 25 des 153 contacts (16%). Nous avons donc cherché des atomes accepteurs d'hydrogène en interaction avec cet hydroxyle mais les interactions avaient préférentiellement lieu avec les hydrogènes de la fonction amine de l'arginine (56%). Une observation des structures tridimensionnelles a montré que l'oxygène de la fonction hydroxyle était visé, plutôt que l'hydrogène, pour la construction d'un lien H.

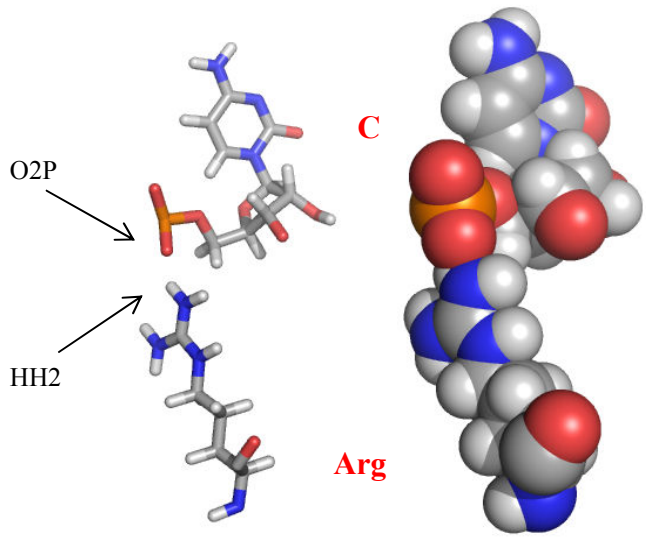



**Figure III-20 : Représentation d'un pont salin entre une arginine (HH2) et une cytidine (O2P) du complexe à ARN nucléaire, 1A9N.[253] Représentation en mode 'bâtons' à gauche et représentation en mode vrais volumes à droite. Imagé générée par le logiciel PyMol.[167]**

### *Arg-G*

Dans les couples Arg-G, de nouveau, les interactions se font très majoritairement via les atomes de la chaîne latérale (Figure III-21) et les hydrogènes de la fonction amine (les atomes de l'arginine les plus fréquemment impliqués avec 67%). 24% des interactions sont des ponts salins et 16% sont des liens H avec les sites accepteurs de la guanine (O6 et N7). 8% des interactions sont des contacts hydrophobes entre l'hydrogène du carbone α du squelette peptidique de l'acide aminé et les atomes d'hydrogène du sucre. 16% des couples Arg-G impliquent les atomes du squelette peptidique. Ce pourcentage est plus élevé que dans les autres couples Arg-nucléotide mais est encore presque deux fois moins élevé que la valeur moyenne des résidus en interaction avec l'ARN (Figure III-11). De nouveau, les atomes d'hydrogène du groupe hydroxyle du ribose sont largement impliqués dans les interactions (10%) mais sans préférence pour un atome précis de l'arginine.

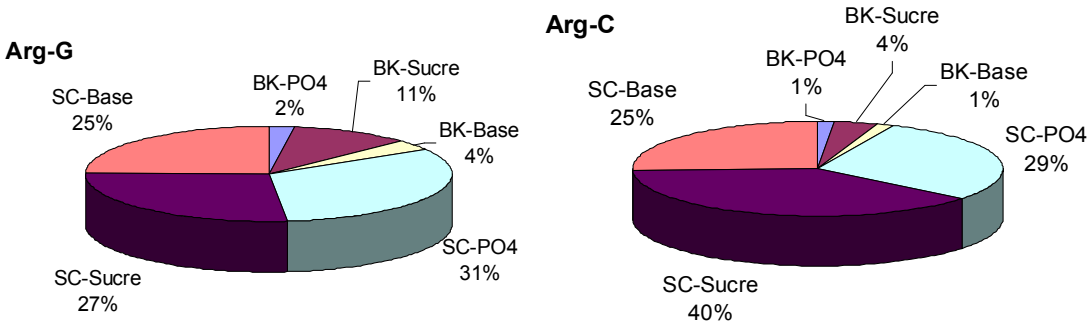



**Figure III-21 : Distribution des interactions dans les couples Arg-G et Arg-C. Les interactions sont différenciées selon le type d'atome impliqué (BK= squelette peptidique ou 'backbone' ; SC = chaîne latérale) : Backbone-phosphate ( ) ; Backbone-sucre ( ) ; Backbone-base ( ) ; Chaîne latérale-phosphate ( ) ; Chaîne latérale-sucre ( ) ; Chaîne latérale-base ( ).**

### *Lys-C*

48% des couples Lys-C impliquent les oxygènes du phosphate (67% sont des ponts salins). Ce pourcentage est deux fois plus élevé que la valeur moyenne pour les complexes protéine-ARN (22% ; Figure III-11, graphique de droite). Les contacts hydrophobes correspondent à 15% des couples et 42% des interactions avec les hydrogènes aliphatiques de la chaîne latérale de la lysine. Finalement, il n'y a que 14 liens H avec la cytidine parmi les 132 couples (10,5%).

### *Asp-G*

Alors que les trois premiers couples favorisent des acides aminés chargés positivement, le quatrième implique un nucléotide régulièrement favorisé (G) et un acide aminé étonnamment favorisé dans les complexes avec l'ARN : l'acide aspartique. 64 des 89 couples détectés impliquent les atomes de la base de la guanosine (72%) ce qui est deux fois plus que la moyenne dans les complexes protéine-ARN (34% ; Figure III-11, graphique de droite). 67% des contacts avec les atomes de la base sont des liens H entre les atomes d'oxygène de la fonction carboxylique de l'acide aspartique et deux atomes d'hydrogène de la guanine (H1 et H2 ; Figure III-22). Dans les double hélices d'acides nucléiques, ces mêmes hydrogènes sont impliqués dans les interactions spécifiques entre une guanine et une cytidine et contribuent à stabiliser par liens H l'hélice de Watson et Crick (Figure I-23).

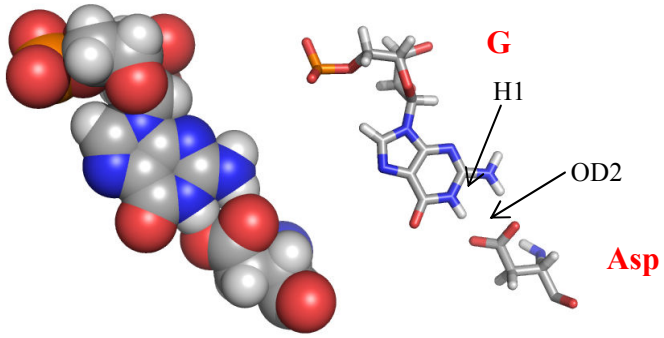



**Figure III-22 : Représentation d'une interaction (lien H) entre un acide aspartique (OD2) et une guanine (H1) d'une ARN-t synthétase, 1H3E.[254] Représentation en mode 'bâtons' (à droite) et représentation en vrais volumes (à gauche). Imagé générée par le logiciel PyMol.[167]**

### *Arg-U*

Les couples Arg-U se comportent de façon semblable aux couples Arg-C. Les ponts salins sont fréquents (27,5%) et les contacts entre les hydrogènes de la fonction amine de l'acide aminé et l'hydrogène du groupe hydroxyle du nucléotide représentent 11% de l'ensemble des couples. Comme pour les couples Arg-C, l'existence de ce type d'interaction pourrait masquer la présence de liens H impliquant l'oxygène du groupement hydroxyle du ribose. Les hydrogènes de la fonction amine sont impliqués dans 17 contacts avec les oxygènes du ribose (15%) mais seulement dans 10 liens H spécifiques avec les sites accepteurs de la base (9% ; O4 et N2).

### *Tyr-U*

Les résultats pour les couples Tyr-U sont assez surprenants. 54% des contacts impliquent les atomes du squelette peptidique protéique et principalement l'atome d'oxygène. Concrètement, 46% des contacts sont des liens H entre l'oxygène du squelette peptidique et l'hydrogène de la fonction amine de l'uracile (H3). Dans les acides nucléiques en conformation de double hélice, cet hydrogène forme un lien H spécifique en combinaison avec l'atome d'azote de l'adénine (N1) et on se retrouve, comme dans le cas des couples Asp-G, face à une interaction spécifique. Comme on peut le voir dans la Figure III-23, ce lien H est peut-être complémenté d'une interaction de type « empilement » ('stacking') entre le cycle de la tyrosine et celui de la guanine, les deux cycles se plaçant de manière presque parallèle.

La deuxième moitié des interactions impliquent principalement le groupement hydroxyle de la tyrosine (23%) et des contacts hydrophobes entre les hydrogènes du ribose et les hydrogènes aromatiques de la tyrosine (15%).

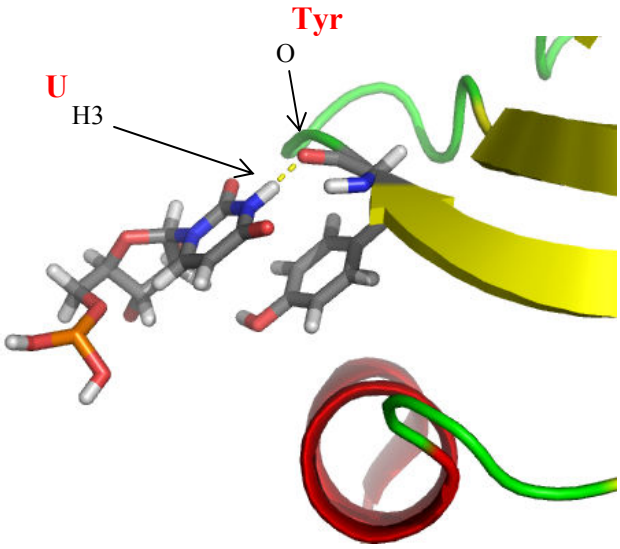



**Figure III-23 : Représentation d'un lien H entre une tyrosine (atome d'oxygène du squelette peptidique - O) et une uracile (atome H3) au cœur du complexe 1M8V.[255] Le lien H est indiqué en pointillés jaunes. La protéine est colorée en fonction de la structure secondaire : hélice en rouge, structures β en jaune et 'random coil' en vert. Image générée par le logiciel PyMol.[167]**

### *Asn-U*

L'asparagine interagit principalement avec l'ARN par le biais de sa fonction amide (79%). Les atomes de l'uridine impliqués sont majoritairement les atomes de la base (56%). Les liens H se réalisent principalement entre les deux atomes accepteurs de l'uracile (O2 et O4 cf. Figure I-23) et les hydrogènes du groupement amine (33%) ainsi qu'entre l'hydrogène

habituellement impliqué dans les liens H spécifiques de l'hélice nucléotidique (H3) et l'oxygène de la fonction amide de l'asparagine (8%).

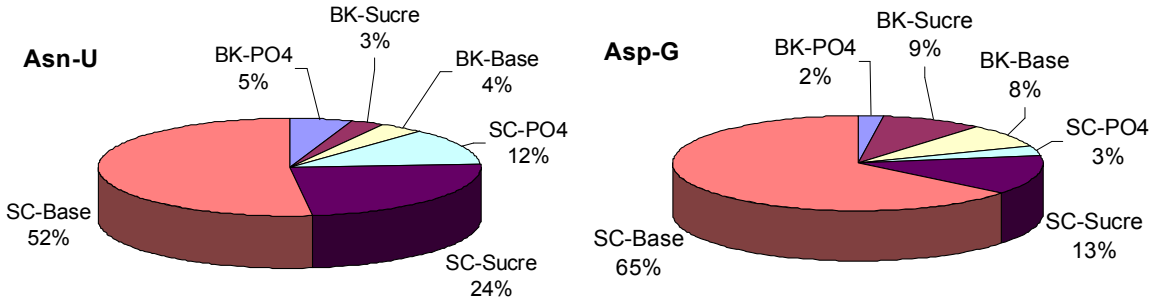



**Figure III-24 : Distribution des interactions dans les couples Asn-U et Asp-G. Les interactions sont différenciées selon le type d'atome impliqué (BK= squelette peptidique ou 'backbone' ; SC = chaîne latérale) : Backbone-phosphate ( ) ; Backbone-sucre ( ) ; Backbone-base ( ) ; Chaîne latérale-phosphate ( ) ; Chaîne latérale-sucre ( ) ; Chaîne latérale-base ( ).**

## III.1.7. Influence de la double hélice d'ADN

Comme décrit au point II.4, chaque nucléotide d'une banque de données reprenant des complexes protéine–double hélice d'ADN de haute résolution a été classé comme faisant partie de l'ADN de type A (A-ADN), de type B (B-ADN), de type Z (Z-ADN) ou comme ne faisant partie d'aucune de ces catégories (---). Cette classification se base sur les angles de torsion δ et χ. La composition des doubles hélices de notre banque de données est présentée dans la Figure III-25, en bleu.

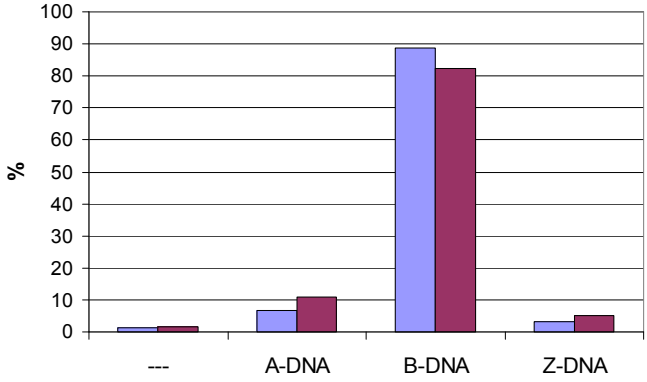



**Figure III-25 : Distribution des types d'ADN dans la banque de donnée de double hélice à haute résolution (<2Å). La distribution pour l'ensemble de la banque est donnée en bleu ( ) et celle pour les nucléotides en interaction est donnée en rouge ( ).**

Le B-ADN est le principal type de structure rencontré et représente 89% des nucléotides. Seulement 7% sont dans une conformation de type A. Si on considère les nucléotides interagissant avec les protéines (Figure III-25 en rouge), le B-ADN est moins fréquent mais toujours largement majoritaire avec 82% des nucléotides. Le A-ADN et le Z-ADN représentent respectivement 11% et 5%. Bien que le A-ADN et le Z-ADN soient plus fréquents dans les sites d'interaction que dans l'ensemble de la banque, le pourcentage très élevé de B-ADN nous empêche de déterminer si un de ces deux types de double hélice est favorisé dans les interactions.

Les acides aminés en interaction avec les deux plus importantes formes d'ADN (A et B) sont distribués de manière relativement différente (Figure III-26). En effet, parce que le B-ADN est le type d'ADN le plus représenté, les acides aminés en interaction avec ce type de structure suivent une distribution similaire à celle de l'ensemble des complexes protéines-ADN (comparer la Figure III-26 en rouge à la Figure III-4 en rouge). Les plus grandes variations sont observées pour l'arginine (17% au lieu de 14,5%) et la lysine (17,5% au lieu de 14%). La comparaison de la distribution des acides aminés en interaction avec le A-ADN et l'ensemble des ADN (comparer la Figure III-26 en bleu à la Figure III-4 en rouge) montre que l'alanine, l'asparagine et l'acide aspartique sont deux fois plus présents (9,5%, 14,5% et 6,5% au lieu de 4,5%, 6,5% et 3%, respectivement) alors que l'arginine et la lysine sont moins fréquents (6% au lieu de 14,5% pour l'arginine et 9% au lieu de 14% pour la lysine). Il est intéressant de remarquer que la distribution des résidus en interaction avec l'A-ADN est plus semblable à la distribution des acides aminés en interaction avec l'ARN (Figure III-5 en rouge) qu'avec la distribution des interactions avec l'ADN total.

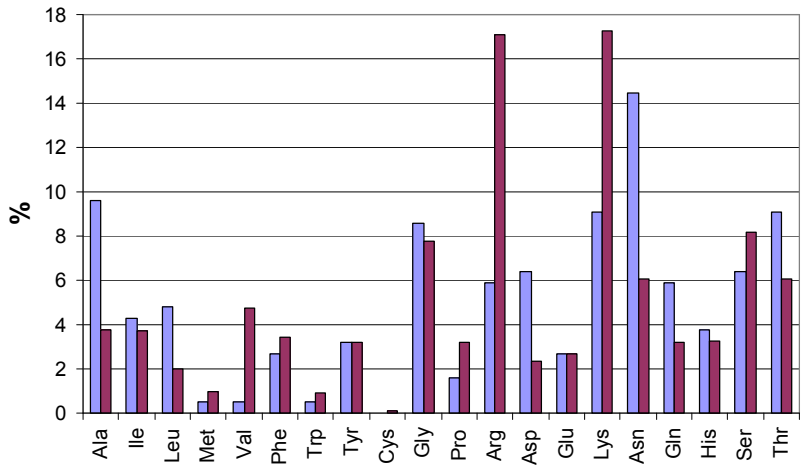



**Figure III-26 : Distribution des acides aminés en interaction avec les doubles hélices d'ADN haute résolution. Les interactions avec l'A-ADN et le B-ADN sont représentées en bleu ( ) et en rouge ( ) respectivement.**

### III.1.8. Résumé des résultats obtenus

Lors des précédents paragraphes, nous avons étudié en détails les propriétés des interfaces entre les protéines et les acides nucléiques. Avant d'étudier les interfaces protéine-protéine, il est utile de résumer les informations obtenues jusqu'ici :

- Les acides aminés chargés positivement sont les plus représentés au sein des interfaces protéine-acide nucléique. Viennent ensuite les résidus polaires (y compris la tyrosine).
- Une tendance identique est retrouvée pour les couples de résidus en interactions. Les couples avec l'arginine et la lysine sont les plus favorisés dans les interfaces.
- La structure, généralement en simple brin, de l'ARN influence les résultats obtenus. Des interactions spécifiques entre las bases nucléotidiques de l'ARN et les acides aminés chargés négativement (principalement l'acide aspartique) ont notamment été mises en évidence.
- Les liens H correspondent au type d'interaction le plus largement retrouvé dans les interfaces protéine-acide nucléique.
- La méthode de classification des doubles hélices d'ADN mise au point durant cette thèse a permis de montrer qu'un enrichissement en ADN de type A au niveau des interfaces est fort probable.

## *III.2. Interactions Protéines - Protéines*

### III.2.1. Caractéristiques générales

La banque de données de complexes protéine-protéine est l'une des plus grandes construites à ce jour et contient 1297 complexes avec une moyenne de 229 résidus par chaîne. Ces complexes sont répartis en 3470 chaînes protéiques qui forment 6175 paires chaîne-chaîne distinctes. La taille minimale d'une chaîne a été fixée à 10 résidus alors que la plus longue en compte 1267. Le nombre de chaînes protéiques par complexe varie de 1 à 10 et la résolution varie de 0,8Å à 2,5Å.

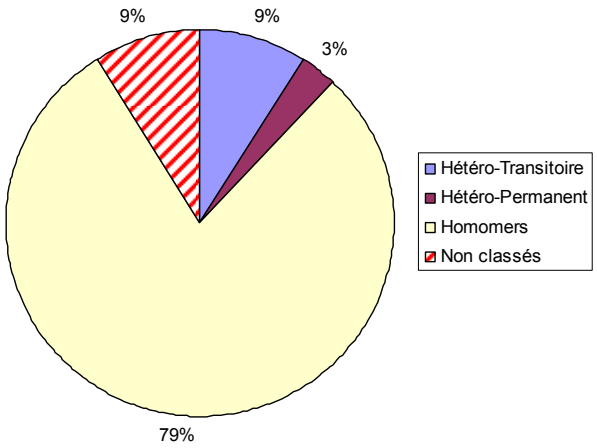


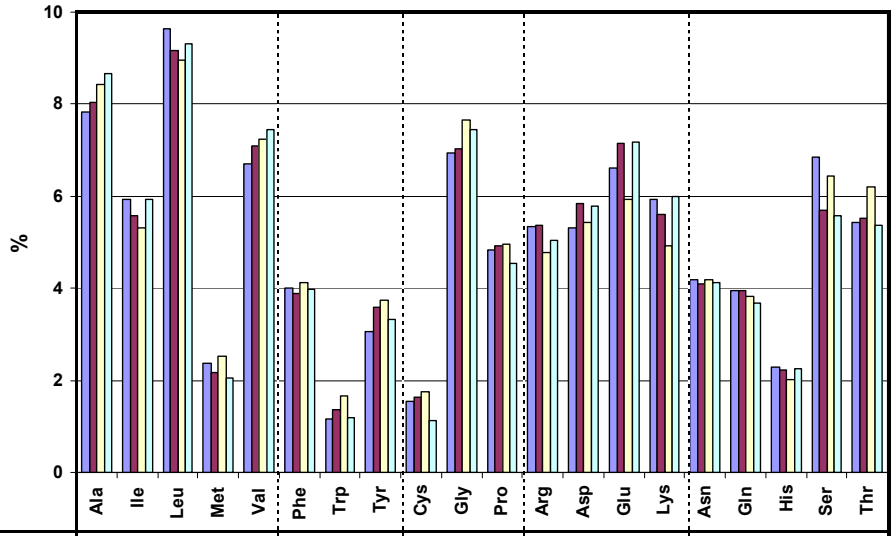

**Figure III-27 : Différents types de complexes protéines-protéines.**

Les interfaces de type homomériques sont les plus fréquentes (4610 - 79%), les interfaces de type hétéromer-transitoires et hétéromer-permanents sont au nombre de 606 et 126, respectivement (9% et 3%).

### III.2.2. Composition des banques et sous-banques de données

#### Analyse de la banque totale

Comme nous l'avons fait pour les complexes protéine-acide nucléique, nous allons dans un premier temps comparer les fréquences des résidus trouvés dans notre banque de données à celles des acides aminés dans la banque de données SwissProt/UniProt.




| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| SwissProt | 32,5 | 8 | 13,5 | 23 | 23 |
| Hetero-Transitoire | 32 | 9 | 13,5 | 24 | 21,5 |
| Hetero-Permanent | 32,5 | 9,5 | 14 | 21 | 23 |
| Homo | 33,5 | 8,5 | 13 | 24 | 21 |
| Prot-Prot Totale | 33 | 8,5 | 13,5 | 23,5 | 21,5 |

**Figure III-28 : Comparaison des fréquences des acides aminés dans la banque de référence (SwissProt/UniProt - ) et dans la banque de complexes protéiques. Les interfaces protéines-protéines sont divisés en complexes de type hétéromer-transitoires ( ), hétéromer-permanents ( ) et homomériques ( ). Les fréquences par familles sont reprises dans le tableau en bas du graphique avec les valeurs obtenues pour l'ensemble de la banque de complexes protéines-protéines.**

Sur la Figure III-28, on peut voir que, globalement, les distributions dans les différents types d'interfaces protéines-protéines et dans la SwissProt/UniProt sont similaires. Ce résultat confirme que notre banque de données est représentative des protéines communément trouvées dans les organismes vivants. On remarque également que les rapports hydrophobe/hydrophile des quatre banques de données montrent un enrichissement en résidus hydrophobes avec des rapports allant de 54/46 à 56/44. Les interfaces les plus hydrophobes sont les interfaces hétéromer-permanentes (56% de résidus aliphatiques, aromatiques et particuliers). Cette particularité provient principalement d'une proportion plus faible de résidus chargés (21% au lieu de 23% dans la SwissProt/UniProt). La composition globale des interfaces hétéromer-transitoires et homomériques est très proche et l'on remarque surtout un enrichissement en acides aminés aliphatiques (33.5% pour les interfaces homomériques et 32% pour les interfaces hétéromer-transitoires).

Pour la suite de l'analyse des résultats, les fréquences pour l'ensemble de la banque de complexes protéine-protéine seront utilisées comme valeurs de références.

**Extraction des résidus en interactions**

Dans cette partie du travail, nous avons voulu réaliser une sélection plus fine des résidus en interactions pour ne conserver que les acides aminés les plus importants. Le critère de distance initial de 5 Å (cf. point III.1.2) a donc été modifié. Le cut-off de distance a été choisi cette fois en fonction du type d'interaction rencontré. Les distances limites ont été déterminées en fonction de deux critères : les données trouvées dans la littérature et l'analyse des courbes du nombre d'interaction en fonction de la distance (cf. Figure III-30 et Figure III-31).

L'approche bibliographique a principalement permis de récolter des valeurs sur diverses sources internet. En effet, les articles scientifiques centrés sur les distances optimales d'interactions ne sont pas nombreux. Pour les ponts salins, les distances trouvées sont variables (2.8, 3.35, 4, 5 et 7Å). Les données sont plus précises pour les liens H où les angles et les distances entre les différents atomes sont bien définis (distance donneur-accepteur de 3.9Å et distance hydrogène-accepteur de 2.5Å).[256] Il en est de même pour les ponts disulfures pour lesquels la distance optimale entre les deux atomes de soufre se situe entre 2 et 3Å. Cette distance est de 3.45 à 4.5Å entre les deux carbones beta et de 5 à 5.9Å entre les deux carbones alpha (cf. Figure III-29). Pour ce qui est des contacts hydrophobes, aucune valeur de référence n'a été trouvée.

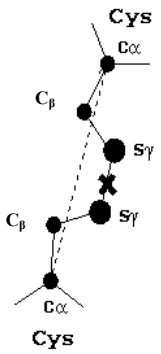



**Figure III-29 : Représentation schématique d'un pont disulfure.**

Dans la Figure III-30 et la Figure III-31, on peut observer la distribution du nombre d'interactions en fonction de la distance pour l'ensemble des atomes de l'échantillon analyse de notre base de données (cf. point III.3.2). Le nombre de données pour les ponts disulfures était trop faible pour réaliser une étude significative.

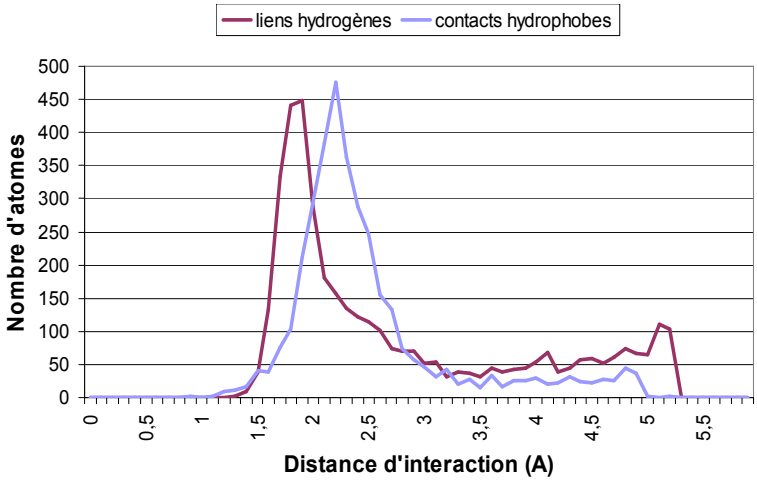



**Figure III-30 : Distribution du nombre d'interactions en fonction de la distance pour les liens H et les interactions hydrophobes.**

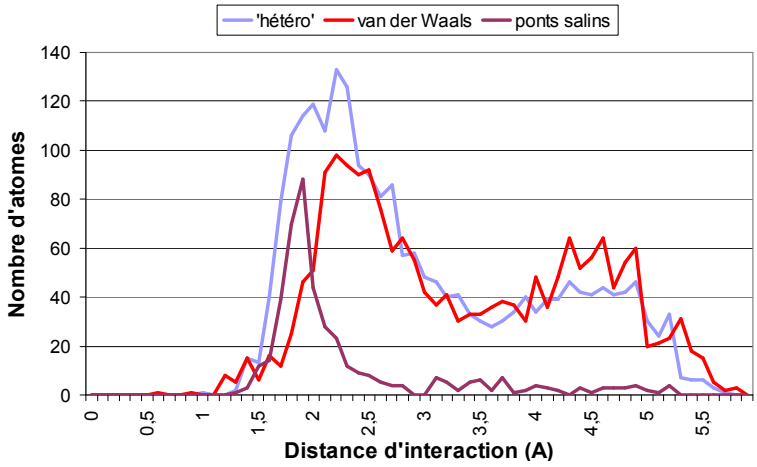
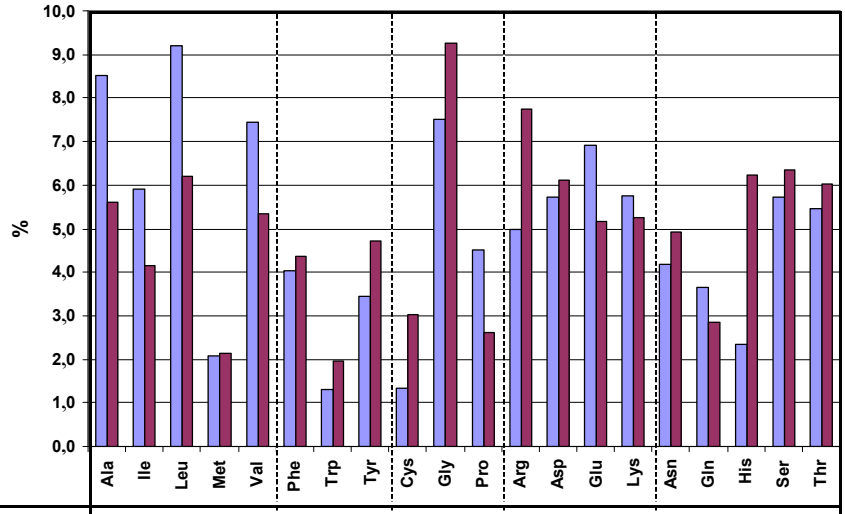



**Figure III-31 : Distribution du nombre d'interactions en fonction de la distance pour les interactions avec les hétéroatomes, les interactions de van der Waals et les ponts salins.**

Les courbes pour les liens H, les contacts hydrophobes et les ponts salins sont clairement unimodales avec une distance optimale à 1.9, 2.2 et 1.9Å, respectivement. Les courbes des interactions de type van der Waals et 'hétéro' sont elles plus difficiles à interpréter. Pour les hétéro-interactions, un deuxième maximum de faible intensité semble présent à environ 4.5Å. Néanmoins, seul le premier maximum sera utilisé pour conserver les interactions les plus fortes. Les interactions de type van der Waals correspondent à des interactions moins spécifiques et elles ne seront donc pas conservées lors de la sélection des résidus en interaction.

### Composition de la banque des résidus en interaction avec les hétéroatomes

Dans nos complexes protéiques, 16,6% des résidus sont en interaction. 13,2% des acides aminés sont en interaction avec d'autres acides aminés alors que les 3,4% restants sont des acides aminés en interaction avec des hétéroatomes (ions principalement, modifications post-traductionnelles…). Ce type d'interactions représente 20% de la totalité des interactions et doit donc être pris en compte.



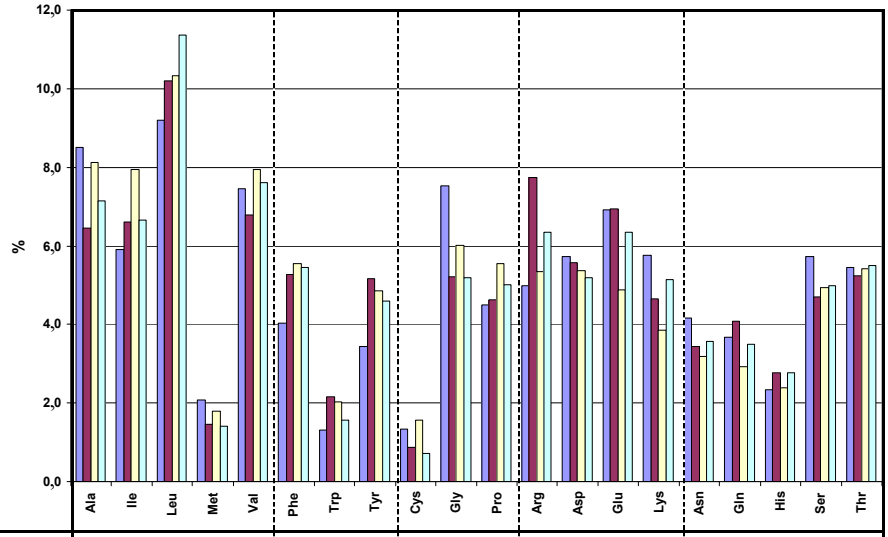

| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| Prot-Prot | 33 | 8,5 | 13,5 | 23,5 | 21,5 |
| Hetero-atomes | 23,5 | 11 | 15 | 24 | 26,5 |

**Figure III-32 : Comparaison des fréquences des acides aminés dans la banque de complexes protéine-protéine totale ( ) et pour les interactions avec les hétéroatomes ( ). Les fréquences par familles sont reprises dans le tableau en bas du graphique.**

Les interactions avec les hétéroatomes sont caractérisées par une faible proportion en résidus aliphatiques (23,5% au lieu de 33% dans la banque totale). Les résidus polaires montrent un comportement inverse avec 26,5% des interactions ce qui est principalement dû à l'histidine qui passe de 2,3% dans la banque totale à 6,2% pour les interactions avec les hétéroatomes. Le même constat est fait pour la cystéine (de 1,3% à 3,0%) et peut-être expliqué par la capacité de ces deux résidus à se lier aux ions comme, par exemple, le zinc (15-20% des interactions). L'arginine et les résidus aromatiques sont également plus fréquents en interactions avec les hétéroatomes que dans la banque totale (7,7% au lieu de 5,0% et 11% au lieu de 8,5%, respectivement).

### Composition de la banque des interactions protéine-protéine

13% des acides aminés (104.782 résidus) sont en interaction directe avec un autre acide aminé. Parmi ces résidus, 9% sont impliqués dans des interfaces de type hétéromer-transitoires, 3% dans des interfaces de type hétéromer-permanentes et 79% dans des interfaces de type homomériques (le reste des interfaces n'a pas pu être classé). Le plus petit groupe (hétéromer-permanent) comporte 3148 interactions ce qui reste suffisant pour une analyse statistique.




| % | Aliphatiques | Aromatiques | Particuliers | Chargés | Polaires |
|---|---|---|---|---|---|
| Prot-Prot Totale | 33 | 8,5 | 13,5 | 23,5 | 21,5 |
| Hetero-Transitoire | 31,5 | 12,5 | 10,5 | 25 | 20,5 |
| Hetero-Permanent | 36 | 12,5 | 13 | 19,5 | 19 |
| Homo | 34 | 11,5 | 11 | 23 | 20,5 |
| Interactions Totales | 33,5 | 12 | 11 | 23 | 20,5 |

**Figure III-33 : Comparaison des fréquences des acides aminés dans la banque totale ( ) et dans les sous-banques de résidus en interaction. Les interactions protéine-protéine sont divisées selon leur appartenance à des interfaces de type hétéromer-transitoire ( ), hétéromer-permanent ( ) et homomérique ( ). Les fréquences par familles sont reprises dans le tableau en bas du graphique avec les valeurs obtenues pour l'ensemble des interactions des complexes protéines-protéines.**

Les résidus aromatiques montrent clairement une préférence pour les sites d'interaction avec en moyenne 12% de participation dans les sites d'interaction des différents types de complexes (à comparer avec les 8,5% trouvés dans l'ensemble de la banque - Figure III-33). La deuxième famille particulièrement favorisée dans les sites d'interaction est celle des résidus aliphatiques (36% au lieu de 33%) mais ce, uniquement dans les complexes de

type hétéromer-permanents et au détriment des résidus chargés et polaires (19,5% et 19% au lieu de 23,5% et 21,5%, respectivement).

Pour détecter plus facilement les acides aminés favorisés dans les sites d'interaction, nous avons calculé leur propension à être en interaction (cf. point II.3.2). Les résultats calculés sont représentés dans la Figure III-34 :

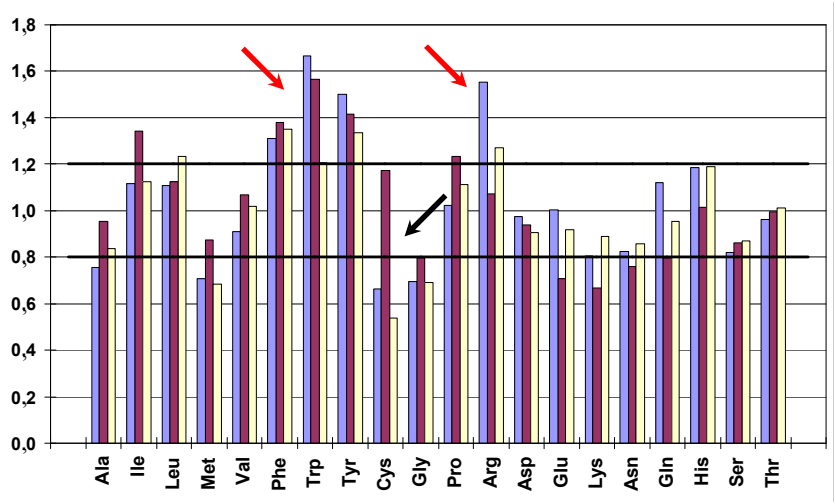



**Figure III-34 : Propensions des acides aminés à interagir. Les valeurs pour les interfaces de type hétéromer-transitoire ( ), hétéromer-permanente ( ) et homomérique ( ) sont représentées.**

Les propensions calculées confirment l'importance des résidus aromatiques dans les trois types d'interfaces (propension de 1,4). Pour les interfaces de type hétéromer-transitoire (en bleu dans la Figure III-34), l'arginine (1,5) est le deuxième résidu le plus favorisé après le tryptophane (1,7) mais avant la tyrosine (1,5) et la phénylalanine (1,3). L'histidine est légèrement favorisée avec une propension de 1,2. La cystéine, la glycine et la méthionine (0,7) sont des acides aminés défavorisés dans les interactions protéine-protéine.

Du côté des interfaces de type hétéromer-permanent (en rouge dans la Figure III-34), les résidus aromatiques sont encore favorisés (1,4) mais à l'inverse des résultats pour les interfaces transitoires, les acides aminés aliphatiques sont légèrement favorisés (1,1), et en particulier, l'isoleucine (1,3). La proline (1,2) et la cystéine (1,2) sont des acides aminés trouvés préférentiellement aux interfaces alors que l'asparagine (0,8) et l'acide glutamique (0,7) y sont défavorisés.

Finalement, l'analyse des interfaces de type homomérique (en jaune dans la Figure III-34), met en évidence l'appauvrissement en acides aminés contenant un atome de soufre (propension de 0,5 pour la cystéine et de 0,7 pour la méthionine). A l'inverse, les résidus aromatiques (1,3), l'arginine (1,3), la leucine (1,2) et l'histidine (1,2) sont des partenaires d'interaction favorisés.

La cystéine montre un comportement particulier : elle est défavorisée dans les complexes de type hétéromer-transitoire (0,7) et homomérique (0,5) mais est favorisée dans les interfaces des complexes de type hétéromer-permanent (1,2). Ce comportement sera discuté dans le Chapitre IV.

### III.2.3. Influence du voisinage des résidus en interaction

Les résultats présentés ci-dessus correspondent aux acides aminés impliqués dans des interactions directes (à courte distance). Ces résidus sont spatialement proches dans la structure tridimensionnelle mais ne sont pas nécessairement proches dans la séquence. Dans ce paragraphe, nous allons nous intéresser plus particulièrement aux résidus séquentiellement proches des résidus en interactions, à leurs 'voisins' pour essayer de déterminer si ceux-ci influencent les résidus en interaction.

**Analyse des résidus voisins des résidus en interaction**

Nous allons étudier les acides aminés en positions +/-1 à +/-5 des acides aminés en interaction. Ils représentent 14,9% de l'ensemble des résidus ce qui est comparable aux 13% de résidus en interaction. La méthionine, la glycine et la cystéine sont favorisées de la position +/-1 à +/-3 et ont des propensions de 1.3, 1.2 et 1.2, respectivement. Les acides aminés aromatiques, qui étaient favorisés dans les sites d'interaction, sont défavorisés au voisinage des résidus en interaction. Ensuite, nous avons comparé les fréquences des résidus situés dans le voisinage des interfaces à celles des résidus en interaction. La glycine est l'acide aminé le plus fréquemment trouvé dans le voisinage des interfaces (9,7%) alors qu'elle n'est pas fréquente dans les sites d'interaction (6,0%) (variation de fréquence de 3,7%, cf. Figure III-35). Inversement, l'arginine qui est un des résidus les plus fréquents dans les sites d'interaction, est le moins fréquent aux abords de ceux-ci (variation de fréquence de -2,3%, cf. Figure III-35).

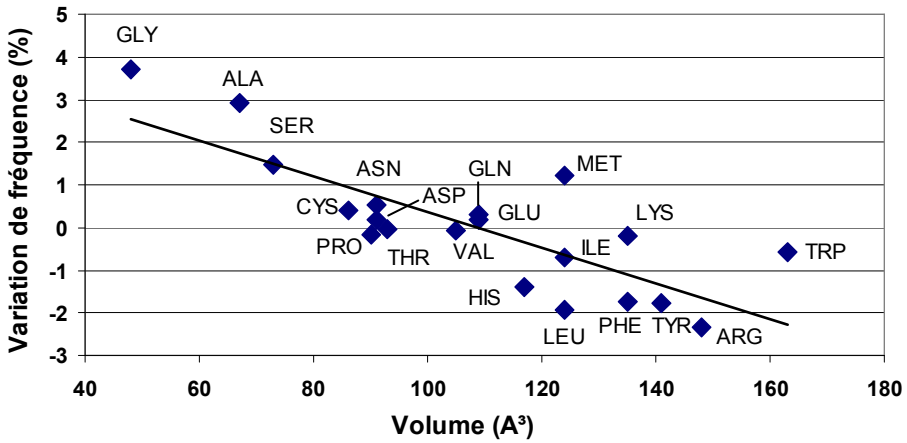



**Figure III-35 : Représentation de la différence entre la fréquence des résidus voisins (position +/-1 d'un résidu en interaction) et la fréquence des résidus en interaction. Ces variations sont données en fonction du volume des acides aminés. La ligne de régression est mise en évidence (y = -0.0421x + 4.5785) et a un coefficient de corrélation de 0.642.**

Il est intéressant de voir que l'on peut corréler la taille des résidus (volume en Å³) à leur occurrence dans le voisinage des acides aminés en interaction ($R^2$ = 0,642 dans la Figure III-35). La corrélation est négative : plus un acide aminé est petit, plus sa probabilité de se trouver près d'un site d'interaction est grande. Le faible encombrement des résidus en position +/-1 des interfaces est probablement nécessaire pour accorder une flexibilité suffisante aux résidus en interaction. Dès lors, ces résidus voisins pourraient avoir un rôle à jouer dans la présentation des résidus en interaction en surface.

Un rôle particulier pourrait être joué par les cystéines. En effet, elles permettraient d'apporter une contrainte conformationnelle aux abords du site d'interaction en étant impliquées dans des ponts disulfures intramoléculaires. Nous avons étudié plus en détail ces cystéines et avons pu montrer que celles-ci étaient plus fréquemment impliquées dans ce type de liaison intramoléculaire dans le voisinage des sites d'interaction qu'ailleurs dans la protéine. En effet, en moyenne dans notre banque de données, 22,8% des cystéines sont impliquées dans des ponts disulfures intramoléculaires alors que cette valeur augmente jusque 32,1% en position +/-5 des acides aminés en interaction.

**Analyse détaillée du voisinage des résidus en interaction les plus favorisés**

Ensuite, nous avons étudié plus en détail le voisinage (position +/-1) des résidus les plus favorisés dans les sites d'interaction : arginine, histidine, leucine, phénylalanine, tryptophane et tyrosine. Pour cela, nous avons calculé deux types de propensions : la première (P1), permet de détecter si un résidu est proche de, par exemple, l'arginine dans la banque totale. La deuxième (P2) permet de détecter ci ce résidu est proche des arginines en interaction. De cette manière, nous pouvons analyser le voisinage des résidus en interaction (P2) en le comparant à leur voisinage habituel (P1).

| | Arg | | His | | Leu | | Phe | | Trp | | Tyr | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 | P1 | P2 |
| ALA | -- | -- | -- | -- | -- | -- | -- | -- | -- | ↓ | -- | -- |
| ILE | -- | -- | -- | 1,10 ↑ | ↓ | ↓ | -- | ↓ | ↓ | ↓ | -- | -- |
| LEU | 1,12 ↑ | 1,12 ↑ | -- | -- | -- | -- | ↓ | ↓ | ↓ | ↓ | -- | ↓ |
| MET | -- | **1,22** ↑ | -- | -- | -- | -- | ↓ | -- | ↓ | ↓ | ↓ | -- |
| VAL | -- | -- | -- | -- | -- | -- | -- | -- | ↓ | ↓ | -- | -- |
| PHE | -- | -- | -- | -- | ↓ | -- | -- | -- | -- | -- | -- | 1,16 ↑ |
| TRP | -- | 1,16 ↑ | **1,20** ↑ | **1,23** ↑ | ↓ | ↓ | -- | -- | **1,24** ↑ | **1,78** ↑ | 1,12 ↑ | **1,28** ↑ |
| TYR | -- | -- | -- | -- | ↓ | ↓ | -- | -- | -- | 1,16 ↑ | -- | 1,18 ↑ |
| CYS | -- | **1,26** ↑ | 1,11 ↑ | -- | -- | -- | -- | -- | -- | ↓ | 1,15 ↑ | **1,22** ↑ |
| GLY | -- | -- | -- | **1,26** ↑ | -- | -- | -- | **1,37** ↑ | -- | **1,23** ↑ | -- | 1,12 ↑ |
| PRO | -- | -- | **1,21** ↑ | -- | -- | -- | -- | -- | ↓ | -- | -- | 1,10 ↑ |
| ARG | -- | -- | -- | ↓ | -- | -- | -- | -- | -- | ↓ | -- | -- |
| ASP | -- | ↓ | ↓ | ↓ | -- | -- | 1,12 ↑ | 1,10 ↑ | **1,20** ↑ | **1,20** ↑ | 1,16 ↑ | 1,13 ↑ |
| GLU | -- | -- | ↓ | ↓ | -- | -- | -- | ↓ | -- | -- | -- | ↓ |
| LYS | -- | ↓ | ↓ | ↓ | -- | -- | ↓ | ↓ | -- | -- | -- | ↓ |
| ASN | ↓ | -- | -- | ↓ | -- | -- | -- | -- | **1,24** ↑ | **1,25** ↑ | -- | 1,10 ↑ |
| GLN | -- | -- | -- | -- | -- | 1,12 ↑ | -- | -- | 1,18 ↑ | 1,19 ↑ | -- | -- |
| HIS | -- | -- | **1,40** ↑ | **1,65** ↑ | -- | -- | -- | -- | 1,13 ↑ | **1,21** ↑ | -- | -- |
| SER | -- | -- | -- | -- | -- | 1,13 ↑ | 1,14 ↑ | 1,14 ↑ | -- | 1,14 ↑ | -- | 1,10 ↑ |
| THR | ↓ | ↓ | -- | -- | -- | 1,11 ↑ | -- | -- | -- | -- | -- | -- |

**Tableau III-2 : Propensions des acides aminés à se trouver au voisinage des 6 résidus les plus favorisés. P1 représente la propension globale d'un résidu à se trouver près d'un des six résidus les plus favorisés alors que P2 représente cette propension dans les sites d'interaction. Les résidus favorisés sont donnés en gras avec une flèche vers le haut ; les résidus légèrement favorisés sont donnés avec une flèche vers le haut. Les résidus défavorisés sont signalés par une flèche vers le bas et les résidus indifféremment trouvés à côté des six résidus les plus favorisés dans les sites d'interaction ou ailleurs dans la protéine sont signalés par un ' -- '.**

Les résultats montrent que la leucine est fréquemment retrouvée en position +/-1 des arginines alors que le tryptophane, la méthionine et la cystéine sont trouvées préférentiellement à proximité d'une arginine en interaction (P2 de 1,16, 1,22 et 1,26, respectivement ; cf. Tableau III-2). Tout comme l'arginine, la cystéine est connue pour

interagir avec les hétéroatomes ce qui pourrait expliquer la présence de ces deux résidus côte à côte dans les sites d'interaction (cf. point III.2.2).

Les histidines forment régulièrement des doublets His-His dans les séquences de protéines (P1=1,40) et d'autant plus au niveau des sites d'interaction (P2=1,65). Les motifs His-His sont impliqués dans les activités enzymatiques,[257] dans les protéines riches en histidines (His-rich proteins - HRP)[258,259] et glycoprotéines (HRG)[260,261] et dans l'activité des peptides riches en histidines.[262,263] Comme nous l'avons signalé dans le paragraphe précédent, la glycine est largement favorisée près des résidus en interaction et, particulièrement, à côté des histidines (P2=1,26). Le tryptophane est aussi fréquemment trouvé près des histidines mais ce, indifféremment du fait qu'elles soient en interaction ou pas (P1=1,20 et P2=1,23). De plus, les quatre acides aminés chargés sont clairement défavorisés aux abords des histidines.

Comme on peut le voir dans le Tableau III-2, il n'existe pas de préférence marquée des acides aminés pour le voisinage des leucines. Seulement la sérine, la glutamine et la thréonine sont légèrement favorisées à côté des leucines en interaction (P2=1,13, 1,12, 1,11, respectivement). Les résidus aromatiques sont quant à eux presque toujours défavorisés, probablement à cause de la grande taille combinée de la leucine et de ces résidus aromatiques.

La phénylalanine est préférentiellement entourée d'une sérine ou d'un acide aspartique (P1/P2= 1,14/1,14 et 1,12/1,10, respectivement). Mais, avec une propension P2 de 1,37, la glycine est l'acide aminé le plus facilement trouvé près des phénylalanines aux interfaces protéine-protéine.

Globalement, les résidus aliphatiques sont défavorisés près du tryptophane alors que l'asparagine, l'acide aspartique, la glutamine et l'histidine y sont favorisés (P1= 1,24, 1,20, 1,18, 1,13, respectivement ; cf. Tableau III-2). La glycine est favorisée aux côtés des tryptophanes en interaction (P2=1,23). Les doublets Trp-Trp sont aussi fréquemment trouvés aux interfaces (P2=1,78) et, comme pour les doublets His-His, les doublets Trp-Trp sont impliqués dans des fonctions biologiques. Par exemple, le domaine PWWP est impliqué dans les interactions protéine-protéine.[264]

Les résidus aromatiques ont tendances à se retrouver autour de la tyrosine (P2=1,21). Les autres principaux voisins de la tyrosine étant l'acides aspartique, la glycine, la proline, l'asparagine, la sérine et, plus fréquemment, la cystéine (P2=1,22). Tous ces acides aminés sont assez petits (<92Å$^3$) et peuvent jouer le rôle de 'résidus charnières'.

### III.2.4. Distribution des types d'interaction

Le Tableau III-3 donne la distribution des types d'interactions aux interfaces protéiques. Les contacts hydrophobes correspondent à 61% des interactions dans les interfaces de type homomériques et pour 57% dans les interfaces de types hétéromériques. Les liens H sont 38% dans les interfaces de type hétéromer-permanent, 36% dans les interfaces hétéromer-transitoires et 34% dans les interfaces homomériques. Les ponts salins représentent 5,4% de toutes les interactions des interfaces homomériques, 6.7% des interfaces de type hétéromer-transitoires et presque deux fois moins (3,8%) dans les interfaces de type hétéromer-permanent. Avec moins de 1%, les ponts disulfures sont les moins représentés. Néanmoins, les interfaces de type hétéromer-permanent ont quatre fois plus de chance de posséder des liens disulfures intermoléculaires que les autres types d'interfaces (Tableau III-3).

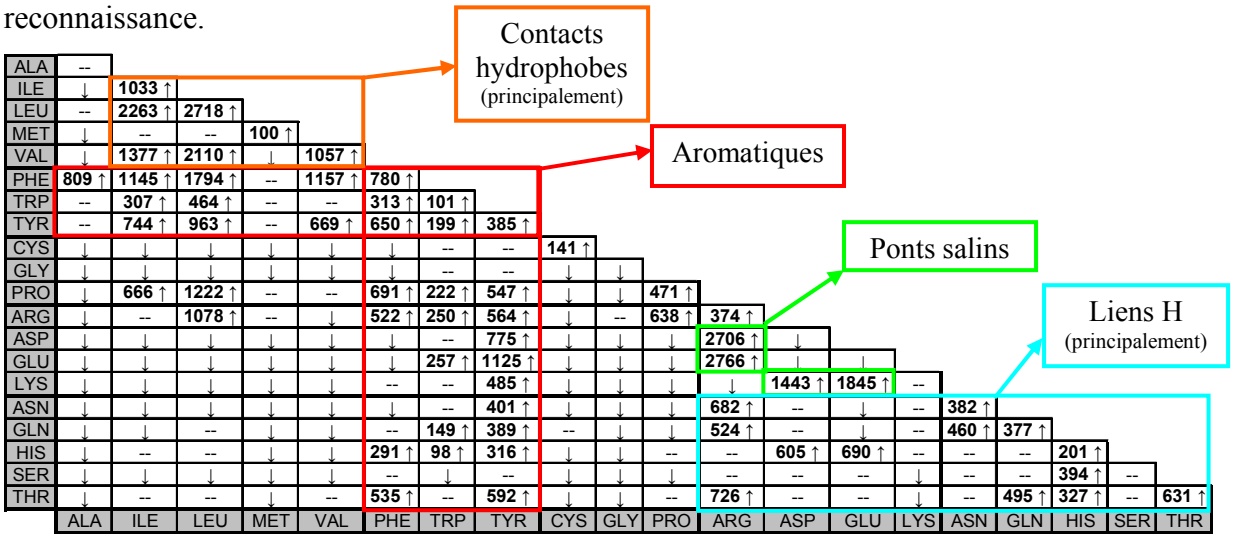
| | Hétéromer-Transitoires % (nombre) | Hétéromer-Permanents % (nombre) | Homomers % (nombre) |
|---|---|---|---|
| Liens-H | 36,0 (3549) | 38,3 (1206) | 33,6 (27726) |
| Hydrophobes | 57,2 (5638) | 57,5 (1811) | 60,9 (50266) |
| Electrostatiques | 6,7 (665) | 3,8 (119) | 5,4 (4468) |
| Disulfures | 0,1 (7) | 0,4 (12) | 0,1 (71) |

**Tableau III-3 : Répartition des différents types d'interactions dans les interfaces protéine-protéine.**

### III.2.5. Matrices d'interactions

Les matrices reprenant tous les couples résidu-résidu sont données dans le Tableau III-4. Les fréquences observées des différents couples ont été comparées aux fréquences attendues si les interactions avaient lieu par chance. Une analyse statistique nous a permis de montrer que ces couples ont une fréquence différente de celle attendue. La valeur de $\chi^2$ statistique (total) est plus élevée que le $\chi^2$ théorique [360,33 pour (20-1)*(20-1) degrés de liberté] pour les trois types d'interfaces: 5.676 pour les hétéromer-transitoires, 1.920 pour les hétéromer-permanentes et 37.848 pour les homomers ; ce qui nous permet de refuser l'hypothèse d'indépendance des acides aminés (cf. point II.3.4). Comme pour les interactions protéine-acide nucléique, ce résultat n'est pas étonnant mais cette étape était nécessaire avant de passer à une analyse plus détaillée des différents couples.

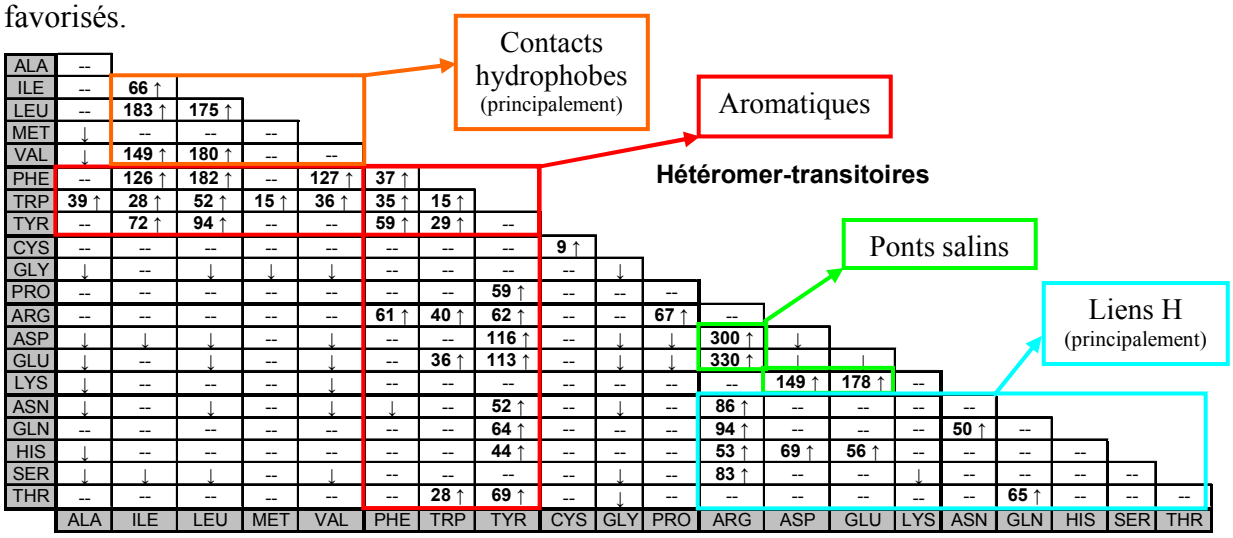
Un test $\chi^2$ de Pearson à un degré de liberté nous a dès lors permis de détecter quels couples contribuaient le plus à cette différence par rapport à une distribution aléatoire. Cette étape permet, en plus de l'observation directe c'est-à-dire de l'observation des couples les plus/moins représentés, de détecter lesquels sont importants pour la spécificité de la reconnaissance.




| | ALA | ILE | LEU | MET | VAL | PHE | TRP | TYR | CYS | GLY | PRO | ARG | ASP | GLU | LYS | ASN | GLN | HIS | SER | THR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALA | -- | | | | | | | | | | | | | | | | | | | |
| ILE | ↓ | 1033 ↑ | | | | | | | | | | | | | | | | | | |
| LEU | -- | 2263 ↑ | 2718 ↑ | | | | | | | | | | | | | | | | | |
| MET | ↓ | -- | -- | 100 ↑ | | | | | | | | | | | | | | | | |
| VAL | ↓ | 1377 ↑ | 2110 ↑ | ↓ | 1057 ↑ | | | | | | | | | | | | | | | |
| PHE | 809 ↑ | 1145 ↑ | 1794 ↑ | -- | 1157 ↑ | 780 ↑ | | | | | | | | | | | | | | |
| TRP | -- | 307 ↑ | 464 ↑ | -- | -- | 313 ↑ | 101 ↑ | | | | | | | | | | | | | |
| TYR | -- | 744 ↑ | 963 ↑ | -- | 669 ↑ | 650 ↑ | 199 ↑ | 385 ↑ | | | | | | | | | | | | |
| CYS | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | -- | -- | 141 ↑ | | | | | | | | | | | |
| GLY | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | -- | -- | ↓ | ↓ | | | | | | | | | | |
| PRO | ↓ | 666 ↑ | 1222 ↑ | -- | -- | 691 ↑ | 222 ↑ | 547 ↑ | ↓ | ↓ | 471 ↑ | | | | | | | | | |
| ARG | ↓ | -- | 1078 ↑ | -- | ↓ | 522 ↑ | 250 ↑ | 564 ↑ | ↓ | -- | 638 ↑ | 374 ↑ | | | | | | | | |
| ASP | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | -- | 775 ↑ | ↓ | ↓ | ↓ | 2706 ↑ | ↓ | | | | | | | |
| GLU | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | 257 ↑ | 1125 ↑ | ↓ | ↓ | ↓ | 2766 ↑ | ↓ | ↓ | | | | | | |
| LYS | ↓ | ↓ | ↓ | ↓ | ↓ | -- | -- | 485 ↑ | ↓ | ↓ | ↓ | ↓ | 1443 ↑ | 1845 ↑ | -- | | | | | |
| ASN | ↓ | ↓ | ↓ | ↓ | ↓ | ↓ | -- | 401 ↑ | ↓ | ↓ | ↓ | 682 ↑ | -- | ↓ | -- | 382 ↑ | | | | |
| GLN | ↓ | ↓ | -- | ↓ | ↓ | -- | 149 ↑ | 389 ↑ | -- | ↓ | ↓ | 524 ↑ | -- | ↓ | -- | 460 ↑ | 377 ↑ | | | |
| HIS | ↓ | -- | -- | ↓ | ↓ | 291 ↑ | 98 ↑ | 316 ↑ | ↓ | ↓ | -- | -- | 605 ↑ | 690 ↑ | -- | -- | -- | 201 ↑ | | |
| SER | ↓ | ↓ | ↓ | ↓ | ↓ | -- | ↓ | -- | ↓ | ↓ | ↓ | -- | -- | -- | ↓ | -- | -- | 394 ↑ | -- | |
| THR | ↓ | -- | -- | ↓ | -- | 535 ↑ | -- | 592 ↑ | ↓ | ↓ | -- | 726 ↑ | -- | -- | ↓ | -- | 495 ↑ | 327 ↑ | -- | 631 ↑ |

**Tableau III-4 : Matrices d'interactions pour l'ensemble des complexes protéine-protéine. Les couples favorisés dans les sites d'interaction sont donnés en gras avec une flèche vers le haut. Les couples défavorisés sont signalés par une flèche vers le bas et les couples indifféremment trouvés dans les sites d'interaction ou ailleurs dans la protéine sont signalés par un ' -- '.**

Globalement (Tableau III-4), les résidus aromatiques sont favorisés en tant que partenaires de la plupart des acides aminés. Les exceptions sont la cystéine, la sérine et la glycine. La tyrosine est le résidu le plus favorisé, devant l'histidine, la phénylalanine et le tryptophane. Les couples entre les résidus de charges opposées (ponts salins) et les couples de résidus hydrophobes (particulièrement l'isoleucine et la leucine) sont aussi particulièrement favorisés.




| | ALA | ILE | LEU | MET | VAL | PHE | TRP | TYR | CYS | GLY | PRO | ARG | ASP | GLU | LYS | ASN | GLN | HIS | SER | THR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALA | -- | | | | | | | | | | | | | | | | | | | |
| ILE | -- | 66 ↑ | | | | | | | | | | | | | | | | | | |
| LEU | -- | 183 ↑ | 175 ↑ | | | | | | | | | | | | | | | | | |
| MET | ↓ | -- | -- | -- | | | | | | | | | | | | | | | | |
| VAL | ↓ | 149 ↑ | 180 ↑ | -- | -- | | | | | | | | | | | | | | | |
| PHE | -- | 126 ↑ | 182 ↑ | -- | 127 ↑ | 37 ↑ | | | | | | | | | | | | | | |
| TRP | 39 ↑ | 28 ↑ | 52 ↑ | 15 ↑ | 36 ↑ | 35 ↑ | 15 ↑ | | | | | | | | | | | | | |
| TYR | -- | 72 ↑ | 94 ↑ | -- | -- | 59 ↑ | 29 ↑ | -- | | | | | | | | | | | | |
| CYS | -- | -- | -- | -- | -- | -- | -- | -- | 9 ↑ | | | | | | | | | | | |
| GLY | ↓ | -- | ↓ | ↓ | ↓ | -- | -- | -- | -- | ↓ | | | | | | | | | | |
| PRO | -- | -- | -- | -- | -- | -- | -- | 59 ↑ | -- | -- | -- | | | | | | | | | |
| ARG | -- | -- | -- | -- | -- | 61 ↑ | 40 ↑ | 62 ↑ | -- | -- | 67 ↑ | -- | | | | | | | | |
| ASP | ↓ | ↓ | ↓ | -- | ↓ | -- | -- | 116 ↑ | -- | ↓ | ↓ | 300 ↑ | ↓ | | | | | | | |
| GLU | ↓ | -- | ↓ | -- | ↓ | -- | 36 ↑ | 113 ↑ | -- | ↓ | ↓ | 330 ↑ | ↓ | ↓ | | | | | | |
| LYS | ↓ | -- | -- | -- | ↓ | -- | -- | -- | -- | -- | -- | -- | 149 ↑ | 178 ↑ | -- | | | | | |
| ASN | ↓ | -- | ↓ | -- | ↓ | ↓ | -- | 52 ↑ | -- | ↓ | -- | 86 ↑ | -- | -- | -- | -- | | | | |
| GLN | -- | -- | -- | -- | -- | -- | -- | 64 ↑ | -- | -- | -- | 94 ↑ | -- | -- | -- | 50 ↑ | -- | | | |
| HIS | ↓ | -- | -- | -- | -- | -- | -- | 44 ↑ | -- | -- | -- | 53 ↑ | 69 ↑ | 56 ↑ | -- | -- | -- | -- | | |
| SER | ↓ | ↓ | ↓ | -- | ↓ | -- | -- | -- | -- | ↓ | -- | 83 ↑ | -- | -- | ↓ | -- | -- | -- | -- | |
| THR | -- | -- | -- | -- | -- | -- | 28 ↑ | 69 ↑ | -- | ↓ | -- | -- | -- | -- | -- | -- | 65 ↑ | -- | -- | -- |

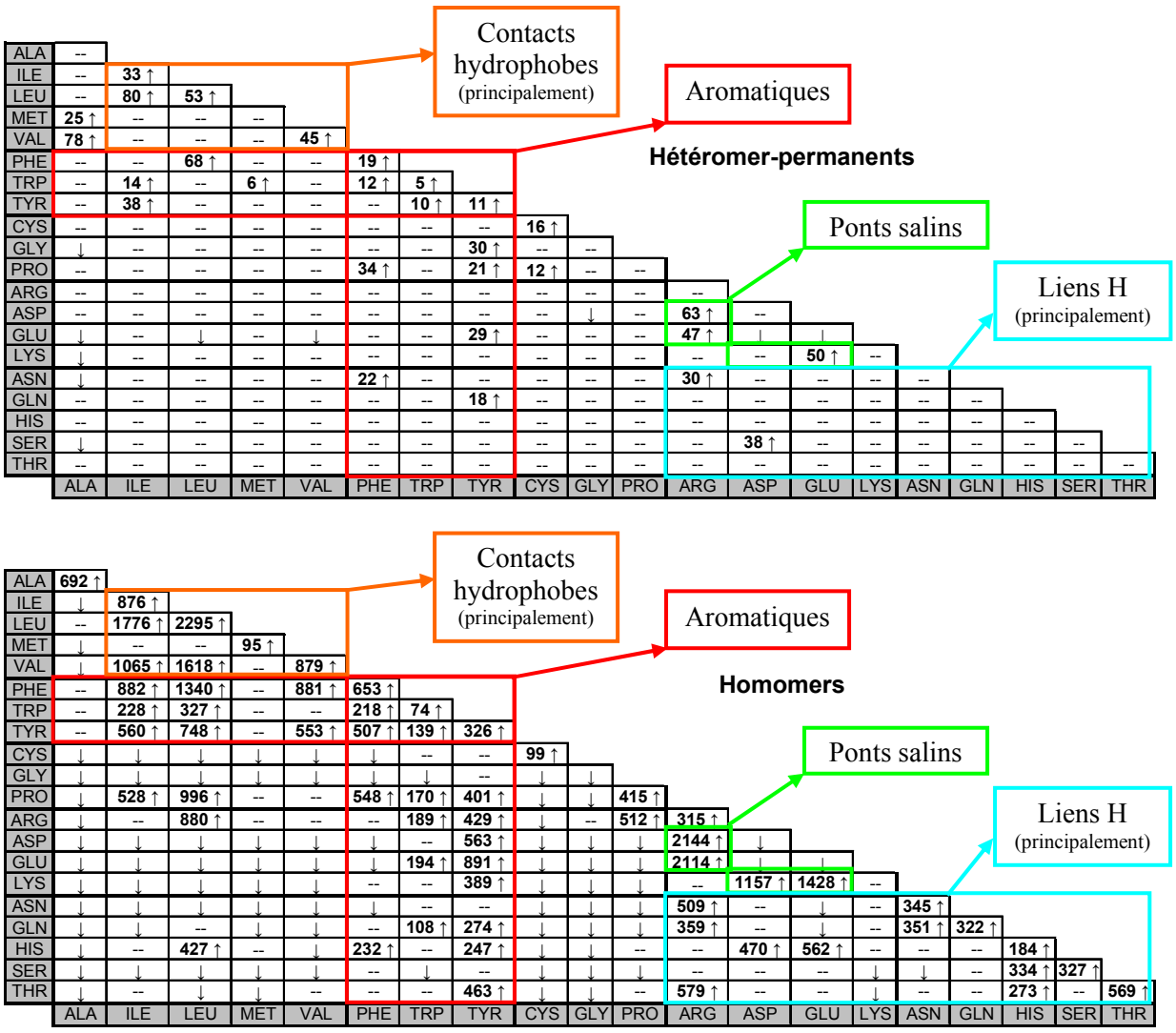



**Tableau III-5 : Matrices d'interactions pour les différents types de complexes protéine-protéine. Les couples favorisés dans les sites d'interaction sont donnés en gras avec une flèche vers le haut. Les couples défavorisés sont signalés par une flèche vers le bas et les couples indifféremment trouvés dans les sites d'interaction ou ailleurs dans la protéine sont signalés par un ' -- '.**

Si l'on observe les différences obtenues en fonction du type d'interface impliqué (Tableau III-5), on voit que les ponts salins et les couples entre et avec les leucines sont largement favorisés dans les interfaces de type hétéromer-transitoires. Ensuite, les couples entre et avec les acides aminés aromatiques et les couples Asp-Tyr sont aussi favorisés. Dans les interfaces de type hétéromer-permanentes, les couples les plus favorisés sont les ponts salins et les ponts disulfures. Finalement, on peut voir que la plupart des couples des interfaces homomériques sont des ponts salins et des contacts hydrophobes entre la leucine, la phénylalanine et l'isoleucine. Le couple Glu-Tyr est aussi favorisé alors que l'acide glutamique est défavorisé dans 14 des 20 couples possibles.

### III.2.6. Couples significativement favorisés

Les couples significativement favorisés ont été classés selon leurs valeurs de $\chi^2$ individuelles et les couples les plus significatifs ont été retenus. Ensuite, nous avons construit des matrices d'interaction au niveau atomique (cf. point II.3.4) pour les résidus les plus favorisés.

**Couples de résidus de charges opposées (ponts salins)**

Les interactions entre des acides aminés de charges opposées sont grandement favorisées dans tous les types de sites d'interaction protéine-protéine. Notre analyse statistique nous a permis de montrer que le couple Arg-Asp est le plus favorisé et est suivi des couples Arg-Glu, Lys-Glu et Lys-Asp.

Les couples avec l'arginine sont les plus nombreux et sont composés pour plus de 60% de ponts salins et pour plus de 30% de liens H. La moitié des liens H détectés impliquent l'hydrogène du troisième atome d'azote de l'arginine (NE) et un des atomes d'oxygène carboxylique de l'acide aspartique ou glutamique. Comme cet atome d'hydrogène peut participer à la délocalisation de la charge positive de l'arginine, les interactions détectées comme liens H peuvent être assimilées à des ponts salins.

Les 1845 couples Lys-Glu contiennent 69% de ponts salins et 15% de liens H. des résultats similaires sont obtenus pour les 1443 couples Lys-Asp avec 73% et 16%, respectivement. Il est intéressant de noter que des contacts hydrophobes impliquant la chaîne latérale de la lysine sont fréquemment trouvés (particulièrement pour les couples Lys-Glu avec 12%) et que ceux-ci y sont deux fois plus nombreux que pour les couples avec l'arginine.

**Couples de résidus aliphatiques (contacts hydrophobes)**

Bien que favorisés dans tous les types d'interfaces, les contacts hydrophobes sont plus fréquemment retrouvés dans les interfaces de type homomérique. Les résultats présentés ci-dessous proviennent de l'analyse des interfaces homomériques et sont fonction des chaînes latérales uniquement. Les atomes du squelette peptidique semblent défavorisés pour les contacts hydrophobes : ils représentent de 5,1% à 8,6% des interactions alors qu'ils correspondent à 15,6% - 18,1% de l'accessibilité des résidus.

Les couples les plus favorisés sont : Leu-Leu > Phe-Phe > Ile-Ile > Phe-Leu > Leu-Ile > Phe-Ile. Ils contiennent plus de 90% de contacts hydrophobes. Les autres types d'interactions retrouvés sont des liens H avec les atomes du squelette peptidique (de 2,5% à 7,6% au lieu de 7,6% en moyenne dans la banque de complexes protéine-protéine complète). La Figure III-36 représente la distribution des contacts hydrophobes des atomes des chaînes latérales des couples les plus favorisés. Cette figure est commentée ci-dessous.

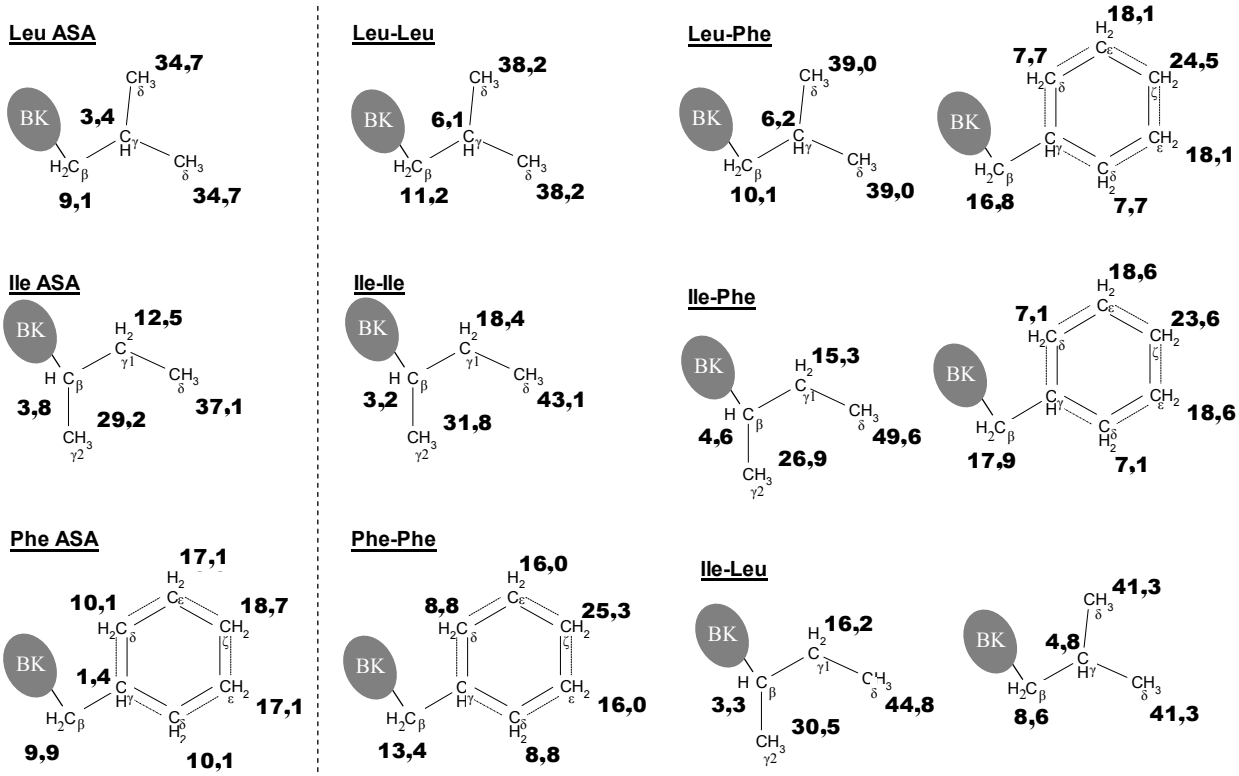



**Figure III-36 : Distribution des interactions atome-atome au sein des contacts hydrophobes entre chaînes latérales des couples de résidus aliphatiques (le 'backbone' est schématisé par une forme grise - BK). Les nombres donnés sont les pourcentages des interactions trouvées pour chaque couple. A gauche, les surfaces accessibles atomiques (ASA) de la leucine, l'isoleucine et la phénylalanine sont données comme base pour comparaison.**

76% des contacts hydrophobes des couples Leu-Leu impliquent les deux groupements méthyles de la leucine et, seulement 11% impliquent les atomes d'hydrogènes en position C-beta. Les atomes d'hydrogènes en C-beta de la phénylalanine représentent quand à eux 13% des contacts hydrophobes des couples Phe-Phe. Les interactions des hydrogènes de la fonction aromatique sont répartis de la manière suivante : 9% pour la position C-beta, 16% pour chacune des positions C-epsilon et 25% pour l'extrémité de la chaîne latérale (C-zeta). Pour

les contacts hydrophobes des couples Ile-Ile, le site le plus favorisé est le groupement méthyle en position delta (43%) suivi des groupements méthyles en position gamma (32%).

Les résultats obtenus pour les couples Leu-Phe confirment les résultats précédents : les sites les plus souvent en interaction de la phénylalanine et de la leucine se trouvent à l'extrémité de la chaîne latérale (position zeta de la phénylalanine et groupement méthyle en delta de la leucine). De plus, les atomes d'hydrogènes en position C-beta de la phénylalanine interagissent principalement (89%) avec les groupements méthyles de la leucine. Inversement, les atomes d'hydrogènes en position beta de la phénylalanine ne montrent pas de préférence claire pour un des atomes de l'isoleucine dans les couples Ile-Phe. Les deux groupements méthyles de la leucine représentent 83% des contacts hydrophobes dans les couples Ile-Leu et l'isoleucine participe fréquemment aux interactions par l'intermédiaire de ses groupements méthyles (45% + 31%) comme elle le fait dans les couples Ile-Ile.

Dans la Figure III-36, les surfaces accessibles (ASA) de la leucine, de l'isoleucine et de la phénylalanine sont données. Ces données proviennent d'une banque de données de 2277 domaines protéiques étudiés par Singh *et al.*[265] La comparaison des ASA atomiques de ce travail et des pourcentages d'interaction peut être utilisée pour détecter des atomes plus souvent impliqués dans les interactions que leur accessibilité ne le laisserait entendre. Pour le cas de la leucine, les groupements méthyles à l'extrémité de la chaîne latérale représentent 39,5% des interactions chacun alors qu'ils correspondent à 34,7% de la totalité de la surface accessible de la leucine. Le même type de résultat est obtenu pour l'isoleucine où le groupement méthyle en delta correspond à 45,8% des interactions pour 37,1% de la surface accessible et pour la phénylalanine pour laquelle le CH en position zeta représente un quart des interactions (24,5%) et 18,7% de l'accessibilité totale. Il est intéressant de noter que le groupement $CH_2$ en beta de la phénylalanine est aussi plus impliqué dans les interactions (16,0%) qu'attendu (9,9% de l'ASA). Inversement, la position delta semble être défavorisée (7,9% au lieu de 10,1%). Ces résultats indiquent que l'accessibilité des atomes de ces acides aminés n'explique pas à elle seule la propension à être en interaction des différents atomes. Il est donc clair que des interactions préférentielles ont lieu au sein des contacts hydrophobes.

**Couples les plus favorisés dans les interfaces de type hétéromer-transitoire**

Après les ponts salins et les contacts hydrophobes, les couples Asp-Tyr et Glu-Tyr sont favorisés dans les interfaces de type hétéromer-transitoire. Les atomes d'oxygènes des fonctions carboxyliques interagissent préférentiellement avec le groupement hydroxyle de la tyrosine (76% des couples Asp-Tyr et 58% des couples Glu-Tyr). Ces interactions ont une

distance optimale d'interaction située à 1.7Å ce qui est légèrement inférieur aux distances optimales des liens H (1,9Å) et des ponts salins (1,85Å) ce qui suggère une plus grande stabilité de ce type de contacts. De plus, le groupement hydroxyle de la tyrosine conserve cet optimum à 1,7Å dans l'ensemble de la banque de complexes de type hétéromer-transitoires (Figure III-37). De la même manière, la distance optimale d'interaction du groupement hydroxyle de la sérine se situe à 1,65Å.

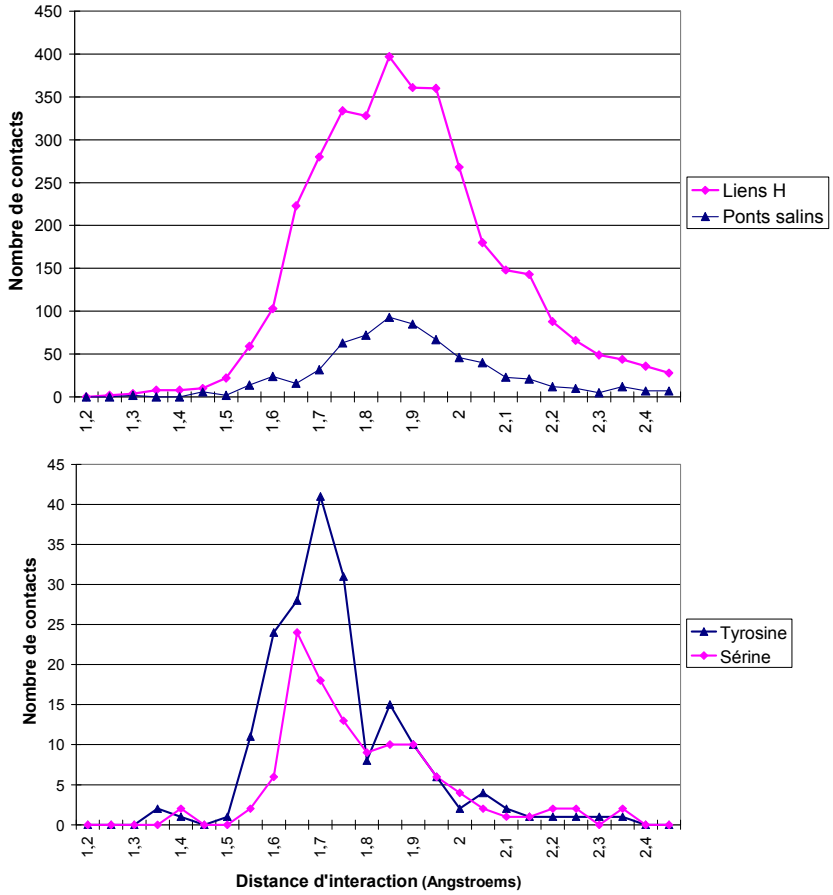



**Figure III-37 : Distribution des interactions en fonction de la distance. Au-dessus, distribution des liens H et des ponts salins dans la banque totale de complexes protéine-protéine. En dessous, distribution des interactions avec le groupement hydroxyle de la tyrosine et de la sérine dans les complexes de type hétéromer-transitoires.**

### Couples les plus favorisés dans les interfaces de type hétéromer-permanent

Les couples les plus favorisés dans les interfaces de type hétéromer-permanent sont les couples entre deux cystéines. 75% de ces couples sont des ponts disulfures. Ce résultat met en évidence le caractère hautement permanent de ce type de complexe. Les ponts salins (Arg-Lys > Lys-Glu) et les contacts hydrophobes (Phe-Leu > Leu-Ile > Ile-Ile) sont aussi favorisés mais dans une moindre mesure.

### III.2.7. Résumé des résultats obtenus

Lors des précédents paragraphes, nous avons étudié en détail les propriétés des interfaces entre protéines. Comme dans le cas des interactions protéine-acide nucléique, il est utile de résumer les informations obtenues sur cette partie de la thèse :

- Une des caractéristiques les plus marquantes des interfaces protéine-protéine est l'enrichissement en acides aminés aromatiques. La tyrosine, la phénylalanine et le tryptophane mais aussi l'arginine sont les résidus les plus favorisés aux interfaces.
- Les interactions électrostatiques et les contacts hydrophobes sont les deux types d'interaction les plus représentés aux interfaces. Les particularités des couples d'acides aminés les plus favorisés aux interfaces ont ensuite été détaillées.
- Une analyse innovante des acides aminés voisins des résidus en interaction a permis de mettre en évidence une corrélation entre la taille des résidus et leurs propensions à se trouver aux abords de l'interface. Plus un acide aminé est petit, plus il aura tendance à se trouver à côté d'un résidu en interaction.
- Notre classification des interfaces en trois types distincts à permis de montrer que les interfaces hétéromer-permanentes sont particulièrement enrichies en résidus hydrophobes et que les interfaces hétéromer-permanentes présentent un taux très élevé de ponts disulfures.

## *III.3. Prédiction des Sites d'Interaction*

### III.3.1. Variables utilisées

Dans le but de développer une méthode de prédiction efficace, nous nous sommes basés sur les informations extraites de l'étude des interfaces réalisées ci-dessus (points III.1 et III.2). Nous avons principalement utilisé les propensions à être en interaction dans les différents types de complexes ainsi que des données dérivées des analyses précédentes comme les doublets de résidus en interaction et les propensions à être en surface. De plus, un maximum de variables directement calculables à partir de la séquence ont été étudiés (prédiction de structure secondaire, 'Receptor Binding Domain'…).

Ci-dessous, les différentes variables utilisées sont décrites en séparant les variables qualitatives des variables quantitatives.

**Variables qualitatives**

- Nom de l'acide aminé.
- Caractère hydrophobe ou hydrophile de l'acide aminé ainsi que sa famille selon la classification présentée au point II.3.1.
- Prédictions des Receptor Binding Domains (RBD - cf. point II.5).

  Les fenêtres de calcul utilisées pour le calcul des RDB sont de 5, 7 et 9 acides aminés. Chaque résidu a donc été défini comme RDB 5, RBD 7 ou RBD 9. De plus, nous avons défini des zones RDB. Dans ces zones, on retrouve les acides aminés prédits comme RBD ainsi que les acides aminés ayant servi au calcul de la valeur (résidu +/- 2, +/-3 ou +/- 4 pour des fenêtres de 5, 7 ou 9 - voir schéma) :

  RBD 5 = _ _ _ _ _ R _ _ _ _ _ → zone RBD 5 = _ _ _ Z Z R Z Z _ _ _
  RBD 7 = _ _ _ _ _ R _ _ _ _ _ → zone RBD 7 = _ _ Z Z Z R Z Z Z _ _
  RBD 9 = _ _ _ _ _ R _ _ _ _ _ → zone RBD 9 = _ Z Z Z Z R Z Z Z Z _

  Chaque tiret représente un acide aminé, les acides aminés détectés comme RBD sont notés par un R et les acides aminés de la zone ayant servi à détecter ces RBD sont notés par un Z. Dans le fichier de données, l'acide aminé est considéré comme RBD ('R' ou 'ZR') ou pas ('_').

- Motifs/domaines protéiques.

  Treize méthodes décrites au point II.6.3 (InterProScan) ont été utilisées et permettent de déterminer si un domaine fonctionnel est connu pour le fragment de séquence considéré. L'acide aminé est décrit comme appartenant à un domaine ('D') ou pas ('_'), le nom du domaine n'est pas utilisé.

- Prédiction d'accessibilité.

  Tous les acides aminés sont définis comme potentiellement exposés en surface de la protéine (e pour 'exposed') ou caché au cœur de celle-ci (b pour 'burried' ou h pour 'hidden') par les deux méthodes présentées au point II.6.2. En plus de cette distinction e/b faite par la méthode NetASA,[243] la seconde méthode (PredAcc)[241] distingue les résidus les plus/moins probablement exposés (E/e) ou cachés (H/h).

  Ces deux méthodes permettent de choisir le seuil d'ASA relative qui permet de désigner un acide aminé comme exposé ou caché. Sous ce seuil d'ASA relative, l'acide aminé sera considéré comme non accessible et au-dessus, comme accessible.

| ASA relative | AA hors interactions | | Acides aminés en interaction | | |
|---|---|---|---|---|---|
| | Nombre | % | Nombre | % | % en interaction |
| 00-05% | **8936** | **29.9** | 200 | 3.9 | 2.2 |
| 05-10% | **2248** | **7.5** | 241 | 4.7 | 10.7 |
| 10-15% | 1716 | 5.7 | **271** | **5.3** | **15.8** |
| 15-20% | 1584 | 5.3 | **314** | **6.1** | **19.8** |
| 20-25% | 1472 | 4.9 | **297** | **5.8** | **20.2** |
| 25-30% | 1397 | 4.7 | **289** | **5.6** | **20.7** |
| 30-35% | 1373 | 4.6 | **336** | **6.5** | **24.5** |
| 35-40% | 1368 | 4.6 | **360** | **7.0** | **26.3** |
| 40-100% | 9794 | 32.8 | **2843** | **55.2** | **29.0** |

**Tableau III-6 : Variation du nombre de résidus en interaction en fonction de l'ASA relative.**

Afin de déterminer quel est le seuil optimal, nous avons analysé notre banque de données. Dans le Tableau III-6, le nombre d'acides aminés en interaction (ou pas) en fonction de différentes valeurs d'ASA relative est donné. Près d'un tiers des acides aminés non impliqués dans les interactions (29,9%), ont une ASA relative inférieure à 5%. Cela correspond aux résidus se trouvant au cœur de la protéine. Le pourcentage d'acides aminés hors interactions par tranche d'ASA relative se stabilise ensuite vers les 5%. Pour les résidus en interactions, le pourcentage augmente en parallèle avec le pourcentage d'ASA relative. Une valeur d'ASA relative de **10%** semble idéale comme seuil car elle permet de conserver 87,1% des acides aminés en interaction tout en éliminant 37,4% des acides aminés qui ne le sont pas (Tableau III-6 - chiffre en gras).

- Prédiction de structure secondaire.

  Les deux méthodes présentées au point II.6.1 ont été utilisées (NPSA et PsiPred). Les résultats obtenus sont la prédiction en structure de type hélice (h), beta (b) et random coil (c) pour chaque résidu.

- Prédiction du désordre des protéines.

  Les résultats de la méthode de prédiction du désordre GloPlot[245] ont été ajoutés aux données. Dans le fichier de données, l'acide aminé est considéré comme appartenant à une zone désordonnée ('D') ou pas ('s').

| Fragment number | AA1 | AA3 | aa_pho-phi | aa_classe | RBD-5 | RBD-7 | RBD-9 | zoneRBD-5 | zoneRBD-7 | zoneRBD-9 | FPrintScan-Domaine | NetASA-Surface-10% | PredAcc-Surface-10% | SS-NPSA | SS-PsiPred | GlobPlot_Disorder_Window |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 18 | V | VAL | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | D | e | e | h | e | s |
| 19 | D | ASP | Hydrophilic | NegativelyCharged | _ | _ | _ | _ | _ | _ | D | e | E | h | e | s |
| 20 | L | LEU | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | D | b | h | h | c | s |
| 21 | F | PHE | Hydrophobic | Aromatic | _ | _ | _ | _ | _ | _ | D | b | h | h | c | s |
| 22 | A | ALA | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | D | b | e | h | c | s |
| 23 | Q | GLN | Hydrophilic | Polar | _ | _ | _ | _ | _ | _ | D | e | E | h | c | s |
| 24 | L | LEU | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | D | e | e | h | c | s |
| 25 | D | ASP | Hydrophilic | NegativelyCharged | _ | _ | _ | _ | _ | _ | D | e | E | h | c | s |
| 26 | L | LEU | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | D | b | e | h | h | s |
| 27 | E | GLU | Hydrophilic | NegativelyCharged | _ | _ | _ | _ | _ | _ | _ | e | E | h | h | s |
| 28 | K | LYS | Hydrophilic | PositivelyCharged | _ | _ | _ | _ | _ | _ | _ | e | E | h | h | s |
| 29 | V | VAL | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | _ | b | h | h | h | s |
| 30 | L | LEU | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | _ | b | e | h | h | s |
| 31 | D | ASP | Hydrophilic | NegativelyCharged | _ | _ | _ | _ | _ | _ | _ | e | E | h | h | s |
| 32 | L | LEU | Hydrophobic | Aliphatic | _ | _ | _ | _ | _ | _ | _ | e | e | c | h | D |
| 33 | C | CYS | Hydrophobic | Sulphurated | _ | _ | _ | _ | _ | _ | _ | b | e | c | c | D |
| 34 | P | PRO | Hydrophobic | Conformational | _ | _ | _ | _ | _ | _ | _ | e | e | c | c | D |
| 35 | N | ASN | Hydrophilic | Polar | _ | _ | _ | _ | _ | _ | D | e | E | c | c | D |
| 36 | S | SER | Hydrophilic | Polar | _ | _ | _ | _ | _ | _ | D | e | e | c | c | D |
| 37 | K | LYS | Hydrophilic | PositivelyCharged | _ | _ | _ | ZR5 | _ | _ | D | e | E | c | c | D |
| 38 | Y | TYR | Hydrophilic | Aromatic | _ | _ | _ | ZR5 | _ | _ | D | e | e | c | c | D |
| 39 | N | ASN | Hydrophilic | Polar | R5 | _ | _ | ZR5 | _ | _ | D | e | E | c | c | D |
| 40 | P | PRO | Hydrophobic | Conformational | _ | _ | _ | ZR5 | _ | _ | D | e | E | c | h | D |
| 41 | E | GLU | Hydrophilic | NegativelyCharged | _ | _ | _ | ZR5 | _ | _ | D | e | E | c | h | D |
| 42 | E | GLU | Hydrophilic | NegativelyCharged | _ | _ | _ | _ | _ | _ | D | e | E | c | h | D |

**Tableau III-7 : Extrait d'un tableau reprenant des variables qualitatives étudiés pour la mise au point d'un modèle de régression logistique.**

**Variables quantitatives**

- Position dans la séquence de l'acide aminé.
- Motifs/domaines protéiques des treize serveurs décrits au point II.6.3 (InterProScan). La valeur attribuée à l'acide aminé est l'indice de confiance de la méthode de prédiction.
- Propensions à être en interaction.

  C'est-à-dire, le rapport entre la fréquence d'un acide aminé dans une sous-banque d'interfaces divisée par la fréquence de ce même acide aminé dans la banque totale (cf. point II.3.2). Les sous-banques sont : résidus en interaction, résidus dans des zones en interaction, résidus en interaction avec des acides aminés, résidus en interaction avec des acides aminés d'interfaces de type hétéromer-transitoire, hétéromer-permanent et homomérique, résidus en interaction avec des nucléotides (ADN et/ou ARN), résidus en interaction avec des hétéroatomes.

  Lors de la mise au point de la méthode de prédiction, ces propensions ont été utilisées telles quelles ou calculées sur différentes fenêtres (3, 5, 7 et 9 résidus) (cf. Tableau III-8).
- Propensions à être en surface.

  Nous allons utiliser la définition des résidus de surface (point II.3.3) pour calculer la propension des acides aminés à se trouver en surface (rapport entre la fréquence en surface et la fréquence dans la banque de données totale).

  Lors de la mise au point de la méthode de prédiction, ces propensions ont été utilisées telles quelles ou calculées sur différentes fenêtres (3, 5, 7 et 9 résidus) (cf. Tableau III-8).
- Propensions à être au cœur de la protéine.

  Ces propensions ont été calculées de la même manière que les propensions à être en surface en gardant cette fois les acides aminés ayant une surface accessible relative inférieure à 10% (cf. Tableau III-8).
- Autres échelles.

  Afin d'augmenter nos chances de trouver une variable significativement corrélée à la présence de résidus en interaction, nous avons collecté un maximum d'échelles attribuant une valeur à chacun des 20 résidus. Ces échelles sont liées à des critères biophysiques (poids moléculaire, volume), à l'hydrophobicité/hydrophilicité, à la polarité, au désordre, à l'encombrement stérique, à la flexibilité… des résidus.

  Comme précédemment, nous avons utilisé ces échelles telles quelles ou calculées sur plusieurs fenêtres de calcul (3, 5, 7 et 9 résidus) (cf. Tableau III-8).

| Acide aminé | Propensions à être en interaction | | | | | | | | | Accessibilité | | Autres propensions | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Toutes interactions | Int. protéine-protéine | Int. avec l'ADN | Int. avec l'ARN | Int. avec hétéroatomes | Int. dans hétéromers | Int. dans hétéromer-transitoires | Int. dans hétéromer-permanents | Int. dans homomomers | Surface | Intérieur | Poids moléculaire | Volume | Polarité (Grantham) | Hydrophobicité (Eisenberg) | Hydrophobicité (Kyte & Doolittle) | Flexibilité (Bhaskaran) | Encombrement (Rose) | Désordre (Deleage-Roux) |
| ALA | 0.78 | 0.82 | 0.71 | 0.59 | 0.66 | 0.74 | 0.76 | 0.95 | 0.84 | 0.79 | 1.40 | 89.00 | 67.00 | 8.10 | 0.62 | 1.80 | 0.36 | 86.60 | -0.28 |
| ARG | 1.34 | 1.28 | 1.77 | 1.72 | 1.55 | 1.33 | 1.55 | 1.07 | 1.27 | 1.36 | 0.33 | 174.00 | 148.00 | 10.50 | -2.50 | -4.50 | 0.53 | 162.20 | -0.18 |
| ASN | 0.93 | 0.87 | 1.40 | 1.46 | 1.18 | 0.90 | 0.82 | 0.76 | 0.86 | 1.23 | 0.57 | 132.00 | 91.00 | 11.60 | -0.78 | -3.50 | 0.46 | 103.30 | 0.48 |
| ASP | 0.94 | 0.91 | 0.61 | 1.33 | 1.07 | 0.93 | 0.98 | 0.94 | 0.91 | 1.30 | 0.45 | 133.00 | 91.00 | 13.00 | -0.90 | -3.50 | 0.51 | 97.80 | 0.46 |
| CYS | 0.93 | 0.58 | 0.62 | 0.66 | 2.28 | 0.76 | 0.67 | 1.17 | 0.54 | 0.50 | 1.93 | 121.00 | 86.00 | 5.50 | 0.29 | 2.50 | 0.35 | 132.30 | -0.13 |
| GLN | 0.93 | 0.97 | 0.91 | 1.23 | 0.77 | 1.05 | 1.12 | 0.80 | 0.95 | 1.30 | 0.45 | 146.00 | 109.00 | 10.50 | -0.85 | -3.50 | 0.49 | 119.20 | -0.06 |
| GLU | 0.88 | 0.91 | 0.49 | 0.75 | 0.75 | 0.90 | 1.00 | 0.71 | 0.92 | 1.39 | 0.29 | 147.00 | 109.00 | 12.30 | -0.74 | -3.50 | 0.50 | 113.90 | -0.27 |
| GLY | 0.80 | 0.69 | 1.13 | 0.90 | 1.23 | 0.69 | 0.69 | 0.80 | 0.69 | 0.97 | 1.05 | 75.00 | 48.00 | 9.00 | 0.48 | -0.40 | 0.54 | 62.90 | 0.67 |
| HIS | 1.48 | 1.18 | 1.31 | 1.49 | 2.65 | 1.13 | 1.18 | 1.02 | 1.19 | 1.14 | 0.74 | 155.00 | 117.00 | 10.40 | -0.40 | -3.20 | 0.32 | 155.80 | 0.14 |
| ILE | 1.05 | 1.13 | 0.66 | 0.72 | 0.70 | 1.16 | 1.12 | 1.34 | 1.13 | 0.61 | 1.72 | 131.00 | 124.00 | 5.20 | 1.40 | 4.50 | 0.46 | 158.00 | -0.52 |
| LEU | 1.11 | 1.22 | 0.42 | 0.70 | 0.68 | 1.17 | 1.11 | 1.12 | 1.23 | 0.68 | 1.58 | 131.00 | 124.00 | 4.90 | 1.10 | 3.80 | 0.37 | 164.10 | -0.44 |
| LYS | 0.88 | 0.87 | 1.68 | 1.60 | 0.91 | 0.82 | 0.81 | 0.67 | 0.89 | 1.46 | 0.15 | 146.00 | 135.00 | 11.30 | -1.50 | -3.90 | 0.47 | 115.50 | -0.05 |
| MET | 0.76 | 0.69 | 1.04 | 0.85 | 1.03 | 0.70 | 0.71 | 0.87 | 0.68 | 0.77 | 1.42 | 149.00 | 124.00 | 5.70 | 0.64 | 1.90 | 0.30 | 172.90 | -0.48 |
| PHE | 1.31 | 1.37 | 0.92 | 0.87 | 1.08 | 1.43 | 1.31 | 1.38 | 1.35 | 0.70 | 1.56 | 165.00 | 135.00 | 5.20 | 1.20 | 2.80 | 0.31 | 194.10 | -0.50 |
| PRO | 0.99 | 1.10 | 0.77 | 0.91 | 0.58 | 1.04 | 1.02 | 1.23 | 1.11 | 1.19 | 0.65 | 115.00 | 90.00 | 8.00 | 0.12 | -1.60 | 0.51 | 92.90 | 1.12 |
| SER | 0.92 | 0.87 | 1.26 | 1.04 | 1.11 | 0.88 | 0.82 | 0.86 | 0.87 | 1.07 | 0.88 | 105.00 | 73.00 | 9.20 | -0.18 | -0.80 | 0.51 | 85.60 | 0.30 |
| THR | 1.02 | 1.00 | 1.27 | 1.05 | 1.10 | 0.98 | 0.96 | 1.00 | 1.01 | 1.06 | 0.90 | 119.00 | 93.00 | 8.60 | -0.05 | -0.70 | 0.44 | 106.50 | 0.15 |
| TRP | 1.35 | 1.31 | 1.17 | 0.93 | 1.51 | 1.70 | 1.66 | 1.57 | 1.20 | 0.83 | 1.32 | 204.00 | 163.00 | 8.00 | 0.81 | -0.90 | 0.31 | 224.60 | -0.26 |
| TYR | 1.37 | 1.37 | 1.22 | 1.26 | 1.37 | 1.51 | 1.50 | 1.41 | 1.33 | 0.95 | 1.10 | 181.00 | 141.00 | 6.20 | 0.26 | -1.30 | 0.42 | 177.70 | 0.08 |
| VAL | 0.95 | 1.01 | 0.74 | 0.52 | 0.72 | 0.95 | 0.91 | 1.07 | 1.02 | 0.63 | 1.68 | 117.00 | 105.00 | 5.90 | 1.10 | 4.20 | 0.39 | 141.00 | -0.71 |

**Tableau III-8 : Extrait d'un tableau reprenant des variables quantitatives étudiées pour la mise au point d'un modèle de régression logistique.**

- Détection des doublets favorisés.

  Nous avons voulu profiter de la taille des bases de données construites pour analyser les combinaisons d'acides aminés consécutifs dans la séquence (doublets). En effet, si l'utilisation de propension des acides aminés à être en interaction a déjà été utilisée, les analyses de doublets de résidus en interaction sont plus rares. Et nous nous sommes donc posé la question de savoir si certaines combinaisons de deux acides aminés pouvaient favoriser une interaction.

  On entend par doublets deux acides aminés se suivant dans la séquence et il existe donc 20 X 20 = 400 possibilité de doublets. Comme expliqué au point II.3.2 : « Définition des zones de résidus en interaction », nous avons défini des résidus en interaction (X) mais aussi des zones de résidus impliqués dans les interactions (Z). Dès lors, les doublets en interactions peuvent être définis de trois manières différentes ce qui est illustré dans la Figure III-38 :

  - Doublet XX : les deux résidus sont en interaction.
  - Doublet X : un seul des deux résidus est en interaction.
  - Doublet ZZ : les deux résidus appartiennent à une zone d'interaction.

  N.B. : les doublets Z n'ont pas été pris en compte car ils correspondent à un élargissement supplémentaire des zones.

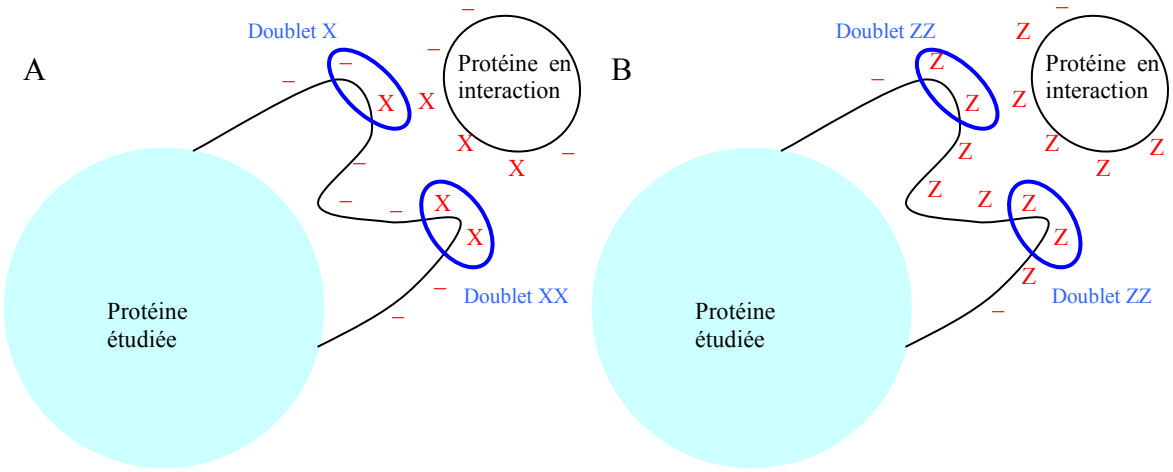



**Figure III-38 : Schéma d'une interaction entre deux protéines et mise en évidence des doublets (cercles bleu). Les X représentent des résidus en interaction et les Z des résidus présents dans les zones en interaction. Dans la partie de gauche (A) sont montrés des doublets de type XX et X. Dans la partie B (zone de résidus en interactions), deux doublets de type ZZ est mis en évidence.**

Ensuite, le nombre de chacun des 400 doublets présents dans la banque de données est comptabilisé ainsi que le nombre de doublets impliqués dans les interactions. Afin de détecter un doublet (dé)favorisé dans les sites d'interaction, les distributions de chaque doublet sont comparées à une distribution de type binomiale. La loi binomiale est la plus adaptée dans ce cas car elle permet d'utiliser des données discontinues et de mimer une distribution aléatoire. Donc, si la propension d'un doublet à se trouver dans un site d'interaction s'éloigne de manière significative de la distribution aléatoire de référence donnée par la loi binomiale (valeur de loi binomiale $< 0,05$), ce doublet est significativement (dé)favorisé dans les sites d'interaction. Il y a en effet moins de 5% de chance que cette propension soit obtenue aléatoirement. Une partie des résultats obtenus pour les doublets ZZ sont présentés dans le Tableau III-9.

| Doublet ZZ | Nombre en interaction (Zones) | Nombre dans l'échantillon | Loi binomiale cumulée | Loi binomiale | Etat du doublet |
|---|---|---|---|---|---|
| **EK** | 37 | 158 | 0,682 | 0,070 | - |
| **EL** | 74 | 251 | 0,997 | 0,002 | Favorisé |
| **EM** | 15 | 68 | 0,556 | 0,116 | - |
| **EN** | 12 | 95 | 0,013 | 0,007 | Défavorisé |
| **EP** | 18 | 81 | 0,563 | 0,106 | - |
| **EQ** | 19 | 89 | 0,482 | 0,101 | - |
| **ER** | 40 | 133 | 0,987 | 0,008 | Favorisé |
| **ES** | 24 | 122 | 0,290 | 0,072 | - |

**Tableau III-9 : Résultats partiels de l'étude des doublets ZZ. En plus du nombre de doublets dans l'ensemble de l'échantillon ou dans les résidus en interactions, sont données les valeurs pour la loi binomiale et la loi binomiale cumulée (probabilité de succès = 0.254). Si la valeur donnée par la loi binomiale est <0.05, le doublet ne suit pas une distribution aléatoire. La loi binomiale cumulée permet de déterminer si le doublet est favorisé ou défavorisé dans les sites d'interaction.**

Pour utiliser ces informations lors de la mise au point de la méthode de prédiction, deux types de données ont été extraites de ces études de doublets. Premièrement, un acide aminé appartenant à un doublet favorisé dans les sites d'interaction (doublets X et doublets ZZ) se voit attribuer une valeur de 1. Chaque acide aminé peut donc avoir une valeur de 0, 1 ou 2 (chevauchement de doublets favorisés). Deuxièmement, les valeurs de loi binomiale cumulées ont été utilisées pour chaque acide aminé (somme des deux valeurs des deux doublets auquel appartient le résidu).

### III.3.2. Création des échantillons d'analyse et de test

Pour les besoins de la mise au point et de la validation de la méthode de prédiction des sites d'interaction, des échantillons ont été construits. A partir de la banque de données (point II.1), un certain nombre de complexes ont été sélectionnés.

**Informations contenues dans la banque de données**

La banque créée a été analysée à l'aide d'un programme créé dans le cadre de cette thèse, **Xplor**, qui fournit des *informations par chaîne* dont notamment le type de chaîne (protéine, ADN/ARN, hétéroatomes, informations SwissProt…) et des informations structurales (CATH et Pfam). Dans le Tableau III-10, on peut en effet retrouver pour chaque complexe (code PDB – colonne 1), toutes les interactions possibles (combinaisons 2 à 2 entre les différentes chaines – colonne 2) et pour chaque chaine (partenaire 1 & 2 – colonnes 3 à 8 et 9 à 15), des informations qui ont été utilisées pour définir le type d'interaction (colonne 16 et 17). Les différents types décrits sont : interface homomérique (homo-permanent), hétéro-permanente et hétéro-transitoire (classification sur base de l'article d'Ofran[180] - cf. point II.1.3).

A l'aide de ces informations, nous allons construire une « image » de la banque (cf. Tableau III-11). Ceci est important afin d'avoir une vue générale de la composition de la banque et d'y détecter un éventuel biais mais aussi afin de réaliser des échantillons de composition optimale. Remarque : Seulement 911 complexes sur les 1475 possèdent assez d'information SwissProt pour déterminer le type d'interaction.

| | | Partenaire 1 (protéine) | | | | | | Partenaire 2 (protéine, acide nucléique, hétéroatome) | | | | | | | Type d'interaction | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| code PDB | Nom-Interaction | Nom partenaire 1 | Nom chaine 1 | Code SwissProt 1 | Nombre acides aminés 1 | Domaine CATH 1 | Domain Pfam 1 | Nom partenaire 2 | Nom chaine 2 | Code SwissProt 2 | Nombre acides aminés 2 | Type de partenaire | Domaine CATH 2 | Domaine Pfam 2 | Homo-Hétéro | Transitoire Permanent |
| 1a0c | 1a0cAD | 1a0cA | A | P19148 | 438 | ALPHA BETA | Xylose isomerase | 1a0cD | D | P19148 | 438 | PROT | ALPHA BETA | Xylose isomerase | homo | permanent |
| 1a0c | 1a0cAC | 1a0cA | A | P19148 | 438 | ALPHA BETA | Xylose isomerase | 1a0cC | C | P19148 | 438 | PROT | ALPHA BETA | Xylose isomerase | homo | permanent |
| 1a0c | 1a0cAB | 1a0cA | A | P19148 | 438 | ALPHA BETA | Xylose isomerase | 1a0cB | B | P19148 | 438 | PROT | ALPHA BETA | Xylose isomerase | homo | permanent |
| 1a0c | 1a0cBC | 1a0cB | B | P19148 | 438 | ALPHA BETA | Xylose isomerase | 1a0cC | C | P19148 | 438 | PROT | ALPHA BETA | Xylose isomerase | homo | permanent |
| 1a0c | 1a0cBD | 1a0cB | B | P19148 | 438 | ALPHA BETA | Xylose isomerase | 1a0cD | D | P19148 | 438 | PROT | ALPHA BETA | Xylose isomerase | homo | permanent |
| 1a0c | 1a0cBA | 1a0cB | B | P19148 | 438 | ALPHA BETA | Xylose isomerase | 1a0cA | A | P19148 | 438 | PROT | ALPHA BETA | Xylose isomerase | homo | permanent |
| 1a0c | 1a0cCB | 1a0cC | C | P19140 | 430 | ALPHA BETA | Xylose isomerase | 1a0cB | B | P19140 | 430 | PROT | ALPHA BETA | Xylose isomerase | homo | permanent |
| 1a14 | 1a14NL | 1a14N | N | P03472 | 388 | BETA | Neuraminidase | 1a14L | L | ___ | 104 | PROT | BETA | 0 | hetero | noSWS |
| 1a14 | 1a14NH | 1a14N | N | P03472 | 388 | BETA | Neuraminidase | 1a14H | H | ___ | 120 | PROT | BETA | 0 | hetero | noSWS |
| 1a14 | 1a14HN | 1a14H | H | ___ | 120 | BETA | 0 | 1a14N | N | P03472 | 388 | PROT | BETA | Neuraminidase | hetero | noSWS |
| 1a14 | 1a14HL | 1a14H | H | ___ | 120 | BETA | 0 | 1a14L | L | ___ | 104 | PROT | BETA | 0 | hetero | noSWS |
| 1a14 | 1a14LN | 1a14L | L | ___ | 104 | BETA | 0 | 1a14N | N | P03472 | 388 | PROT | BETA | Neuraminidase | hetero | noSWS |
| 1a14 | 1a14LH | 1a14L | L | ___ | 104 | BETA | 0 | 1a14H | H | ___ | 120 | PROT | BETA | 0 | hetero | noSWS |
| 1a2x | 1a2xAB | 1a2xA | A | P02586 | 159 | ALPHA | EF hand | 1a2xB | B | P02643 | 47 | PROT | _ | Troponin | hetero | transitoire |
| 1a2x | 1a2xBA | 1a2xB | B | P02643 | 47 | _ | Troponin | 1a2xA | A | P02586 | 159 | PROT | ALPHA | EF hand | hetero | transitoire |

**Tableau III-10 : Extrait d'un fichier contenant des informations par interaction.**

| | Transitoire / Permanent | Homo / Hetero | Nombre d'interactions | % | Nombre de complexes |
|---|---|---|---|---|---|
| DNA/RNA | / | / | 277 | 7,5 | 97 |
| Protéine | Permanent | hetero | 126 | 3,4 | 27 |
| Protéine | Permanent | homo | 2702 | 72,8 | 644 |
| Protéine | Transitoire | hetero | 606 | 16,3 | 143 |
| Total | | | 3711 | 100 | 911 |

**Tableau III-11 : Caractérisation de la banque de données utilisée.**

### Création des échantillons

Différentes stratégies peuvent être adoptées pour extraire un échantillon représentatif d'une base de données. Classiquement, on crée des échantillons à fraction sondée constante c'est-à-dire que l'on prélève des complexes de chaque type de façon proportionnelle afin d'avoir un échantillon à l'image de la base de données. Dans notre cas, cette méthodologie nous amènerait à récolter un grand pourcentage de complexes homomériques (~70% - cf. Tableau III-11) or ceci est dû à un biais de la PDB. En effet, ce type de complexe est globalement plus aisé à cristalliser mais cette surreprésentation de complexes homomériques ne représente pas une réalité biologique (bien que ce type de complexe soit fréquent).

Une deuxième solution pour créer un échantillon consiste à prendre un nombre identique de chaque type de complexe. Dans ce cas, la proportion de certains complexes, comme par exemple les complexes hétéromer-permanents serait exagérément élevée par rapport à la quantité de ce type de complexe que l'on peut s'attendre à retrouver dans les systèmes biologiques.

Une autre manière de travailler aurait été de choisir des proportions représentatives des complexes trouvés dans la cellule (ou dans un type cellulaire donné). Bien qu'il soit probable que la plupart des protéines agissent par l'intermédiaire de complexes multi-protéiques (70% chez la levure),[266] les publications abordant ce sujet sont rares et les chiffres donnés ne correpondent pas exactement aux valeurs recherchées.

Nous avons donc finalement décidé de choisir la proportion de type de complexes dans les échantillons de manière arbitraire. Ceci nous permet dans un premier temps d'augmenter la proportion de complexes hétéromer-transitoires qui sont les interactions ayant le plus d'intérêt scientifique. En effet, les interactions de ce type correspondent à des complexes de type : anticorps-antigène, enzyme-inhibiteur, protéine signal… Dans un deuxième temps, nous allons garder une proportion élevée de complexes homomériques car ils représentent une fraction importante des complexes protéiques. Finalement, le pourcentage

de complexes protéine-acide nucléique sera légèrement revu à la hausse et quelques complexes hétéromer-permanents seront conservés.

Les conditions choisies pour la création des échantillons sont les suivantes :

- 26 interfaces de type protéine-acide nucléique (21%).
- 12 interfaces de type hétéromer-permanent (9%).
- 46 interfaces de type homomérique (35%).
- 46 interfaces de type hétéromer-transitoire (35%).

Chaque complexe de la banque de données s'est vu attribuer un numéro aléatoire qui a permis de classer l'ensemble des complexes de manière aléatoire. Les complexes sont ensuite placés dans les échantillons en les sélectionnant dans l'ordre de classement aléatoire et en veillant à atteindre les conditions données ci-dessus. Deux échantillons ont été construits, un 'échantillon analyse' (122 complexes) qui va permettre de mettre au point notre modèle de prédiction et un 'échantillon test' (126 complexes) qui va permettre de valider le modèle construit sur base de l'échantillon analyse et donc d'estimer la qualité et la robustesse de notre méthode.

### III.3.3. Méthode de prédiction

**Régression logistique**

Dans le cadre de notre travail, les informations ont été recueillies sous formes de variables numériques quantitatives (propensions à être en interaction p.ex.) et qualitatives (ex. motif/domaine). La réponse à obtenir est elle, de type oui/non (booléen – 1/0) en fonction d'un résidu prédit en interaction ou pas. Dans ce cas, la régression logistique est une solution adaptée à la combinaison et à la sélection des variables.

Dans un premier temps, les variables les plus significatives sont retenues par une sélection de type « pas à pas » ('stepwise'). Tout d'abord, toutes les variables significatives sont incluses dans le modèle ('forward selection'). Ensuite, ces variables ne sont conservées que si elles restent significatives en présence des autres variables c'est-à-dire si elles apportent une information complémentaire. Pour être significative, une variable doit obtenir une valeur de niveau de signification (α) de 0,05.

Dans un deuxième temps, les variables significatives vont être combinées dans une équation afin de permettre le calcul du Logit (Logit - Figure III-39). Dans cette équation, les valeurs des variables quantitatives sont multipliées par le coefficient correspondant alors que

pour les variables qualitatives, une valeur différente est attribuée à chaque modalité (forme) que la variable peut prendre. Ensuite, la méthode statistique construit une courbe de probabilité à partir de la valeur du Logit calculée. Cette probabilité est calculée selon l'équation : Probabilité = Exp(Logit) / (1 + Exp(Logit)). On obtient donc pour chaque acide aminé, une valeur de probabilité (P). La courbe est optimisée pour qu'une P>0.5 corresponde à un acide aminé en interaction (1 ou oui), et inversement pour une P<0.5. Dans la suite de ce travail, la probabilité calculée sera appelée 'score' du résidu considéré. Une description plus détaillée de la régression logistique peut être trouvé dans le livre de Agresti.[267]

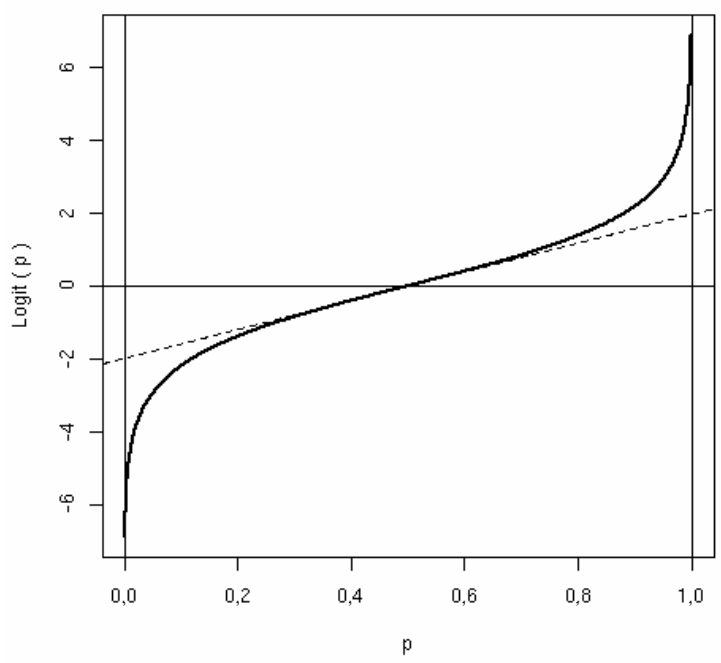



**Figure III-39 : Exemple de fonction Logit [Logit = ln(p/(1-p))].**

### Évaluation des résultats

Lors de l'optimisation de la méthode de prédiction, il est essentiel de pouvoir évaluer la méthode sur base de paramètres pertinents. Premièrement, les notions de vrais/faux positifs/négatifs doivent être reprécisées. Remarque : afin d'éviter tout problème face aux termes utilisés en littérature anglaise, la traduction des paramètres choisis est également fournie.

- TP = 'True Positive' = Vrai Positif = Prédiction « en interaction » correcte
- TN = 'True Negative' = Vrai Négatif = Prédiction « non en interaction » correcte
- FP = 'False Positive' = Faux Positif = Prédiction « en interaction » incorrecte

- FN = ‘False Negative’ = Faux Négatif = Prédiction « non en interaction » incorrecte

Les trois principaux paramètres utilisés dans cette thèse sont :

- La spécificité a été définie comme le rapport du nombre de prédictions correctes sur le nombre de résidus prédits en interaction. Une spécificité de 100% signifie que toutes les prédictions de résidus en interaction réalisées sont correctes.

  → **Spécificité = TP / (TP + FP)** aussi appelée ‘Accuracy’.

  Remarque : en général, la spécificité (**‘specificity’**) est égale à : TN / (TN + FP).

- La sensibilité est définie comme le nombre de résidus correctement prédits sur le nombre total de résidus en interaction. Une sensibilité de 100% signifie que l’on a détecté tous les résidus en interaction.

  → **Sensibilité = TP / (TP + FN)** aussi appelée ‘Recall’.

- Le **coefficient de corrélation de Matthews** (**MCC** - ‘Matthews Correlation Coefficient’). Ce facteur varie entre -1 et +1, entre désaccord et accord total. Une valeur de 0 signifie une prédiction aléatoire.

  → $$MCC = \frac{t_p t_n - f_p f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}}$$

### III.3.4. Mise au point de la régression logistique

Les variables utilisées comme données d’entrée du système sont décrites au point III.3.1 et ont été calculées pour l’ensemble des acides aminés de l’échantillon analyse et test (voir ci-dessus). L’ensemble de ces variables a été analysé par le système de régression logistique et lors de la mise au point du modèle, nous avons recherché une spécificité optimale.

**Sélection des variables**

Lors de tests préliminaires, il est apparu que le modèle construit n’était pas robuste : les résultats obtenus sur l’échantillon analyse étaient tout à fait satisfaisants mais les résultats de validation sur l’échantillon test étaient moins bons. En effet, la spécificité calculée lors de la validation était inférieure de 11% par rapport à celle calculée sur l’échantillon analyse

(47,4% au lieu de 58%). Les causes de ce problème peuvent être multiples mais la plus probable est le sur-apprentissage qui a déjà été rencontré lors de la mise au point d'autres méthodes de prédiction.[187,196,200,210] En effet, face à un problème complexe (et c'est le cas de la prédiction des sites d'interaction), beaucoup de systèmes statistiques 'connaissent par cœur' les données utilisées et ont tendance à créer un modèle propre à ces données. La solution proposée par le modèle n'arrive donc pas à résoudre le problème dans son ensemble.

Pour palier à ce problème et selon le principe du rasoir d'Occam, nous avons réduit le nombre de variables retenues par le modèle. Par exemple, un des premiers modèles construits contenait 24 variables quantitatives et 9 variables qualitatives ce qui donne un total de 33 variables utilisées par le modèle. Ce nombre peut être suffisant pour permettre au système de trouver une solution propre à l'échantillon (sur-apprentissage). Afin d'arriver à réduire le nombre de variables, nous avons recherché les variables robustes.

Les variables robustes ont été définies selon deux critères. Premièrement, nous avons recherché les variables qui étaient retenues dans au moins 3 des 4 modèles construits : modèles basés sur les résidus en interaction (X) ou sur les résidus dans les zones d'interactions (Z) et modèles basés sur l'ensemble des séquences ou sur les séquences comprenant de 50 à 300 résidus (cf. sections suivantes). Deuxièmement, les variables trouvées doivent avoir des coefficients multiplicateurs (cf. point III.3.3) de même signe et de même ordre de grandeur (variation maximale du coefficient tolérée = facteur 3 - p.ex. « 0.1*volume » dans un premier modèle et « 0.3*volume » dans un second).

Il faut remarquer que lors de cette étape, la robustesse du modèle est obtenue au détriment d'une baisse de sensibilité et d'une légère baisse de spécificité du modèle.

**Fraction de résidus en interaction**

Un des facteurs influençant l'efficacité de la méthode est le pourcentage de résidus en interaction dans les séquences. En effet, lorsque ce pourcentage est faible, le modèle construit a tendance à avoir une sensibilité trop faible car la fraction de résidus prédits comme étant en interaction est très faible (et parfois nulle) et la spécificité du modèle est également affectée. Les meilleurs résultats ont donc été obtenus pour des modèles de régression logistique utilisant les zones de résidus en interaction (cf. point II.3.2). L'utilisation des zones de résidus en interaction permet, par exemple pour l'échantillon analyse, de passer de 16,2% de résidus en interaction à 31,5% de résidus dans les zones en interaction et, par la même occasion, de passer d'une spécificité de 31,4% à 56,5% (cf. Tableau III-12, résultats 'Sur X' et 'Sur Z').

**Longueur de séquence**

Lors de la suite des analyses, il s'est avéré que la qualité des prédictions dépendait de la longueur de la séquence. Pour les séquences de petite taille, certaines variables comme la position du résidu dans la séquence ou les variables calculées sur des fenêtres de calcul de plusieurs acides aminés deviennent moins significatives et limitent l'efficacité du modèle. Pour les séquences de grande taille, c'est le faible pourcentage de résidu en interaction (11,3% dans les séquences de plus de 300 résidus au lieu de 16,2% en moyenne dans l'échantillon analyse) qui influence négativement le modèle. En effet, le modèle prédit très peu de résidus comme étant en interaction et la sensibilité du modèle en est affectée. Lors de la mise au point de la méthode, nous avons donc conservé uniquement les séquences ayant une longueur comprise entre 50 et 300 acides aminés ce qui permet d'avoir une sensibilité qui passe de 3,5% à 14,7% (sur l'échantillon test). Cette restriction à des séquences de 50 à 300 résidus devra être rappelée à chaque utilisation de la méthode de prédiction et, bien que des séquences au-delà de ces limites puissent être étudiées, l'efficacité de la méthode sur celles-ci ne pourra être garantie.

Une remarque importante doit être faite ici : les mauvais résultats obtenus pour les séquences de grande taille sont également expliqués par le manque de caractérisation fonctionnelle de certaines de ces séquences. En effet, le faible taux de résidus en interaction pour les protéines de grande taille est, dans certains cas, dû au fait que l'on ne connait pas tous les sites d'interaction de ces protéines ou que ceux-ci ne sont pas contenus dans la structure 3D.

### III.3.5. Analyse du modèle de prédiction

Le modèle de régression logistique donnant les meilleurs résultats a été construit sur base des séquences de 50 à 300 acides aminés de l'échantillon analyse. Ce modèle a été construit en considérant les zones de résidus en interaction et contient un nombre réduit de variables appelées variables robustes (cf. point III.3.4). Les principales caractéristiques du modèle final de prédiction des sites d'interaction protéiques sont décrites ci-dessous.

**Variables significatives**

Parmi les propensions calculées à partir des fréquences des différents types de résidus dans les banques de données, 2 ont été retenues dans le modèle final :

- Propension à être en interaction avec l'ADN sur une fenêtre de 9 résidus.
- Propension à être en interaction avec les hétéroatomes sur une fenêtre de 5 résidus.

D'autres valeurs quantitatives ont été retenues :

- Position du résidu dans la séquence.
- Propensions à se trouver dans des doublets X favorisés.

Parmi les variables qualitatives retenues, nous avons :

- Le nom de l'acide aminé (code 1 lettre).
- La prédiction PredAcc à être en surface.
- La prédiction de structure secondaire selon PsiPred.

Les recherches InterProScan de motifs/domaines protéiques ont aussi été retenues :

- Domaine PRINTS.
- Domaine SMART.
- Domaine PROSITE.

Finalement, les RBD calculés sur une fenêtre de 9 résidus avec prise en compte de la fenêtre de calcul (zone RBD 9) font également partie des variables significatives retenues.

Le modèle construit contient donc 4 des 110 variables quantitatives et 7 des 29 variables qualitatives (11 variables au lieu de 33 variables que dans le premier modèle construit). Ces variables couvrent plusieurs caractéristiques importantes des acides aminés aux interfaces : la capacité d'interagir des résidus, leur accessibilité, leur structure secondaire, leur position dans la séquence, leur localisation dans un domaine fonctionnel connu. La pertinence de ces variables est discutée dans le prochain chapitre (point IV.2).

**Résultats fournis par le modèle**

Pour réaliser une prédiction sur une protéine, notre méthode va tout d'abord calculer les valeurs des 11 variables du modèle pour chaque acide aminé de la protéine. Ensuite, le calcul du score est réalisé et permet notamment de fournir une sortie graphique comme celle présentée sur la Figure III-40. Les résidus ayant un score supérieur à 0,5 sont prédits comme étant impliqués dans une interaction.

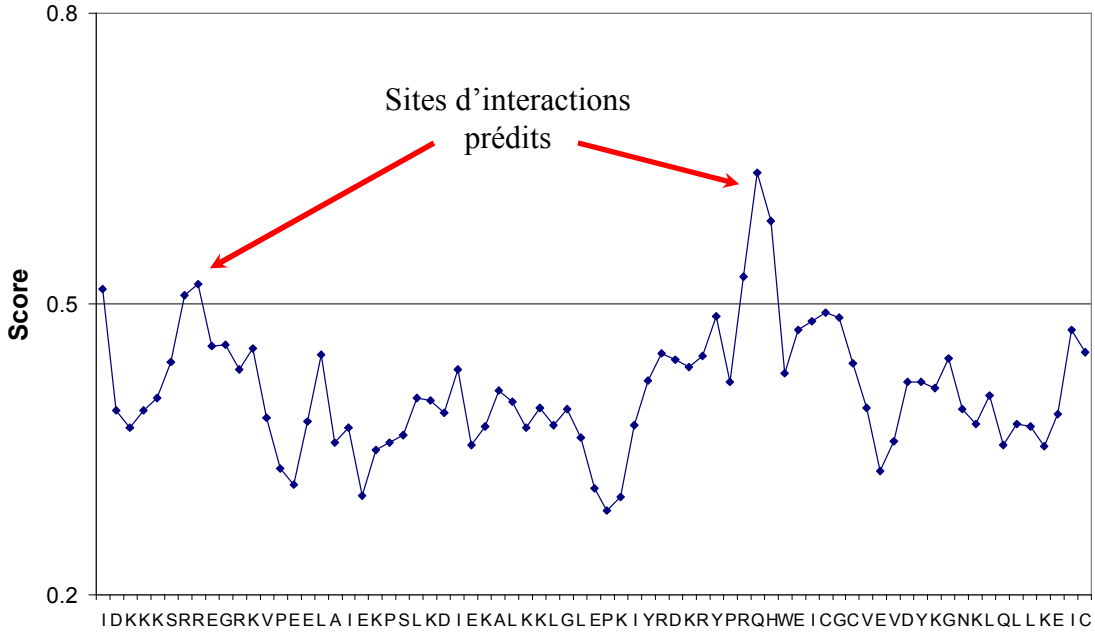



**Figure III-40 : Exemple de prédiction obtenue par le modèle de régression logistique sur la chaîne protéique du complexe 1LNG.[268] Les acides aminés ayant un score supérieur à 0,5 sont prédits comme étant en interaction.**

## III.3.6. Qualité des résultats

### Qualité de la méthode de prédiction

La qualité des résultats de notre modèle est estimée par le calcul de la spécificité (pourcentage de prédictions correctes), la sensibilité (pourcentage de résidus en interaction détectés) et le coefficient de corrélation de Matthews (MCC) (cf. point III.3.3). Ces valeurs sont reprises dans le tableau suivant pour les résultats obtenus sur l'échantillon ayant été utilisé pour mettre au point la méthode (échantillon analyse), sur un échantillon indépendant (échantillon test) et les résultats qui auraient été obtenus par une prédiction aléatoire. Les valeurs aléatoires ont été calculées par sélection aléatoire, dans l'échantillon test, d'un nombre de résidus égal au nombre de résidus prédits par le modèle et ce, à 15 reprises.

Dans le Tableau III-12, on peut voir que les résultats de la validation sur l'échantillon test sont proches de ceux obtenus directement sur l'échantillon analyse et que notre modèle est donc robuste. La méthode de prédiction est optimale pour la prédiction des zones de résidus en interaction avec une spécificité de plus de 51%. La sensibilité est de l'ordre de 15% alors que le coefficient de corrélation est significativement différent de zéro. La valeur de ces résultats sera discutée plus en détail dans le prochain chapitre mais il est d'ores et déjà important de rappeler ici que le modèle a été optimisé pour obtenir une spécificité maximale.

En résumé et en moyenne pour une protéine donnée, notre méthode va prédire 10 % des acides aminés de la séquence comme étant en interaction. De ces 10%, un peu plus de la moitié (51%) sont réellement des acides aminés en interaction et environ 15% des acides aminés en interaction sont détectés.

| | | Echantillon analyse | Echantillon indépendant | Aléatoire (écart-type) |
|---|---|---|---|---|
| Sur X | Spécificité (%) | 31,4 | 25,1 | 17,5 (+/-0,9) |
| | Sensibilité (%) | 14,9 | 15,1 | 10,0 (+/-0,5) |
| | MCC | 9 | 6,8 | 0,0 (+/-0,8) |
| Sur Z | Spécificité (%) | 56,5 | **51,1** | 36,7 (+/-1) |
| | Sensibilité (%) | 14 | **14,7** | 10,0 (+/-0,5) |
| | MCC | 11,7 | **10,4** | 0,0 (+/-0,7) |
| | Nb séquences | 116 | 146 | 146*15 |

**Tableau III-12 : Résultats des prédictions faites par le modèle final de régression logistique. MCC = 'Matthews Correlation Coefficient', X = résidus en interaction et Z = résidus dans une zone en interaction.**

En pratique, les résultats obtenus diffèrent d'une protéine à l'autre :

- Pour 41 séquences sur les 146 de l'échantillon test, notre méthode de prédiction atteint une spécificité remarquable de plus de 70%. Presqu'un tiers des séquences sont donc prédites avec une qualité élevée : 16,6% des acides aminés de ces séquences sont prédits en interaction et 70% de ces acides aminés prédits sont réellement des acides aminés en zone d'interaction. Parmi ces 41 séquences, un taux de spécificité maximal (100%) est atteint pour 19 séquences.
- Pour 15% des séquences (22 séquences), la méthode de prédiction s'avère inefficace car aucun acide aminé n'est prédit comme étant en interaction. Pour palier à ce problème, le cut-off de probabilité de la régression logistique peut-être diminué afin de pouvoir réaliser une prédiction (cut-off optimal = 0,5 voir point III.3.3 et Figure III-40). Néanmoins, quand le cut-off est diminué, la prédiction sur ces 22 séquences est moins précise avec dans notre cas, une spécificité de 38.6% (sur les zones) pour un cut-off de 0,4. Ce résultat est insuffisant car très faiblement différent d'un résultat aléatoire (cf. Tableau III-12) et de plus, pour 3 séquences, la méthode ne prédit toujours pas de résidus en interaction. La diminution du cut-off doit donc être réalisée avec prudence.

**Exemples de prédiction**

Sur la Figure III-41, les résultats obtenus pour la 'Signal Recognition Particule' en interaction avec l'ARN dans le complexe 1LNG[268] sont donnés. On peut voir que les deux sites prédits correspondent bien à des sites d'interaction et que seul le troisième site n'est pas prédit.

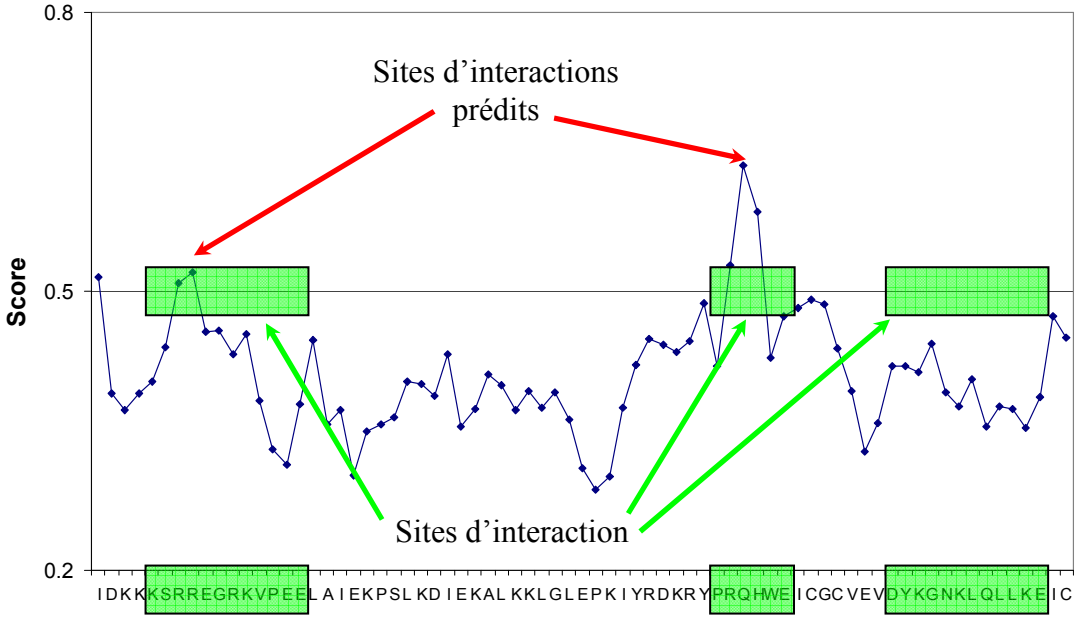



**Figure III-41 : Exemple de prédiction obtenue par le modèle de régression logistique sur la chaîne protéique du complexe 1LNG.[268] Les acides aminés en interaction dans la structure 3D sont signalés par des cadres verts et les résidus ayant un score supérieur à 0,5 sont ceux prédits comme étant en interaction.**

Les prédictions réalisées sur la 'Signal Recognition Particule' (A) ainsi que les séquences de la 5'-désoxy-ribonucléotidase YfbR (B), de la protéine Rab7 (C) et d'un inhibiteur de lysozyme de type C (D) sont représentés en 3D dans la Figure III-42. Sur ces exemples, on voit que les prédictions ciblent bien les sites d'interactions et que peu de prédictions ne correspondent pas à l'interface.

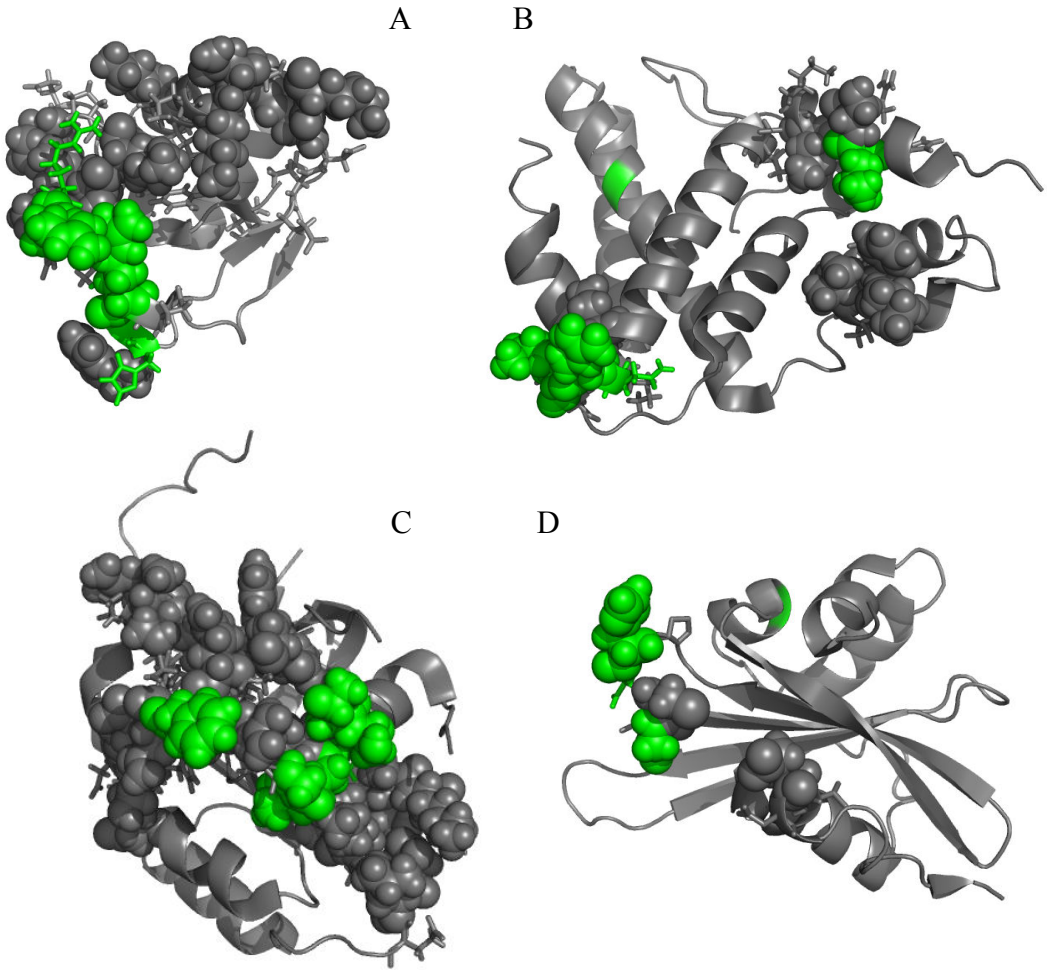



**Figure III-42 : Exemple de prédictions d'acides aminés en interaction. Les acides aminés en interaction (X) sont représentés en volume réel, les résidus dans les zones en interaction (Z) sont représentés en mode bâtonnets et le reste de la protéine est représenté en mode ruban. Les acides aminés prédits sont mis en évidence en vert. A : complexe 1LNG,[268] chaîne A en interaction avec de l'ARN au sein de la 'Signal Recognition Particule' ; B : interaction homo-dimérique de la chaîne A de 1WPH ; C : le site d'interaction avec la guanosine di-phosphate et l'interaction hétéro-transitoire avec un résidu de la deuxième chaîne protéique du complexe 1VG0[269] sont détectés sur la chaîne A ; D : prédiction des interactions au sein du trimère de l'inhibiteur de lysozyme de type C (1XS0[270] chaîne A). Images générées par le logiciel PyMol.[167]**

Au sein des séquences qui ont un taux de spécificité maximal (100%) se trouve le complexe ribonucléase-inhibiteur. Sur la séquence de la chaîne inhibitrice, 5 acides aminés sont correctement prédits comme étant en interaction et la sensibilité est de 25% (l'interface étant constituée de 20 acides aminés) (cf. Figure III-43).

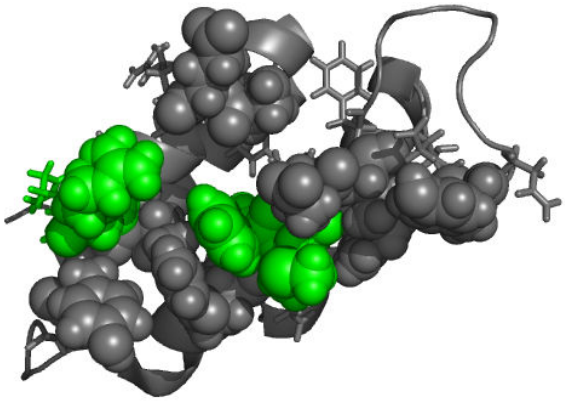


**Figure III-43 : Représentation du complexe ribonucléase-inhibiteur (1v74).[271] Exemple de prédictions d'acides aminés en interaction. Les acides aminés en interaction (X) sont représentés en volume réel, les résidus dans les zones en interaction (Z) sont représentés en mode bâtonnets et le reste de la protéine est représenté en mode ruban. Les acides aminés prédits sont mis en évidence en vert. Image générée par le logiciel PyMol.[167]**

# IV. DISCUSSION GÉNÉRALE

## *IV.1. Analyse atomique des interactions protéiques*

La première partie de cette thèse avait pour but de décrire le plus précisément possible les caractéristiques des sites d'interaction protéique. Pour cela, nous avons construit des banques de structures de complexes protéine-protéine et protéine-acide nucléique de haute résolution et contenant des séquences non-homologues. Nous avons extrait 131.531 interactions de 1.297 complexes protéine-protéine, 7.671 interactions de 139 complexes protéines-ADN et 3.367 interactions de 49 complexes protéines-ARN. Ces banques de complexes ont été comparées à la banque de séquences protéiques Swiss-Prot/UniProt et nous avons pu montrer que les complexes utilisés sont représentatifs des protéines couramment rencontrées en biologie.

Cette première partie du travail a été réalisée avec deux objectifs principaux. Premièrement, mettre à jour les connaissances, au niveau atomique, sur les interfaces protéiques par l'utilisation d'une des plus grandes banques de données construite à ce jour. Et deuxièmement, extraire des informations pertinentes sur les interfaces afin de les utiliser pour la mise au point d'une méthode de prédiction des sites d'interaction. Les principaux résultats obtenus sont discutés ci-dessous.

**Vers un modèle des interfaces protéiques ?**

Comme nous l'avons expliqué au point I.4.3 - « État des lieux », les deux théories les plus couramment reprises sont la théorie du O-ring et celle des hot-spots. Le modèle élémentaire du O-ring[87] peut être fiable mais uniquement pour des interfaces suffisamment grandes que pour permettre la mise en place d'un tel anneau. De leur côté, les hot-spots[103] semblent être retrouvés dans un grand nombre d'interfaces mais ne donnent pas d'idée générale de ce que pourrait être une interface modèle. Ces deux théories peuvent être combinées de manière assez élégante[111] pour décrire les sites d'interaction mais le modèle proposé n'en n'est que plus restrictif. Pour rappel et en résumé, le modèle combiné situerait les résidus énergétiquement importants (hot-spots) au centre d'une zone complètement isolée du solvant grâce aux acides aminés de l'anneau du O-ring. Ces deux modèles peuvent donc être utilisés pour se donner une idée de ce que pourrait être une interface type mais il est clair que les interfaces protéiques sont très variables et se laissent difficilement schématiser d'une seule et unique manière.

Pour illustrer la diversité des interfaces, prenons par exemple le cas de l'entérotoxine LT-IIb de *Escherichia coli* (1TII).[272] Cette protéine est constituée de deux sous-unités : une sous-unité dimérique (sous-unité A) et une sous-unité homo-pentamérique (sous-unité B en bleu dans la Figure IV-1). La sous-unité A est constituée de la sous-unité A1 et du 'linker' A2 (en orange et rouge dans la Figure IV-1, respectivement). La sous-unité A catalyse l'ADP-ribosylation d'une sous-unité de la protéine G[272] alors que la sous-unité homo-pentamérique est responsable de la liaison aux récepteurs de surface des cellules épithéliales du petit intestin (récepteurs gangliosides).[273]

L'analyse des interactions de la sous-unité A1 met en évidence la diversité des interactions qui peuvent être trouvées au sein d'une seule séquence. En effet, la sous-unité A1 est premièrement impliquée dans une interaction de type hétéromer-transitoire avec la sous-unité B, interaction notamment réalisée par l'intermédiaire d'un pont salin (cadre 1 dans la Figure IV-1). Dans un deuxième temps, la sous-unité A1 interagit avec le linker A2 de manière hétéromer-permanente, cette interaction étant caractérisée par la présence d'un pont disulfure (cadre 2 dans la Figure IV-1). Par ailleurs, les 5 chaînes polypeptidiques du pentamère interagissent entre elles de manière permanente (homomérique) et la présence de grands patchs hydrophobes aux interfaces stabilise la sous-unité B (cadre 3 dans la Figure IV-1). La structure de type « AB5-like » de ce complexe nous fournit donc un exemple concret sur la diversité que peuvent adopter les interfaces protéiques.

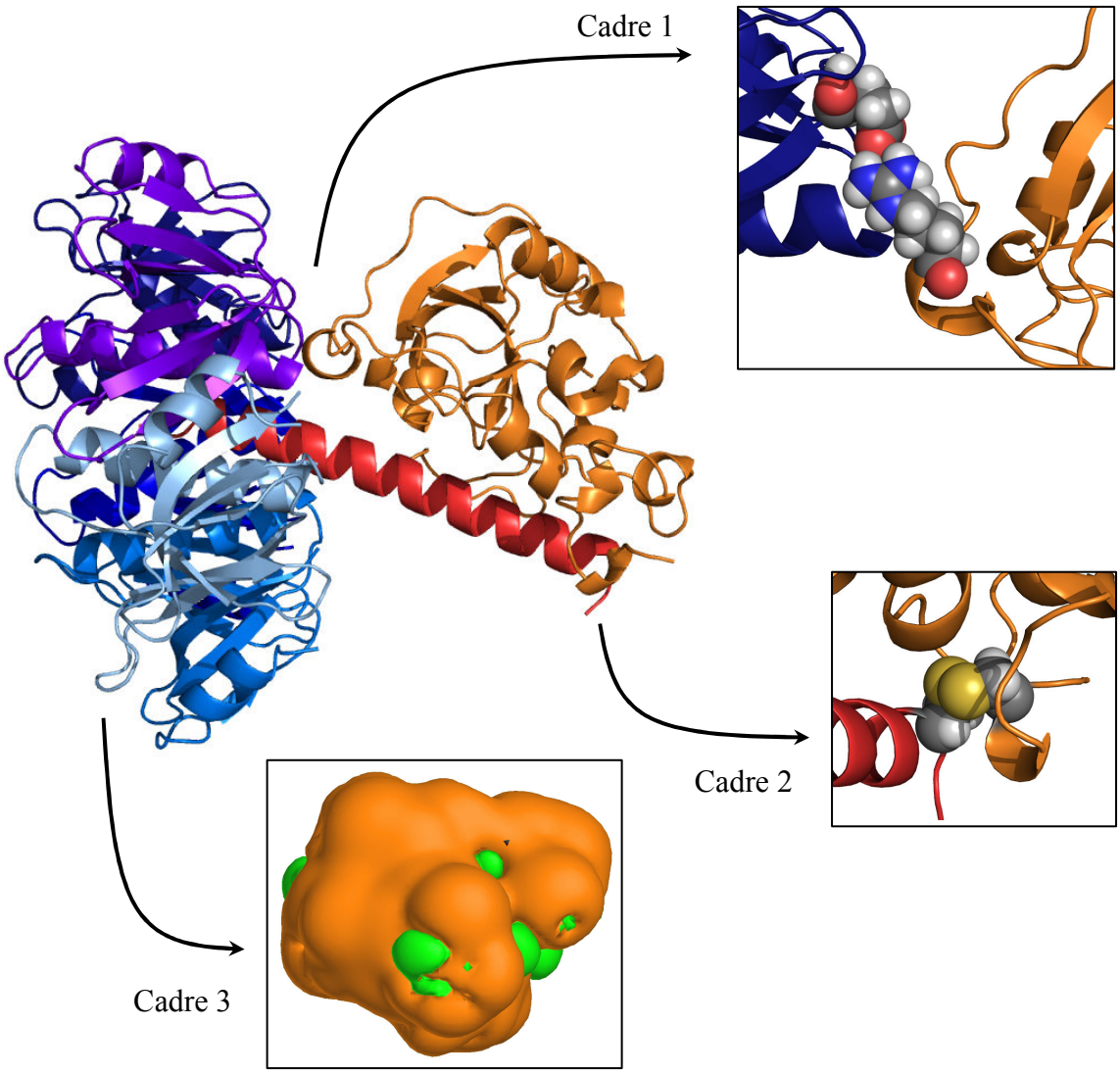



**Figure IV-1 : Représentation 3D du complexe entérotoxine LT-IIb. La sous-unité A est composée de la partie A1 (en orange) et du linker A2 (en rouge), et la sous-unité B est constituée de 5 chaînes polypeptidiques (niveaux de bleu). Cadre 1 : pont salin entre la sous-unité A1 et la sous-unité B. Cadre 2 : pont disulfure entre la sous-unité A1 et le 'linker' A2. Les atomes de soufre impliqués dans le pont disulfure sont représentés en jaune. Cadre 3 : vue MHP[15] de l'interface au sein du pentamère de la sous-unité B (surfaces hydrophobes en orange et hydrophiles en vert). Images générées par les logiciels PyMol[167] et YAGME.**

### Caractéristiques communes des interfaces

Comme expliqué dans la section précédente, la diversité des interfaces protéiques est un fait certain et il semble que la diversité de celles-ci soit à l'échelle du nombre de complexes protéiques retrouvés dans les organismes vivants (1.485 dans cette thèse). Dans certains cas, une protéine d'un complexe peut être impliquée dans plusieurs types d'interactions et cette variabilité limite notre capacité à mettre en évidence les caractéristiques

propres aux interfaces. Néanmoins, certaines caractéristiques globales ont été mises en évidence dans cette thèse et sont décrites ci-dessous.

La concentration élevée en résidus aromatiques (phénylalanine, tryptophane et tyrosine) aux interfaces est l'une des propriétés principales de notre banque de données. En effet, au sein des interfaces protéiques, la propension globale à être en interaction des acides aminés aromatiques est de 1,4. De plus, les couples de résidus avec les acides aminés aromatiques sont presque toujours favorisés. Les seules exceptions sont les interactions des aromatiques avec la cystéine, la sérine et la glycine. L'importance des acides aminés aromatiques a été montrée dans plusieurs travaux.[86,107,274] Il a également été montré que les couples entre ce type de résidus sont impliqués dans les interactions entre acides aminés éloignés dans la séquence[275] et qu'ils permettent la différenciation des structures secondaires durant le reploiement protéique ('folding').[276] Cette capacité à favoriser les interactions à longue distance au sein d'une même séquence est certainement également responsable de l'importance des résidus aromatiques au sein des interfaces.

Une autre caractéristique des interfaces protéiques est la présence très fréquente d'arginine. Nos résultats montrent que l'arginine est l'un des acides aminés les plus favorisés dans tous les types d'interfaces et particulièrement au sein des interactions protéine-acide nucléique. De plus, les couples Arg-Tyr sont presque toujours favorisés dans les interfaces protéine-protéine ce qui est en accord avec les résultats de Bahadur *et al*.[277] Il a également été proposé que les interactions cations-pi et particulièrement les couples Arg-Tyr participent à la spécificité des interactions protéine-protéine.[278] Finalement, Rajamani *et al*.[116] ont suggéré que l'arginine et les résidus aromatiques sont utilisés comme résidus d'ancrage aux interfaces protéine-protéine.

Les ponts salins entre résidus de charges opposées ainsi que les contacts hydrophobes sont largement impliqués dans les interactions protéiques. En particulier, l'enrichissement en contacts hydrophobes a déjà été montré pour les protéines oligomériques[225,279-282] mais aussi pour les interfaces transitoires.[220] Ces deux types d'interaction ont la particularité de faire intervenir les résidus aromatiques et l'arginine, et nous avons montré qu'ils sont favorisés dans l'ensemble de nos interfaces.

Récemment, une nouvelle étude sur les propriétés des interfaces a été réalisée par Negi & Braun.[283] Ils ont calculé les propensions des acides aminés à se situer aux interfaces à partir de 72 complexes protéiques et, comme dans notre banque de données, ils ont mis en évidence

les fortes propensions des acides aminés aromatiques, de l'arginine, de la cystéine et de l'histidine. Finalement, les résultats obtenus lors de cette thèse sont également en accord avec les résultats obtenus sur la composition des hot-spots. En effet, l'implication dans les hot-spots du tryptophane,[284] des résidus aromatiques[115] et des couples de résidus aromatiques[108] a été démontrée. Dans leur revue parue en 2007, Moreira *et al*.[112] ont confirmé l'importance du tryptophane, de l'arginine et de la tyrosine dans les hot-spots.

| | Propension à être en interaction | Propension à être dans les hot-spots |
|---|---|---|
| **ALA** | 0,82 | --- |
| **ARG** | **1,28** | **2,47** |
| **ASN** | 0,87 | 0,93 |
| **ASP** | 0,91 | **1,67** |
| **CYS** | 0,58 | 0.00 |
| **GLN** | 0,97 | 0,58 |
| **GLU** | 0,91 | 0,68 |
| **GLY** | 0,69 | 0,45 |
| **HIS** | **1,18** | **1,49** |
| **ILE** | **1,13** | **1,79** |
| **LEU** | **1,22** | 0,01 |
| **LYS** | 0,87 | **1,17** |
| **MET** | 0,69 | 0,54 |
| **PHE** | **1,37** | 0,56 |
| **PRO** | 1,10 | **1,25** |
| **SER** | 0,87 | 0,21 |
| **THR** | 1,00 | 0,28 |
| **TRP** | **1,31** | **3,91** |
| **TYR** | **1,37** | **2,29** |
| **VAL** | 1,01 | 0.00 |

**Tableau IV-1 : Propensions à se trouver dans les interfaces protéine-protéine et dans les hot-spots.[103] Les acides aminés ayant les propensions les plus élevées sont mis en évidence en gras.**

Deux hypothèses peuvent être émises suite à l'observation de cette corrélation entre hot-spots et résidus favorisés aux interfaces :

- La distribution retrouvée dans les hot-spots est une conséquence de la présence élevée de certains résidus dans les interfaces. Dans ce cas, c'est la propension de certains résidus à se trouver dans les interfaces qui implique que les mêmes résidus soient également retrouvés dans les hot-spots.
- La forte proportion de certains résidus aux interfaces serait guidée par leur importance dans l'énergie de liaison. Dès lors, les acides aminés des hot-spots sont favorisés aux interfaces ce qui conduit à une augmentation de la fréquence de ces résidus dans les interfaces.

Pour tenter de déterminer laquelle de ces hypothèses est la plus vraisemblable, nous avons comparé les propensions des acides aminés à se trouver dans les interfaces de nos complexes à celles à se trouver dans les hot-spots.[103] Le Tableau IV-1 confirme que des tendances communes sont retrouvées par les deux méthodes (arginine, tryptophane, tyrosine mais aussi histidine et isoleucine) mais la corrélation entre les valeurs obtenues est loin d'être parfaite. La deuxième hypothèse est à privilégier et il semble donc que ce soit l'implication énergétique des ces acides aminés qui les favorise aux interfaces. Cette corrélation entre hot-spots et résidus favorisés aux interfaces démontre que notre manière de travailler, par une sélection optimisée des résidus en interaction est efficace. Elle permet en effet de mettre en évidence les acides aminés les plus importants énergétiquement pour les interactions tout en évitant de devoir passer par une étape de construction de banques de mutants alanine.

Remarque : les conclusions de la comparaison faite dans le Tableau IV-1 doivent néanmoins être modérées car les distributions dans les banques de référence sont différentes ce qui influence les valeurs de propensions. Ces différences sont dues à la grande variation dans le nombre de sites étudiés : l'analyse des hot-spots a été réalisée sur base de 2325 mutants alanine ce qui est très loin de notre banque de 290.013 acides aminés

**Spécificités des différentes sous-banques d'interfaces**

Il est largement admis que les interactions homomériques sont plus hydrophobes que les interactions hétéromériques.[86] Néanmoins, Ofran & Rost[99] ont montré que lorsque les complexes transitoires et permanents sont séparés, cette distinction disparait. Au sein de nos interfaces protéine-protéine, nous avons montré que les interfaces hétéromer-transitoires sont légèrement plus hydrophiles que la Swiss-Prot/UniProt (banque de référence) alors que les interfaces les plus hydrophobes sont les interfaces hétéromer-permanentes. Il apparait donc que ce soit bien le caractère permanent qui confère une plus grande hydrophobicité aux interfaces alors que les interfaces transitoires sont elles, plus hydrophiles.

Dans les interfaces hétéromer-transitoires, l'importance des résidus aromatiques (propension de 1,4) et de l'arginine est confirmée. De plus, l'histidine (1,2) est également favorisée tout comme dans les complexes transitoires utilisés par Ansari & Helms.[220]

Au sein des interfaces homomériques, l'histidine mais aussi la leucine (1,2) sont également favorisées. Néanmoins, malgré le grand nombre de complexes utilisés et bien qu'une propension supérieure à 1,2 soit significative, aucun résidu n'obtient de propension particulièrement élevée (plus de 1,3). Finalement, comme expliqué dans la section précédente,

les homodimères sont régulièrement considérés comme étant stabilisés par des interactions hydrophobes.[225,279-281] Cette tendance est retrouvée dans notre banque d'interfaces homomériques, spécialement pour la leucine mais de manière moins marquée pour les autres acides aminés hydrophobes.

Ensuite, nous avons montré que les interfaces hétéromer-permanentes sont particulièrement enrichies en isoleucine et que la cystéine y adopte un comportement particulier. En effet, alors qu'elle est défavorisée dans les interfaces hétéromer-transitoires (0,7) et homomériques (0,5), la cystéine est favorisée dans les interfaces hétéromer-permanentes (1,2). Ceci peut être expliqué par le fait que les interfaces de ce type de complexes ne savent pas toujours mettre en place de patchs hydrophobes suffisamment grands pour se stabiliser. En effet, ce type d'interface est régulièrement constitué de chaînes intimement liées ou « entremêlées ». Dès lors, le caractère permanent des interfaces est assuré par la mise en place de ponts disulfures : les cystéines en interaction y sont quatre fois plus impliquées dans des ponts disulfures que dans les autres types d'interfaces protéine-protéine.

Dans les résultats ci-dessus, nous avons montré que certaines tendances peuvent être tirées pour différencier les interfaces hétéromers/homomers et transitoires/permanentes. Néanmoins, les caractéristiques mises en évidence ne sont pas aussi nettes que celles présentées par Ofran & Rost en 2003[99] qui montrent dans leur travail que la composition des séquences suffit à différencier les types d'interfaces. Il est possible que la composition et la taille de la banque de données utilisée (deux tiers de la nôtre) ou que notre manière optimisée d'extraire les acides aminés en interaction soient responsables de ces différences. Finalement, il ne faut pas perdre de vue qu'il existe un continuum entre les différents types d'interfaces et que la stabilité des complexes est également influencée par les conditions physiologiques et l'environnement cellulaire (cf. Nooren & Thornton).[96]

Au sein des interfaces protéine-acide nucléique, les acides aminés hydrophiles représentent plus de 60% des résidus impliqués. L'arginine et la lysine possèdent les propensions les plus élevées à interagir. L'importance de ces deux résidus ainsi que celle de l'asparagine et de la sérine a été montrée précédemment pour une banque de donnée comprenant 25 complexes protéines-ARN et 20 chaînes ribosomiales.[136] Dans notre travail nous avons montré que les propensions élevées des acides aminés chargés positivement sont dues à leur capacité à interagir avec la charge négative des phosphates nucléotidiques, ce qui a aussi été montré d'autres études.[136,138] En effet, les ponts salins représentent 8% des contacts dans les complexes avec l'ADN et 7% avec l'ARN. De plus, les ponts salins ont été montrés

comme étant importants pour la pré-orientation des protéines en interaction,[119] pour la stabilisation des complexes et pour la détermination des sites d'interaction avec l'ADN.[209]

Pour approfondir notre analyse, nous avons ensuite observé comment les acides aminés interagissaient spécifiquement avec les bases nucléotidiques. En effet, il a été montré[130,197] que les liens H impliquant les bases nucléotidiques sont très souvent retrouvés dans ce type d'interfaces. Nos résultats montrent que pour les interactions avec l'ADN, l'arginine a une propension encore plus élevée à interagir avec les bases qu'avec le reste du nucléotide (2,6 au lieu de 1,9) alors qu'en interaction avec l'ARN, l'arginine interagit préférentiellement avec les groupements phosphates. Par contre, la lysine interagit préférentiellement avec les phosphates dans les interfaces avec l'ADN et l'ARN. De plus, comme décrit dans d'autres travaux,[130,285] les couples Arg-G sont largement favorisés dans les complexes protéine-ADN ce qui est explicable par la présence de deux sites accepteurs d'hydrogène sur le sillon majeur de la guanine. Dans les complexes protéine-ARN, les couples Arg-G sont également favorisés même si le nombre de liens H avec les bases y est deux fois moins élevé.

L'importance de l'arginine dans les interactions protéine-acide nucléique est donc due à sa capacité à interagir à la fois avec les groupements phosphate par le biais de ponts salins et avec les bases nucléotidiques par le biais de liens H spécifiques. Les interactions de l'arginine avec les nucléotides sont représentées par 25,4% (ARN) à 32,0% (ADN) de ponts salins et par 14,4% (ARN) à 20,9% (ADN) de liens H avec les bases. Ceci correspond à un rapport entre pont salins et liens H avec les bases nucléotidiques de 1,7 dans les interfaces protéine-ARN et de 1,5 dans les interfaces protéine-ADN.

De leur côté, les acides aminés polaires, spécialement l'asparagine (2,2) et l'histidine (2,1), interagissent plus fréquemment avec les bases. Les couples impliquant les acides aminés polaires sont favorisés et environ 60% de ces interactions correspondent à des liens H. Dans les complexes avec l'ADN, nous avons mis en évidence les couples suivants : His-G > Asn-C > Asn-T > Ser-T > Thr-T > Thr-C > His-T > Thr-G > Asn-G > Ser-G. Dans une autre étude,[286] seulement 5 de ces dix couples ont été détectés comme étant favorisés. Pour les complexes avec l'ARN, les couples favorisé contenant des résidus polaires sont classés comme suit : Asn-U > His-G > His-U.

Comme nous l'avons expliqué dans la section précédente, les acides aminés aromatiques sont favorisés dans tous les types d'interfaces protéiques. Parmi les acides aminés aromatiques, la tyrosine est particulièrement favorisée dans les complexes protéine-acide nucléique et elle y adopte un comportement semblable à celui d'un acide aminé polaire.

En effet, nous avons montré que les couples Tyr-G et Tyr-T (ADN) ainsi que le couple Tyr-U (ARN) sont favorisés et forment principalement des liens H. La forte proportion de liens H impliquant la tyrosine a aussi été montrée dans les interfaces protéine-protéine et notamment pour les couples Tyr-Asp et Tyr-Glu.

Finalement, l'acide aspartique a une propension à interagir avec l'ARN de 1,2 et, si on ne prend en compte que les interactions avec les bases, cette valeur augmente à 2,2. Le couple Asp-G est le principal responsable de cette propension élevée et est largement favorisé dans les interactions protéine-ARN. Cette interaction surprenante entre un acide aminé chargé négativement et un nucléotide (chargé également négativement de part la présence des groupements phosphate), est due à la possibilité des deux atomes d'oxygène de la chaîne latérale de créer un lien H avec les sites donneurs d'hydrogène en position 1 et 2 de la guanine (50% des interactions). Ce mécanisme a également été proposé par Jeong *et al.*[133] Ces atomes d'hydrogène sont habituellement impliqués dans des liens H avec la base nucléotidique complémentaire (C), participant de cette manière à la stabilité de la double hélice. Le même mécanisme est utilisé par l'acide glutamique mais, dans notre banque de donnée, l'acide glutamique n'est pas significativement favorisé quand il interagit avec l'ARN. Néanmoins, l'acide glutamique n'est pas défavorisé comme on aurait pu l'imaginer pour un acide aminé chargé négativement et comme il a été observé pour les complexes protéine-ADN.

**Voisinage des acides aminés en interaction**

Les acides aminés en interaction sont influencés par leur voisinage que ce soit au niveau séquentiel ou au niveau structural. Dans cette thèse, nous avons réalisé une étude originale des acides aminés aux positions +/-1 à +/-5 des résidus en interaction et nous avons montré que les résidus voisins significativement favorisés se trouvent aux positions +/-1 à +/-3. Nous avons également montré que les variations de fréquences entre le voisinage et l'interface sont inversement corrélées au volume des acides aminés. Cette corrélation suggère que les acides aminés du voisinage ont un rôle à jouer dans la présentation et l'accessibilité des sites d'interaction. En particulier, la glycine, l'alanine et la sérine, qui sont les 3 acides aminés naturels les moins volumineux, montrent les plus grandes propensions à se trouver en position +/-1. De plus, les doublets d'histidines et de tryptophanes sont régulièrement trouvés aux abords des interfaces.

Peu d'études[120,199,287] sur le voisinage des sites d'interaction ont été publiées dans la littérature. Dans deux de ces études,[120,287] les résultats montrent une présence favorisée de

cystéines aux positions +/-4 et +/-6 des sites d'interaction alors que nous le faisons aux positions +/-1 à +/-3. Cette différence provient probablement de différences dans la définition des sites d'interaction. En effet, Kini et ses collaborateurs[120,287] ont utilisé de très courts fragments de séquences nommés 'Minimum Recognition Sites' (MRS) alors que nous avons utilisé tous les acides aminés en contact direct avec un acide aminé d'une autre chaîne. Par contre, les prolines sont également décrites[120,287] comme jouant un rôle structural aux abords des sites d'interaction (positions +/-1 et +/-2) ce qui n'est pas confirmé dans notre analyse. Nous ne montrons en effet pas de propension particulière pour la proline que ce soit aux abords des sites d'interaction ou même à l'intérieur de ceux-ci. Finalement, nos résultats sont en parfait accord avec ceux de Terribilini *et al*.[199] Lors de la mise au point de leur méthode de prédiction des sites de liaison à l'ARN en 2006, ils ont montré que la glycine et, dans une moindre mesure, la sérine étaient particulièrement favorisées en position +/-1 des résidus en interaction.

**Influence de la structure des acides nucléiques**

Si on étudie les interactions du point de vue des nucléotides, on observe que la distribution des types d'atomes impliqués est différente selon que les protéines interagissent avec de l'ADN ou de l'ARN. Dans les complexes protéine-ADN, 47% des interactions font intervenir les atomes du phosphate et seulement 24% impliquent les atomes de la base. Ces résultats sont semblables à ceux obtenus par Nadassy *et al.*[138] dans un travail basé sur 65 complexes d'ADN en double hélice (43% avec les phosphates et 27% avec les bases). Dans les complexes protéine-ARN, 22% des contacts font intervenir les atomes du phosphate, 43% les atomes du ribose et 36% les atomes de la base. La nature dynamique de l'ARN et les structures diverses qu'il peut prendre,[288] ainsi que la composition de notre banque de données (dans laquelle seulement 15 des 49 structures contiennent plus d'une chaîne d'ARN alors que 118 des 139 complexes avec l'ADN contiennent des doubles hélices d'ADN) peuvent expliquer le fait qu'il y ait 1,5 fois plus de contacts avec les atomes de la base dans les complexes avec l'ARN que dans ceux avec l'ADN.

Nous nous sommes également intéressés à l'influence de la conformation de la double hélice d'ADN. En effet, il a été montré que certaines protéines liant l'ADN interagissent avec des zones d'ADN « courbé » ('bent DNA').[147,149] La déformabilité de l'ADN joue clairement un rôle dans la reconnaissance pour de nombreux complexes[145,146] et est connue sous le nom de reconnaissance indirecte. Nous avons donc décidé d'analyser cette déformabilité en

étudiant l'influence de la structure de l'ADN sur les interactions. Dickerson & Ng[157] et Varagson *et al.*[159] discutent dans leurs travaux des transitions de l'ADN double brins de type B vers un ADN double brins de type A et suggèrent que le type d'hélice influence les interactions avec les protéines. Pour étudier de tels effets, nous avons classé chaque nucléotide d'une banque de données haute résolution d'ADN en double hélice selon leur type (ADN de type A, B et Z) sur base de leurs angles de torsion $\chi$ et $\delta$.[235]

Malgré une très importante proportion d'ADN de type B dans notre sous-banque de complexes, nos résultats suggèrent que l'ADN de type A serait plus fréquent dans les sites d'interaction protéine-ADN que l'ADN classique de type B. Ceci est en accord avec une étude de structure réalisée par Tolstorukov *et al.*[142] et avec les résultats expérimentaux obtenus par Elrod-Erickson *et al.*[164] Ces derniers ont montré que les nucléotides du complexe en doigt de zinc Zif268 adoptent une conformation intermédiaire entre l'ADN de type A et de type B au niveau du site de liaison et, que la mise en présence d'un ADN en conformation de type B conduit à une liaison moins efficace.[164] Dans nos résultats, la distribution des acides aminés en interaction avec l'ADN de type B est fort proche de celle des interactions avec l'ADN total. Par contre, les acides aminés en interaction avec l'ADN de type A suivent une distribution plus proche des résidus en interaction avec l'ARN. En particulier, peu d'acides aminés chargés positivement sont impliqués et l'acide aspartique est favorisé dans les sites d'interaction. Ceci est en corrélation avec les résultats montrant une plus grande accessibilité des bases nucléotidiques de l'ADN quand les acides aminés interagissent avec des nucléotides de type A.[142]

## *IV.2. Prédiction des sites d'interaction*

Dans la première partie de cette thèse, nous avons récolté un grand nombre d'informations sur les caractéristiques des sites d'interaction. Une partie de ces informations a ensuite été utilisée pour mettre au point une méthode de prédiction des sites d'interaction protéiques. En effet, suite aux séquençages de génomes de différents organismes,[18] un intérêt nouveau est porté à la prédiction de ces sites afin d'aider à la détermination des fonctions des protéines codées par ces gènes. Dans cette thèse, nous nous sommes particulièrement intéressés à la prédiction des sites d'interaction en se basant uniquement sur la séquence des protéines. D'autres travaux (cf. Tableau I-1 et Tableau I-2) se sont concentrés sur la mise au point de méthodes de prédiction utilisant les structures 3D et/ou l'homologie de séquence. Les résultats obtenus sont satisfaisants avec des optima de 58%[186] de spécificité et même jusqu'à 72%[31] en calculant les résultats sur base des acides aminés de surface uniquement. Les résultats sont encore meilleurs dans le cas des complexes protéine-acide nucléique (jusque ~80%).[206] Néanmoins, l'intérêt pour une méthode de prédiction efficace en l'absence de structure 3D, ni même de séquences homologues, reste du premier ordre. En effet, pour un grand nombre de protéines, seule la séquence de la protéine est connue et les méthodes citées ci-dessus sont dès-lors inutilisables. La méthode mise au point dans cette thèse est discutée ci-dessous.

Pour mettre au point notre méthode de prédiction, nous avons utilisé un modèle construit par régression logistique. Comparé aux systèmes d'intelligence artificielle tels les réseaux neuronaux (NN) et les machines à vecteurs de support (SVM), ce type de modèle a le principal avantage de conserver une transparence par rapport aux variables choisies et à leur importance respective.[289] Les variables d'entrée pour la construction du modèle ont été présentées au point III.3.1 et 11 des 139 variables ont été retenues dans le modèle final. Les variables choisies dans le modèle final couvrent plusieurs caractéristiques importantes des acides aminés aux interfaces : la capacité d'interagir des acides aminés, des variables liées à leur accessibilité et à leur structure secondaire, leur position dans la séquence et leur situation dans un domaine fonctionnel connu.

Dans la plupart des méthodes de prédiction existantes, les profils de séquences et la surface accessible sont utilisés. Étant donné que nous nous sommes basés uniquement sur la séquence, ces variables n'étaient pas utilisables directement mais il n'est pas étonnant de voir

la prédiction de surface accessible faire partie des variables retenues dans notre modèle. Ce type de prédiction a déjà été utilisé[208] et parfois, les prédictions de structures secondaires ont aussi été utilisés[182,208,290] tout comme dans notre modèle. Certaines propensions des acides aminés à se trouver aux interfaces ont aussi été retenues. Ceci n'est pas étonnant car les propensions sont des variables directement liées aux sites d'interaction et ce type de propension a déjà été largement utilisé.[182,183,190,197,206,209,283] La propension de doublets en interaction a elle aussi été retenue et apporte un degré de discrimination supplémentaire à la prédiction.[198,206] Parmi les variables retenues, deux ont été calculées sur une fenêtre de 9 résidus et une sur une fenêtre de 5 résidus. Les fenêtres de 9 résidus ont déjà été montrées comme idéales pour étudier les interactions car elles seraient composées d'au moins 5 résidus en interaction.[180,202] D'autres travaux ce sont basés sur l'utilisation de fenêtre de 5[199] ou 11[201] acides aminés. Finalement, les 'Receptor Binding Domain' et trois serveurs de recherche de motifs/domaines protéiques ont été retenus. Ces variables significatives sont spécifiques à notre méthode de prédiction et permettent d'inclure une première prédiction des sites d'interaction par le biais des RBD[179] ainsi que la détection de motifs/domaines fonctionnels connus par le biais des serveurs Prints,[291] Smart[292] et Prosite.[293]

Le nombre de variables retenues peut paraître faible par rapport à la quantité de variables fournies en entrée au système de régression logistique (11/139). Néanmoins, de nombreuses méthodes publiées dans la littérature utilisent également un nombre réduit de variables (3 seulement pour certaines méthodes).[31,188,189,201] De plus, en utilisant un nombre réduit de variables, on limite les risques de sur-apprentissage. En effet, comme expliqué au point III.3.4, le problème de la prédiction des sites d'interaction est loin d'être simple et les modèles statistiques ont tendance à « connaître par cœur » l'échantillon fourni. Cette « connaissance » de l'échantillon est rendue possible par la combinaison d'un nombre de variables suffisamment élevé que pour décrire de manière unique, et non globale, les sites d'interaction. En utilisant un nombre réduit de variables hautement significatives, on augmente donc la robustesse du modèle.

Notre méthode de prédiction a été validée sur un échantillon de séquences totalement indépendantes de l'échantillon ayant permis de mettre au point le modèle. A notre connaissance, trois groupe de travail ont tenté de prédire les sites d'interaction protéine-protéine en se basant uniquement sur la séquence : Kini & Evans en 1996,[178] Gallet *et al*. en 2000[179] et Ofran & Rost en 2003.[180] La qualité des deux premières méthodes, bien qu'estimée dans les articles de référence, est difficilement comparable à la nôtre. En effet, les données

fournies par les auteurs sont partielles (pas d'indication sur les faux positifs p.ex.) et il est donc impossible de calculer certains paramètres (spécificité p.ex.). Nous avons donc comparé nos résultats à ceux obtenus par Ofran & Rost[180]sur base de l'utilisation de complexes hétéromer-transitoires. Le Tableau IV-2 résume cette comparaison :

| | | Notre méthode | Aléatoire (écart-type) | Ofran & Rost[180] | Aléatoire |
|---|---|---|---|---|---|
| 1 | Spécificité (%) | 25,1 | 17,5 (+/-0,9) | 20 | 14 |
| | Sensibilité (%) | 15,1 | 10,0 (+/-0,5) | 30 | 22 |
| 2 | Spécificité (%) | 51,1 | 36,7 (+/-1,0) | 42 | 38 |
| | Sensibilité (%) | 14,7 | 10,0 (+/-0,5) | 30 | 22 |

**Tableau IV-2 : Comparaison des résultats des prédictions faites par notre méthode et celle de Ofran & Rost[180] ; 1 : prédictions sur les résidus en interaction/valeurs minimales ; 2 : prédictions sur les zones de résidus en interaction/valeurs maximales.**

Premièrement, il est important de rappeler que la qualité de nos résultats, comme ceux de Ofran & Rost,[180] est très probablement sous-estimée. En effet, une protéine interagit régulièrement avec plusieurs partenaires et les complexes utilisés dans ce travail ne contiennent parfois qu'une partie des partenaires des protéines étudiées. Ceci est d'autant plus vrai que, dans certains cas, les protéines mettent en place des systèmes de coopération dont les différentes interactions peuvent être séparées dans le temps. Il est donc fort probable que certains sites prédits et considérés comme des faux positifs correspondent en fait à des interactions non encore identifiées et/ou non-reprises dans la structure étudiée. Il est néanmoins très intéressant de comparer les résultats obtenus par les deux méthodes. Les résultats obtenus sont significativement différents de résultats aléatoires et ceux obtenus par notre méthode sont plus spécifiques que ceux obtenus par Ofran & Rost[180] (51.1% au lieu de 42%). Néanmoins, la sensibilité y est plus faible (~15% au lieu de 30%) mais le pourcentage de résidus prédits par notre méthode l'est aussi (10% au lieu de 22%). Il est important de signaler que le pourcentage de résidus en interactions retrouvés dans notre banque de données influence grandement la qualité du modèle et c'est pourquoi nous avons calculés nos résultats sur base des zones de résidus en interaction (deuxième résultats [2] dans le tableau Tableau IV-2). Ce problème a déjà été rapporté par d'autres auteurs[186,188] mais est difficilement évitable car il correspond à une propriété intrinsèque des interfaces protéiques.

A ce stade, il est important de rappeler que nous avons expressément demandé à notre méthode d'avoir une spécificité élevée préférentiellement à sa sensibilité. Au lieu d'essayer de détecter l'ensemble des acides aminés en interaction (sensibilité), il nous a semblé préférable d'avoir une grande certitude quand à la fiabilité des prédictions faites. En effet, une spécificité élevée signifie que les résidus prédits sont réellement des acides aminés en interaction.

Dans leur travail, Ofran & Rost[180] ont utilisé uniquement des interfaces hétéromer-transitoires ce qui a facilité la mise au point de leur méthode. En effet, leur base de données contient des interfaces moins diversifiées en comparaison à nos échantillons reprenant tous les types d'interfaces. La sélection d'un seul type d'interfaces simplifie le problème à résoudre par le système statistique mais limite également le champ d'application de leur méthode.

Pour démontrer la fiabilité de notre méthode, quelques exemples de prédictions faites sur les protéines de notre échantillon de validation ont été décrits au point III.3.6. Ces exemples démontrent que l'utilisation des prédictions lors de mutagénèses dirigées, aura comme conséquence un gain de temps considérable en vue de déterminer et/ou de moduler la fonction de la protéine d'intérêt.

# V. CONCLUSIONS ET PERSPECTIVES

L'objectif de cette thèse de doctorat était à la fois d'améliorer nos connaissances sur les sites d'interactions par une analyse de structures 3D et de développer une nouvelle méthode de prédiction des sites d'interaction protéiques à partir de la structure primaire. En effet, un nombre de plus en plus grand de complexes protéiques sont disponibles et une mise à jour des connaissances sur les caractéristiques de leurs interfaces est d'actualité. De plus, suite au séquençage de génomes d'organismes de référence, nous sommes entrés à grands pas dans l'ère de la post-génomique et un nombre de plus en plus important de séquences de protéines nécessite une étude fonctionnelle efficace. A cette fin, des méthodes expérimentales à haut débit ont été mises au point et les méthodes informatiques sont devenues un outil de choix. Une méthode de prédiction des sites d'interaction protéiques serait donc d'une grande utilité pour l'étude de ces protéines aux fonctions inconnues.

Nous avons tout d'abord analysé en détail les caractéristiques des interfaces protéine-protéine et protéine-acide nucléique sur base de banques de données de structures 3D. Lors de cette analyse, nous avons mis en évidence l'importance de l'arginine et des résidus aromatiques (phénylalanine, tryptophane et tyrosine) au sein des interfaces. Ces acides aminés sont spécifiques de l'ensemble des interfaces étudiées et ont été décrits expérimentalement comme hot-spots énergétiques (excepté la phénylalanine). La combinaison de l'implication structurale dans les contacts atome-atome de ces résidus et dans l'énergie de liaison leur donne une importance primordiale pour l'efficacité des interactions. Parallèlement, nous avons aussi montré que le voisinage des sites d'interaction est enrichi en acides aminés de petite taille et que la structure des acides nucléiques a une grande influence sur les acides aminés impliqués dans les interactions.

Ensuite, afin d'affiner notre description des complexes protéiques, nous avons étudié les couples de résidus significativement favorisés aux interfaces tout en différenciant quatre types d'interfaces : hétéromer-transitoires, hétéromer-permanentes, homomériques et celles impliquant les acides nucléiques. Cette classification nous a permis de mettre en évidence des caractéristiques plus spécifiques à ces classes d'interfaces comme le taux de ponts disulfures dans les interfaces hétéromer-permanentes, l'importance des contacts hydrophobes dans les complexes homomériques et l'importance des ponts salins et des liens H dans les interfaces nucléotidiques. Il pourrait également être intéressant de mettre en place une nouvelle classification qui se baserait par exemple sur des critères biochimiques. Cette classification par « famille » de complexes (anticorps-antigène, transduction du signal, cycle cellulaire…) est utilisée depuis de longues années mais le problème majeur consiste à appliquer cette

classification à un nombre très élevé de complexes et donc à automatiser un minimum la classification. Par ailleurs, il pourrait également être intéressant de profiter de notre banque de données pour analyser des sous-banques particulières : interfaces subissant des changements de conformation importants, complexes permettants d'étudier les problèmes de compétition, complexes à hautre résolution en vue d'analyser le rôle de l'eau, complexes contenant des modifications post-traductionnelles, etc.

Dans cette thèse, nous avons montré que l'analyse de structures 3D de complexes protéiques permet de récolter des informations pertinentes sur la composition des interfaces. En plus des interactions protéine-protéine et protéine-acide nucléique, un autre type de complexe est impliqué dans les systèmes biologiques : les complexes de protéines membranaires. Les structures 3D de protéines membranaires sont difficiles à obtenir étant donné que celles-ci sont très souvent insolubles dans l'eau et donc difficilement cristallisables. La taille d'une banque de complexes membranaire serait donc réduite[294] mais une nouvelle analyse détaillée de ce type de complexes selon notre méthodologie pourrait permettre une meilleure compréhension de leur fonctionnement.

Tous les résultats de l'analyse de structures et d'autres, prédits directement à partir de la séquence, ont été soumis à un système de régression logistique afin de mettre au point une méthode de prédiction des sites d'interaction. Les variables les plus significatives ont été extraites et couvrent plusieurs caractéristiques importantes des acides aminés aux interfaces : la capacité d'interagir des acides aminés, des variables liées à l'accessibilité et à la structure secondaire, la position des résidus dans la séquence ainsi que leur situation dans un domaine fonctionnel connu. Le modèle final a été optimisé pour des séquences d'une longueur comprise entre 50 et 300 résidus et afin d'obtenir une spécificité maximale. La méthode de prédiction a été validée sur un échantillon indépendant de protéines et la spécificité obtenue (~51%) est significativement meilleure que des résultats aléatoires. De plus, nos prédictions sont globalement meilleures que celles obtenues par la seule autre méthode de prédiction se basant uniquement sur la séquence existant à ce jour et dont la qualité était directement comparable à la nôtre.

Des améliorations pourront être amenées afin d'augmenter l'efficacité de la méthode et plusieurs voies sont envisageables. Premièrement, dans les cas où cela est possible, la méthode créée pourrait être combinée à d'autres qui utilisent les informations d'homologie de séquence et/ou de regroupement phylogénétique ou, notre méthode pourrait être adaptée pour tenir compte de ces informations quand elles sont disponibles. Il serait également possible

d'utiliser des outils de prédiction des sites de modifications post-traductionnelles[295] afin d'éliminer certains faux positifs. Deuxièmement, au lieu d'améliorer une méthode globale, efficace pour tout type de protéine, nous pourrions également créer plusieurs méthodes qui seraient spécifiques à un type de protéine particulier. Troisièmement, comme nous l'avons montré dans l'introduction de cette thèse (cf. point I.5), les méthodes d'intelligence artificielle sont de plus en plus utilisées. Dans ce travail, nous avons utilisé une méthode qui laisse une certaine « transparence » aux résultats obtenus et qui nous a permis de mettre en évidence les caractéristiques les mieux corrélées aux interfaces. Cette connaissance va nous permettre de gagner du temps et d'augmenter notre efficacité lors de la mise en place d'un réseau neuronal pour lequel des tests préliminaires sont en cours.

Les résultats obtenus par notre méthode de prédiction sont très encourageants et, bien que la méthode puisse encore être améliorée, la prédiction des acides aminés en interaction sur base de la séquence des protéines est d'ores et déjà possible.

Comme signalé dans l'introduction de cette thèse (point I.4.1), la compréhension des interactions entre protéines et la prédiction des sites impliqués ouvrent de nombreuses voies dont deux principales qui dépendent l'une de l'autre : la compréhension de la biologie cellulaire et la production de nouveaux médicaments.

L'analyse réalisée ici a permis d'améliorer nos connaissances sur les caractéristiques essentielles des interfaces. Cette connaissance doit nous aider à mieux appréhender les phénomènes physiologiques se déroulant dans la cellule. De plus, la méthode de prédiction mise au point doit nous aider à détecter plus rapidement les cascades d'interactions nécessaires au bon fonctionnement de la cellule et les interactions susceptibles d'être impliquées dans certaines maladies.[296] En effet, la prédiction des sites d'interactions permet de guider efficacement les méthodes expérimentales d'analyse des interactions protéiques (travaux de mutagénèse dirigée p.ex.).

D'un point de vue pratique, la production de médicaments sera facilitée par une identification plus aisée des cibles pharmaceutiques potentielles. Une fois les cibles pharmaceutiques détectées avec précision grâce à l'identification des acides aminés de l'interface, le design de médicaments modulant les interactions et/ou la mise au point de molécules imitant le partenaire de l'interaction sera simplifiée.

# VI. RÉFÉRENCES BIBLIOGRAPHIQUES

## *VI.1. Publications personnelles*

Une grande partie des résultats présentés dans le chapitre sur les interactions entre protéines et acides nucléiques (III.1) ont été publiés dans : Delsaux,N., Lejeune,D., Charloteaux,B., Thomas,A. & Brasseur,R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins* **61**, 258-271 (2005).[297]

La présentation des résultats sur les interactions entre protéines sont en cours de rédaction pour soumission. Protein-protein recognition: a statistical analysis of atomic interactions. *En préparation*.

Durant cette thèse, un article a été réalisé en collaboration avec le professeur P. Talmud du département de médecine, division de génétique cardiovasulaire, UCL de Londres : *In vitro* effects on lipoprotein lipase activity of *ApoA5* mutations associated with severe hypertriglyceridemia. Cet article a été soumis à 'Journal of Biological Chemistry'.

## *VI.2. Publications citées*


1. Karp,G. Biologie Cellulaire Moléculaire : concepts et expériences., pp. 30-78 (De Boeck Université, Paris, Bruxelles,1998).

2. Creighton,T.E. PROTEINS : Structures and Molecular Properties. W. H. Freeman and Company, (1997).

3. Edison,A.S. Linus Pauling and the planar peptide bond. *Nat. Struct. Biol.* **8**, 201-202 (2001).

4. Chou,K.C. Prediction of Tight Turns and Their Types in Proteins. *Anal. Biochem.* **286**, 1-16 (2000).

5. Blackburn & Gait. Nucleic acids in chemistry and biology. Oxford University Press, New York (1996).

6. WATSON,J.D. & CRICK,F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).

7. Herbert,A. & Rich,A. The biology of left-handed Z-DNA. *J. Biol. Chem.* **271**, 11595-11598 (1996).

8. Olson,W.K. *et al.* A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* **313**, 229-237 (2001).

9. Dickerson,R.E. Definitions and nomenclature of nucleic acid structure components. *Nucleic Acids Res.* **17**, 1797-1803 (1989).

10. Lafontaine,I. & Lavery,R. High-speed molecular mechanics searches for optimal DNA interaction sites. *Comb. Chem. High Throughput. Screen.* **4**, 707-717 (2001).

11. Lu,X.J. & Olson,W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31**, 5108-5121 (2003).

12. Nagaswamy,U., Voss,N., Zhang,Z. & Fox,G.E. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res.* **28**, 375-376 (2000).

13. Mandel-Gutfreund,Y., Margalit,H., Jernigan,R.L. & Zhurkin,V.B. A role for CH...O interactions in protein-DNA recognition. *J. Mol. Biol.* **277**, 1129-1140 (1998).

14. Thomas,A., Benhabiles,N., Meurisse,R., Ngwabije,R. & Brasseur,R. Pex, analytical tools for PDB files. II. H-Pex: noncanonical H-bonds in alpha-helices. *Proteins* **43**, 37-44 (2001).

15. Brasseur,R. Differentiation of lipid-associating helices by use of three-dimensional molecular hydrophobicity potential calculations. *J. Biol. Chem.* **266**, 16120-16127 (1991).

16. Fauchere,J.L., Quarendon,P. & Kaetterer,L. Estimating and representing hydrophobicity potential. *J. Mol. Graph.* **6**, 203-206 (1988).

17. Bock,J.R. & Gough,D.A. Predicting protein--protein interactions from primary structure. *Bioinformatics.* **17**, 455-460 (2001).

18. Binnewies,T.T. *et al.* Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct. Integr. Genomics* **6**, 165-185 (2006).

19. The yeast genome directory. *Nature* **387**, 5 (1997).

20. Goffeau,A. 1996: a vintage year for yeast and Yeast. *Yeast* **12**, 1603-1605 (1996).

21. Goffeau,A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563-567 (1996).

22. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).

23. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).

24. Lander,E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

25. Venter,J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).

26. Liolios,K., Tavernarakis,N., Hugenholtz,P. & Kyrpides,N.C. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.* **34**, D332-D334 (2006).

27. Stites,W.E. Protein-Protein Interactions: Interface Structure, Binding Thermodynamics, and Mutational Analysis. *Chem. Rev.* **97**, 1233-1250 (1997).

28. Golemis,E.A., Tew,K.D. & Dadke,D. Protein interaction-targeted drug discovery: evaluating critical issues. *Biotechniques* **32**, 636-638 (2002).

29. Svedberg,T. & Fahraeus,R. A new method for the determination of the molecular weight of proteins. *J. Am. Chem. Soc.* **48**, 430-438 (1926).

30. Svedberg,T. Mass and size of protein molecules. *Nature* **123**, 871 (1929).

31. Fariselli,P., Pazos,F., Valencia,A. & Casadio,R. Prediction of protein--protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356-1361 (2002).

32. Cusick,M.E., Klitgord,N., Vidal,M. & Hill,D.E. Interactome: gateway into systems biology. *Hum. Mol. Genet.* **14 Spec No. 2**, R171-R181 (2005).

33. Hoffman,M.M. *et al.* AANT: the Amino Acid-Nucleotide Interaction Database. *Nucleic Acids Res.* **32 Database issue**, D174-D181 (2004).

34. Prabakaran,P. *et al.* Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics.* **17**, 1027-1034 (2001).

35. Kumar,M.D. *et al.* ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* **34**, D204-D206 (2006).

36. Gromiha,M.M. *et al.* ProTherm: Thermodynamic Database for Proteins and Mutants. *Nucleic Acids Res.* **27**, 286-288 (1999).

37. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. & Gilson,M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **35**, D198-D201 (2007).

38. Fischer,T.B. *et al.* The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics.* **19**, 1453-1454 (2003).

39. Chiu,W.L., Sze,C.N., Ip,L.N., Chan,S.K. & Au-Yeung,S.C. NTDB: Thermodynamic Database for Nucleic Acids. *Nucleic Acids Res.* **29**, 230-233 (2001).

40. Spirin,S., Titov,M., Karyagina,A. & Alexeevski,A. NPIDB, a Database of Nucleic Acids Protein Interactions. *Bioinformatics.* **23**, 3247-3248 (2007).

41. Thorn,K.S. & Bogan,A.A. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics.* **17**, 284-285 (2001).

42. Fields,S. & Song,O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-246 (1989).

43. Marcotte,E.M., Xenarios,I. & Eisenberg,D. Mining literature for protein-protein interactions. *Bioinformatics.* **17**, 359-363 (2001).

44. Xenarios,I. & Eisenberg,D. Protein interaction databases. *Curr. Opin. Biotechnol.* **12**, 334-339 (2001).

45. Villalobos,V., Naik,S. & Piwnica-Worms,D. Current state of imaging protein-protein interactions in vivo with genetically encoded reporters. *Annu. Rev. Biomed. Eng* **9**, 321-349 (2007).

46. Salwinski,L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32**, D449-D451 (2004).

47. Xenarios,I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289-291 (2000).

48. Mewes,H.W., Hani,J., Pfeiffer,F. & Frishman,D. MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.* **26**, 33-37 (1998).

49. Mewes,H.W. *et al.* MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **34**, D169-D172 (2006).

50. Chatr-aryamontri,A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572-D574 (2007).

51. Peri,S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363-2371 (2003).

52. Kerrien,S. *et al.* IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561-D565 (2007).

53. Hermjakob,H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452-D455 (2004).

54. von Mering,C. *et al.* STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35**, D358-D362 (2007).

55. Stark,C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535-D539 (2006).

56. Edwards,A.M. *et al.* Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529-536 (2002).

57. Leach,S., Gabow,A., Hunter,L. & Goldberg,D.S. Assessing and combining reliability of protein interaction sources. *Pac. Symp. Biocomput.* 433-444 (2007).

58. Uetz,P. *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).

59. Ito,T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A* **98**, 4569-4574 (2001).

60. Jeong,H., Mason,S.P., Barabasi,A.L. & Oltvai,Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41-42 (2001).

61. Greener,M. From Worm to Fly, Y2H takes on. *The Scientist* **19**, 18-19 (2005).

62. Giot,L. *et al.* A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-1736 (2003).

63. Li,S. *et al.* A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-543 (2004).

64. Rual,J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).

65. Date,S.V. & Stoeckert,C.J., Jr. Computational modeling of the Plasmodium falciparum interactome reveals protein function on a genome-wide scale. *Genome Res.* **16**, 542-549 (2006).

66. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. & Ouzounis,C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90 (1999).

67. Pazos,F., Helmer-Citterich,M., Ausiello,G. & Valencia,A. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271**, 511-523 (1997).

68. Dohkan,S., Koike,A. & Takagi,T. Support Vector Machines for Predicting Protein-Protein Interactions. *Genome Informatics* **14**, 502-503 (2003).

69. Gomez,S.M., Noble,W.S. & Rzhetsky,A. Learning to predict protein-protein interactions from protein sequences. *Bioinformatics.* **19**, 1875-1881 (2003).

70. Martin,S., Roe,D. & Faulon,J.L. Predicting protein-protein interactions using signature products. *Bioinformatics.* **21**, 218-226 (2005).

71. Wu,X. *et al.* SPIDer: Saccharomyces protein-protein interaction database. *BMC. Bioinformatics.* **7 Suppl 5**, S16 (2006).

72. Pitre,S. *et al.* PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC. Bioinformatics.* **7**, 365 (2006).

73. Roberts,P.M. Mining literature for systems biology. *Brief. Bioinform.* **7**, 399-406 (2006).

74. Chothia,C. & Janin,J. Principles of protein-protein recognition. *Nature* **256**, 705-708 (1975).

75. Li,H., Li,J., Tan,S.H. & Ng,S.K. Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data. *Pac. Symp. Biocomput.* 312-323 (2004).

76. Aytuna,A.S., Gursoy,A. & Keskin,O. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics.* **21**, 2850-2855 (2005).

77. Grabowski,M., Joachimiak,A., Otwinowski,Z. & Minor,W. Structural genomics: keeping up with expanding knowledge of the protein universe. *Curr. Opin. Struct. Biol.* **17**, 347-353 (2007).

78. Jones,S. & Thornton,J.M. Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31-65 (1995).

79. Larsen,T.A., Olson,A.J. & Goodsell,D.S. Morphology of protein-protein interfaces. *Structure.* **6**, 421-427 (1998).

80. Janin,J. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure. Fold. Des* **7**, R277-R279 (1999).

81. Rodier,F., Bahadur,R.P., Chakrabarti,P. & Janin,J. Hydration of protein-protein interfaces. *Proteins* **60**, 36-45 (2005).

82. Fernandez,A. & Scheraga,H.A. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl. Acad. Sci. U. S. A* **100**, 113-118 (2003).

83. Reichmann,D., Phillip,Y., Carmi,A. & Schreiber,G. On the contribution of water-mediated interactions to protein-complex stability. *Biochemistry* **47**, 1051-1060 (2008).

84. Sundberg,E.J. & Mariuzza,R.A. Luxury accommodations: the expanding role of structural plasticity in protein-protein interactions. *Structure. Fold. Des* **8**, R137-R142 (2000).

85. Janin,J. & Chothia,C. The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027-16030 (1990).

86. Jones,S. & Thornton,J.M. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A* **93**, 13-20 (1996).

87. Lo,C.L., Chothia,C. & Janin,J. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177-2198 (1999).

88. Robertson,A.D. Intramolecular interactions at protein surfaces and their impact on protein function. *Trends Biochem. Sci.* **27**, 521-526 (2002).

89. Wodak,S.J. & Janin,J. Structural basis of macromolecular recognition. *Adv. Protein Chem.* **61**, 9-73 (2002).

90. Goh,C.S., Milburn,D. & Gerstein,M. Conformational changes associated with protein-protein interactions. *Curr. Opin. Struct. Biol.* **14**, 104-109 (2004).

91. Grunberg,R., Leckner,J. & Nilges,M. Complementarity of structure ensembles in protein-protein binding. *Structure (Camb. )* **12**, 2125-2136 (2004).

92. Ming,D. & Wall,M.E. Interactions in native binding sites cause a large change in protein dynamics. *J. Mol. Biol.* **358**, 213-223 (2006).

93. Heifetz,A. & Eisenstein,M. Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking. *Protein Eng* **16**, 179-185 (2003).

94. Xu,D., Lin,S.L. & Nussinov,R. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol.* **265**, 68-84 (1997).

95. Gunasekaran,K. & Nussinov,R. How different are structurally flexible and rigid binding sites? Sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding. *J. Mol. Biol.* **365**, 257-273 (2007).

96. Nooren,I.M. & Thornton,J.M. Diversity of protein-protein interactions. *EMBO J.* **22**, 3486-3492 (2003).

97. Lu,H., Lu,L. & Skolnick,J. Development of unified statistical potentials describing protein- protein interactions. *Biophys. J.* **84**, 1895-1901 (2003).

98. Jones,S., Marin,A. & Thornton,J.M. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng* **13**, 77-82 (2000).

99. Ofran,Y. & Rost,B. Analysing six types of protein-protein interfaces. *J. Mol. Biol.* **325**, 377-387 (2003).

100. Nooren,I.M. & Thornton,J.M. Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.* **325**, 991-1018 (2003).

101. Chakrabarti,P. & Janin,J. Dissecting protein-protein recognition sites. *Proteins* **47**, 334-343 (2002).

102. Young,L., Jernigan,R.L. & Covell,D.G. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717-729 (1994).

103. Bogan,A.A. & Thorn,K.S. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1-9 (1998).

104. Kortemme,T. & Baker,D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. U. S. A* **99**, 14116-14121 (2002).

105. Clackson,T. & Wells,J.A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383-386 (1995).

106. DeLano,W.L. Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **12**, 14-20 (2002).

107. Ma,B., Elkayam,T., Wolfson,H. & Nussinov,R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. U. S. A* **100**, 5772-5777 (2003).

108. Halperin,I., Wolfson,H. & Nussinov,R. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure (Camb. )* **12**, 1027-1038 (2004).

109. Keskin,O., Ma,B. & Nussinov,R. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.* **345**, 1281-1294 (2005).

110. Li,X., Keskin,O., Ma,B., Nussinov,R. & Liang,J. Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J. Mol. Biol.* **344**, 781-795 (2004).

111. Li,Y., Huang,Y., Swaminathan,C.P., Smith-Gill,S.J. & Mariuzza,R.A. Magnitude of the hydrophobic effect at central versus peripheral sites in protein-protein interfaces. *Structure (Camb. )* **13**, 297-307 (2005).

112. Moreira,I.S., Fernandes,P.A. & Ramos,M.J. Hot spots--a review of the protein-protein interface determinant amino-acid residues. *Proteins* **68**, 803-812 (2007).

113. Bloom,J.D. & Adami,C. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC. Evol. Biol.* **3**, 21 (2003).

114. Caffrey,D.R., Somaroo,S., Hughes,J.D., Mintseris,J. & Huang,E.S. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13**, 190-202 (2004).

115. Reddy,B.V. & Kaznessis,Y.N. A quantitative analysis of interfacial amino acid conservation in protein-protein hetero complexes. *J. Bioinform. Comput. Biol.* **3**, 1137-1150 (2005).

116. Rajamani,D., Thiel,S., Vajda,S. & Camacho,C.J. Anchor residues in protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A* **101**, 11287-11292 (2004).

117. Mihalek,I., Res,I. & Lichtarge,O. On Itinerant Water Molecules and Detectability of Protein-Protein Interfaces through Comparative Analysis of Homologues. *J. Mol. Biol.* **369**, 584-595 (2007).

118. Nicola,G. & Vakser,I.A. A simple shape characteristic of protein-protein recognition. *Bioinformatics.* **23**, 789-792 (2007).

119. Drozdov-Tikhomirov,L.N., Linde,D.M., Poroikov,V.V., Alexandrov,A.A. & Skurida,G.I. Molecular mechanisms of protein-protein recognition: whether the surface placed charged residues determine the recognition process? *J. Biomol. Struct. Dyn.* **19**, 279-284 (2001).

120. Kini,R.M. & Evans,H.J. A hypothetical structural role for proline residues in the flanking segments of protein-protein interaction sites. *Biochem. Biophys. Res. Commun.* **212**, 1115-1124 (1995).

121. Res,I. & Lichtarge,O. Character and evolution of protein-protein interfaces. *Phys. Biol.* **2**, S36-S43 (2005).

122. Keskin,O., Ma,B., Rogale,K., Gunasekaran,K. & Nussinov,R. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys. Biol.* **2**, S24-S35 (2005).

123. Reichmann,D., Rahat,O., Cohen,M., Neuvirth,H. & Schreiber,G. The molecular architecture of protein-protein binding sites. *Curr. Opin. Struct. Biol.* **17**, 67-76 (2007).

124. Luscombe,N.M., Austin,S.E., Berman,H.M. & Thornton,J.M. An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**, 1-10 (2000).

125. Michael,G.M., Siebers,J.G., Selvaraj,S., Kono,H. & Sarai,A. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.* **337**, 285-294 (2004).

126. Paillard,G. & Lavery,R. Analyzing protein-DNA recognition mechanisms. *Structure. (Camb. )* **12**, 113-122 (2004).

127. Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. & Lathrop,R.H. DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics.* **18 Suppl 1**, S22-S30 (2002).

128. Ahmad,S., Kono,H., Arauzo-Bravo,M.J. & Sarai,A. ReadOut: structure-based calculation of direct and indirect readout energies and specificities for protein-DNA recognition. *Nucleic Acids Res.* **34**, W124-W127 (2006).

129. Seeman,N.C., Rosenberg,J.M. & Rich,A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U. S. A* **73**, 804-808 (1976).

130. Luscombe,N.M., Laskowski,R.A. & Thornton,J.M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29**, 2860-2874 (2001).

131. Jones,S., van Heyningen,P., Berman,H.M. & Thornton,J.M. Protein-DNA interactions: A structural analysis. *J. Mol. Biol.* **287**, 877-896 (1999).

132. Mandel-Gutfreund,Y., Schueler,O. & Margalit,H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.* **253**, 370-382 (1995).

133. Jeong,E., Kim,H., Lee,S.W. & Han,K. Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes. *Mol. Cells* **16**, 161-167 (2003).

134. Pabo,C.O. & Sauer,R.T. Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053-1095 (1992).

135. Morozova,N., Allers,J., Myers,J. & Shamoo,Y. Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics.* **22**, 2746-2752 (2006).

136. Treger,M. & Westhof,E. Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recognit.* **14**, 199-214 (2001).

137. Jayaram,B. & Jain,T. The role of water in protein-DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 343-361 (2004).

138. Nadassy,K., Wodak,S.J. & Janin,J. Structural features of protein-nucleic acid recognition sites. *Biochemistry* **38**, 1999-2017 (1999).

139. Reddy,C.K., Das,A. & Jayaram,B. Do water molecules mediate protein-DNA recognition? *J. Mol. Biol.* **314**, 619-632 (2001).

140. Coulocheri,S.A., Pigis,D.G., Papavassiliou,K.A. & Papavassiliou,A.G. Hydrogen bonds in protein-DNA complexes: where geometry meets plasticity. *Biochimie* **89**, 1291-1303 (2007).

141. Pabo,C.O. & Sauer,R.T. Protein-DNA recognition. *Annu. Rev. Biochem.* **53**, 293-321 (1984).

142. Tolstorukov,M.Y., Jernigan,R.L. & Zhurkin,V.B. Protein-DNA hydrophobic recognition in the minor groove is facilitated by sugar switching. *J. Mol. Biol.* **337**, 65-76 (2004).

143. Jones,S., Daley,D.T., Luscombe,N.M., Berman,H.M. & Thornton,J.M. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.* **29**, 943-954 (2001).

144. Selvaraj,S., Kono,H. & Sarai,A. Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *J. Mol. Biol.* **322**, 907-915 (2002).

145. Bao,G. Mechanics of biomolecules. *Journal of Mechanics and Physics of Solids* **50**, 2237-2274 (2002).

146. Changela,A., Perry,K., Taneja,B. & Mondragon,A. DNA manipulators: caught in the act. *Curr. Opin. Struct. Biol.* **13**, 15-22 (2003).

147. Dickerson,R.E. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.* **26**, 1906-1926 (1998).

148. Sarai,A. & Kono,H. Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 379-398 (2005).

149. Dickerson,R.E. & Chiu,T.K. Helix bending as a factor in protein/DNA recognition. *Biopolymers* **44**, 361-403 (1998).

150. El Hassan,M.A. & Calladine,C.R. Two distinct modes of protein-induced bending in DNA. *J. Mol. Biol.* **282**, 331-343 (1998).

151. Zhang,Y., Xi,Z., Hegde,R.S., Shakked,Z. & Crothers,D.M. Predicting indirect readout effects in protein-DNA interactions. *Proc. Natl. Acad. Sci. U. S. A* **101**, 8337-8341 (2004).

152. ElHassan,M.A. & Calladine,C.R. Conformational characteristics of DNA: Empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Philosophical Transactions of the Royal Society of London Series A-Mathematical Physical and Engineering Sciences* **355**, 43-100 (1997).

153. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. & Zhurkin,V.B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A* **95**, 11163-11168 (1998).

154. Steffen,N.R. *et al.* The role of DNA deformation energy at individual base steps for the identification of DNA-protein binding sites. *Genome Inform. Ser. Workshop Genome Inform.* **13**, 153-162 (2002).

155. Nekludova,L. & Pabo,C.O. Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc. Natl. Acad. Sci. U. S. A* **91**, 6948-6952 (1994).

156. Wang,A.H., Fujii,S., van Boom,J.H. & Rich,A. Molecular structure of the octamer d(G-G-C-C-G-G-C-C): modified A-DNA. *Proc. Natl. Acad. Sci. U. S. A* **79**, 3968-3972 (1982).

157. Dickerson,R.E. & Ng,H.L. DNA structure from A to B. *Proc. Natl. Acad. Sci. U. S. A* **98**, 6986-6988 (2001).

158. Ng,H.L., Kopka,M.L. & Dickerson,R.E. The structure of a stable intermediate in the A <--> B DNA helix transition. *Proc. Natl. Acad. Sci. U. S. A* **97**, 2035-2039 (2000).

159. Vargason,J.M., Henderson,K. & Ho,P.S. A crystallographic map of the transition from B-DNA to A-DNA. *Proc. Natl. Acad. Sci. U. S. A* **98**, 7265-7270 (2001).

160. Varnai,P., Djuranovic,D., Lavery,R. & Hartmann,B. Alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res.* **30**, 5398-5406 (2002).

161. Garvie,C.W. & Wolberger,C. Recognition of specific DNA sequences. *Mol. Cell* **8**, 937-946 (2001).

162. Prabakaran,P. *et al.* Classification of protein-DNA complexes based on structural descriptors. *Structure.* **14**, 1355-1367 (2006).

163. Beamer,L.J. & Pabo,C.O. Refined 1.8 A crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.* **227**, 177-196 (1992).

164. Elrod-Erickson,M., Rould,M.A., Nekludova,L. & Pabo,C.O. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure.* **4**, 1171-1180 (1996).

165. Parraga,A., Bellsolell,L., Ferre-D'Amare,A.R. & Burley,S.K. Co-crystal structure of sterol regulatory element binding protein 1a at 2.3 A resolution. *Structure.* **6**, 661-672 (1998).

166. Somers,W.S. & Phillips,S.E. Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. *Nature* **359**, 387-393 (1992).

167. DeLano,W.L. The PyMOL molecular graphics system. *Unpublished work* (2002).

168. Draper,D.E. Themes in RNA-protein recognition. *J. Mol. Biol.* **293**, 255-270 (1999).

169. Perez-Canadillas,J.M. & Varani,G. Recent advances in RNA-protein recognition. *Curr. Opin. Struct. Biol.* **11**, 53-58 (2001).

170. Handa,N. *et al.* Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature* **398**, 579-585 (1999).

171. Lewis,H.A. *et al.* Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* **100**, 323-332 (2000).

172. Allain,F.H., Howe,P.W., Neuhaus,D. & Varani,G. Structural basis of the RNA-binding specificity of human U1A protein. *EMBO J.* **16**, 5764-5772 (1997).

173. Hicks,J.M. & Hsu,V.L. The extended left-handed helix: a simple nucleic acid-binding motif. *Proteins* **55**, 316-329 (2004).

174. Shanahan,H.P., Garcia,M.A., Jones,S. & Thornton,J.M. Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.* **32**, 4732-4741 (2004).

175. Yu,X., Cao,J., Cai,Y., Shi,T. & Li,Y. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J. Theor. Biol.* **240**, 175-184 (2006).

176. Privalov,P.L. *et al.* What drives proteins into the major or minor grooves of DNA? *J. Mol. Biol.* **365**, 1-9 (2007).

177. Luscombe,N.M., Laskowski,R.A. & Thornton,J.M. NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. *Nucleic Acids Res.* **25**, 4940-4945 (1997).

178. Kini,R.M. & Evans,H.J. Prediction of potential protein-protein interaction sites from amino acid sequence. Identification of a fibrin polymerization site. *FEBS Lett.* **385**, 81-86 (1996).

179. Gallet,X., Charloteaux,B., Thomas,A. & Brasseur,R. A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* **302**, 917-926 (2000).

180. Ofran,Y. & Rost,B. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.* **544**, 236-239 (2003).

181. Koike,A. & Takagi,T. Prediction of Protein Interaction Sites and Protein-Protein Interaction Pairs Using Support Vector Machines. *Genome Informatics* **14**, 500-501 (2003).

182. Hoskins,J., Lovell,S. & Blundell,T.L. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci.* **15**, 1017-1029 (2006).

183. Murakami,Y. & Jones,S. SHARP2: protein-protein interaction predictions using patch analysis. *Bioinformatics.* **22**, 1794-1795 (2006).

184. Kufareva,I., Budagyan,L., Raush,E., Totrov,M. & Abagyan,R. PIER: protein interface recognition for structural proteomics. *Proteins* **67**, 400-417 (2007).

185. Porollo,A. & Meller,J. Prediction-based fingerprints of protein-protein interactions. *Proteins* **66**, 630-645 (2007).

186. Koike,A. & Takagi,T. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel* **17**, 165-173 (2004).

187. Res,I., Mihalek,I. & Lichtarge,O. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics.* **21**, 2496-2501 (2005).

188. Chen,H. & Zhou,H.X. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins* **61**, 21-35 (2005).

189. Zhou,H.X. & Shan,Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins* **44**, 336-343 (2001).

190. Bradford,J.R. & Westhead,D.R. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics.* **21**, 1487-1494 (2005).

191. Bordner,A.J. & Abagyan,R. Statistical analysis and prediction of protein-protein interfaces. *Proteins* **60**, 353-366 (2005).

192. Chung,J.L., Wang,W. & Bourne,P.E. Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins* **62**, 630-640 (2006).

193. Wang,B. *et al.* Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett.* **580**, 380-384 (2006).

194. Wang,B., Wong,H.S. & Huang,D.S. Inferring protein-protein interacting sites using residue conservation and evolutionary information. *Protein Pept. Lett.* **13**, 999-1005 (2006).

195. Dong,Q., Wang,X., Lin,L. & Guan,Y. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC. Bioinformatics.* **8**, 147 (2007).

196. Li,M.H., Lin,L., Wang,X.L. & Liu,T. Protein-protein interaction site prediction based on conditional random fields. *Bioinformatics.* **23**, 597-604 (2007).

197. Mandel-Gutfreund,Y. & Margalit,H. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Res.* **26**, 2306-2312 (1998).

198. Ahmad,S., Gromiha,M.M. & Sarai,A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics.* **20**, 477-486 (2004).

199. Terribilini,M. *et al.* Prediction of RNA binding sites in proteins from amino acid sequence. *RNA.* **12**, 1450-1462 (2006).

200. Kuznetsov,I.B., Gou,Z., Li,R. & Hwang,S. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins* **64**, 19-27 (2006).

201. Wang,L. & Brown,S.J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.* **34**, W243-W248 (2006).

202. Yan,C. *et al.* Predicting DNA-binding sites of proteins from amino acid sequence. *BMC. Bioinformatics.* **7**, 262 (2006).

203. Jeong,E., Chung,I.F. & Miyano,S. A neural network method for identification of RNA-interacting residues in protein. *Genome Inform.* **15**, 105-116 (2004).

204. Bhardwaj,N. & Lu,H. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.* **581**, 1058-1066 (2007).

205. Ahmad,S. & Sarai,A. PSSM-based prediction of DNA binding sites in proteins. *BMC. Bioinformatics.* **6**, 33 (2005).

206. Kim,O.T., Yura,K. & Go,N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **34**, 6450-6460 (2006).

207. Ho,S.Y., Yu,F.C., Chang,C.Y. & Huang,H.L. Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method. *Biosystems* **90**, 234-241 (2007).

208. Ofran,Y., Mysore,V. & Rost,B. Prediction of DNA-binding residues from sequence. *Bioinformatics.* **23**, i347-i353 (2007).

209. Jones,S., Shanahan,H.P., Berman,H.M. & Thornton,J.M. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* **31**, 7189-7198 (2003).

210. Tjong,H. & Zhou,H.X. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.* **35**, 1465-1477 (2007).

211. Eisenberg,D., Weiss,R.M. & Terwilliger,T.C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **299**, 371-374 (1982).

212. De Loof,H., Rosseneu,M., Brasseur,R. & Ruysschaert,J.M. Use of hydrophobicity profiles to predict receptor binding domains on apolipoprotein E and the low density lipoprotein apolipoprotein B-E receptor. *Proc. Natl. Acad. Sci. U. S. A* **83**, 2295-2299 (1986).

213. Vapnik,V.N. The nature of statistical learning theory. Springer, New York (1995).

214. Berman,H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).

215. Thomas,A., Bouffioux,O., Geeurickx,D. & Brasseur,R. Pex, analytical tools for PDB files. I. GF-Pex: basic file to describe a protein. *Proteins* **43**, 28-36 (2001).

216. Wang,G. & Dunbrack,R.L., Jr. PISCES: a protein sequence culling server. *Bioinformatics.* **19**, 1589-1591 (2003).

217. Henrick,K. & Thornton,J.M. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358-361 (1998).

218. Word,J.M., Lovell,S.C., Richardson,J.S. & Richardson,D.C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735-1747 (1999).

219. Srinivasan,R. & Rose,G.D. LINUS: A Hierarchic Procedure to Predict the Fold of a Protein. *Proteins* **22**, 81-99 (1995).

220. Ansari,S. & Helms,V. Statistical analysis of predominantly transient protein-protein interfaces. *Proteins* **61**, 344-355 (2005).

221. Samanta,U., Bahadur,R.P. & Chakrabarti,P. Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng* **15**, 659-667 (2002).

222. Kim,W.K., Henschel,A., Winter,C. & Schroeder,M. The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS. Comput. Biol.* **2**, 1151-1164 (2006).

223. Pang,P.S., Jankowsky,E., Wadley,L.M. & Pyle,A.M. Prediction of functional tertiary interactions and intermolecular interfaces from primary sequence data. *J. Exp. Zoolog. B Mol. Dev. Evol.* **304**, 50-63 (2005).

224. Jiang,L., Gao,Y., Mao,F., Liu,Z. & Lai,L. Potential of mean force for protein-protein interaction studies. *Proteins* **46**, 190-196 (2002).

225. Tsai,C.J., Lin,S.L., Wolfson,H.J. & Nussinov,R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53-64 (1997).

226. Levy,E.D., Pereira-Leal,J.B., Chothia,C. & Teichmann,S.A. 3D complex: a structural classification of protein complexes. *PLoS. Comput. Biol.* **2**, 1395-1406 (2006).

227. Liang,S., Zhang,C., Liu,S. & Zhou,Y. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.* **34**, 3698-3707 (2006).

228. Lee,B. & Richards,F.M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400 (1971).

229. Shrake,A. & Rupley,J.A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351-371 (1973).

230. Lins,L., Thomas,A. & Brasseur,R. Analysis of accessible surface of residues in proteins. *Protein Sci.* **12**, 1406-1417 (2003).

231. Valdar,W.S. & Thornton,J.M. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42**, 108-124 (2001).

232. de Vries,S.J. & Bonvin,A.M. Intramolecular surface contacts contain information about protein-protein interface regions. *Bioinformatics.* **22**, 2094-2098 (2006).

233. Bode,W., Papamokos,E. & Musil,D. The high-resolution X-ray crystal structure of the complex formed between subtilisin Carlsberg and eglin c, an elastase inhibitor from the leech Hirudo medicinalis. Structural analysis, subtilisin structure and interface geometry. *Eur. J. Biochem.* **166**, 673-692 (1987).

234. Dagnelie,P. Théorie et méthodes statistiques. Presses agronomiques, Gembloux (1975).

235. Lu,X.J., Shakked,Z. & Olson,W.K. A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **300**, 819-840 (2000).

236. Marrec-Fairley,M. *et al.* Differential functionalities of amphiphilic peptide segments of the cell-septation penicillin-binding protein 3 of Escherichia coli. *Mol. Microbiol.* **37**, 1019-1031 (2000).

237. Kini,R.M. & Evans,H.J. A novel approach to the design of potent bioactive peptides by incorporation of proline brackets: antiplatelet effects of Arg-Gly-Asp peptides. *FEBS Lett.* **375**, 15-17 (1995).

238. Sundstrom,M. *et al.* Crystal structure of an antagonist mutant of human growth hormone, G120R, in complex with its receptor at 2.9 A resolution. *J. Biol. Chem.* **271**, 32197-32203 (1996).

239. McGuffin,L.J., Bryson,K. & Jones,D.T. The PSIPRED protein structure prediction server. *Bioinformatics.* **16**, 404-405 (2000).

240. Combet,C., Blanchet,C., Geourjon,C. & Deleage,G. NPS@: network protein sequence analysis. *Trends Biochem. Sci.* **25**, 147-150 (2000).

241. Mucchielli-Giorgi,M.H., Hazout,S. & Tuffery,P. PredAcc: prediction of solvent accessibility. *Bioinformatics.* **15**, 176-177 (1999).

242. Mucchielli-Giorgi,M.H., Tuffery,P. & Hazout,S. Prediction of solvent accessibility of amino-acid residues: critical aspects. *Theo. Chem. Abstracts.* **101**, 186-193 (1999).

243. Ahmad,S. & Gromiha,M.M. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics.* **18**, 819-824 (2002).

244. Bullock,T.L., Breddam,K. & Remington,S.J. Peptide aldehyde complexes with wheat serine carboxypeptidase II: implications for the catalytic mechanism and substrate specificity. *J. Mol. Biol.* **255**, 714-725 (1996).

245. Linding,R., Russell,R.B., Neduva,V. & Gibson,T.J. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701-3708 (2003).

246. Skordalakes,E. & Berger,J.M. Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell* **114**, 135-146 (2003).

247. Schwartz,T., Behlke,J., Lowenhaupt,K., Heinemann,U. & Rich,A. Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nat. Struct. Biol.* **8**, 761-765 (2001).

248. Bogden,C.E., Fass,D., Bergman,N., Nichols,M.D. & Berger,J.M. The structural basis for terminator recognition by the Rho transcription termination factor. *Mol. Cell* **3**, 487-493 (1999).

249. Hopcroft,N.H., Wendt,A.L., Gollnick,P. & Antson,A.A. Specificity of TRAP-RNA interactions: crystal structures of two complexes with different RNA sequences. *Acta Crystallogr. D. Biol. Crystallogr.* **58**, 615-621 (2002).

250. Shimizu,T. *et al.* Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition. *EMBO J.* **16**, 4689-4697 (1997).

251. Cheetham,G.M., Jeruzalmi,D. & Steitz,T.A. Structural basis for initiation of transcription from an RNA polymerase-promoter complex. *Nature* **399**, 80-83 (1999).

252. Raghunathan,S., Kozlov,A.G., Lohman,T.M. & Waksman,G. Structure of the DNA binding domain of E. coli SSB bound to ssDNA. *Nat. Struct. Biol.* **7**, 648-652 (2000).

253. Price,S.R., Evans,P.R. & Nagai,K. Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **394**, 645-650 (1998).

254. Yaremchuk,A., Kriklivyi,I., Tukalo,M. & Cusack,S. Class I tyrosyl-tRNA synthetase has a class II mode of cognate tRNA recognition. *EMBO J.* **21**, 3829-3840 (2002).

255. Thore,S., Mayer,C., Sauter,C., Weeks,S. & Suck,D. Crystal structures of the Pyrococcus abyssi Sm core and its complex with RNA. Common features of RNA binding in archaea and eukarya. *J. Biol. Chem.* **278**, 1239-1247 (2003).

256. McDonald,I.K. & Thornton,J.M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777-793 (1994).

257. Sherwood,A.L., Upchurch,D.A., Stroud,M.R., Davis,W.C. & Holmes,E.H. A highly conserved His-His motif present in alpha1-->3/4fucosyltransferases is required for optimal activity and functions in acceptor binding. *Glycobiology* **12**, 599-606 (2002).

258. Wellems,T.E. & Howard,R.J. Homologous genes encode two distinct histidine-rich proteins in a cloned isolate of Plasmodium falciparum. *Proc. Natl. Acad. Sci. U. S. A* **83**, 6065-6069 (1986).

259. Knapp,B., Nau,U. & Hundt,E. Conservation of antigen components from two recombinant hybrid proteins protective against malaria. *Infect. Immun.* **61**, 892-897 (1993).

260. Zehnder,J.L. & Leung,L.L. Histidine-rich glycoprotein: is there a role in hemostasis or immune function? *J. Lab Clin. Med.* **125**, 682-683 (1995).

261. Lijnen,H.R. & Collen,D. Interaction of heparin with histidine-rich glycoprotein. *Ann. N. Y. Acad. Sci.* **556**, 181-185 (1989).

262. Pichon,C., Goncalves,C. & Midoux,P. Histidine-rich peptides and polymers for nucleic acids delivery. *Adv. Drug Deliv. Rev.* **53**, 75-94 (2001).

263. Kichler,A., Mason,A.J. & Bechinger,B. Cationic amphipathic histidine-rich peptides for gene delivery. *Biochim. Biophys. Acta* **1758**, 301-307 (2006).

264. Stec,I., Nagl,S.B., van Ommen,G.J. & den Dunnen,J.T. The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation? *FEBS Lett.* **473**, 1-5 (2000).

265. Singh,Y.H., Gromiha,M.M., Sarai,A. & Ahmad,S. Atom-wise statistics and prediction of solvent accessibility in proteins. *Biophys. Chem.* **124**, 145-154 (2006).

266. Gavin,A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147 (2002).

267. Agresti,A. An Introduction to Categorical Data Analysis. Wiley-Interscience, Hoboken (2007).

268. Hainzl,T., Huang,S. & Sauer-Eriksson,A.E. Structure of the SRP19 RNA complex and implications for signal recognition particle assembly. *Nature* **417**, 767-771 (2002).

269. Rak,A. *et al.* Structure of the Rab7:REP-1 complex: insights into the mechanism of Rab prenylation and choroideremia disease. *Cell* **117**, 749-760 (2004).

270. Abergel,C. *et al.* Structure and evolution of the Ivy protein family, unexpected lysozyme inhibitors in Gram-negative bacteria. *Proc. Natl. Acad. Sci. U. S. A* **104**, 6394-6399 (2007).

271. Graille,M., Mora,L., Buckingham,R.H., van Tilbeurgh,H. & de Zamaroczy,M. Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. *EMBO J.* **23**, 1474-1482 (2004).

272. van den Akker,F. *et al.* Crystal structure of a new heat-labile enterotoxin, LT-IIb. *Structure.* **4**, 665-678 (1996).

273. Yao,G., Wolinski,J. & Zabielski,R. Effect of Escherichia coli heat-labile enterotoxin on the myoelectric activity of the duodenum in weaned pigs. *J. Vet. Med. A Physiol Pathol. Clin. Med.* **51**, 106-112 (2004).

274. Keskin,O., Tsai,C.J., Wolfson,H. & Nussinov,R. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci.* **13**, 1043-1055 (2004).

275. Thomas,A., Meurisse,R., Charloteaux,B. & Brasseur,R. Aromatic side-chain interactions in proteins. I. Main structural features. *Proteins* **48**, 628-634 (2002).

276. Chelli,R., Gervasio,F.L., Procacci,P. & Schettino,V. Inter-residue and solvent-residue interactions in proteins: a statistical study on experimental structures. *Proteins* **55**, 139-151 (2004).

277. Bahadur,R.P., Chakrabarti,P., Rodier,F. & Janin,J. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.* **336**, 943-955 (2004).

278. Crowley,P.B. & Golovin,A. Cation-pi interactions in protein-protein interfaces. *Proteins* **59**, 231-239 (2005).

279. Bahadur,R.P., Chakrabarti,P., Rodier,F. & Janin,J. Dissecting subunit interfaces in homodimeric proteins. *Proteins* **53**, 708-719 (2003).

280. Saha,R.P., Bahadur,R.P. & Chakrabarti,P. Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface. *J. Proteome. Res.* **4**, 1600-1609 (2005).

281. Miller,S., Lesk,A.M., Janin,J. & Chothia,C. The accessible surface area and stability of oligomeric proteins. *Nature* **328**, 834-836 (1987).

282. Glaser,F., Steinberg,D.M., Vakser,I.A. & Ben Tal,N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins* **43**, 89-102 (2001).

283. Negi,S.S. & Braun,W. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. *J. Mol. Model.* **13**, 1157-1167 (2007).

284. Samanta,U. & Chakrabarti,P. Assessing the role of tryptophan residues in the binding site. *Protein Eng* **14**, 7-15 (2001).

285. Raman,B. *et al.* N(omega)-arginine dimethylation modulates the interaction between a Gly/Arg-rich peptide from human nucleolin and nucleic acids. *Nucleic Acids Res.* **29**, 3377-3384 (2001).

286. Kono,H. & Sarai,A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **35**, 114-131 (1999).

287. Kini,R.M. *et al.* Flanking proline residues identify the L-type Ca2+ channel binding site of calciseptine and FS2. *Biochemistry* **37**, 9058-9063 (1998).

288. Szymanski,M., Barciszewska,M.Z., Zywicki,M. & Barciszewski,J. Noncoding RNA transcripts. *J. Appl. Genet.* **44**, 1-19 (2003).

289. Zhou,H.X. & Qin,S. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics.* **23**, 2203-2209 (2007).

290. Jeong,E., Chung,I.F. & Miyano,S. Prediction of Residues in Protein-RNA Interaction Sites by Neural Networks. *Genome Informatics* **14**, 506-507 (2003).

291. Attwood,T.K. *et al.* PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* **31**, 400-402 (2003).

292. Letunic,I. *et al.* SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.* **34**, D257-D260 (2006).

293. Hulo,N. *et al.* The 20 years of PROSITE. *Nucleic Acids Res.* **36**, D245-D249 (2008).

294. White,S.H. The progress of membrane protein structure determination. *Protein Sci.* **13**, 1948-1949 (2004).

295. Zhou,F., Xue,Y., Yao,X. & Xu,Y. A general user interface for prediction servers of proteins' post-translational modification sites. *Nat. Protoc.* **1**, 1318-1321 (2006).

296. Zhou,H.X. Improving the understanding of human genetic diseases through predictions of protein structures and protein-protein interaction sites. *Curr. Med. Chem.* **11**, 539-549 (2004).

297. Lejeune,D., Delsaux,N., Charloteaux,B., Thomas,A. & Brasseur,R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins* **61**, 258-271 (2005).

# VII. ANNEXES

## *VII.1. Annexe 1*

Masse et volume de van der Waals des 20 acides aminés naturels. La masse donnée correspond à la masse de l'acide aminé non-ionisé moins la masse d'une molécule d'eau.

| Code 1 lettre | Code 3 lettres | Masse (Dalton) | Volume de van der Waals ($Å^3$) |
|---|---|---|---|
| A | Ala | 71.09 | 67 |
| R | Arg | 156.19 | 148 |
| N | Asn | 114.1 | 96 |
| D | Asp | 115.09 | 91 |
| C | Cys | 103.15 | 86 |
| Q | Gln | 128.14 | 114 |
| E | Glu | 129.12 | 109 |
| G | Gly | 57.05 | 48 |
| H | His | 137.14 | 117 |
| I | Ile | 113.16 | 124 |
| L | Leu | 113.16 | 124 |
| K | Lys | 128.17 | 135 |
| M | Met | 131.19 | 124 |
| F | Phe | 147.18 | 135 |
| P | Pro | 97.12 | 90 |
| S | Ser | 87.08 | 73 |
| T | Thr | 101.11 | 93 |
| W | Trp | 186.21 | 163 |
| Y | Tyr | 163.18 | 141 |
| V | Val | 99.14 | 105 |
| Moyenne pondérée | | 119.4 | 161 |

## *VII.2. Annexe 2*

Codes PDB des complexes utilisés dans cette thèse. Les codes soulignés dans le tableau des complexes protéine-ADN correspondent à la sous-banque de structures 3D hautes résolutions contenant une double hélice d'ADN.

| Complexes protéine-ADN | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1a0a | 1cez | 1ecr | 1gdt | 1ic8 | 1kdh | 1mw8 | 1qna | 2bpa |
| 1a1v | 1cfF7 | 1efa | 1gt0 | 1ign | 1ku7 | 1mwi | 1qpi | 2cgp |
| 1a3q | 1ckt | 1egw | 1gu4 | 1j1v | 1kx5 | 1n6q | 1qpz | 2drp |
| 1a73 | 1cl8 | 1esg | 1gxp | 1j75 | 1l3l | 1nh2 | 1qrv | 2hdd |
| 1ais | 1cw0 | 1ewn | 1h6f | 1jb7 | 1l3s | 1nkp | 1qum | 2irf |
| 1am9 | 1d02 | 1ewq | 1h9d | 1je8 | 1llm | 1nlw | 1r2z | 2pjr |
| 1awc | 1dc1 | 1exj | 1hao | 1jey | 1lmb | 1noy | 1rep | 2up1 |
| 1bB01 | 1ddn | 1eyg | 1hcr | 1jfi | 1lq1 | 1odh | 1skn | 3hts |
| 1b3t | 1dew | 1f0v | 1hio | 1jj4 | 1lrr | 1oe4 | 1t7p | 3pvi |
| 1bc8 | 1dfm | 1f44 | 1hlv | 1jmc | 1lwy | 1orn | 1tc3 | 6cro |
| 1bdt | 1dhH3 | 1f4k | 1hwt | 1jt0 | 1m07 | 1oup | 1tro | 6mht |
| 1bg1 | 1diz | 1fiu | 1i3j | 1jx4 | 1m5r | 1p4e | 1tup | |
| 1bl0 | 1dmu | 1fok | 1i6j | 1k3x | 1mhd | 1p71 | 1ubd | |
| 1bpy | 1dp7 | 1fzp | 1i7d | 1k4t | 1mjo | 1p7h | 1vas | |
| 1brn | 1dsz | 1g38 | 1i8m | 1k78 | 1mnn | 1puf | 1zme | |
| 1c8c | 1e3o | 1g9z | 1iaw | 1kc6 | 1mus | 1pv4 | 2bop | |

| Complexes protéine-ARN | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1a9n | 1c0a | 1e7k | 1g1x | 1hq1 | 1k8w | 1mms | 1qf6 | 2fmt |
| 1apg | 1cxo | 1ec6 | 1gax | 1i6u | 1knz | 1mzp | 1qtq | |
| 1asy | 1ddl | 1f7u | 1gtf | 1il2 | 1kq2 | 1n35 | 1rmv | |
| 1av6 | 1dfu | 1f8v | 1h2c | 1j1u | 1lng | 1n78 | 1ser | |
| 1b23 | 1di2 | 1feu | 1h3e | 1jbr | 1m8v | 1nb7 | 2a8v | |
| 1b7f | 1e6t | 1ffy | 1h4s | 1jid | 1m8x | 1ooa | 2bbv | |

| Complexes protéine-protéine (1) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12as | 1dpg | 1g8l | 1jig | 1mtp | 1p4n | 1rre | 1udz | 1wmz |
| 1a0c | 1dqa | 1g8q | 1jiw | 1mty | 1p57 | 1rrm | 1ueh | 1wnf |
| 1a0t | 1dqe | 1g99 | 1jk4 | 1muc | 1p5t | 1rsg | 1uf3 | 1wom |
| 1a12 | 1dqi | 1ga6 | 1jke | 1mvc | 1p5v | 1rtw | 1uf5 | 1wph |
| 1a14 | 1dqp | 1gcq | 1jkg | 1mvf | 1p60 | 1rw0 | 1ugp | 1wpn |
| 1a16 | 1dtd | 1gd0 | 1jkm | 1mw5 | 1p6o | 1rwi | 1ugx | 1wpo |
| 1a2x | 1dtj | 1ggg | 1jl0 | 1mwq | 1p6z | 1rx0 | 1uh5 | 1wq1 |
| 1a4i | 1duv | 1ggx | 1jlj | 1mwv | 1p9e | 1rxd | 1uhv | 1wr8 |
| 1a4p | 1dwk | 1gk2 | 1jly | 1mww | 1p9l | 1rxz | 1ui5 | 1wty |
| 1a4q | 1dxe | 1gk9 | 1jm6 | 1mx1 | 1p9o | 1rya | 1uii | 1wu9 |
| 1a64 | 1dxr | 1gkb | 1jmt | 1mxr | 1pb0 | 1ryd | 1uik | 1wuw |
| 1a6j | 1dyo | 1gl2 | 1jnp | 1my7 | 1pby | 1rzh | 1uix | 1wvg |
| 1a7t | 1dys | 1gl4 | 1jnr | 1mz9 | 1pcx | 1s0a | 1uj2 | 1wvi |
| 1a92 | 1e0b | 1gnw | 1jo0 | 1mzh | 1pdk | 1s14 | 1ukk | 1wwj |
| 1aa7 | 1e0t | 1go4 | 1joc | 1n0q | 1pe0 | 1s2d | 1ukv | 1www |
| 1ade | 1e19 | 1goi | 1jps | 1n0w | 1pe1 | 1s2k | 1uli | 1wwz |
| 1ae9 | 1e1h | 1got | 1jqb | 1n1b | 1pfb | 1s3e | 1ulk | 1wx1 |

| Complexes protéine-protéine (2) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1agq | 1e2k | 1gpm | 1jql | 1n1c | 1pin | 1s4k | 1um0 | 1wy1 |
| 1ahs | 1e44 | 1gpu | 1jr8 | 1n1e | 1pix | 1s4n | 1unn | 1wy2 |
| 1aj8 | 1e4y | 1gq6 | 1js1 | 1n1j | 1pjh | 1s57 | 1upa | 1x6i |
| 1ajs | 1e54 | 1gqi | 1js3 | 1n2a | 1pju | 1s5a | 1upk | 1x79 |
| 1aoc | 1e5d | 1gqp | 1js8 | 1n2d | 1pk6 | 1s5d | 1uqt | 1x7d |
| 1aoh | 1e5r | 1gto | 1jsd | 1n2f | 1pl8 | 1s5p | 1us7 | 1x7o |
| 1apy | 1e6c | 1gu7 | 1jsu | 1n46 | 1pmm | 1s95 | 1usc | 1x7y |
| 1aq0 | 1e6i | 1gut | 1jtd | 1n69 | 1poi | 1s96 | 1usi | 1x99 |
| 1aqc | 1e7l | 1gux | 1jub | 1n7f | 1ppv | 1s98 | 1uso | 1x9i |
| 1at3 | 1e7w | 1gve | 1juh | 1n7h | 1pq1 | 1s99 | 1usu | 1x9z |
| 1atz | 1e8g | 1gvj | 1jv1 | 1n8k | 1pqh | 1s9r | 1ut7 | 1xa3 |
| 1au1 | 1e8i | 1gvn | 1jva | 1n98 | 1pqz | 1sby | 1uth | 1xb2 |
| 1avq | 1e96 | 1gxj | 1jw9 | 1na6 | 1psr | 1sbz | 1uty | 1xew |
| 1avw | 1e9g | 1gxr | 1jx2 | 1nbu | 1pt7 | 1sc3 | 1uv7 | 1xfo |
| 1awi | 1eaj | 1gyg | 1jxh | 1nbw | 1ptm | 1sce | 1uvc | 1xfs |
| 1axi | 1ebf | 1gyk | 1jy2 | 1nc7 | 1pvg | 1scj | 1uvq | 1xg0 |
| 1ay7 | 1ebl | 1gyo | 1jya | 1nd4 | 1pvm | 1sd4 | 1uwk | 1xg5 |
| 1aym | 1ecf | 1gyx | 1jyo | 1nd6 | 1pvn | 1sed | 1uwz | 1xg7 |
| 1ayo | 1ecs | 1h18 | 1jys | 1ng5 | 1pwb | 1sei | 1uxa | 1xgs |
| 1azs | 1ed8 | 1h1y | 1jz8 | 1nh0 | 1pwx | 1sff | 1uyr | 1xiw |
| 1b00 | 1eej | 1h2k | 1jzd | 1ni4 | 1pxy | 1sfl | 1uzb | 1xkl |
| 1b07 | 1eeo | 1h2s | 1jzt | 1nki | 1pym | 1sfx | 1uzv | 1xkq |
| 1b0n | 1eer | 1h32 | 1k1x | 1nkz | 1pyo | 1sg4 | 1uzx | 1xly |
| 1b25 | 1eex | 1h3f | 1k2x | 1nlf | 1pyt | 1sgj | 1v18 | 1xm3 |
| 1b2p | 1ef8 | 1h3o | 1k32 | 1nln | 1pzg | 1sgm | 1v1a | 1xnf |
| 1b34 | 1efn | 1h4g | 1k3b | 1nlq | 1q08 | 1sh8 | 1v25 | 1xo1 |
| 1b35 | 1efu | 1h4r | 1k3r | 1nnw | 1q0e | 1sj1 | 1v3v | 1xoc |
| 1b3a | 1eg5 | 1h54 | 1k3s | 1np3 | 1q0q | 1slu | 1v4n | 1xor |
| 1b4f | 1ehi | 1h59 | 1k3y | 1npe | 1q15 | 1smt | 1v4v | 1xpj |
| 1b5e | 1ehk | 1h5w | 1k4m | 1nq7 | 1q16 | 1smx | 1v58 | 1xqh |
| 1b5f | 1ejd | 1h6p | 1k5n | 1nqj | 1q1a | 1snd | 1v5v | 1xr4 |
| 1b5p | 1ek6 | 1h6w | 1k66 | 1nrj | 1q1g | 1snr | 1v6z | 1xrh |
| 1b62 | 1ek9 | 1h8e | 1k7w | 1ns5 | 1q3o | 1spp | 1v74 | 1xrk |
| 1b66 | 1el5 | 1h8g | 1k8k | 1nsz | 1q4u | 1sqe | 1v7b | 1xs0 |
| 1b67 | 1el6 | 1hbn | 1ka9 | 1ntv | 1q5y | 1sqi | 1v7z | 1xsj |
| 1b77 | 1elu | 1hdf | 1kae | 1nu7 | 1q67 | 1sqs | 1v84 | 1xsq |
| 1b7g | 1emu | 1hdm | 1kbj | 1nul | 1q6o | 1squ | 1v8b | 1xtg |
| 1b7y | 1eo6 | 1hei | 1kcf | 1nuu | 1q7f | 1sr4 | 1v8d | 1xtt |
| 1b8a | 1eo9 | 1hf2 | 1kek | 1nvr | 1q7l | 1ss4 | 1v9y | 1xu1 |
| 1b8g | 1eoi | 1hgx | 1khq | 1nvv | 1q7s | 1ssq | 1va0 | 1xu2 |
| 1b8z | 1ep3 | 1hi9 | 1kht | 1nw1 | 1q8f | 1stm | 1vc1 | 1xu9 |
| 1b93 | 1epx | 1hk9 | 1kic | 1nww | 1q8r | 1stz | 1vdk | 1xva |
| 1b9m | 1etx | 1hl9 | 1kiy | 1nxm | 1q98 | 1su2 | 1vdw | 1xvh |

| Complexes protéine-protéine (3) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1b9w | 1eu3 | 1hm6 | 1kjn | 1nxu | 1q9u | 1suw | 1ve9 | 1xvs |
| 1baz | 1eud | 1hqk | 1kjq | 1nyt | 1qb5 | 1svm | 1vet | 1xy7 |
| 1bc5 | 1euh | 1hqs | 1kko | 1nzy | 1qbz | 1svv | 1vf6 | 1xzp |
| 1bcs | 1euj | 1hrk | 1knv | 1o0s | 1qc7 | 1sw6 | 1vfj | 1xzw |
| 1bd0 | 1euv | 1hru | 1kny | 1o12 | 1qci | 1sz2 | 1vg0 | 1y0b |
| 1bdf | 1ev7 | 1hw1 | 1kpe | 1o1h | 1qd1 | 1t01 | 1vgg | 1y0e |
| 1bdy | 1ewk | 1hw5 | 1kps | 1o26 | 1qd6 | 1t06 | 1vgt | 1y0h |
| 1bis | 1ex0 | 1hwx | 1kqf | 1o58 | 1qdn | 1t0h | 1vgy | 1y0u |
| 1bkp | 1ex2 | 1hx1 | 1kqp | 1o5x | 1qex | 1t0q | 1vh4 | 1y0z |
| 1bo4 | 1ext | 1hx6 | 1ksh | 1o63 | 1qf8 | 1t0t | 1vh5 | 1y13 |
| 1bou | 1exz | 1hx8 | 1kso | 1o66 | 1qfh | 1t15 | 1vhq | 1y14 |
| 1bpl | 1eyq | 1hxh | 1kta | 1o69 | 1qft | 1t33 | 1vhv | 1y1o |
| 1brw | 1eyv | 1hyo | 1kug | 1o6a | 1qfx | 1t3g | 1vhw | 1y2o |
| 1btk | 1ez3 | 1hzd | 1kut | 1o6s | 1qg3 | 1t3i | 1vhy | 1y3t |
| 1bvn | 1ezf | 1i0d | 1kxp | 1o6z | 1qg6 | 1t3u | 1vhz | 1y44 |
| 1bw0 | 1ezg | 1i0r | 1kyf | 1o75 | 1qge | 1t4a | 1vi0 | 1y4j |
| 1byf | 1ezi | 1i1q | 1kz8 | 1o7j | 1qgj | 1t4b | 1via | 1y4t |
| 1byk | 1f06 | 1i1r | 1kzq | 1o7k | 1qgr | 1t4f | 1vic | 1y60 |
| 1c02 | 1f08 | 1i31 | 1l4d | 1o7n | 1qhf | 1t4h | 1vim | 1y71 |
| 1c1d | 1f0k | 1i36 | 1l5b | 1o81 | 1qi9 | 1t5b | 1vio | 1y7r |
| 1c1y | 1f1u | 1i4d | 1l5w | 1o8b | 1qjb | 1t5r | 1vj2 | 1y7y |
| 1c3c | 1f2t | 1i4u | 1l6r | 1o8u | 1qjc | 1t61 | 1vje | 1y8t |
| 1c4q | 1f34 | 1i4y | 1l6x | 1o97 | 1qjj | 1t6b | 1vjl | 1y8x |
| 1c5e | 1f3m | 1i73 | 1l7a | 1o9i | 1qkr | 1t6s | 1vjn | 1y97 |
| 1c77 | 1f3v | 1i7n | 1l8a | 1o9r | 1qks | 1t7r | 1vjq | 1y9b |
| 1c7g | 1f46 | 1i7w | 1l8b | 1oa8 | 1qkz | 1t8t | 1vk0 | 1y9i |
| 1c8u | 1f4q | 1i8d | 1l8d | 1oac | 1ql0 | 1t92 | 1vke | 1y9w |
| 1c9k | 1f5m | 1ia9 | 1lb6 | 1oah | 1qla | 1ta3 | 1vki | 1yac |
| 1cg2 | 1f5v | 1iar | 1ldd | 1oai | 1qlw | 1taf | 1vkm | 1yav |
| 1cgh | 1f60 | 1ibv | 1lfa | 1oao | 1qmh | 1tc6 | 1vkn | 1yb4 |
| 1ci4 | 1f6b | 1id1 | 1lh0 | 1obb | 1qmv | 1tdt | 1vkp | 1yb5 |
| 1ckm | 1f6f | 1idp | 1lj2 | 1obf | 1qo0 | 1te5 | 1vkv | 1ybk |
| 1cku | 1f74 | 1ig0 | 1lk9 | 1obw | 1qo3 | 1tej | 1vl2 | 1ybx |
| 1cli | 1f80 | 1igq | 1lkk | 1obx | 1qo7 | 1tii | 1vl4 | 1yco |
| 1cnz | 1f86 | 1ii7 | 1lm5 | 1oc0 | 1qop | 1tiq | 1vl7 | 1ydh |
| 1coz | 1f89 | 1ijy | 1lm8 | 1ock | 1qoz | 1tk4 | 1vla | 1ydm |
| 1cp2 | 1f8m | 1ik9 | 1ln0 | 1od5 | 1qpb | 1tk9 | 1vlg | 1ydw |
| 1cq3 | 1f9z | 1im8 | 1lp1 | 1of8 | 1qq5 | 1tki | 1vlj | 1ydy |
| 1cqx | 1fbt | 1in0 | 1lq9 | 1ofu | 1qqg | 1tl9 | 1vlr | 1yem |
| 1cru | 1fc4 | 1inl | 1lqa | 1ogd | 1qqp | 1tlu | 1vlt | 1yf2 |
| 1cs1 | 1fd3 | 1ipe | 1lqt | 1oh0 | 1qsd | 1tmq | 1vm6 | 1yfs |
| 1ct4 | 1fe0 | 1ips | 1lsh | 1ohe | 1qsm | 1tn6 | 1vm7 | 1ygy |
| 1ct9 | 1fhw | 1iq6 | 1lss | 1ohz | 1qtn | 1to2 | 1vme | 1ykd |
| 1cun | 1fiq | 1iq8 | 1lt8 | 1oi4 | 1qu9 | 1to6 | 1vmf | 1yki |

| Complexes protéine-protéine (4) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1cvr | 1fj2 | 1iqd | 1lua | 1oi6 | 1qv9 | 1tr8 | 1vp2 | 1ypq |
| 1cxz | 1fjh | 1iqy | 1luc | 1oih | 1qvc | 1tu1 | 1vp4 | 1zpd |
| 1czf | 1fjr | 1ird | 1luq | 1oj5 | 1qve | 1tu5 | 1vpg | 2arc |
| 1d09 | 1fle | 1irq | 1lw1 | 1ok7 | 1qvz | 1tu7 | 1vpj | 2beq |
| 1d0c | 1flm | 1itb | 1lwd | 1oki | 1qw9 | 1tvx | 1vpm | 2bez |
| 1d0q | 1flt | 1itu | 1lwj | 1ol5 | 1qxr | 1twd | 1vps | 2bji |
| 1d1g | 1fm0 | 1iu1 | 1lxn | 1olz | 1qyc | 1twi | 1vpz | 2bkq |
| 1d1q | 1fm9 | 1iu8 | 1m0w | 1omo | 1qyn | 1tx4 | 1vq3 | 2dyn |
| 1d2f | 1fn8 | 1iug | 1m1e | 1omz | 1qz7 | 1txg | 1vqu | 2e2a |
| 1d2v | 1fn9 | 1iuj | 1m1f | 1on2 | 1r0r | 1txk | 1vr4 | 2ebo |
| 1d3y | 1fns | 1ix9 | 1m1z | 1on3 | 1r12 | 1ty9 | 1vyb | 2fib |
| 1d4t | 1fo0 | 1iyb | 1m2d | 1one | 1r17 | 1tz0 | 1vzy | 2gsa |
| 1d4v | 1fpo | 1iye | 1m2t | 1onr | 1r1d | 1tzj | 1w07 | 2hrv |
| 1d4x | 1fqj | 1iz9 | 1m3k | 1onw | 1r28 | 1tzy | 1w23 | 2mpr |
| 1d8h | 1fr2 | 1j1b | 1m41 | 1oo0 | 1r31 | 1u00 | 1w27 | 2prg |
| 1d8l | 1fs0 | 1j1j | 1m4i | 1ooe | 1r3j | 1u07 | 1w2i | 2psp |
| 1d8w | 1fs1 | 1j1y | 1m4r | 1ooy | 1r42 | 1u08 | 1w2y | 2pva |
| 1d9c | 1fs5 | 1j2g | 1m4u | 1oqj | 1r4p | 1u0f | 1w30 | 2tnf |
| 1dbf | 1fsg | 1j2j | 1m4z | 1oqq | 1r61 | 1u0s | 1w3i | 2tps |
| 1dbq | 1ftr | 1j2r | 1m55 | 1oqw | 1r7a | 1u11 | 1w44 | 2trc |
| 1dbw | 1fux | 1j30 | 1m56 | 1or4 | 1r8e | 1u19 | 1w5n | 2vub |
| 1dci | 1fvk | 1j34 | 1m5w | 1or7 | 1r8g | 1u1i | 1w6n | 3chb |
| 1dd3 | 1fxk | 1j3v | 1m6k | 1or8 | 1r8j | 1u1w | 1w6s | 3csu |
| 1ddv | 1fxw | 1j3w | 1m6p | 1orj | 1r8s | 1u1z | 1w6u | 3daa |
| 1ddz | 1g0o | 1j5w | 1m6s | 1ors | 1rdq | 1u2m | 1w98 | 3eip |
| 1deb | 1g0s | 1j71 | 1m93 | 1oru | 1reg | 1u5k | 1w9a | 3erd |
| 1dek | 1g1b | 1j79 | 1m98 | 1ory | 1rew | 1u5u | 1w9c | 3fiv |
| 1deu | 1g1c | 1j8d | 1mfg | 1osp | 1rg9 | 1u7b | 1wa5 | 3hbi |
| 1df9 | 1g1j | 1jat | 1mgq | 1osy | 1rgx | 1u7e | 1wb4 | 3lyn |
| 1dgf | 1g1s | 1jb2 | 1mhq | 1otf | 1rj4 | 1u7i | 1wc3 | 3ygs |
| 1di1 | 1g29 | 1jbo | 1miz | 1otg | 1rk8 | 1u7k | 1wdc | 4ubp |
| 1dj0 | 1g2o | 1jcd | 1mk4 | 1otv | 1rkg | 1u7n | 1wdd | 5hpg |
| 1djl | 1g2q | 1jd0 | 1mka | 1ou0 | 1rkt | 1u8s | 1wdj | 7ahl |
| 1djt | 1g31 | 1jdh | 1mkf | 1ouw | 1rku | 1u8v | 1wdv | |
| 1dk0 | 1g3k | 1je0 | 1mkk | 1ovn | 1rm6 | 1uad | 1weh | |
| 1dku | 1g4y | 1jek | 1mkz | 1ozh | 1rmw | 1uan | 1wiw | |
| 1dle | 1g57 | 1jet | 1mna | 1p1c | 1ro7 | 1ubk | 1wkq | |
| 1dmh | 1g64 | 1jfl | 1mpx | 1p1j | 1roz | 1uc2 | 1wkr | |
| 1dok | 1g6u | 1jh6 | 1mqk | 1p1x | 1rp0 | 1uc7 | 1wmg | |
| 1dow | 1g73 | 1jhf | 1mr8 | 1p2j | 1rq2 | 1uc8 | 1wmh | |
| 1dp5 | 1g8e | 1ji7 | 1msp | 1p32 | 1rqp | 1udv | 1wms | |

## *VII.3. Annexe 3.1*

Nomenclature (PDB) des atomes des 20 acides aminés naturels.

| | *Aliphatiques* | | | | |
|---|---|---|---|---|---|
| | **Alanine** | **Isoleucine** | **Leucine** | **Méthionine** | **Valine** |
| **BK** | N | N | N | N | N |
| | CA | CA | CA | CA | CA |
| | C | C | C | C | C |
| | O | O | O | O | O |
| **SC** | CB | CB | CB | CB | CB |
| | H | CG1 | CG | CG | CG1 |
| | | CG2 | CD1 | SD | CG2 |
| | | CD1 | CD2 | CE | H |
| | | H | H | H | HA |
| | | HA | HA | HA | HB |
| | | HB | 1HB | 1HB | 1HG1 |
| | | 1HG1 | 2HB | 2HB | 2HG1 |
| | | 2HG1 | HG | 1HG | 3HG1 |
| | | 1HG2 | 1HD1 | 2HG | 1HG2 |
| | | 2HG2 | 2HD1 | 1HE | 2HG2 |
| | | 3HG2 | 3HD1 | 2HE | |
| | | 1HD1 | 1HD2 | 3HE | |
| | | 2HD1 | 2HD2 | | |
| | | 3HD1 | 3HD2 | | |

| | *Chargés* | | | |
|---|---|---|---|---|
| | **Arginine** | **Aspartate** | **Glutamate** | **Lysine** |
| **BK** | N | N | N | N |
| | CA | CA | CA | CA |
| | C | C | C | C |
| | O | O | O | O |
| **SC** | CB | CB | CB | CB |
| | CG | CG | CG | CG |
| | CD | OD1 | CD | CD |
| | NE | OD2 | OE1 | CE |
| | CZ | H | OE2 | NZ |
| | NH1 | HA | H | H |
| | NH2 | 1HB | HA | HA |
| | H | 2HB | 1HB | 1HB |
| | HA | | 2HB | 2HB |
| | 1HB | | 1HG | 1HG |
| | 2HB | | 2HG | 2HG |
| | 1HG | | | 1HD |
| | 2HG | | | 2HD |
| | 1HD | | | 1HE |
| | 2HD | | | 2HE |
| | HE | | | 1HZ |
| | 1HH1 | | | 2HZ |
| | 2HH1 | | | |
| | 1HH2 | | | |
| | 2HH2 | | | |

| | *Aromatiques* | | |
|---|---|---|---|
| | **Phénylalanine** | **Tryptophane** | **Tyrosine** |
| **BK** | N | N | N |
| | CA | CA | CA |
| | C | C | C |
| | O | O | O |
| **SC** | CB | CB | CB |
| | CG | CG | CG |
| | CD1 | CD1 | CD1 |
| | CD2 | CD2 | CD2 |
| | CE1 | NE1 | CE1 |
| | CE2 | CE2 | CE2 |
| | CZ | CE3 | CZ |
| | H | CZ2 | OH |
| | HA | CZ3 | H |
| | 1HB | CH2 | HA |
| | 2HB | H | 1HB |
| | HD1 | HA | 2HB |
| | HD2 | 1HB | HD1 |
| | HE1 | 2HB | HD2 |
| | HE2 | HD1 | HE1 |
| | HZ | HE1 | HE2 |
| | | HE3 | HH |
| | | HZ2 | |
| | | HZ3 | |
| | | HH2 | |

| | *Polaires* | | | | |
|---|---|---|---|---|---|
| | **Asparagine** | **Glutamine** | **Histidine** | **Sérine** | **Thréonine** |
| **BK** | N | N | N | N | N |
| | CA | CA | CA | CA | CA |
| | C | C | C | C | C |
| | O | O | O | O | O |
| **SC** | CB | CB | CB | CB | CB |
| | CG | CG | CG | OG | OG1 |
| | OD1 | CD | ND1 | H | CG2 |
| | ND2 | OE1 | CD2 | HA | H |
| | H | NE2 | CE1 | 1HB | HA |
| | HA | H | NE2 | 2HB | HB |
| | 1HB | HA | H | HG | HG1 |
| | 2HB | 1HB | HA | | 1HG2 |
| | 1HD2 | 2HB | 1HB | | 2HG2 |
| | 2HD2 | 1HG | 2HB | | 3HG2 |
| | | 2HG | HD1 | | |
| | | 1HE2 | HD2 | | |
| | | 2HE2 | HE1 | | |
| | | | HE2 | | |

| | *Particuliers* | | |
|---|---|---|---|
| | **Cystéine** | **Glycine** | **Proline** |
| **BK** | N | N | N |
| | CA | CA | CA |
| | C | C | C |
| | O | O | O |
| **SC** | CB | H | CB |
| | SG | 1HA | CG |
| | H | 2HA | CD |
| | HA | | HA |
| | 1HB | | 1HB |
| | 2HB | | 2HB |
| | HG | | 1HG |
| | | | 2HG |
| | | | 1HD |
| | | | 2HD |

## *VII.4. Annexe 3.2*

Nomenclature (PDB) des atomes des 5 nucléotides.

| | Cytidine | Adénosine | Thymidine | Guanosine | Uracile |
|---|---|---|---|---|---|
| PO4 | P | P | P | P | P |
| | O1P | O1P | O1P | O1P | O1P |
| | O2P | O2P | O2P | O2P | O2P |
| Sucre | O5* | O5* | O5* | O5* | O5* |
| | C5* | C5* | C5* | C5* | C5* |
| | C4* | C4* | C4* | C4* | C4* |
| | O4* | O4* | O4* | O4* | O4* |
| | C3* | C3* | C3* | C3* | C3* |
| | O3* | O3* | O3* | O3* | O3* |
| | C2* | C2* | C2* | C2* | C2* |
| | C1* | C1* | C1* | C1* | O2* |
| | 1H5* | 1H5* | 1H5* | 1H5* | C1* |
| | 2H5* | 2H5* | 2H5* | 2H5* | H5* |
| | H4* | H4* | H4* | H4* | H5* |
| | H1* | H1* | H1* | H1* | H4* |
| | H3* | H3* | H3* | H3* | H1* |
| | 1H2* | 1H2* | 1H2* | 1H2* | H3* |
| | 2H2* | 2H2* | 2H2* | 2H2* | H2* |
| | | | | | HO2 |
| Base | N1 | N1 | N9 | N9 | N1 |
| | C2 | C2 | C8 | C8 | C2 |
| | O2 | O2 | N7 | N7 | O2 |
| | N3 | N3 | C5 | C5 | N3 |
| | C4 | C4 | C6 | C6 | C4 |
| | N4 | O4 | N6 | O6 | O4 |
| | C5 | C5 | N1 | N1 | C5 |
| | C6 | C5M | C2 | C2 | C6 |
| | H6 | C6 | N3 | N2 | H6 |
| | H5 | H6 | C4 | N3 | H5 |
| | 1H4 | 1H5M | H8 | C4 | H3 |
| | 2H4 | 2H5M | 1H6 | H8 | |
| | | 3H5M | 2H6 | H1 | |
| | | H3 | H2 | 1H2 | |
| | | | | 2H2 | |
| | ADN | | | | ARN |