Utilisation de la tessellation de Voronoï pour l'étude des complexes protéine-protéine

THÈSE

présentée et soutenue publiquement le 7 avril 2006

pour l'obtention du

Doctorat de l'Université Paris-Sud

(École Doctorale Innovation Thérapeutique)

par

Julie Bernauer

Composition du jury

Président :	P[r] Joël Janin	IBBMC, Université Paris-Sud
Rapporteurs :	D[r] Frédéric Cazals	Projet Geometrica, INRIA Sophia Antipolis
	P[r] Olivier Lichtarge	Computational Biology, Baylor College of Medicine, TX, USA
Directrice de thèse :	D[r] Anne Poupon	IBBMC, CNRS/Université Paris-Sud
Examinateur :	P[r] Alain Denise	LRI, Université Paris-Sud

Équipe de Génomique Structurale
Laboratoire d'Enzymologie et de Biochimie Structurale — UPR 9063
Institut de Biochimie et de Biophysique Moléculaire et Cellulaire — UMR 8619

Remerciements

Je tiens tout d'abord à remercier Anne Poupon qui m'a fait confiance, m'a formée, encadrée, encouragée et ... supportée pendant ces quatre années. Je la remercie pour son soutien et sa disponibilité au jour le jour mais aussi pour ses conseils et tous les bons moments au laboratoire, ainsi qu'en dehors.

Merci à Joël Janin, pour son accueil en tant que directeur du LEBS, mais également pour toute l'aide scientifique qu'il m'a apportée.

Je remercie Lucienne Letellier et Michel Desmadril, directeurs de l'IBBMC de m'avoir accueillie chaleureusement.

Je remercie Frédéric Cazals et Olivier Lichtarge, d'avoir accepté d'être rapporteurs de ce travail, et Alain Denise d'avoir bien voulu en être examinateur.

Je remercie aussi tous mes collaborateurs qui m'ont soutenue et fait confiance tout au long de ce travail : Jérôme Azé au Laboratoire de Recherche en Informatique à Orsay, Marie-Hélène Mucchielli-Giorgi au Centre de Génétique Moléculaire à Gif-sur-Yvette, Alexandre Bonvin au Bijvoet Center à Utrecht (Pays-Bas), Frédéric Cazals, Mariette Yvinec et Jean-Daniel Boissonnat à l'INRIA Sophia Antipolis et Sébastien Graziani, Ursula Liebl et Hannu Myllykallio.

Je n'oublie pas Herman van Tilbeurgh et toute l'équipe de génomique structurale qui m'ont accompagnée ainsi que les membres des deux laboratoires qui m'ont soutenue tout au long de ce travail. J'ai une pensée particulière pour Éric, Laurent, Françoise, Sabrina, Nicolas, Marc, Mark, Sophie, Lionel, Ranjit, Charles, Zhou, Bruno, Karine, Yuxing, Solange, Virginie, Sylvie, Pierre, Philippe, Vincent, Michael, Isabelle et Inès mais aussi pour tous les « indispensables » des deux laboratoires : Jérôme, Éric, Joëlle, Marie-Hélène, Stéphanie, Pierrette, Sophie, Diana, Willy, Anne-Pascale, Jocelyne, les Annie, Mireille et Pascal.

Je remercie également les membres du Plan Pluriformation (PPF) Bioinformatique et Genomique, pour leur soutien et leur sympathie, ainsi que mes collègues d'enseignement à l'Université.

Merci à ma famille et mes amis, en particulier Chantal, Jacques, Charlotte, Jacqueline et Denis.

Enfin, mais non des moindres, je remercie Kilian pour son soutien au quotidien et ses relectures patientes.

Que tous ceux que j'ai oubliés me pardonnent...

Sommaire

Introduction

La fonction biologique d'une protéine dépend souvent de l'interaction de celle-ci avec un ou plusieurs partenaires. La caractérisation de ces interactions dans les systèmes cellulaires et la compréhension de ces mécanismes est donc un des thèmes principaux de la biologie à l'heure actuelle. Les études récentes à grande échelle ont en effet montré que, dans des organismes modèles tels que la levure, on peut montrer l'existence de plus de 15 000 complexes. Ces interactions, souvent associées à de grands assemblages macromoléculaires, jouent un rôle clé dans la fonction cellulaire.

Cependant, ces assemblages ne sont souvent pas connus expérimentalement, comme le prouve le nombre relativement faible de complexes protéine-protéine dans la *Protein Data Bank (PDB)*[14] (environ 400 complexes pour 27 000 entrées fin 2004). Même si certaines initiatives de génomique structurale s'attachent spécifiquement à la résolution de structures d'assemblages macromoléculaires, l'échantillon de l'interactome dont on connaît la structure reste et restera faible, en raison de toutes les contraintes que la résolution expérimentale impose. Quant au nombre de structures des partenaires isolés, il croît exponentiellement, permettant une bonne connaissance de ces protéines.

Pour toutes ces raisons, l'étude informatique de ces interactions est nécessaire. En particulier, les méthodes d'amarrage ou *docking*, capables de prédire de façon fiable des modèles structuraux de complexes à partir des composants pris séparément, peuvent jouer un rôle important dans la compréhension de ces phénomènes.

Les algorithmes d'amarrage comportent deux étapes successives : d'abord, un grand nombre de modèles est généré, puis une fonction de score est utilisée pour les classer et, si possible, déterminer la conformation native. Cette fonction de score doit prendre en compte à la fois la complémentarité géométrique des deux molécules, mais aussi les propriétés physico-chimiques des surfaces en interaction.

Dans ce travail, je me suis intéressée à la seconde étape de l'algorithme : la mise au point d'une fonction de score rapide et fiable. Ce travail utilise une construction géométrique connue depuis le début du XX^{e} siècle : la tessellation de Voronoï. Le diagramme de Voronoï doit son nom à l'étude qu'en a faite en dimension n le mathématicien ukrainien Georgy Fedoseevich Voronoï en 1908 [220]. Cette construction, qui permet d'associer à chacun des points considérés sa « zone d'influence » a été utilisée dès 1854, c'est-à-dire avant d'être caractérisée et étudiée par Voronoï, par John Snow pour modéliser la répartition de l'épidémie de choléra de Londres, prouvant que les personnes atteintes se trouvaient plus près du point d'eau de Broad Street que de toutes les autres pompes du quartier de Soho. Son utilisation a ensuite été très importante en géophysique et en

météorologie pour analyser la distribution spatiale de données, telles que les mesures de pluviométrie. En particulier, les polygones de Voronoï sont parfois appelés polygones de Thiessen, en hommage au météorologue Alfred H. Thiessen qui s'y est beaucoup intéressé au début du XX^{e} siècle. Le diagramme de Voronoï recouvre également la notion de maille de Wigner-Seitz (ou première zone de Brillouin) d'un réseau cristallin définie en physique du solide. À l'heure actuelle, on trouve des tessellations de Voronoï dans de très nombreux domaines d'application (chimie, réseaux et télécommunications, reconstruction/compression d'images, transport, économie...) dont la biologie. En particulier, les tessellations de Voronoï ou de Laguerre se sont avérées être de bons modèles mathématiques de la structure des protéines. Cette formalisation permet en effet de faire une bonne description de l'empilement et des propriétés structurales des résidus.

Dans la première partie de ce manuscrit, je montrerai comment cette construction peut être envisagée pour modéliser les interfaces protéine-protéine et comment nous l'avons utilisée pour la mise au point d'une fonction de score pour l'amarrage protéine-protéine. Pour cela, nous avons utilisé plusieurs méthodes d'apprentissage statistique que nous présenterons. Nous avons aussi appliqué cette méthode, mise au point pour l'amarrage, pour discriminer entre dimères biologiques et cristallographiques avec de bons résultats.

Dans une deuxième partie, je présenterai l'étude structurale d'une protéine tétramérique, la thymidylate synthase X. Cette protéine, découverte récemment, joue un rôle primordial dans la synthèse de l'ADN de la plupart des organismes procaryotes. C'est pourquoi la compréhension de son mécanisme au niveau moléculaire est importante : elle est une cible antibiotique de choix. Cette étude, menée en collaboration avec une équipe de biochimistes de l'Institut de Génétique et Microbiologie, a aussi permis de mieux comprendre le mécanisme de fonctionnement de cette protéine.

Première partie

Mise en place d'une fonction de score pour le docking protéine-protéine

Chapitre 1

Introduction

Sommaire

1.1 Structure des protéines

Une protéine est une macromolécule biologique constituée d'un enchaînement linéaire d'acides aminés reliés par une liaison peptidique. La protéine est le résultat de l'expression d'un gène, porté par l'ADN (acide désoxyribonucléique), qui est d'abord transcrit en ARN (acide ribonucléique) messager, lui-même traduit en protéine par le ribosome (voir figure 1.1).

Dans les conditions physiologiques, la protéine se replie, adoptant une conformation compacte spécifique. Ce repliement peut être spontané, ou bien se faire grâce à des protéines spécialisées : les chaperonnes. Dans certains cas, une maturation peut se produire par l'ajout de glucides ou bien par le clivage de certaines parties de la protéine. Ensuite la protéine, mature et repliée, est soit libérée dans le cytoplasme, soit dirigée vers une membrane, soit encore excrétée dans le milieu extérieur.

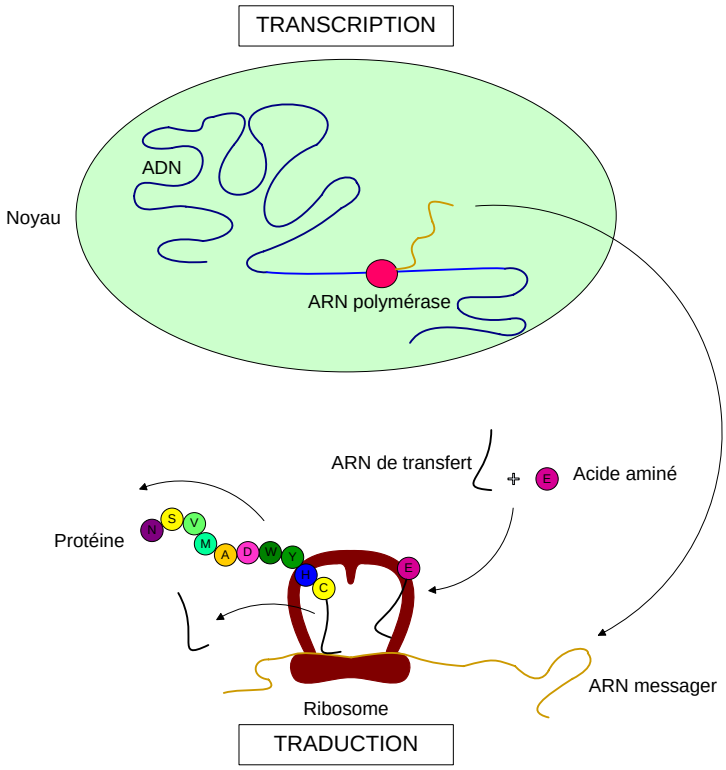


FIG. 1.1 – Du gène à la protéine. *Le gène (en bleu clair), appartenant à une molécule d'ADN, est transcrit en ARN messager (en beige) par un enzyme : l'ARN polymérase. Puis l'ARN messager est traduit en protéine dans le ribosome. À cette traduction participent notamment les ARN de transfert.*

Tout changement, qu'il soit dans l'enchaînement des acides aminés (par mutagenèse ou par modification chimique), ou dans les conditions du milieu (pH, force ionique, température, présence d'agents chimiques), peut modifier le repliement de la protéine, ses interactions et moduler son activité.

Le repliement des protéines est imposé par les interactions entre les différents acides aminés qui les composent, d'une part, et entre ces acides aminés et le solvant, d'autre part. Cependant, toutes les interactions ne sont pas parfaitement connues et, de plus, la complexité de ce phénomène est telle qu'il est à l'heure actuelle impossible d'en décrire le déroulement. La structure adoptée par une protéine et ses interactions avec différents partenaires peuvent être déterminés expérimentalement, mais, même avec l'essor des projets de génomique structurale qui ont mis en place des stratégies de résolution de structure à haut débit, cette détermination peut être longue et difficile. Étant donné le nombre de séquences de protéines et d'interactions potentielles connues actuellement, il n'est pas envisageable de déterminer expérimentalement toutes les structures correspondantes. Il faut donc se tourner vers la modélisation, qui tente de prévoir le repliement des protéines à partir des séquences, mais aussi, de prévoir la conformation du complexe à partir des

structures de partenaires, déterminées séparément.

On définit plusieurs niveaux d'organisation de la structure des protéines :
- la structure primaire, ou séquence, est l'enchaînement des acides aminés ;
- la structure secondaire est le résultat d'interactions à courte distance (essentiellement des liaisons hydrogènes entre atomes du squelette, qui ne dépendent qu'en partie de la nature des chaînes latérales des acides aminés impliqués). Certains segments de la protéine adoptent ainsi une conformation périodique d'angles dièdres successifs ;
- la structure tertiaire est la forme fonctionnelle repliée d'une chaîne protéique. Elle résulte de l'assemblage selon une topologie déterminée des structures secondaires ;
- la structure quaternaire est l'association de plusieurs chaînes protéiques (identiques ou non).

1.1.1 Les acides aminés

Un acide aminé est une molécule constituée d'un carbone asymétrique (appelé carbone α ou C_α) lié à un groupement carboxyle COOH, un groupement aminé NH_2, un hydrogène H et un radical R aussi appelé chaîne latérale.

Selon la conformation du carbone α, on parle d'acide aminé D ou L (voir figure 1.2).

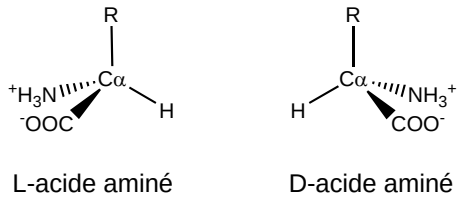


FIG. 1.2 – Formes D et L d'un acide aminé. *Ces formes, non superposables, sont l'image l'une de l'autre dans un miroir.*

Les appellations D et L ont été données à l'origine pour désigner les composés dextrogyres et lévogyres. Cependant, la conformation D ou L ne permet pas de prévoir les propriétés optiques d'un acide aminé. Chaque acide aminé doit son nom à la nature de son radical. Les acides aminés naturels les plus courants sont au nombre de 20, tous de configuration L (voir figure 1.3). Certaines protéines contiennent un petit nombre d'acides aminés modifiés tels que l'hydroxyproline ou la sélénocystéine.

La nature du radical (aussi appelé chaîne latérale) confère à chaque acide aminé des propriétés physico-chimiques particulières (hydrophobie, polarité, acidité, flexibilité, encombrement stérique...). Ces propriétés permettent le repliement protéique, garantissent la stabilité de la protéine, et permettent son activité biochimique.

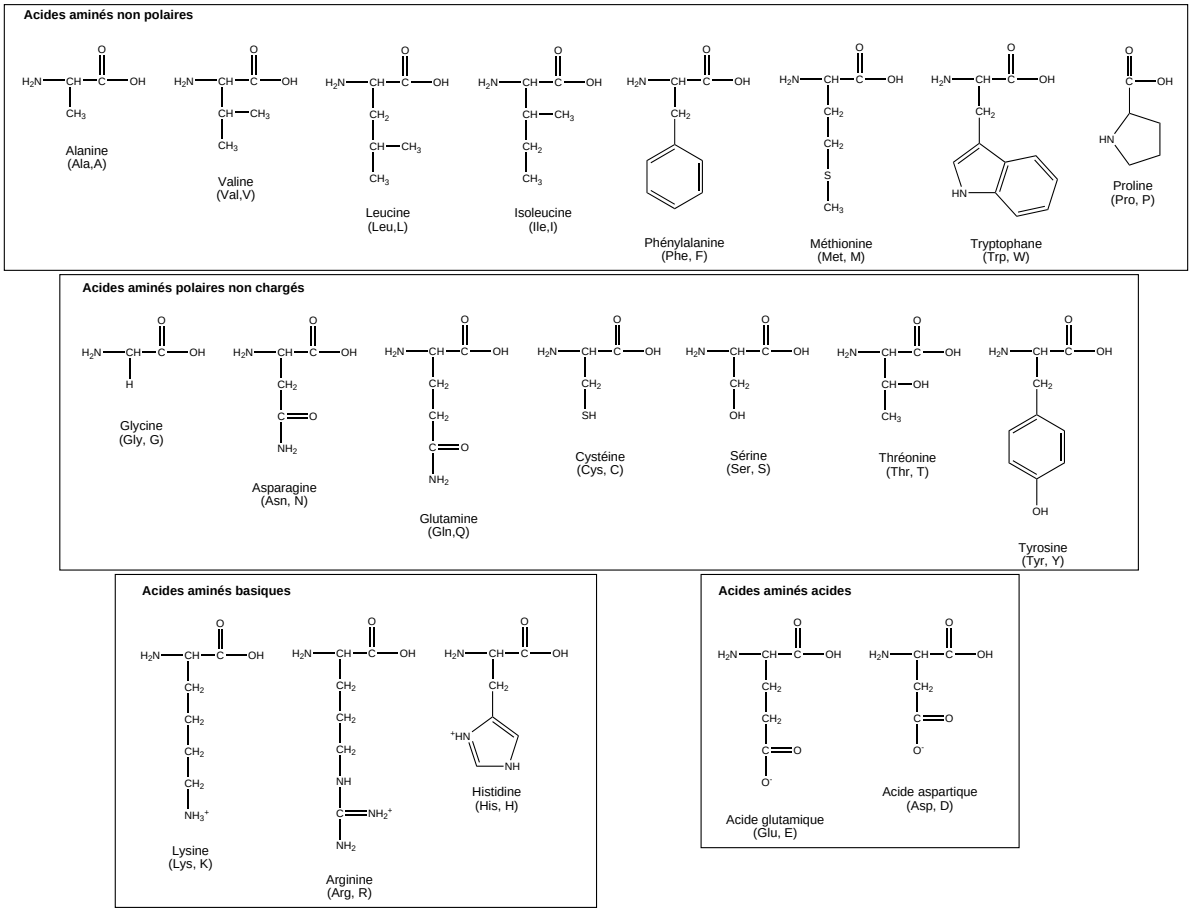


FIG. 1.3 – Les 20 acides aminés usuels

1.1.2 La structure primaire ou séquence

1.1.2.1 Définition

La structure primaire d'une protéine, également appelée chaîne polypeptidique, est l'enchaînement de ses acides aminés. Lors de la traduction, le groupement acide d'un acide aminé est lié au groupement aminé de l'acide aminé suivant ; cette liaison est appelée liaison peptidique. La nature de cette liaison impose certaines contraintes spatiales : en particulier, le C et le O du groupement carboxyle du premier acide aminé, ainsi que le N et le C_α de l'acide aminé suivant sont coplanaires (figure 1.4).

Les liaisons covalentes établies lors de la traduction ne sont généralement pas modifiées. Les exceptions peuvent être la coupure de certaines parties de la chaîne protéique, l'établissement de ponts disulfure entre deux cystéines ou encore la liaison avec des glucides lors de la maturation. Une protéine comprend entre 30 et 30 000 acides aminés, la moyenne se situant autour de 330 [240]. On appelle squelette de la protéine, la chaîne des N, C_α, C, et O de tous les acides aminés qui la constituent.

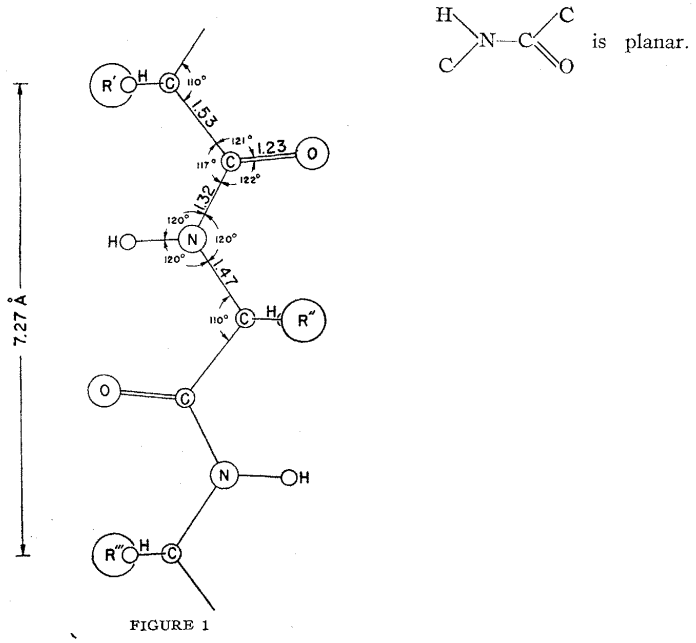


FIGURE 1
Dimensions of the polypeptide chain.

FIG. 1.4 – Géométrie de la liaison peptidique telle que décrite par Linus Pauling [171].

1.1.2.2 Détermination

La détermination de la séquence d'une protéine est une étape indispensable de son étude. En effet, la séquence donne non seulement des informations sur le repliement de la protéine (surtout s'il est possible de détecter des similitudes avec des protéines dont la structure est connue), mais également sur sa fonction et sa localisation cellulaire.

Cette détermination se fait généralement de manière indirecte, par traduction de la séquence du gène. De nombreuses techniques peuvent être utilisées pour le séquençage à grande échelle [97].

Cette méthode ne permet cependant pas de connaître les évènements post-traductionnels, en particulier les substitutions, les glycosylations ou les délétions d'une partie de la chaîne. Pour connaître directement la structure primaire d'une protéine de petite taille (moins de 150 acides aminés), il est possible d'utiliser la spectroscopie de masse [208]. La séquence peut également être découpée, par digestion enzymatique ou chimique, en tronçons de longueur inférieure à 30 acides aminés, dont la séquence est déterminée par micro-séquençage.

1.1.3 La structure secondaire

1.1.3.1 Définition

On appelle structure secondaire régulière une partie de la chaîne adoptant une conformation périodique. Ces structures sont stabilisées par des réseaux de liaisons hydrogène entre les acides aminés non voisins dans la chaîne polypeptidique.

Les premières structures secondaires ont été définies par Linus Pauling en 1951 [170, 171] : l'hélice α et le brin β. Ce sont deux structures secondaires très largement majoritaires dans les protéines. Ces deux organisations moléculaires minimisent les gênes stériques et les répulsions électrostatiques entre les chaînes latérales et maximisent le nombre de liaisons hydrogène, elles sont donc très largement favorisées [114].

Dans une protéine, en moyenne la moitié des acides aminés est impliquée dans des structures secondaires régulières. L'autre moitié des résidus se trouve dans des boucles qui relient entre elles les structures secondaires régulières. La longueur moyenne d'un brin β est de 5 acides aminés, celle d'une hélice α est de 12 acides aminés, celle d'une boucle est de 6 acides aminés [48].

L'hélice α est stabilisée par des liaisons hydrogène entre acides aminés appartenant à la même hélice, acides aminés distants seulement de 3,5 résidus en moyenne dans la chaîne polypeptidique (figure 1.5). L'hélice α est stable, même isolée.

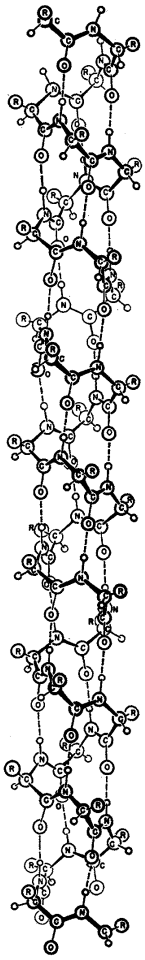


FIGURE 2
The helix with 3.7 residues per turn.

FIG. 1.5 – Hélice α telle que représentée dans l'article original de Linus Pauling [171].

Au contraire, le brin β n'est stable qu'associé à au moins un autre brin β, formant ainsi un feuillet (figure 1.6). Les brins d'un feuillet peuvent être tous dans le même sens,

on parle alors de brins parallèles, ou bien être positionnés alternativement dans un sens et dans l'autre, on parle alors de brins antiparallèles. Les liaisons hydrogène assurant la

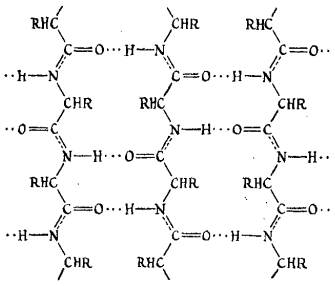


FIGURE 1

Diagrammatic representation of a hydrogen-bonded layer structure of polypeptide chains with alternate chains oppositely oriented.

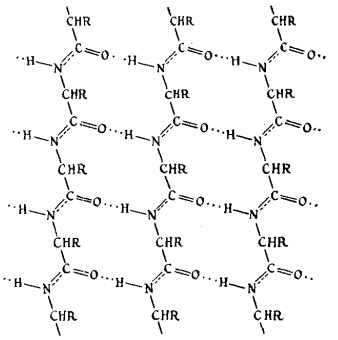


FIGURE 2

Diagrammatic representation of a hydrogen-bonded layer structure of polypeptide chains with all chains similarly oriented (the pleated sheet).

FIG. 1.6 – Brins β tels que représentés dans l'article original de Linus Pauling [170].

cohésion entre eux des acides aminés s'établissent entre acides aminés distants dans la séquence.

La quantité et l'agencement des structures secondaires régulières conduisent à classer les protéines en cinq catégories : tout-α, tout-β, α/β (alternance d'hélices α et de brins β), α+β (structures contenant des hélices α et des brins β mais n'alternant pas), et « autres » [158, 167].

1.1.3.2 Détermination à partir d'une solution de protéine

À partir d'une solution de protéine purifiée, il est possible d'estimer la composition globale en structures secondaires régulières (nombre d'acides aminés participant à des brins β ou à des hélices α) par des méthodes spectroscopiques :

- dichroïsme circulaire vibrationnel ou ultra-violet [169] ;
- spectroscopie infrarouge [179] ;
- spectroscopie Raman [5] ;
- analyse des déplacements chimiques en RMN (résonance magnétique nucléaire) [230].

Cependant, le seul moyen de déterminer avec précision la position dans la structure tertiaire de ces structures secondaires régulières reste la détermination complète de la structure tertiaire.

1.1.3.3 Détermination à partir des coordonnées atomiques

Même lorsqu'on dispose des coordonnées atomiques d'une protéine, il n'est pas évident d'attribuer les structures secondaires régulières. Bien évidemment, il ne s'agit plus ici de déterminer le nombre et la nature des structures secondaires régulières, mais plutôt de déterminer la position exacte de leurs extrémités dans la séquence. Il existe de nombreux

programmes permettant de réaliser l'attribution des structures secondaires, à savoir : dire à quel type de structure secondaire participe chaque acide aminé. La comparaison de ces programmes montre que les résultats obtenus par les différentes méthodes peuvent être assez différents au niveau des limites de chaque structure secondaire.

1.1.3.4 Prédiction

La prédiction des structures secondaires est une étape intéressante de l'étude d'une protéine. En effet, elle peut permettre d'émettre des hypothèses sur la nature du repliement, aider à localiser des résidus du site actif, ou encore donner une hypothèse quant à la localisation de la protéine dans la cellule (en particulier pour les protéines membranaires).

Il existe un certain nombre de logiciels de prédiction de structure secondaire, fondés sur des méthodes différentes [78, 191]. Les prédictions obtenues sont désormais exactes à plus de 75%, comme le montrent les résultats de l'expérience *CASP* (voir paragraphe 1.1.4.3 page 11).

1.1.4 La structure tertiaire

1.1.4.1 Définition

La structure tertiaire est la description du repliement d'une chaîne polypeptidique en sa forme fonctionnelle, ainsi que des liaisons covalentes apparues après la traduction (essentiellement les ponts disulfure), la présence éventuelle d'ions ou de cofacteurs plus complexes (hème, flavine adénine dinucléotide ou FAD...). Les structures tertiaires sont très variées et très complexes.

Les chaînes polypeptidiques de grande taille (plus de 200 acides aminés) se replient souvent en plusieurs régions fonctionnelles. On parle de domaine si ces unités fonctionnelles adoptent un repliement stable lorsqu'elles sont isolées.

Lorsque deux séquences protéiques présentent plus de 30% d'identité de séquence, elles adoptent le même repliement [45, 194]. En dessous de ce seuil, il est difficile de prévoir, par les méthodes classiques d'alignement de séquence, si deux protéines vont adopter la même structure tertiaire. De plus, certaines protéines adoptent des repliements similaires sans présenter d'identité de séquence détectable ; c'est le cas notamment de la superfamille des immunoglobulines [90].

Le repliement repose principalement sur des interactions à courte distance. Ces interactions ont lieu, d'une part, entre les acides aminés enfouis dans la protéine, et, d'autre part, entre les acides aminés de la surface et les molécules du solvant [190]. Ces interactions sont des liaisons hydrogène, des ponts salins ou des liaisons de type Van der Waals.

1.1.4.2 Détermination

La première structure de protéine résolue a été celle de la myoglobine [118] par cristallographie aux rayons X. À l'heure actuelle, la *Protein Data Bank* (PDB) [14, 16], banque de données des structures tridimensionnelles des protéines, contient plus de 33 000 fichiers, dont environ 28 000 correspondent à des structures résolues par cristallographie et 5 000 à des structures résolues par RMN (dans sa version *PDB 2004 archives release*

#1). D'autres méthodes de résolution de structure peuvent aussi être utilisées, mais elles restent pour l'instant moins efficaces.

Ces méthodes, même si leurs performances se sont beaucoup améliorées, en particulier avec l'apparition des projets de génomique structurale, restent tributaires de conditions expérimentales restrictives. La cristallographie nécessite l'obtention de cristaux diffractants, ce qui demande beaucoup de matériel et de travail. Quant à la RMN, même si la contrainte du cristal est supprimée, elle ne peut s'appliquer que sur des protéines relativement petites (moins de 300 résidus) et il faut obtenir une quantité importante de solution de protéine pure à plus de 95%. Étant donné le nombre de séquences connues à l'heure actuelle, il n'est donc pas envisageable de résoudre toutes les structures correspondantes.

Par exemple pour la cristallographie X, selon la protéine et la qualité du cristal, on connaît la structure avec une résolution plus ou moins bonne. À basse résolution (supérieure à 3 Å), on connaît le squelette de la protéine et les structures secondaires. À moyenne résolution, on peut observer les interactions entre acides aminés, en particulier les liaisons hydrogène et les interactions de type Van der Waals. A haute résolution (moins de 1,5 Å), on peut déterminer avec précision longueurs et angles des liaisons, l'hydratation, et les mouvements atomiques autour des positions d'équilibre.

1.1.4.3 Prédiction

Modélisation par homologie Lorsqu'on peut établir une similitude entre la séquence dont on cherche la structure et une séquence dont la structure tridimensionnelle est connue, il est possible de construire un modèle de la structure recherchée.

Un modèle obtenu de cette manière est d'autant plus précis que l'identité de séquence entre le support et la séquence à modéliser est forte. Pour de faibles taux d'identité, on ne connaît avec précision, dans le modèle, que les acides aminés strictement conservés, et les parties ne comportant pas de longues insertions/délétions. Le modèle obtenu n'est donc pas l'équivalent d'une structure déterminée par des méthodes physiques. Cependant, il rend souvent compte du comportement du site actif ou encore des parties de la protéine nécessaires à son repliement ou à son interaction avec des partenaires.

Les méthodes d'enfilage Les méthodes d'enfilage, ou *threading*, permettent de tester la compatibilité d'une séquence avec un repliement [211]. Dans ce cas, pour une séquence donnée, on cherche parmi les structures connues, celle qui est la plus compatible avec la séquence dont on dispose.

La modélisation *ab initio* La finalité des techniques de modélisation *ab initio* est de prédire la structure d'une protéine à partir de sa seule séquence. De nombreux modèles de calculs sont utilisés, faisant appel par exemple à la dynamique moléculaire. Mais, même si les progrès sont conséquents, les résultats sont très variables, comme l'atteste l'expérience *CASP* [24] ou l'état du projet *folding@home* [1].

[1] `http://folding.stanford.edu/`

Évaluation des prédictions : l'expérience *CASP*[2] L'expérience *CASP* (*Critical Assessment of Methods of Protein Structure Prediction*), qui a lieu tous les deux ans depuis 1994, a pour objectif de tester les méthodes de prédiction de structure. Des protéines, dont la structure vient d'être résolue mais pas encore publiée, sont proposées aux prédicteurs. Ceux-ci doivent tenter de prédire, selon la catégorie, la structure *ab initio*, la structure par homologie ou la structure secondaire. Les dernières évaluations des prédictions [55, 123, 154] montrent d'importants progrès dans la prédiction de structure *ab initio* (voir figure 1.7).

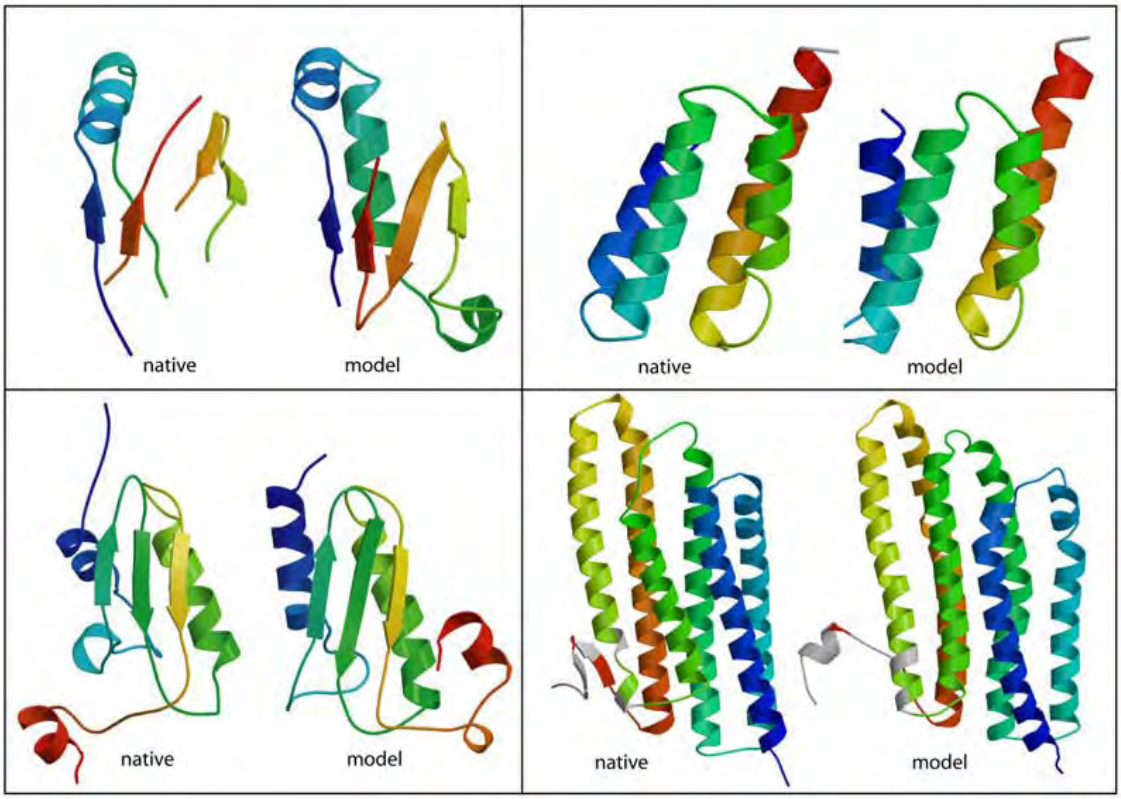


FIG. 1.7 – Prédictions issues de la dernière session de *CASP*. *Prédictions de structures obtenues à l'aide du logiciel Rosetta [26] pour* CASP6. *Image originale en couverture du journal* PROTEINS : Structure, Function, and Bioinformatics*, volume 61 du 26 septembre 2005.*

1.1.5 La structure quaternaire

1.1.5.1 Définition

La structure quaternaire est la géométrie de l'association de plusieurs sous-unités protéiques. Certaines protéines, comme la thymidylate synthase X (voir deuxième partie) ne sont fonctionnelles que sous forme d'oligomères. Il existe des oligomères formés de sous-unités identiques, comme par exemple le tétramère de la thymidylate synthase X, et des

[2] `http://predictioncenter.org/`

oligomères réunissant des sous-unités différentes, comme les histones. On parlera alors de complexes. Enfin, certaines protéines forment des polymères, constitués d'un très grand nombre de sous-unités, comme les polymères actine/myosine dans les muscles.

L'association de ces sous-unités est stabilisée par des interactions à courte distance, similaires à celles qui assurent la stabilité de la structure tertiaire (essentiellement des liaisons hydrogène, des ponts salins et des interactions hydrophobes) [46, 109].

1.1.5.2 Détermination

L'existence d'oligomères de chaînes protéiques, qu'elles soient ou non identiques, peut être déterminée par filtration sur gel ou la centrifugation analytique par exemple, mais aussi par des méthodes de biochimie et de biologie moléculaire plus poussées et qui peuvent être utilisées de manière systématique, telles que l'analyse double-hybride [100], l'analyse par TAP-tag (ou FLAP-tag) couplée à la spectrométrie de masse [82, 96]. La géométrie de l'association peut être déterminée à basse résolution par chromatographie sur gel, diffusion des rayons X ou des neutrons aux petits angles, ou encore par microscopie électronique.

La connaissance de l'interaction au niveau des acides aminés peut se faire, soit directement par la détermination de la structure par cristallographie aux rayons X, soit par l'étude des interactions par RMN, ou encore indirectement, par mutagenèse dirigée ou modification chimique sélective des chaînes latérales de certains acides aminés. Mais, en plus des contraintes associées aux deux méthodes vues précédemment, s'ajoutent les contraintes inhérentes aux complexes, telles que la taille, mais aussi, et surtout, leur instabilité. En effet, pour pouvoir être étudié d'un point de vue structural, un complexe doit être stable dans les conditions requises. Or, de très nombreux complexes sont transitoires. Ainsi, même s'il est désormais possible d'obtenir la structure de nombreuses protéines isolées de plus en plus rapidement, la résolution des structures de complexes reste difficile.

1.1.5.3 Prédiction

Le premier modèle de complexe protéine-protéine (trypsine/inhibiteur) a été réalisé en 1972 [19]. C'est en 1978 qu'est apparu le premier algorithme d'amarrage [231]. Les procédures d'amarrage utilisent les coordonnées atomiques des deux protéines partenaires, génèrent un grand nombre de conformations et leur attribuent un score [232]. Cette modélisation est en général assimilée à la recherche de modes d'association complémentaires entre deux molécules de forme prédéfinie. Un certain degré de flexibilité peut parfois être pris en compte, mais en général, l'amarrage protéine-protéine est principalement envisagé dans une approche d'association de corps rigides.

Ces méthodes s'appliquent à des protéines différentes, mais peuvent aussi être envisagées pour déterminer l'état d'oligomérisation d'une protéine. Elles peuvent prendre en compte les symétries connues comme pour les protéines virales [13, 49, 175, 196], mais aussi utiliser des études plus fines des interfaces [18, 9, 176, 241].

1.2 Les complexes protéine-protéine

1.2.1 Fonctions

Au niveau moléculaire, la fonction d'une protéine est souvent subordonnée à l'interaction avec un certain nombre de partenaires. Les complexes interviennent à de nombreux niveaux et la compréhension de leur mécanisme de formation/association permet de mieux comprendre de nombreux processus. Pour se rendre compte de leur importance, on peut citer des assemblages tels que le ribosome, les anticorps/antigène, les capsides virales ou encore les microtubules. Ainsi, la fonction d'une protéine ne peut être envisagée sans tenir compte des interactions.

1.2.2 Détection expérimentale

Les interactions protéine-protéine sont présentes partout et en grand nombre, c'est pourquoi de nouvelles méthodes expérimentales d'analyse systématique sont développées [107]. Deux types sont présentés dans la suite, les méthodes d'analyse par double-hybride et celles utilisant des marqueurs.

1.2.2.1 Le double-hybride sur la levure

La première méthode utilisée pour étudier les interactions protéine-protéine à grande échelle dans la levure a été l'analyse par double hybride. Cette technique, mise au point en 1989, permet la détection indirecte de l'interaction, car celle-ci induit la formation d'un complexe moléculaire activant un gène rapporteur [72] (figure 1.8). Cependant, dans cette détection, le nombre de faux positifs (interactions détectées mais non présentes) et de faux négatifs (interactions présentes non détectées) est très important. C'est donc une méthode relativement peu fiable, à moins de refaire un grand nombre de fois ces expériences, en plus d'expériences complémentaires, ce qui est relativement coûteux et long dans une approche génomique.

De plus, cette méthode ne peut détecter dans sa forme originelle que des complexes binaires. Or, la détection et la caractérisation de complexes multiprotéiques est très importante.

Deux études sur la levure utilisent le double-hybride pour la détection systématique [100, 216]. Il est toutefois très difficile de comparer ces études entre elles, en raison principlament des problèmes de fiabilité dus aux contraintes expérimentales.

1.2.2.2 Utilisation de marqueurs (*TAP-tag* et *FLAP-tag*)

Deux autres études ont été menées sur la levure *S. cerevisiae* [82, 96] pour identifier les complexes cellulaires et comprendre leur rôle dans la cellule eucaryote. Des centaines de séquences codantes de levure ont été fusionnées à des cassettes d'ADN codant pour des marqueurs de purification. Puis, les souches de levure ont été cultivées, chacune exprimant une protéine cible marquée, et soumises à une procédure dans laquelle les complexes entiers, contenant la protéine marquée, ont été purifiés. Ensuite, les complexes ont été

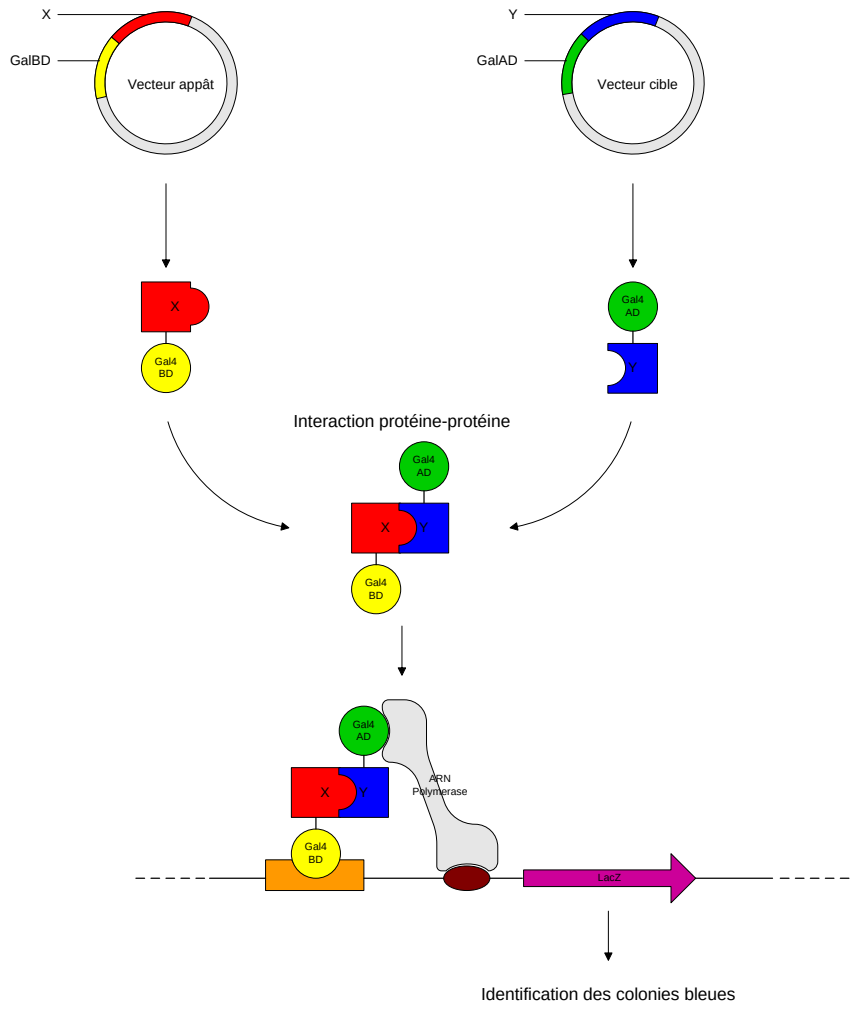


FIG. 1.8 – Schéma de principe de détection des interactions protéine-protéine par double-hybride chez la levure. *La protéine Gal4 est l'activateur naturel des différents gènes intervenant dans le métabolisme du galactose. Elle agit en se fixant sur des séquences appelées UASG (*Upstream Activating Sequence GAL*) qui régulent la transcription. Les protéines étudiées (X et Y), partenaires potentiels d'interaction, sont fusionnées, l'une au domaine de fixation de Gal4 sur l'ADN (domaine DBD ou* DNA binding domain*), et l'autre au domaine de Gal4 activant la transcription (domaine AC ou* Activation Domain*). C'est ce qui donne à ce système le nom de double-hybride. Quand il y a interaction en X et Y, les domaines DBD (*DNA Binding Domain*) et AD (*Activation domain*) sont associés et forment un activateur de transcription DBD-X/Y-AD. C'est cet activateur hybride qui va se lier à l'ADN au niveau des séquences qui contrôlent le gène rapporteur (les séquences UASG), permettant la transcription du gène par l'ARN polymérase II. Il suffit ensuite d'observer le produit du gène rapporteur, pour voir si un complexe s'est formé entre les protéines X et Y. Souvent, le gène LacZ est inséré dans l'ADN de la levure juste après le promoteur Gal4, de façon à ce que, si l'interaction a lieu, le gène LacZ, qui code pour la β-galactosidase, soit produit. Sur un substrat approprié, la β-galactosidase devient bleue, ce qui permet de déterminer simplement si l'interaction a lieu.*

fractionnés par électrophorèse sur gel et leurs composants identifiés par spectrométrie de masse.

Il est très difficile de comparer les résultats obtenus par ces études car les jeux utilisés ne sont pas identiques et le protéome complet de la levure n'a pas pu être analysé. Globalement, ces études donnent des résultats en accord avec celles réalisées précédemment, mais dans le détail, les résultats et la complétude des données ne permettent pas de conclure.

De plus, il est aussi pour les mêmes raisons difficile de comparer les études utilisant des marqueurs et les études de double-hybride présentées précédemment.

L'ensemble de ces études expérimentales permet de prédire environ 15 000 complexes potentiels pour le génome de la levure. Parmi ces 15 000, beaucoup s'avèreront être de faux positifs et il est certain qu'il existe également un grand nombre de faux négatifs.

1.2.3 Les méthodes d'amarrage protéine-protéine

1.2.3.1 Le problème

Le but des méthodes d'amarrage protéine-protéine est de prédire la structure d'un complexe à partir des structures ou modèles des partenaires isolés (figure 1.9). Le problème

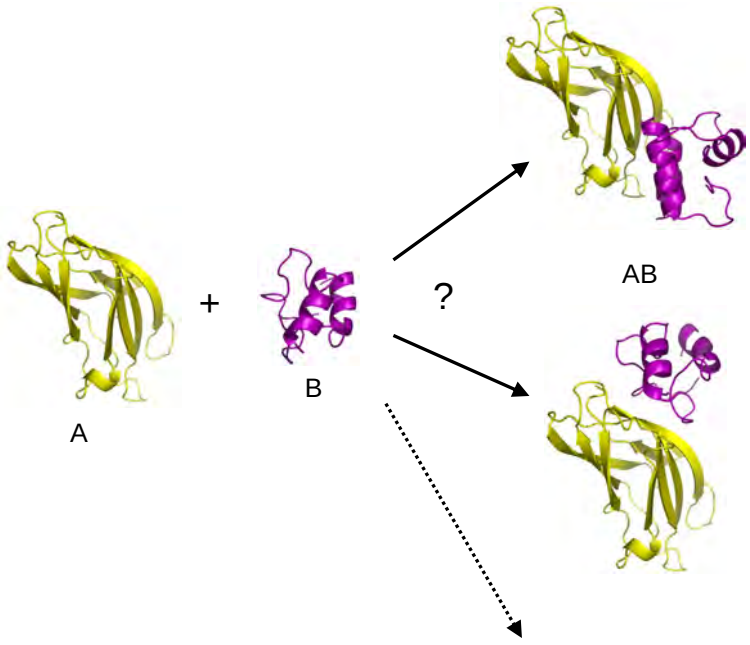


FIG. 1.9 – Le problème de l'amarrage. *Comment associer la protéine A et la protéine B ? Des configurations AB obtenues, laquelle est susceptible d'exister* in vivo *?*

se divise en deux étapes : d'abord, on explore l'espace pour obtenir toutes les conformations possibles et ensuite, on trie ces conformations en espérant classer en premier la conformation native observée expérimentalement.

Avec une approximation de corps rigides, si on considère la protéine comme une sphère de 15 Å de rayon à la surface de laquelle les propriétés atomiques sont décrites sur une grille d'un Angström, une recherche systématique présente 10^9 modes distincts d'association [53]. La question est ensuite de déterminer, parmi ces modes d'association, lequel est le mode natif.

Pour pouvoir accéder aux changements de conformation et aux mouvements des chaînes latérales, le modèle doit être de type « soft », c'est-à-dire que les molécules doivent pouvoir légèrement s'interpénétrer et on doit considérer les molécules comme des ensembles de sphères articulées. Ainsi, il est possible de traiter aussi bien les molécules issues de résolution de structures de protéines seules, c'est-à-dire non-liées (*unbound*), ou complexées c'est-à-dire liées (*bound*).

1.2.3.2 Les algorithmes

Le premier algorithme, inventé par Shoshana Wodak et Joël Janin [108, 231] à partir des travaux de Cyrus Levinthal [130] réalise une recherche de l'espace sur six degrés de liberté (5 rotations et une translation) pour amener les deux molécules en contact une fois leur orientation fixée, et attribue un score simple en fonction de la surface de contact. Pour gagner du temps sur le calcul de la surface, une approximation à partir du modèle de Levitt [131] est réalisée. Cet algorithme a été amélioré en 1991 à l'aide d'une minimisation d'énergie [42].

D'autres types d'algorithmes, utilisant la complémentarité de surface, ont été mis en œuvre à partir d'une description en points critiques définis comme « trous et bosses » (*knobs and holes*) [53, 128, 235], les solutions données correspondant à une concordance de groupes de quatre points critiques, laquelle est identifiée grâce à une triangulation de surface comme définie par M. Connolly en 1985 [52]. Cette méthode a été beaucoup améliorée en 1991 par H. Wang, avec la modélisation de la surface à l'aide d'une grille [222].

En 1992, un programme utilisant ces grilles pour les petites molécules a été modifié par I. Kuntz et ses collaborateurs, pour s'appliquer aux complexes protéine-protéine et a permis d'obtenir de bons résultats [150, 202] tout en générant de nombreux faux-positifs.

Des algorithmes de vision par ordinateur (*computer vision*) à partir de hachage géométrique ont ensuite étendu la méthode des « trous et bosses ». En 1993 a été développé un algorithme qui fait correspondre des propriétés de surface à partir de triplets de points critiques qui sont stockés dans des tables de hachage [75, 141]. Cette méthode, très efficace pour les molécules de type lié, est très sensible aux faibles variations de surface, ce qui la rend rapidement inefficace pour les molécules de type non-lié.

1.2.3.3 La transformation de Fourier

Parmi toutes les méthodes de complémentarité de surface, celle utilisant la transformation de Fourier rapide (*Fast Fourier Transform* ou FFT), apparue dès 1991 [110], est l'une des plus simples et des plus utilisées [11, 33, 39, 40, 117, 124, 153, 206, 229]. Une grille cubique est tracée, et, à chaque point, on attribue un poids qui est négatif et im-

portant si le point est situé à l'intérieur de la protéine A, nul s'il est à l'extérieur et 1 s'il est proche de la surface ; on fait de même pour la protéine B.

Le produit est donc important et positif (défavorable) si les deux volumes moléculaires s'interpénètrent, et négatif (favorable) pour les points qui appartiennent à la surface d'une molécule et au volume de l'autre. Lorsque la molécule A est translatée par rapport à la molécule B, le score peut être rapidement calculé par transformation de Fourier rapide (FFT), si la grille de A est identique à la grille de B. La grille doit donc être redéfinie à chaque nouvelle orientation pour que la recherche soit complète.

Cette approche présente de nombreux avantages : les poids peuvent contenir des informations sur les propriétés physico-chimiques de la surface, et la résolution peut être ajustée en limitant le nombre de termes de Fourier calculés dans la somme.

Les résultats obtenus par cette méthode sont relativement bons [12, 37, 145], mais le temps de calcul associé est trop important pour une approche à grande échelle.

1.2.3.4 Nouveaux algorithmes de docking et partitionnement du problème

Ces dernières années, en particulier grâce à l'expérience de docking CAPRI (*Critical Assessment of PRediction of Interactions*) [102, 103, 104, 105, 233], plusieurs nouvelles méthodes ont vu le jour [205]. Cette expérience est un test à l'aveugle des algorithmes de docking protéine-protéine qui doivent prédire le mode d'association de deux protéines à partir de leur structure tridimensionnelle. La structure du complexe, résolue expérimentalement, n'est dévoilée aux participants et publiée qu'à l'issue des soumissions.

Les nouvelles méthodes d'amarrage utilisent des techniques très variées telles que le hachage géométrique [74, 99, 163, 164, 165, 193, 197, 234], les algorithmes génétiques [80], les harmoniques sphériques [187], la dynamique moléculaire [31, 121, 210], la minimisation Monte-Carlo [87, 199], ou encore des méthodes de minimisation d'énergie ou de détection d'interfaces dirigées par des données biologiques [61, 151, 219] (voir paragraphe 2.1 page 23).

Le domaine de recherche a beaucoup progressé et l'une des conclusions de cette expérience est que l'on dispose à l'heure actuelle d'algorithmes de recherche de complémentarité de surfaces performants [149]. Cependant, la deuxième étape du processus d'amarrage, à savoir le tri des configurations putatives obtenues par une fonction de score, reste à améliorer, car la seule méthode réellement performante à l'heure actuelle est l'expertise humaine. Les fonctions énergétiques classiquement utilisées ayant montré leurs limites [41, 50, 69, 70, 91, 94, 132], de nouvelles fonctions de score statistiques sont apparues. Essentiellement basées sur les propriétés physico-chimiques des atomes, elles ont tout d'abord été utilisées pour le repliement et l'amarrage de petites molécules, puis adaptées à l'amarrage protéine-protéine [58, 237, 238, 239].

1.3 La tessellation de Voronoï et ses dérivés dans l'étude des complexes protéine-protéine

La première utilisation connue de cette construction est la modélisation de la répartition de l'épidémie de choléra de Londres par John Snow en 1854, dans laquelle est

démontrée que la fontaine au centre de l'épidémie est celle de Broad Street, en plein coeur du quartier de Soho. Depuis lors, les applications utilisant cette construction sont nombreuses : en météorologie d'abord, par A.H. Thiessen, en cristallographie par F. Seitz et E. Wigner, qui ont aussi donné leur nom à cette construction ; mais aussi en physiologie (analyse de la répartition des capillaires dans les muscles), métallurgie (modélisation de la croissance des grains dans les films métalliques), robotique (recherche de chemin en présence d'obstacles) et bien d'autres.

La tessellation de Voronoï, ainsi que les autres tessellations qui en ont été dérivées (figure 1.10), sont aussi beaucoup utilisées en biologie, où elles permettent de nombreuses représentations des structures des protéines [178].

Étant donné un ensemble de points appelés centroïdes, la tessellation de Voronoï divise l'espace en maximum autant de régions qu'il y a de points (voir paragraphe 2.2 page 28). Chaque région, appelée cellule de Voronoï, est un polyèdre qui peut être considéré comme la zone d'influence du point autour duquel est tracée la cellule.

1.3.1 Structure des protéines et méthodes dérivées de la tessellation de Voronoï

1.3.1.1 Constructions

Dans le cadre de l'analyse structurale des protéines, la tessellation de Voronoï a été utilisée pour la première fois par Richards en 1974 [185] pour évaluer, dans une protéine globulaire, les volumes des atomes, définis par les volumes de leurs polyèdres de Voronoï. Dans cette étude, Richards est le premier à proposer une solution à deux problèmes que l'on retrouve dans toutes les études qui utilisent cette construction. Tout d'abord, les atomes exposés au solvant ayant peu de voisins, leurs cellules de Voronoï sont grandes et ont un volume très grand, peu représentatif de leurs propriétés. Ensuite, cette construction considère tous les atomes comme équivalents, sans tenir compte de leur nature chimique.

Pour résoudre le premier problème, Richards a placé des molécules d'eau sur un réseau cubique entourant la protéine et a relaxé leurs positions. Cette méthode a été ensuite affinée par Gerstein et ses collaborateurs [84, 85, 212, 213, 214]. D'autres méthodes ont été proposées telles que :

- prendre en considération uniquement les atomes ayant une cellule de Voronoï de volume « raisonnable » [181] ;
- placer les molécules d'eau en utilisant la dynamique moléculaire [30, 35] ;
- utiliser une représentation d'union de sphères [146] ;
- utiliser un mélange entre la représentation en diagramme de puissance et la représentation en union de sphères (figure 1.10).

Pour résoudre le problème des poids des atomes, Richards a proposé d'introduire des poids lors du placement des plans dans la construction de Voronoï. Cette méthode, appelée *méthode B de Richards*, a été très utilisée. Elle manque de rigueur mathématique car on trouve des volumes non attribués entre les cellules, l'intersection des plans n'étant plus réduite à un point. Cependant, Richards a montré que ce volume mort, bien que non nul, est petit en comparaison des volumes des atomes. Cette méthode a été de nombreuses fois améliorée [73, 186], jusqu'à utiliser le diagramme de Laguerre [83] (voir paragraphe

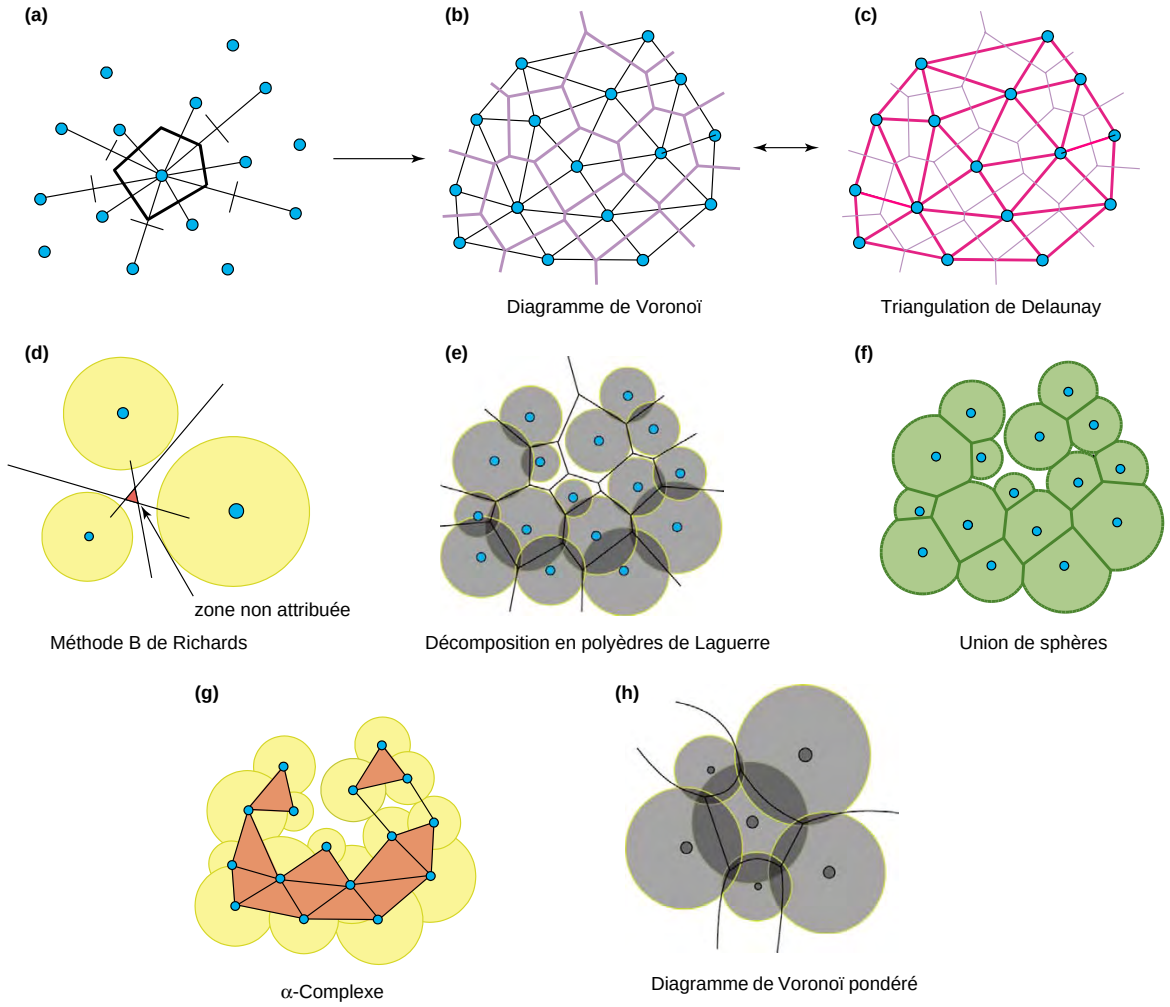


FIG. 1.10 – Tessellation de Voronoï et constructions dérivées *(a) Construction d'une cellule de Voronoï : on trace la médiatrice entre un point donné et chacun des autres points et ensuite, on considère le plus petit polyèdre défini par ces médiatrices ; c'est la cellule de Voronoï de ce même point. (b) On obtient le diagramme de Voronoï (en violet) en répétant l'opération pour tous les points de l'ensemble. (c) La triangulation de Delaunay contient les arêtes roses et les triangles ainsi définis. C'est le dual du diagramme de Voronoï. (d) Dans la méthode de Richards, on ne considère pas la médiatrice, mais on définit une droite perpendiculaire au segment qui coupe celui-ci en fonction des poids attribués à chacun des atomes. Cela laisse une zone non attribuée. (e) Si on remplace les droites précédentes par les plans radicaux des sphères, on obtient à nouveau un pavage de l'espace : le diagramme de puissance ou tessellation de Laguerre. (f) L'intersection du diagramme de Laguerre et des sphères donne ce qu'on appelle l'union des sphères. (g) On définit une région restreinte comme une boule restreinte à sa région de Voronoï. L'α-complexe correspond alors aux arêtes et aux triangles définis par l'intersection de deux ou trois régions restreintes. L'α-shape est le domaine de l'α-complexe. (h) La surface de division d'un diagramme de Voronoï pondéré est définie par l'ensemble des points dont la distance aux deux points de référence est égale au rayon de la sphère correspondante plus une constante. Cette surface n'est pas plane, mais le diagramme correspondant est un pavage de l'espace.*

2.2.1.4 page 31) ou le diagramme de Voronoï dit pondéré [54, 86], dans lequel les faces des cellules ne sont plus planes (figure 1.10).

Une analyse formelle de toutes ces applications a été réalisée par Edelsbrunner et ses collaborateurs [65, 66, 67, 136, 137]. En plus des utilisations des tessellations de Voronoï/Delaunay/Laguerre, ils mettent en place la notion d'α-shape pour les protéines : c'est un sous-ensemble des segments issus de la tessellation de Delaunay qui sont contenus dans le volume de la protéine (figure 1.10). Cela permet de modéliser l'intérieur de la protéine et de détecter les vides et les cavités [138].

1.3.1.2 Mesures

Toutes ces constructions ont permis de montrer que la tessellation de Voronoï est un bon modèle mathématique de la structure des protéines. Elle permet en particulier de montrer que les protéines sont des objets compacts, c'est-à-dire que la densité d'atomes à l'intérieur d'une protéine est comparable à celle observée dans les cristaux de petites molécules [85, 93]. De même, une analyse où les points considérés pour la construction (appelés centroïdes) sont les centres géométriques des acides aminés a permis de montrer que les protéines sont aussi des objets compacts au sens des modèles classiques des matières condensées en physique [4, 207].

Elle a aussi servi à l'analyse des cavités dans les structures [10, 65, 135, 172], à l'étude de propriétés mécaniques des protéines [120, 168, 195], à la mise en place de potentiels empiriques pour l'affinement de modèles structuraux [22, 32, 79, 122, 133, 139, 157, 226, 242], ou encore à la détection des hélices transmembranaires [1].

De telles méthodes ont également été utilisées pour détecter les cavités des protéines susceptibles d'interagir, mais aussi pour ajuster les ligands dans les poches ou encore étudier les interactions protéine-ADN [23, 25, 56, 160, 161]. Des études utilisant le modèle B de Richards ou la construction de Laguerre ont montré qu'à l'interface protéine-ADN et protéine-protéine, la densité de l'empilement est la même qu'à l'intérieur de la protéine pour la grande majorité des complexes [144].

1.3.2 Amarrage et tessellation de Voronoï

Dans ce travail, j'ai utilisé un modèle basé sur la tessellation de Voronoï pour étudier la structure de complexes protéine-protéine. L'objectif de cette étude était de mettre au point une méthode d'amarrage protéine-protéine efficace, rapide, et n'utilisant pas l'information biologique, même quand celle-ci est disponible. Nous avons choisi cette représentation, car, en plus d'être une bonne représentation des structures protéiques et de l'empilement, elle est très maniable d'un point de vue informatique. Dans un premier temps, nous nous sommes limités à la seconde étape de l'amarrage : trier les solutions issues de l'exploration afin d'en extraire la bonne conformation. Pour cela, nous avons étudié les complexes de structures connues et des conformations non-natives à l'aide de la tessellation de Voronoï. Ensuite, à l'aide de méthodes d'apprentissage statistiques, nous avons construit une fonction de score permettant de distinguer les complexes natifs des complexes non-natifs.

Le fait de ne pas utiliser l'information biologique peut sembler illogique. Cependant, le temps nécessaire à la recherche et à la prise en compte de cette information empêcherait l'approche systématique qui est l'un de nos objectifs.

Chapitre 2

Méthodes et logiciels

Sommaire

2.1 Algorithmes d'amarrage

Dans ce travail, nous nous sommes intéressés uniquement à la fonction de score. Pour générer les conformations possibles des différents complexes, nous avons utilisé deux algorithmes différents. Le premier algorithme appelé *DOCK*, est aussi le premier algorithme

à avoir été mis au point en 1985 [108]. Il permet une exploration systématique sans tenir compte des informations biologiques. Le second, *HADDOCK* [61], est un des algorithmes plus récents, et utilise les informations biologiques dont on dispose sur les structures pour guider la recherche. Nous ne détaillerons pas ici les algorithmes basés sur la transformée de Fourier que nous n'avons pas pu utiliser, soit en raison de durée du calcul, soit en raison de problèmes stériques dans leurs résultats.

2.1.1 L'exploration géométrique : *DOCK*

2.1.1.1 Représentation des partenaires à amarrer et des interactions

Les partenaires potentiels sont représentés en utilisant le modèle simplifié de Levitt [131]. Dans ce modèle, chaque résidu est réduit à un simple centre d'interaction en remplaçant chaque résidu par une sphère centrée sur le centre de masse de sa chaîne latérale et du carbone α. Les rayons des sphères associés à chaque type de résidu sont ceux décrits dans une étude précédente de S. Wodak et J. Janin [231].

Les potentiels énergétiques utilisés sont ceux mis au point par M. Levitt en 1976 [131] pour des interactions non covalentes entre représentations simplifiées de résidus. Ils permettent de représenter les répulsions à courte distance entre les sphères. Ce sont des fonctions 6-8 de type « mou » du rapport x de la somme des rayons des sphères et de la distance centre à centre :

$$E_{ij} = \epsilon_{ij}(1 + 3x^8 - 4x^6) \quad \text{avec} \quad x = \frac{r_i + r_j}{d_ij} \quad \text{et} \quad x < 1 \tag{2.1}$$

E_{ij} est donc strictement positif pour des distances centre à centre d_{ij} plus petites que la somme des rayons et nulle pour des distances plus grandes. Sa dérivée par rapport à d_{ij} est continue, on peut donc l'utiliser comme potentiel. Le paramètre ϵ_{ij} a été calibré dans une étude précédente de S. Wodak et J. Janin [231]. Les contributions des interactions non covalentes décrites par l'équation précédente sont additionnées sur toutes les paires des résidus des deux partenaires pour donner une contribution répulsive E_{NB} à l'énergie libre d'association.

Les contributions qui permettent de stabiliser le complexe sont représentées de façon différente. Tout d'abord, la surface de l'interface entre les partenaires est définie comme la réduction de la surface accessible au solvant due à l'association :

$$S_I = S_1 + S_2 - A_{12} \tag{2.2}$$

Avec S_1 et S_2 surfaces accessibles au solvant de chacun des partenaires pris séparément et A_{12}, surface du complexe accessible au solvant. Pour ramener la composante d'énergie attractive correspondant à la surface de l'interface S_I à la même échelle que E_{NB}, S. Wodak et J. Janin utilisent une corrélation linéaire établie par R.B. Herrmann [95] et C. Chothia [43, 44] entre les énergies libres hydrophobes et les aires accessibles au solvant. Ainsi, on utilise :

$$E_S = -\gamma S_I \tag{2.3}$$

comme contribution attractive à l'énergie libre d'association (avec $\gamma = 25$ cal.mol^{-1}.Å^{-2} [44]).

2.1.1.2 Le système de coordonnées angulaires

La position du partenaire 1 par rapport au partenaire 2 est déterminée par 6 paramètres. Selon l'étude de C. Levinthal et collaborateurs [130], *DOCK* travaille sur 5 angles et une distance (figure 2.1).

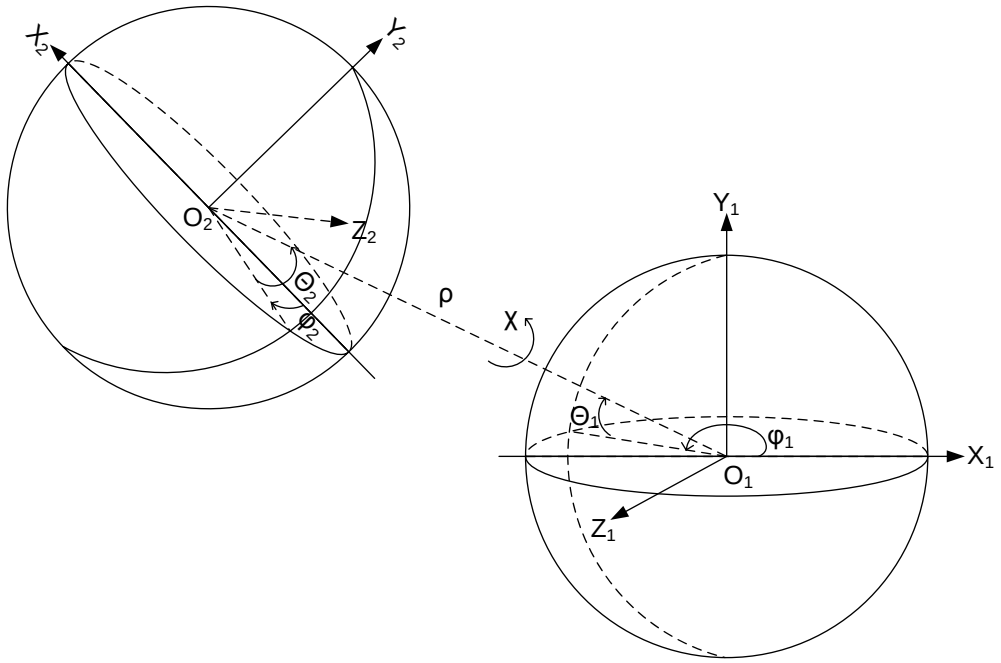


FIG. 2.1 – Le système de coordonnées utilisé par *DOCK*. θ_1 *et* φ_1 *sont la latitude et la longitude du centre* O_2 *dans le référentiel associé à la molécule 1 ;* θ_2 *et* φ_2 *sont celles du centre de la molécule 1 dans le référentiel de la molécule 2 ;* χ *est une rotation axiale autour de* O_1O_2 *;* ρ *est la distance centre à centre. La distance et les 5 angles fixent la position et l'orientation d'une molécule relativement à l'autre.*

À partir d'une position donnée, la molécule 2 est donnée par :

1. Une rotation de ϕ_1 autour de l'axe y des coordonnées ;
2. Une rotation de θ_1 autour de l'axe z courant ;
3. Une translation ρ le long de l'axe x courant ;
4. Une rotation de χ autour du même axe x ;
5. Une rotation de θ_2 autour de l'axe z courant ;
6. Une rotation de ϕ_2 autour de l'axe y courant.

Si on travaille sur des molécules symétriques (homodimères), le nombre de paramètres peut être réduit.

2.1.1.3 L'algorithme d'amarrage

Ici, il consiste simplement à mettre en contact les deux molécules tout en gardant leurs orientations fixées.

La molécule 2 est translatée le long de la droite passant par les deux centres, jusqu'à ce qu'elle touche la molécule 1. À ce moment, il y a deux résidus i et j qui satisfont la condition :

$$d_{ij} = s(r_i + r_j) \text{ avec } s < 1 \tag{2.4}$$

où r_i et r_j sont les rayons des résidus et s un facteur ajustable qui permet la pénétration. Pour une représentation satisfaisante de sphères « molles » représentant les résidus, on prend $s = 0,75$.

L'équation suivante donne plusieurs valeurs de translation ξ possibles (définie ici pour simplifier sur l'axe x) :

$$(x_i - x_j + \xi)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 = s^2(r_i + r_j)^2 \tag{2.5}$$

x_i doit donc être calculé pour toutes les paires de résidus entre les deux molécules. À la plus petite valeur réelle de ξ, correspond le premier contact ayant lieu par translation le long de x. Si les surfaces sont convexes, l'amarrage s'arrête ici. Sinon, il est possible de s'approcher plus quand les surfaces sont concaves (figure 2.2).

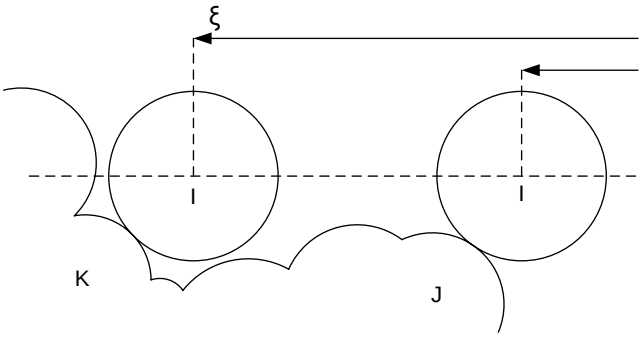


FIG. 2.2 – *DOCK* : amarrage dans les parties concaves.*Pour simplifier, la molécule 2 est dessinée comme étant une simple sphère I, se déplaçant le long du segment reliant les centres (en pointillés). I s'approche de la molécule 1 à partir de la droite et touche d'abord le résidu J. Avançant vers la gauche, I passe à travers le résidu J pour atteindre sa position native, correspondant à une translation ξ.*

2.1.1.4 Évaluation et affinement des énergies d'interaction

La surface de l'interface S_I obtenue dans une conformation donnée est évaluée, de même que l'énergie d'interaction non-liée E_{NB}. Si aucune pénétration n'était autorisée, c'est-à-dire $s = 1$, d'après l'équation (2.1), E_{NB} serait nulle. Avec $s = 0,75$, quelques paires de résidus se chevauchent et E_{NB} n'est pas nulle. Comme E_{NB} et E_S sont des fonctions analytiques de r_{ij}, distances entre résidus (la surface de l'interface S_I étant calculée par approximation analytique), l'algorithme de gradient conjugué de Fletcher et Reeves peut être utilisé [77]. Les poids relatifs du terme répulsif E_{NB} et du terme attractif E_S dépendent de la valeur γ (équation (2.3)) que C. Chothia a calibrée à 25 cal.mol^{-1} et qui est la valeur utilisée dans *DOCK*.

Le tri des différentes configurations obtenues réalisé par *DOCK* est donc très simple. De plus, pour avoir un résultat fiable, une étape de vérification manuelle des résultats est nécessaire.

2.1.2 Utilisation des données biologiques avec *HADDOCK*

2.1.2.1 Contraintes d'interaction ambiguës *AIRs (Ambiguous Interaction Restraints)*

Présentation Les données biologiques utilisées dans *HADDOCK* sont définies comme des contraintes d'interaction ambiguës *AIRs*. Ces contraintes peuvent provenir de n'importe quel type d'information expérimentale disponible sur les résidus impliqués dans l'interaction intermoléculaire. La distinction est faite entre résidus « actifs » et « passifs ».

Dans le cas de données de titrage par RMN, les résidus actifs sont les résidus qui présentent une perturbation de déplacement chimique significative à la formation du complexe et ceux qui ont une grande accessibilité au solvant dans la protéine isolée (plus de 50% d'accessibilité relative calculée avec *NACCESS*[3]). Le seuil à définir pour savoir si une perturbation de déplacement chimique est significative doit être ajusté par l'utilisateur en fonction du complexe étudié. Les résidus passifs sont ceux qui présentent une perturbation de déplacement chimique faible et/ou qui sont voisins de résidus actifs en surface et qui ont une grande accessibilité au solvant (>50%).

Dans le cas de données de mutagenèse dirigée, les résidus actifs sont ceux dont la mutation empêche la formation du complexe, ainsi que ceux exposés au solvant.

Définition Une contrainte d'interaction ambiguë est définie comme une distance intermoléculaire ambiguë d_{iAB} avec une valeur maximale de 3 Å entre un atome m_{iA} d'un résidu actif i d'une protéine A et chaque atome n_{kB} des résidus actifs et passifs k (N_{res} au total) de la protéine B (et inversement pour la protéine B). La distance effective d_{iAB}^{eff} pour chaque contrainte est calculée en utilisant l'équation :

$$d_{iAB}^{eff} = \left(\sum_{m_{iA}=1}^{N_{atoms}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{atoms}} \frac{1}{d_{m_{iA}n_{kB}}^{6}} \right)^{-\frac{1}{6}} \tag{2.6}$$

où N_{atoms} représentent tous les atomes d'un résidu donné et N_{res} la somme des résidus actifs et passifs pour une protéine donnée. Ainsi, les résidus passifs n'ont pas de contrainte d'interaction ambiguë avec la protéine partenaire, mais peuvent satisfaire les contraintes des résidus actifs de la protéine partenaire.

Une mise à l'échelle en $1/r^6$ est réalisée, non pas par analogie avec les contraintes *NOE (Nuclear OverHauser Effect)* traditionnellement utilisées en RMN, mais pour mimer la partie attractive d'un potentiel Lennard-Jones et assurer que les contraintes d'interaction ambiguës sont satisfaites quand les deux protéines sont en contact. La contrainte de 3 Å représente un compromis entre les distances de Van der Waals hydrogène-hydrogène et atome lourd-atome lourd.

[3]http ://wolf.bi.umist.ac.uk/naccess/nacwelcome.html

L'utilisation de contraintes ambiguës permet à *HADDOCK* de rechercher toutes les configurations possibles autour du site actif défini par des données biochimiques et/ou biophysiques telles que les données de perturbation de déplacement chimique RMN ou de mutagenèse et de trouver les paires de résidus en interaction les plus favorables entre tous les résidus, qu'ils soient actifs ou passifs.

2.1.2.2 Protocole d'amarrage

HADDOCK utilise des procédures traditionnelles de calculs de structure RMN pour effectuer sa recherche.

Les calculs de structure et l'affinement ont été implémentés dans la suite logicielle *CNS (Crystallography & NMR System)* [28, 29] et l'automatisation utilise *ARIA (Ambiguous Restraints for Iterative Assignment)* [143]. Les énergies inter- et intramoléculaires sont évaluées en utilisant un potentiel électrostatique complet et des termes d'énergie de Van der Waals avec un seuil à 8,5 Å.

La procédure se compose de trois étapes :

1. Les orientations sont rendues aléatoires et une minimisation d'énergie en corps rigides est réalisée ;
2. Un recuit simulé dans l'espace des angles de torsion est effectué ;
3. Un affinement dans l'espace cartésien avec un solvant explicite termine la procédure.

Pendant les étapes de recuit simulé et d'affinement avec le solvant, les résidus à l'interface peuvent bouger pour améliorer l'empilement. Les structures finales sont groupées en utilisant le *RMSD* (voir paragraphe 2.3.2.1 page 42) du squelette de la protéine à l'interface et analysées en fonction de leurs énergies d'interaction et de leur surfaces enfouies moyennes.

À la fin de cette procédure, l'utilisateur choisit parmi les groupes de conformations, ceux qui lui semblent les plus vraisemblables.

2.2 Le diagramme de Voronoï

Étant donné un ensemble de points centroïdes, la tessellation « classique » de Voronoï divise l'espace en régions, appelées cellules de Voronoï, centrées sur ces points [166].

Ici, après avoir défini les constructions que nous avons utilisées (diagramme de Voronoï, tessellation de Delaunay et diagramme de Laguerre), nous présenterons une nouvelle méthode de représentation des structures protéiques qui sera utilisée pour générer une fonction de score.

2.2.1 Définitions[4]

2.2.1.1 Diagramme de Voronoï

Soit un ensemble fini de points de $\mathbb{R}^d$, $E = \{p_1, \ldots, p_n\}$. À chaque p_i, on associe sa région de Voronoï $V(p_i)$ qui est constituée des points de $\mathbb{R}^d$ plus proches de p_i que des autres éléments de E :

$$V(p_i) = \{x \in \mathbb{R}^d : \|x - p_i\| \leq \|x - p_j\|, \forall j \leq n\} \tag{2.7}$$

Soit Π_{ij} le plan[5] médiateur de p_i et p_j, et π_{ij}^i celui des deux demi-espaces limités par π_{ij} qui contient p_i. $V(p_i)$ est l'intersection des demi-espaces π_{ij}^i, $j \neq i$, c'est-à-dire :

$$V(p_i) = \bigcap_{j \neq i} \pi_{ij}^i \tag{2.8}$$

Cette intersection contient le point p_i et n'est donc pas vide. $V(p_i)$ est un polyèdre convexe, éventuellement non borné.

On appelle *diagramme de Voronoï* de E l'ensemble des régions de Voronoï et leurs faces (figure 2.3).

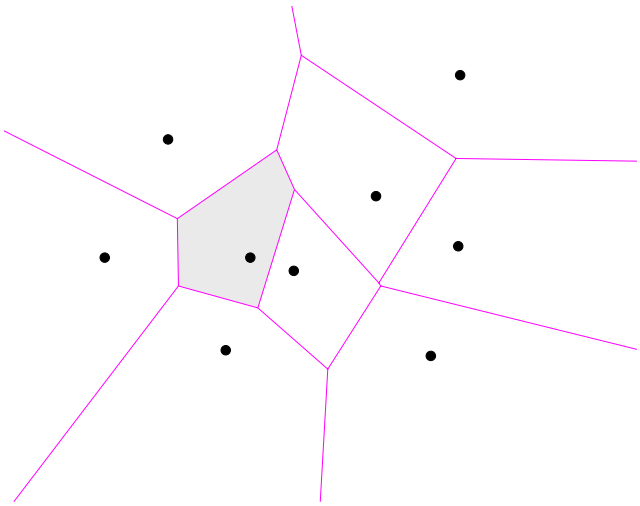


FIG. 2.3 – Diagramme de Voronoï avec, en gris, une des cellules du diagramme.

Comme tout point de $\mathbb{R}^d$ appartient à au moins une région de Voronoï, le diagramme est un *pavage* de $\mathbb{R}^d$. Si un point appartient à $k \geq 1$ régions, il appartient à une face commune à ces k régions qu'on appellera *face du diagramme*. Un point d'une telle face est à égale distance et plus proche de k points de E que des autres points.

[4]Pour plus de détails sur ces constructions, on se reportera à l'ouvrage de J.D. Boissonnat et M. Yvinec [20] dont les notations ont été utilisées ici et dont les paragraphes suivants sont largement inspirés.

[5]Pour éviter les confusions dans la suite, on adopte, pour parler des objets de $\mathbb{R}^d$, le vocabulaire de l'espace à trois dimensions.

2.2.1.2 Triangulation de Delaunay

Soit E un ensemble de n points de $\mathbb{R}^d$. On appelle *triangulation* de E un ensemble de tétraèdres dont les sommets sont les points de E et vérifiant :

- l'intersection de deux tétraèdres est soit vide, soit une face commune aux deux tétraèdres ;
- les tétraèdres pavent l'enveloppe convexe de E.

On appelle *faces* de la triangulation les tétraèdres (enveloppes convexes de k points affinement indépendants) ainsi que leur sous-faces.

On appelle *puissance* d'un point par rapport à une sphère σ de centre c et de rayon r le réel :

$$\sigma(x) = (x - c) \cdot (x - c) - r^2 \tag{2.9}$$

La sphère σ est alors définie par l'ensemble des points x tels que : $\sigma(x) = 0$. On dit qu'une sphère σ *englobe* un point y si l'intérieur de la boule limitée par la sphère contient le point, ce qui équivaut à $\sigma(y) < 0$.

Soit E un ensemble de n points de $p_1, \ldots, p_n$ de $\mathbb{R}^d$. On appelle *triangulation de Delaunay* de E une triangulation de E dont tous les tétraèdres peuvent être circonscrits par une sphère n'englobant aucun des p_i (figure 2.4).

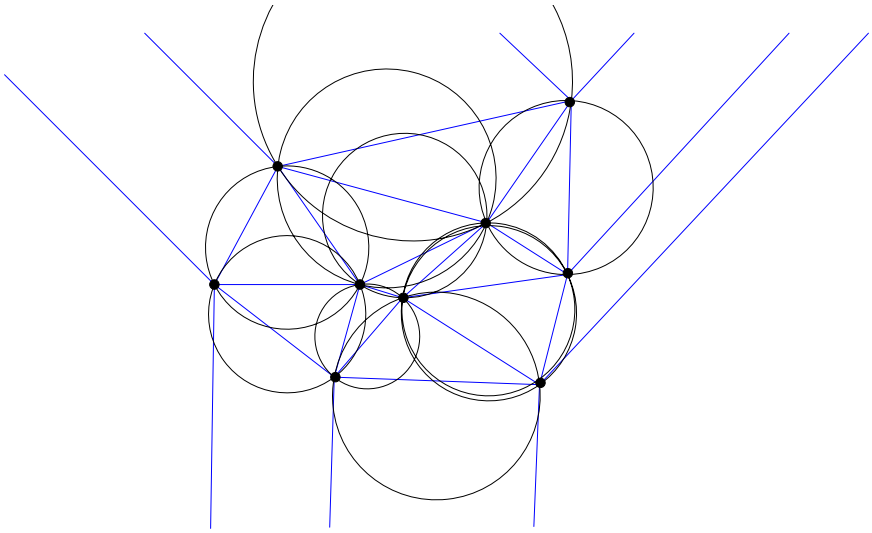


FIG. 2.4 – Tessellation de Delaunay

On appelle *sphère de Delaunay* une sphère circonscrite à un tétraèdre d'une triangulation de Delaunay, et *boule de Delaunay*, la boule délimitée par une telle sphère.

2.2.1.3 Propriétés

Dans le cas présenté dans la suite, puisque tous les points sont en position générique (c'est à dire qu'il n'y pas plus de 4 points cosphériques et pas de points alignés), la tessellation de Delaunay et le diagramme de Voronoï sont uniques et les tétraèdres ne

sont pas plats. La tessellation de Delaunay est le dual du diagramme de Voronoï (figure 2.5).

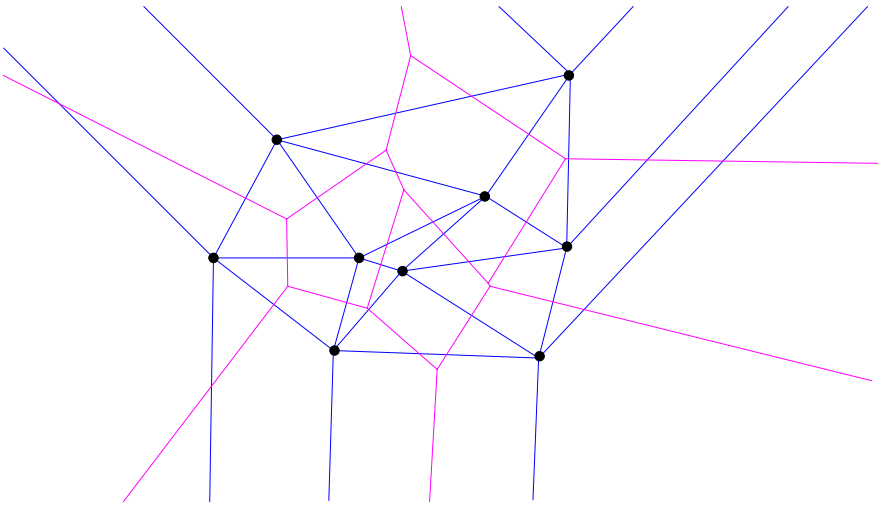


FIG. 2.5 – Diagramme de Voronoï et tessellation de Delaunay correspondante

2.2.1.4 Diagramme de Laguerre

Soit un ensemble fini de sphères de $\mathbb{R}^d$, $E = \sigma_1, \ldots, \sigma_n$. On note c_i le centre de σ_i et r_i son rayon. A chaque σ_i, on associe la région $L(\sigma_i)$ constituée des points de $\mathbb{R}^d$ dont la puissance par rapport à σ_i (voir équation 2.9) est plus petite que la puissance par rapport aux autres sphères de E :

$$L(\sigma_i) = \{x \in \mathbb{R}^d : \sigma_i(x) \leq \sigma_j(x), 1 \leq j \leq n\} \tag{2.10}$$

Régions et diagramme de Laguerre L'ensemble des points qui ont la même puissance par rapport à deux sphères σ_i et σ_j est un plan, noté ρ_{ij} et appelé *plan radical* de σ_i et σ_j. ρ_{ij} est orthogonal à la droite joignant les centres de σ_i et σ_j. On note ϱ_{ij}^i celui des deux demi-espaces limités par r_{ij} constitué des points dont la puissance par rapport à σ_i est plus petite que par rapport à σ_j. $L(\sigma_i)$ est l'intersection des demi-espaces ϱ_{ij}^i, $j \neq i$. Si cette intersection n'est pas vide, c'est un polyèdre convexe, éventuellement non borné. On appelle *régions de Laguerre* les $L(\sigma_i)$ qui ne sont pas vides.

On appelle diagramme de Laguerre de E l'ensemble des régions de Laguerre et leurs faces (figure 2.6). On note qu'il est possible qu'une sphère σ_i ne soit pas entièrement incluse dans de sa région de Laguerre.

Quand tous les rayons des sphères sont égaux, le diagramme de Laguerre des sphères s'identifie au diagramme de Voronoï de leurs centres.

Dualité Comme pour le diagramme de Voronoï, on peut définir une triangulation duale du diagramme de Laguerre, celle-ci est appelée triangulation régulière (figure 2.6).

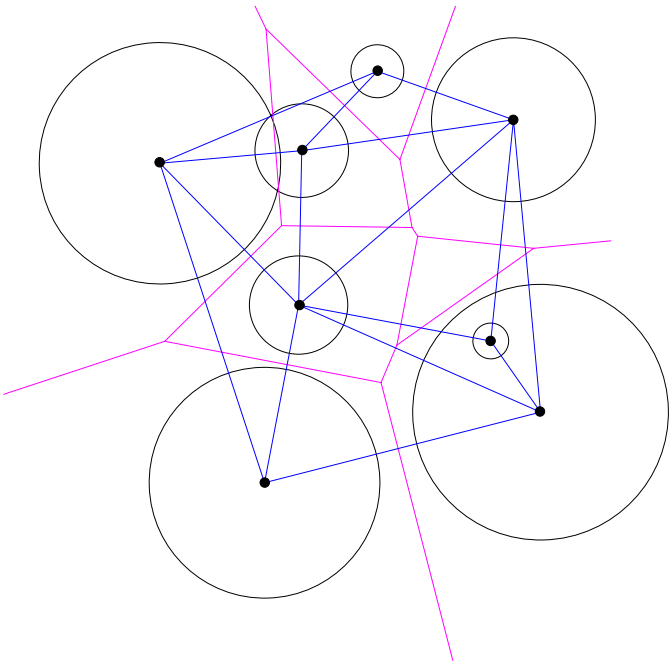


FIG. 2.6 – Diagramme de Laguerre (en rose) et son dual, la triangulation régulière (en bleu). *On remarque sur le dessin qu'une des sphères (la plus petite) ne se trouve pas dans sa région de Laguerre. Il arrive aussi que la région de Laguerre associée à une sphère « n'existe pas ».*

2.2.2 Méthodes de construction

En biologie structurale, les constructions des tessellations de Voronoï sont souvent réalisées de façon naïve [63]. Ici, nous présentons la méthode de construction naïve, mais aussi la construction que nous avons utilisée par la suite, utilisée dans la bibliothèque *CGAL* [21]. Cette dernière méthode, en plus d'offrir une construction optimale en temps, permet de s'affranchir d'éventuels problèmes d'erreur numérique, gérés par la bibliothèque.

2.2.2.1 Structure de données

Le diagramme de Voronoï (Laguerre) étant le dual de la triangulation de Delaunay (régulière), il n'est pas nécessaire de coder les deux structures, une seule suffit. On choisit donc de ne coder que la triangulation, car toutes ses faces sont des simplexes (tétraèdres dans l'espace), plus simples à coder.

On considère un ensemble de points E de $\mathbb{R}^d$. Un sommet s de la triangulation est représenté par un pointeur sur le point correspondant $\texttt{point}(s)$. Chaque sommet pointe sur un tétraèdre incident. Dans chaque tétraèdre de la triangulation sont stockés ses $d+1$ sommets, $s_0, \ldots, s_d$, qui sont des pointeurs sur $d+1$ points de E et des pointeurs sur ses $d+1$ tétraèdres adjacents, $t_0, \ldots, t_d$. Le tétraèdre t_i est celui qui n'a pas s_i comme sommet. Les tétraèdres qui ont une facette sur l'enveloppe convexe de E notée $\text{conv}(E)$ n'ont pas $d+1$ tétraèdres adjacents. De façon à traiter de manière unifiée tous les tétraèdres, la

bibliothèque *CGAL* utilise l'adjonction d'un point fictif (sans coordonnées) noté ∞. On peut ainsi représenter tous les tétraèdres obtenus en joignant le point ∞ à toutes les facettes de l'enveloppe convexe de E.

2.2.2.2 Prédicats et constructeurs

Prédicats Les prédicats sont des tests numériques élémentaires qui sont utilisés par un algorithme. Les deux prédicats les plus importants pour construire une triangulation de Delaunay à partir de laquelle sera déduite la triangulation de Voronoï sont les prédicats `orientation` et `dans_la_sphère`.

Orientation Le prédicat `orientation` permet de déterminer l'orientation d'un tétraèdre dans $\mathbb{R}^3$, ou plus généralement d'un d-simplexe de $\mathbb{R}^d$. Le simplexe $p_0 \dots p_d$ est orienté positivement, négativement ou est dégénéré (le sous-espace affine engendré par ses sommets est de dimension $< d$) selon que le déterminant de la matrice de dimension $d+1$ suivante est positif, négatif ou nul :

$$\texttt{orientation}(p_0, \dots, p_d) = \text{signe} \begin{vmatrix} 1 & \dots & 1 \\ p_0 & \dots & p_d \end{vmatrix} = \text{signe}|p_1 - p_0 \dots p_d - p_0| \tag{2.11}$$

Soit, dans l'espace à trois dimensions, en appelant x_i, y_i et z_i les coordonnées :

$$\texttt{orientation}(p_0, p_1, p_2, p_3) = \text{signe} \begin{vmatrix} 1 & 1 & 1 & 1 \\ x_0 & x_1 & x_2 & x_3 \\ y_0 & y_1 & y_2 & y_3 \\ z_0 & z_1 & z_2 & z_3 \end{vmatrix} = \text{signe} \begin{vmatrix} x_1 - x_0 & x_2 - x_0 & x_3 - x_0 \\ y_1 - y_0 & y_2 - y_0 & y_3 - y_0 \\ z_1 - z_0 & z_2 - z_0 & z_3 - z_0 \end{vmatrix} \tag{2.12}$$

L'interprétation de ce prédicat est la suivante : si h est un hyperplan passant par $p_0, \dots, p_{d-1}$ et $\pi(p)$ la projection du point p dans $x_d = 0$. Si h n'est pas vertical et si on a : $\texttt{orientation}(\pi(p_0), \dots, \pi(p_{d-1})) > 0$ alors $\texttt{orientation}(p_0, \dots, p_d)$ est positif, négatif ou nul selon que le point p_d est au-dessus, en dessous ou sur h.

On dit qu'une représentation d'une triangulation est *orientée* si tous les tétraèdres sont orientés positivement. On ne travaille par la suite que sur des triangulations orientées.

Dans la sphère Le prédicat `dans_la_sphère` permet de déterminer si un point p_{d+1} est ou non englobé par la sphère circonscrite à $d+1$ points $p_0, \dots, p_d$. La triangulation étant orientée, p_{d+1} est à l'intérieur, sur la surface, ou à l'extérieur de la sphère circonscrite à $p_0, \dots, p_d$ selon que le déterminant de la matrice de dimension $d+2$ suivante est négatif, nul ou positif :

$$\texttt{dans\_la\_sphère}(p_0, \dots, p_{d+1}) = \text{signe} \begin{vmatrix} 1 & \dots & 1 \\ p_0 & \dots & p_{d+1} \\ \|p_0\|^2 & \dots & \|p_{d+1}\|^2 \end{vmatrix} \tag{2.13}$$

Soit, dans l'espace à trois dimensions, en appelant x_i, y_i et z_i les coordonnées :

$$\begin{aligned}&\texttt{dans\_la\_sphère}(p_0,p_1,p_2,p_3,p_4)\\&=\text{signe}\begin{vmatrix}1 & 1 & 1 & 1 & 1\\ x_0 & x_1 & x_2 & x_3 & x_4\\ y_0 & y_1 & y_2 & y_3 & y_4\\ z_0 & z_1 & z_2 & z_3 & z_4\\ x_0^2+y_0^2+z_0^2 & x_1^2+y_1^2+z_1^2 & x_2^2+y_2^2+z_2^2 & x_3^2+y_3^2+z_3^2 & x_4^2+y_4^2+z_4^2\end{vmatrix}\end{aligned} \tag{2.14}$$

Pour obtenir la triangulation régulière permettant de construire le diagramme de Laguerre, il suffit d'utiliser l'analogue du prédicat `dans_la_sphère`, qu'on peut appeler `test_puissance`, et qui est :

$$\texttt{test\_puissance}(\sigma_0,\ldots,\sigma_{d+1})=\text{signe}\begin{vmatrix}1 & \ldots & 1\\ p_0 & \ldots & p_{d+1}\\ \|p_0\|^2-r_0^2 & \ldots & \|p_{d+1}\|^2-r_{d+1}^2\end{vmatrix} \tag{2.15}$$

σ_i étant définie comme la sphère de centre p_i et de rayon r_i.

Remarque sur les prédicats Ils jouent un rôle central dans l'algorithme, car ils conditionnent les branchements. Une erreur numérique dans l'évaluation d'un prédicat peut entraîner des conséquences fatales (par exemple des tétraèdres non fermés). Il faut donc évaluer leur signe avec précision pour éviter tout problème d'erreur numérique.

Constructeurs Pour construire le diagramme de Voronoï (ou de Laguerre) explicitement à partir de la triangulation de Delaunay (ou de la triangulation régulière), il faut calculer les coordonnées des sommets du diagramme. Le sommet de Voronoï $v(p_0,\ldots,p_d)$ à égale distance et plus proche de $p_0,\ldots,p_d$ que des autres points de E, est le point x qui vérifie :

$$(x-p_0)\cdot(x-p_0)=\ldots=(x-p_d)\cdot(x-p_d) \tag{2.16}$$

et qui correspond au système :

$$\begin{aligned}2(p_0-p_1)\cdot x&=\|p_0\|^2-\|p_1\|^2\\ &\vdots\\ 2(p_0-p_d)\cdot x&=\|p_0\|^2-\|p_d\|^2\end{aligned} \tag{2.17}$$

soit en notant $O=(p_0-p_1,\ldots,p_0-p_d)$ la matrice qui intervient dans le prédicat d'orientation et $L=(l_1,\ldots,l_d)^t$,avec $l_i=\|p_0\|^2-\|p_i\|^2$

$$x=\frac{1}{2}O^{-1}L \tag{2.18}$$

Dans le plan[6] , en notant x_i, y_i les coordonnées, on obtient donc :

[6]Pour des raisons de clarté, cette équation est exceptionnellement illustrée dans $\mathbb{R}^2$ et non dans $\mathbb{R}^3$. On peut à partir de ce résultat, deviner aisément les coordonnées dans l'espace.

$$v(p_0,p_1,p_2) = \frac{1}{2D}\left(\begin{vmatrix} 1 & 1 & 1 \\ y_0 & y_1 & y_2 \\ x_0^2+y_0^2 & x_1^2+y_1^2 & x_2^2+y_2^2 \end{vmatrix}, \begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2+y_0^2 & x_1^2+y_1^2 & x_2^2+y_2^2 \end{vmatrix}\right) \tag{2.19}$$

$$\text{avec } D = \begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \end{vmatrix} \tag{2.20}$$

2.2.2.3 Construction naïve

La construction dite « naïve » est très simple à mettre en oeuvre. Pour chacun des points p_i considérés, on cherche les points p_j plus proches de p_i que de tous les autres points. On teste si le centre du cercle circonscrit ne contient pas d'autre point (grâce au prédicat `dans_la_sphère`) et on construit le sommet de Voronoï correspondant (grâce au constructeur vu précédemment). Le problème de cette construction est sa complexité. En effet, dans l'espace, si n est le nombre points considérés, la complexité est au moins cubique ($O(n^3)$).

2.2.2.4 Construction incrémentale randomisée

Méthode incrémentale

Insertion On suppose qu'on a une triangulation de Delaunay notée $Del(E)$ et qu'on veut insérer un nouveau point x, c'est-à-dire calculer $Del(E \cup \{x\})$.

Si ce point x est dans l'enveloppe convexe de E, c'est-à-dire $x \in \text{conv}(E)$, alors par définition de la triangulation de Delaunay, il faut supprimer tous les tétraèdres de $Del(E)$ en *conflit* avec x, c'est-à-dire ceux dont la sphère circonscrite englobe x puis retrianguler le trou ainsi créé. La triangulation se fait en construisant les nouveaux tétraèdres $\text{conv}(x,f), f \in F$, où F est l'ensemble des facettes des bords du trou. Ensuite, il reste à mettre à jour la structure de données en mettant à jour les relations d'adjacence.

Si le point à insérer n'appartient pas à l'enveloppe convexe de E, c'est-à-dire $x \notin \text{conv}(E)$, il faut identifier les facettes de $\text{conv}(E)$ en conflit avec x, c'est-à-dire celles dont les plans support séparent x de l'intérieur de $\text{conv}(E)$. Si on a bien au préalable associé à chaque facette de $\text{conv}(E)$ un tétraèdre (f,∞), en posant :

$$\texttt{dans\_la\_sphère}(\infty,p_1,\ldots,p_d) = \texttt{orientation}(p_1,\ldots,p_d) \tag{2.21}$$

l'insertion de x se fait alors de la même façon que dans le cas ou $x \in \text{conv}(E)$.

On peut remarquer que la mise à jour de la triangulation utilise comme seules opérations numériques les prédicats `orientation` et `dans_la_sphère`.

Localisation La méthode d'insertion décrite dans le paragraphe précédent implique qu'il faut pouvoir localiser le point à insérer. Étant donné un point x, on veut identifier un tétraèdre qui le contient. Si x n'appartient pas à l'enveloppe convexe de E, on veut identifier une face de $\text{conv}(E)$ qui sépare x de l'intérieur de $\text{conv}(E)$.

Pour cela, on peut, à partir d'un sommet s de la triangulation, parcourir les tétraèdres coupés par le segment de droite $p(s)x$. Pour des distributions de points uniformes, ou si on connaît un sommet proche de x, cela est beaucoup plus efficace que l'examen de tous les tétraèdres.

Si on autorise les sauts, on peut rendre cette méthode plus rapide en construisant une structure de données hiérarchique D constituée de h niveaux [60]. Chaque niveau i de la hiérarchie est la triangulation de Delaunay d'un sous-ensemble de E_i avec $E_h \subset \ldots \subset E_1 = E$. Dans la structure de données, on établit un lien entre deux sommets de $Del(E_{i+1})$ et $Del(E_i)$ qui pointent sur le même point de E. Si on a identifié le sommet s_{i+1} le plus proche de x dans la triangulation $Del(E_{i+1})$, la structure de données permet d'accéder directement au somment s_i dans $Del(E_i)$ tel que $\texttt{point}(s_{i+1})$=$\texttt{point}(s_i)$. La marche dans $Del(E_i)$ suit le segment $\texttt{point}(s_i)x$. La marche dans $Del(E_1)$ termine la localisation.

Randomisation Dans le pire des cas, la complexité de l'algorithme incrémental décrit précédemment est $O(n^3)$. Si on arrive à ce que le pire cas ne se produise que très rarement, on peut obtenir une complexité moyenne sous-quadratique. On peut introduire des tirages aléatoires dans l'algorithme décrit précédemment : pour décider quel point de E_i est retenu dans E_{i+1} et aussi pour choisir l'ordre d'insertion des points.

En dimension d, on peut montrer dans ce cas, que l'espérance du temps d'exécution est $O\left(n \log n + n^{\lceil \frac{d}{2} \rceil}\right)$ et l'espérance de la place mémoire requise est $O\left(n^{\lceil \frac{d}{2} \rceil}\right)$. Cet algorithme est optimal dans le cas le pire.

En pratique, pour la modélisation moléculaire, la complexité de la triangulation est linéaire. Avec la bibliothèque CGAL, la complexité de la triangulation est linéaire, et son calcul a une complexité de $O(n \log n)$. Dans notre cas, grâce à la bibliothèque *CGAL*, nous construisons la triangulation de Delaunay de 10 000 points en 1 seconde sur une machine standard[7].

2.2.3 Application aux protéines

2.2.3.1 Triangulations

Quels points trianguler ? Dans le cadre du problème de l'amarrage, nous avons choisi de travailler avec un seul point par résidu, le centre géométrique de la chaîne latérale et du Cα. En effet, ce choix permet tout d'abord de travailler sur un nombre réduit de points, mais aussi de s'affranchir des mouvements des chaînes latérales. En effet, même lorsque la chaîne latérale bouge, ce qui est fréquent à la surface de la protéine, le centre géométrique n'est souvent pas trop affecté. La cellule de Voronoï correspondante est donc presque identique.

La triangulation de Delaunay pour les protéines a été le plus souvent effectuée à partir des carbones α [63, 157, 204, 221]. Nous avons choisi de construire les diagrammes de Voronoï à partir de carbones α, et également à partir des « centres géométriques » des acides aminés. Les différents résultats (voir paragraphe 3.1.3 page 65), nous ont fait préférer les « centres géométriques » des acides aminés. Ceux-ci ont été définis comme

[7] Cette machine dispose d'un processeur de type *Pentium IV* cadencé à 3,06 GHz

centres de gravité des atomes de la chaîne latérale et du carbone α, les atomes d'hydrogène n'étant pas pris en compte. Le centre géométrique ainsi obtenu est donc très proche du centre de masse de l'acide aminé.

Ce choix est à rapprocher d'une étude réalisée par S. Karlin débutée en 1994 [116, 115] dans laquelle trois types de distances sont étudiées :

- entre carbones α ;
- entre centres de gravité des acides aminés sans prendre en compte les carbones α ;
- entre centres de gravité de tous les atomes du résidu (chaîne latérale et squelette).

Ces trois types de distances sont utilisés pour mesurer la distance entre deux résidus dans un ensemble de structures tridimensionnelles de protéines connues. Cette étude statistique montre que les distances entre centres de gravité des chaînes latérales sont très sensibles aux interactions électrostatiques et hydrophobes, mais très peu aux contraintes stériques, contrairement aux distances entre les centres de gravité de tous les atomes de chaque résidu. Le fait de prendre en compte le carbone α dans le calcul du point qui va « représenter » chaque acide aminé, permet donc d'obtenir des propriétés intermédiaires. S. Karlin et ses collaborateurs montrent également que les distances entre carbones α sont largement décorrélées, à la fois des interactions et des contraintes stériques.

Problème de la surface et fermeture des cellules Pour « fermer » la triangulation de Voronoï (paragraphe 2.2 page 28), on peut utiliser un point fictif situé à l'infini. Si l'on ferme la triangulation à la surface de la protéine avec une construction de ce type, on risque, lorsque la protéine n'est pas vraiment globulaire, d'avoir une très mauvaise description de l'empilement pour les résidus situés juste en dessous de la surface.

Afin de contourner ce problème, nous avons décidé, en accord avec une étude précédemment publiée [207] de « plonger » la protéine dans un « solvant ». En pratique, dans un rayon de 30 Å autour de la protéine, on dispose des sphères de 6,5 Å de diamètre (figure 2.7). Ces sphères sont placées sur un réseau du même type que celui de l'eau, la relaxation de ce réseau étant optionnelle. Elles ont pour but d'imiter le mieux possible un acide aminé, de façon à ce que le diagramme de Voronoï obtenu corresponde à la forme qu'aurait la surface de la protéine en présence d'un partenaire. Le diamètre 6,5 Å a été choisi de manière à ce que le volume des cellules correspondant à ces sphères soit voisin de la moyenne du volume des cellules correspondant à des acides aminés. Afin de s'affranchir du biais qu'introduit cette construction, les cellules en contact avec le solvant ne seront pas prises en compte dans les études statistiques effectuées par la suite.

Quel type de diagramme choisir : Voronoï ou Laguerre ? Comme signalé précédemment, le diagramme de Voronoï n'est pas pondéré. Or nous utilisons un seul point par résidu, ces points devraient donc avoir des poids très différents. Nous avons donc dans un premier temps construit également le diagramme de Laguerre pour chacune des structures étudiées, puisqu'il permet de prendre en compte ces différences de poids, et est donc, *a priori*, mieux adapté.

Les rayons des sphères utilisées pour la construction du diagramme de Laguerre ont été choisis de façon à ce que celles-ci aient un volume moyen, pour un type d'acide aminé

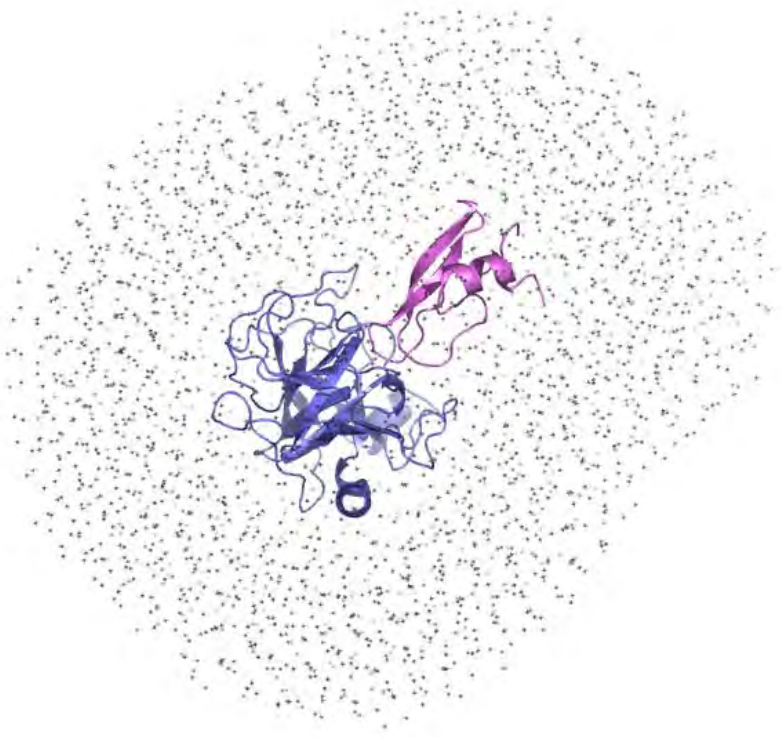


FIG. 2.7 – « Sphère de solvant » autour du complexe 1p2k. *Les points gris représentent les centres des sphères de diamètres 6,5 Å placées tout autour de la protéine sur un réseau du même type que celui de l'eau.*

donné, correspondant à son volume tel que déterminé par l'étude atomique de J. Pontius [177].

Cependant, la construction de Laguerre pose un certain nombre de problèmes, en particulier la « disparition » de certains résidus qui rend l'interprétation biologique plus difficile (voir paragraphe 3.1.2 page 64) et donc n'a pas été utilisée pour la mise en place de la fonction de score d'amarrage.

2.3 Paramètres pour l'apprentissage

2.3.1 Mesures issues de la construction de Voronoï

Pour la suite, nous procédons uniquement à des constructions de Voronoï en utilisant les centres géométriques de la chaîne latérale et du carbone α.

2.3.1.1 Définitions

La tessellation de Voronoï pour un complexe protéine-protéine (figure 2.8) permet de poser les définitions suivantes :

– deux résidus sont voisins au sens de Voronoï si leur cellules de Voronoï partagent une face commune ;
– un résidu appartient au cœur de la protéine si tous ses voisins sont des résidus de la même protéine ;
– un résidu appartient à la surface de la protéine si au moins un de ses voisins est de type « solvant » ;
– un résidu appartient à l'interface protéine-protéine si un au moins de ses voisins appartient à l'autre protéine ;
– un résidu appartient au cœur de l'interface s'il appartient à l'interface et qu'aucun de ses voisins n'est de type « solvant » ;
– les faces des cellules partagées par les résidus des deux protéines constituent l'interface.

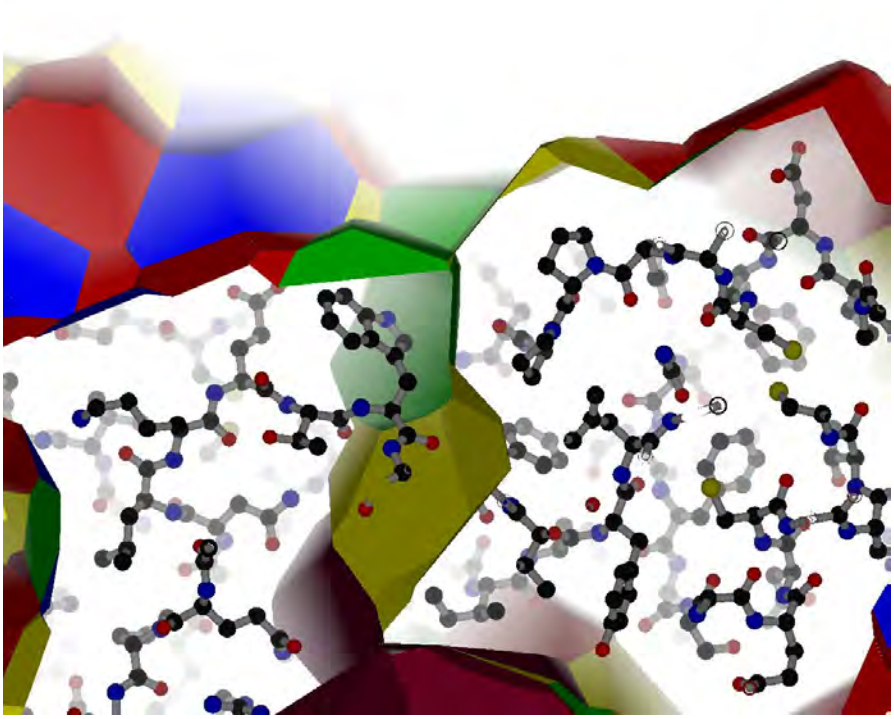


FIG. 2.8 – Vue de partielle de l'interface de Voronoï. *Les facettes des cellules de Voronoï sont colorées en fonction des propriétés physico-chimiques des résidus (voir paragraphe 2.3.1.3 page 41).*

.

La définition du voisinage au sens de Voronoï est intéressante, car elle n'implique pas de distance seuil (*cutoff*) traditionnellement utilisée en biologie structurale pour définir le voisinage. Ainsi, deux résidus voisins dans le cadre d'une définition avec seuil, ne le sont pas forcément au sens de Voronoï (figure 2.9).

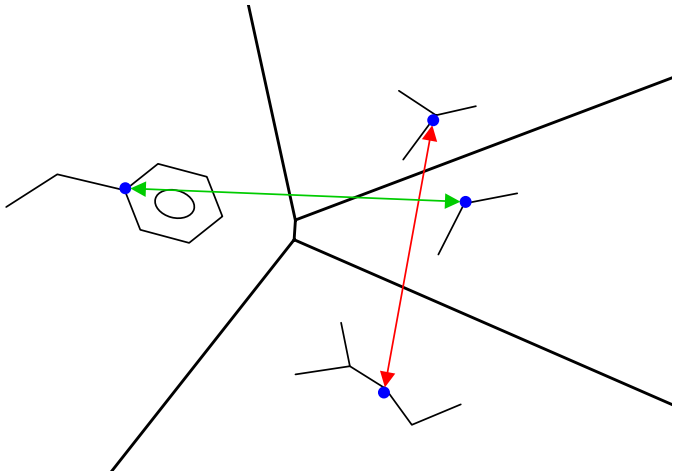


FIG. 2.9 – Définition du voisinage au sens de Voronoï. *Dans une définition classique avec une distance seuil, les deux résidus reliés par la flèche rouge pourraient être considérés comme voisins et pas ceux reliés par la flèche verte. Au sens de Voronoï, les deux résidus reliés par la flèche verte sont voisins et ceux reliés par la flèche rouge ne le sont pas.*

2.3.1.2 Mesures

À partir de cette construction et des définitions associées, de nombreuses mesures sont possibles. Pour l'apprentissage, nous avons choisi de travailler avec les mesures suivantes :

- la surface de l'interface (1 paramètre) ;
- le nombre de résidus dans le coeur de l'interface (1 paramètre) ;
- le volume de Voronoï (figure 2.10) de chaque type de résidu (20 paramètres) ;
- la fréquence d'apparition de chaque type de résidu à l'interface (20 paramètres) ;
- la fréquence des paires de résidus en contact (210 paramètres) ;
- les distances de paires entre résidus (210 paramètres).

FIG. 2.10 – Cellule de Voronoï de la tyrosine 104 du complexe 1A0O. *Le volume de Voronoï associé au résidu est le volume du polyèdre représenté en vert.*

.

Pour chaque complexe putatif, on dispose donc au total de 462 mesures.

Dans le cadre de l'apprentissage effectué par la suite, vu la taille de l'échantillon d'apprentissage (voir paragraphe 2.4.2 page 48), il n'a pas été possible d'utiliser l'ensemble des mesures prises individuellement. Nous avons donc décidé de les regrouper en fonction de leurs propriétés physico-chimiques.

2.3.1.3 Regroupement des variables

Les propriétés physico-chimiques des résidus peuvent être décrites de nombreuses façons [17, 190]. Dans cette étude nous nous sommes basés sur ces descriptions pour composer 6 groupes (figure 2.11) :

- les hydrophobes aromatiques : phénylalanine, tyrosine, tryptophane ;
- les hydrophobes non aromatiques : isoleucine, leucine, valine, méthionine ;
- les polaires anioniques : acide aspartique, acide glutamique ;
- les polaires cationiques : histidine, arginine, lysine ;
- les polaires anioniques ou cationiques : asparagine, glutamine ;
- les petits : alanine, cystéine, glycine, sérine, proline, thréonine.

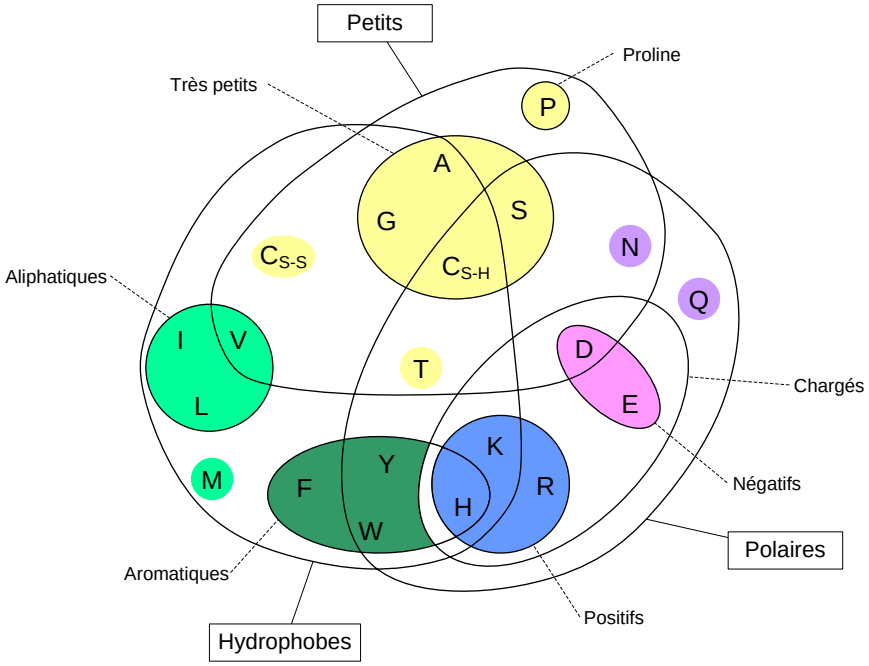


FIG. 2.11 – Regroupement des acides aminés en fonction de leurs propriétés physico-chimiques. *Chaque couleur représente un groupe et est assortie à la couleur utilisée pour représenter les faces de Voronoï associées au résidu.*

.

2.3.1.4 Paramètres finaux

Avec les groupes définis ci-dessus, le nombre de paramètres utilisés dans l'apprentissage est réduit à 84 :

- la surface de l'interface (1 paramètre) ;
- le nombre de résidus dans le coeur de l'interface (1 paramètre) ;
- le volume de Voronoï de chaque type de résidu (20 paramètres) ;
- la fréquence d'apparition de chaque type de résidu à l'interface (20 paramètres) ;
- la fréquence des paires de résidus en contact (21 paramètres) ;
- les distances de paires entre résidus (21 paramètres).

Il est évident que les paramètres que nous avons choisis ne sont pas les seuls possibles, et que les regroupements que nous avons effectués ne sont pas les seuls possibles non plus. Cependant, une étude exhaustive du choix des paramètres n'était pas envisageable dans le cadre de cette thèse.

2.3.2 Mesures pour l'évaluation des résultats

Pour évaluer la qualité d'une prédiction par rapport à la structure cristallographique, un certain nombre de mesures, utilisant des algorithmes usuels, sont effectuées sur le modèle. Tout d'abord, le complexe doit présenter une surface d'interaction raisonnable [8, 9, 144], ne pas présenter d'atomes trop proches (collisions, encore appelées *clashes*), mais aussi présenter des résidus à l'interface conformes à ceux observés dans la structure cristallographique [148]. Dans cette section, nous présenterons comment ces mesures ont été effectuées.

2.3.2.1 Mesures générales de biologie structurale

Surface de l'interface L'aire de la surface de l'interface d'un complexe protéine-protéine peut être définie comme la réduction de la surface accessible au solvant due à l'association (voir équation 2.2 paragraphe 2.1.1 page 24) [144]. Ces mesures sont en général effectuées grâce à l'algorithme de Lee et Richards [127] qui a été implémenté de nombreuses façons et amélioré par la suite [203, 51, 68]. D'autres méthodes de mesure existent[8], telles que la somme des surfaces des facettes du diagramme de Voronoï atomique [34] ou la triangulation de la surface [236].

De nombreuses implémentations/améliorations des algorithmes précédents ont été réalisées. Après une brève présentation de l'algorithme de Lee et Richards, nous présenterons rapidement deux programmes usuels de mesure.

L'algorithme de Lee et Richards [127] L'algorithme de Lee et Richards procède en plusieurs étapes :

- la protéine est découpée en « tranches » dont l'épaisseur est choisie par l'utilisateur. Un bon compromis entre temps de calcul et fiabilité du résultat est de 0,25 Å. En effet, plus les tranches sont fines, meilleure est la précision, mais le temps de calcul augmente très rapidement ;
- dans chaque tranche, on trace la surface de Van der Waals de chaque atome ;
- on fait « rouler » une sphère sonde sur l'extérieur de la tranche (voir figure 2.12), pour chaque section de surface de Van der Waals, on détermine la surface de contact

[8] À ce sujet, on pourra se reporter à l'excellente revue en ligne de Michael Connolly `http://www.netsci.org/Science/Compchem/feature14.html`

avec la sphère. Le rayon de la sphère est choisi par l'utilisateur, en général, on utilise 1,4 Å, ce qui correspond à la taille d'une molécule d'eau ;
– pour chaque atome, la surface est intégrée sur les différentes tranches dans lesquelles il est présent.

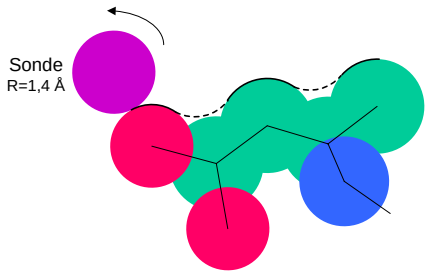


FIG. 2.12 – Principe du calcul de l'algorithme de Lee et Richards. *On fait rouler une sphère sonde pour chaque section de la surface et on mesure la surface de contact (ici représentée par les lignes continues vertes).*

Grâce à cet algorithme, on peut obtenir la surface accessible sur la surface entière, mais aussi par résidu.

Implémentations De nombreuses implémentations et améliorations de l'algorithme précédent ont été réalisées et permettent de calculer la surface accessible au solvant d'une molécule (protéine ou ADN/ARN) dont on connaît la structure au format PDB. Les logiciels les plus utilisés sont :
– *Naccess (Atomic Solvent Accessible Area Calculations)*[9] : écrit en 1992 par Simon Hubbard et Janet Thornton, il utilise la méthode de Lee et Richards ;
– *DSSP (Database of Secondary Structure Assignments)* [114] : programme qui permet de donner les structures secondaires associées à un fichier PDB mais qui donne aussi les surfaces accessibles au solvant selon la méthode de Lee et Richards, avec un algorithme d'intégration amélioré.

Détection des collisions Un complexe ne doit pas contenir d'atomes trop proches, car ce n'est pas vraisemblable d'un point de vue chimique. De plus, pour la suite, il nous faut pouvoir déterminer les atomes/résidus à l'interface selon la définition traditionnellement utilisée par les cristallographes.

On peut considérer que deux résidus sont en contact à l'interface s'ils appartiennent chacun à un partenaire différent et ont deux atomes lourds (c'est-à-dire deux atomes n'étant pas des hydrogènes) à moins de 5 Å l'un de l'autre.

On dit que deux résidus à l'interface sont trop proches et sont en collision s'ils appartiennent chacun à un partenaire différent et ont deux atomes lourds à moins de 3 Å l'un de l'autre. Ils peuvent parfois être plus proches, mais c'est cette mesure qui a été retenue.

Il existe bien sûr d'autres définitions, mais ce sont celles-ci qui sont utilisées dans l'évaluation des résultats *CAPRI*, c'est pourquoi nous les utiliserons par la suite.

[9] http ://wolf.bms.umist.ac.uk/naccess/

Programme utilisé Les mesures précédentes sont facilement calculables. Une implémentation très simple et rapide à utiliser est le programme *contact*[10] distribué avec la suite *CCP4*, très utilisée en cristallographie.

Mesure de l'écart à la structure attendue *Root Mean Square Deviation (RMSD)* En biologie structurale, on mesure souvent l'écart entre deux structures par le *RMSD (Root Mean Square Deviation)*. Cette mesure de l'écart-type (obtenue en Å), est réalisée sur un ensemble d'atomes des résidus alignés entre les deux structures superposées au mieux selon la formule suivante pour chaque atome i (N étant le nombre total d'atomes et x et y les coordonnées des atomes dans chaque structure) :

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|x(i) - y(i)\|^2} \tag{2.22}$$

Plus les structures sont différentes, plus le *RMSD* est élevé. Le *RMSD* est une mesure que fournissent la plupart des logiciels de visualisation après avoir fait un alignement structural, en général avec l'algorithme de W. Kabsch [112, 113].

Dès que le *RMSD* est supérieur à quelques angströms, ce n'est plus une bonne mesure de l'écart entre deux structures. Cependant, comme aucune autre mesure ne peut représenter correctement l'écart entre deux structures, c'est le *RMSD* qui est utilisé.

2.3.2.2 Mesures plus spécifiques aux complexes utilisées dans le cadre de l'expérience *CAPRI*

Dans le cadre de l'évaluation des résultats de l'expérience *CAPRI*, de nouvelles mesures ont été définies pour les complexes [148]. On compare à chaque fois une configuration prédite à une structure cristallographique.

Dans le cadre des mesures utilisées pour *CAPRI*, on distingue les deux protéines du complexe considéré : on appelle R (comme Récepteur) la plus longue des protéines et L (comme Ligand) la plus petite. On appelle épitopes les zones de chaque protéine qui interagissent.

Écart de position du « ligand » Ces mesures se font en deux étapes (figure 2.13).

1. On superpose d'abord les récepteurs de la structure et du modèle et on mesure les *RMSD* de la chaîne principale sur le ligand :
   - avec tous les résidus, on appelle cette mesure L_{rms} (en Å),
   - avec uniquement les résidus à l'interface (résidus qui ont des atomes à moins de 10 Å de R dans la structure cristallographique), on appelle cette mesure I_{rms} (en Å) ;
2. Puis, à partir de cette position, on superpose les ligands pour obtenir :
   - θ_L, angle de déplacement de l'orientation (en degrés),
   - d_L déplacement du centre moléculaire (en Å).

[10] http ://www.ccp4.ac.uk/dist/html/contact.html

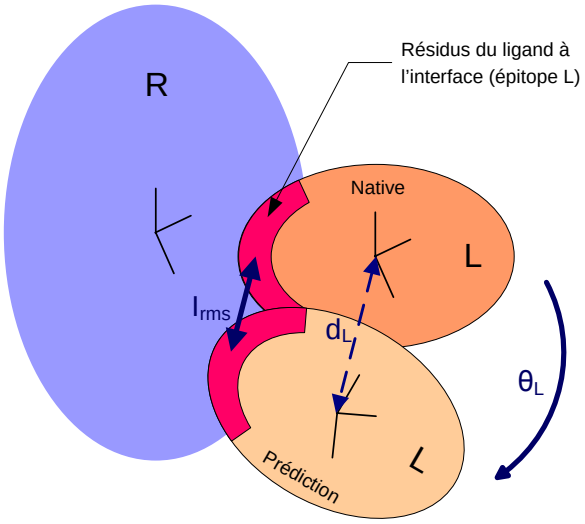


FIG. 2.13 – Mesures *CAPRI* pour l'évaluation des prédictions : mesure de l'écart de position du « ligand ». *Ces mesures L_{rms}, I_{rms}, θ_L et d_L permettent d'obtenir une bonne description de la géométrie prédite de l'interaction.*

Proportion de l'interface correctement prédite Pour ces mesures, on superpose uniquement les récepteurs, et on étudie trois critères (figure 2.14) :

1. La prédiction des épitopes :
   - on compare les résidus de R (respectivement L) ayant des atomes à moins de 5 Å d'atomes de L (resp. R) dans la structure cristallographique avec les mêmes résidus dans la structure prédite,
   - on obtient les fractions de résidus à l'interface correctement prédits : f_R pour R et f_L pour L ;
2. La prédiction des paires de résidus en contact :
   - on compte :
     (a) les paires de résidus R/L avec des atomes à moins de 5 Å de distance dans la structure cristallographique : N_{nc},
     (b) les mêmes paires dans le complexe prédit : N_{pc} ;
   - on obtient :
     (a) la fraction des contacts natifs : $f_{nc} = paires\ correctement\ prédites/N_{nc}$,
     (b) la fraction des contacts non-natifs : $f_{fc} = paires\ incorrectement\ prédites/N_{fc}$;
3. Le rejet des collisions :
   - on compte le nombre de paires de résidus R/L avec des atomes à moins de 3 Å de distance dans la structure prédite, on l'appelle : N_{bad},
   - on calcule la moyenne $\overline{N_{bad}}$ et l'écart-type σ sur un ensemble de prédictions pour une cible donnée, et on rejette les prédictions avec : $N_{bad} > \overline{N_{bad}} + 2\sigma$.

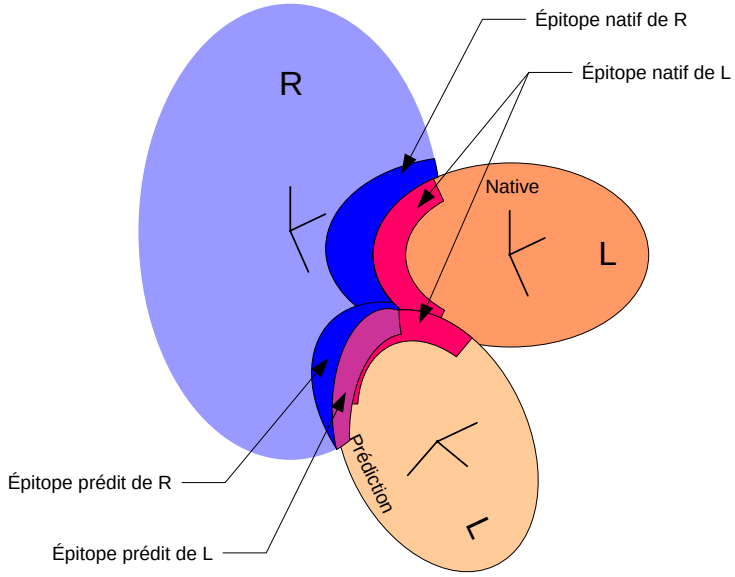


FIG. 2.14 – Mesures *CAPRI* pour l'évaluation des prédictions : proportion de l'interface correctement prédite. *Dans cet exemple, la fraction de résidus en contact correctement prédits est élevée pour L,* $f_L \approx 0,5$ *mais elle est faible pour R,* $f_R < 0,1$ *et aucun contact de paire n'est correctement prédit* $f_{nc} = 0$.

2.3.3 Visualisation

Pour pouvoir présenter à l'expert des résultats de modélisation, comparer les prédictions aux structures expérimentales et observer la modélisation effectuée, nous avons utilisé des logiciels de visualisation moléculaire. Plusieurs ont été utilisés et sont présentés ici en fonction de leurs possibilités.

2.3.3.1 Logiciels les plus utilisés

Pymol[11] C'est un des logiciels les plus complets et les plus diffusés depuis 2001. Écrit en Python, il est très modulaire et dispose d'une API très complète de manipulation des formats usuels et de calculs simples (*RMSD*, alignement...). La plupart des figures de ce manuscrit ont été réalisées avec ce logiciel, en particulier grâce à ses options de rendu performantes.

Molscript[12] Ce logiciel plus ancien (1997), est un des premiers à avoir été capable d'intégrer des objets OpenGL standards tels que des triangles de façon très simple (proche de la syntaxe VRML). Moins convivial d'utilisation que Pymol, il est très performant et a été utilisé pour les premiers essais de tracé des cellules de Voronoï dans cette étude. Associé à Raster3D[13] [7], il permet d'obtenir images avec un rendu de bonne qualité.

[11] `http://pymol.sourceforge.net/`
[12] `http://www.avatar.se/molscript/`
[13] `http://www.bmsc.washington.edu/raster3d/`

XmMol[14] **[215]** Très maniable, XmMol est un logiciel qui permet très facilement de manipuler/visualiser les squelettes des protéines. Permettant un aperçu/tracé aisé des polyèdres à partir des centres de gravité des acides aminés, nous l'avons utilisé au fur et à mesure du développement pour vérifier les tracés et observer les éventuels problèmes de construction (fermeture des cellules...).

2.3.3.2 Autres logiciels

Chimera[15] **[173]** Le principal intérêt de ce logiciel est de pouvoir traiter des molécules de très grande taille. Très utilisé pour la visualisation des capsides de virus, il a surtout été utilisé ici pour visualiser les molécules difficilement observables avec Pymol. En effet, sa maniabilité et ses options le rendent moins intéressant que Pymol dans ce type d'étude.

VMD[16] **[98]** Très utilisé en dynamique moléculaire, ce logiciel dispose de nombreuses options associées. Très performant, Pymol lui a été préféré uniquement à cause de l'API associée et de la flexibilité du langage Python.

Turbo Frodo[17] **[192] et O**[18] Ces deux logiciels sont dédiés à la reconstruction de molécules dans l'enveloppe de densité électronique issue des expériences de cristallographie X. Ils ont été utilisés dans la deuxième partie de ce travail.

2.4 Constitution de l'échantillon d'apprentissage

2.4.1 La banque de données de structures

La *Protein Data Bank* ou *PDB* est la banque de données qui contient toutes les données structurales (tridimensionnelles) associées aux protéines et aux acides nucléiques [14]. Ces données, principalement obtenues par cristallographie aux rayons X et résonance magnétique nucléaire (RMN), sont mises à disposition du public par l'intermédiaire d'un site internet.

Mise en place en 1971 par le *Brookhaven National Laboratory*, la *Protein Data Bank* a été transférée en 1998 au *Research Collaboratory for Structural Bioinformatics (RCSB)*[19] qui est un regroupement de plusieurs laboratoires de recherche américains. En 2003, trois organisations membres, le RCSB aux USA, le MSD-EBI en Europe et la PDBj au Japon, s'associent pour former la *Worldwide Protein Data Bank*[20] [15]. Les données issues de cette source sont unifiées et de nouveaux modes de représentation y sont associés [228].

[14] http://condor.urbb.jussieu.fr/logiciels/XmMol.html
[15] http://www.cgl.ucsf.edu/chimera/
[16] http://www.ks.uiuc.edu/Research/vmd/
[17] http://www.afmb.univ-mrs.fr/-TURBO-
[18] http://xray.bmc.uu.se/~alwyn/
[19] http://www.rcsb.org
[20] http://www.wwpdb.org

À sa création, en 1977, la *Protein Data Bank* contenait 7 structures protéiques. Depuis, le nombre de structure n'a cessé d'augmenter et à l'heure actuelle, avec l'essor des projets de génomique structurale, on voit apparaître plus de 2000 nouvelles structures par an. Dans la version figée de décembre 2004 (*PDB release #1*), la *Protein Data Bank* contenait plus de 27 000 fichiers de coordonnées atomiques.

Le format PDB et les formats qui en sont dérivés sont très complets [227]. Cependant pour des raisons historiques, le format défini au départ n'a pas été toujours suivi par les dépositaires de structures. Initialement, dans le format, il est prévu de pouvoir savoir si une structure est celle d'un complexe par un champ (noté `REMARK 900`). Mais aucune extraction systématique non redondante n'est possible à travers ce champ.

2.4.2 Les complexes binaires non redondants de la *Protein Data Bank*

La dernière étude complète sur les assemblages protéiques de la *Protein Data Bank* fait état de plus de 400 structures de complexes [64]. Ces données, très redondantes, ne permettent pas de déterminer un jeu de complexes qui convienne pour une étude statistique.

Dans un premier temps, nous avons donc travaillé sur des listes de complexes protéine-protéine bien connues et étudiées [38, 144], puis, nous avons décidé de réaliser une extraction exhaustive des complexes binaires dont la structure a été déterminée à haute résolution. Cette extraction sera dans un proche avenir complétée par l'extraction des complexes non-binaires selon la même stratégie.

Pour réaliser cette extraction, nous avons mis en place la procédure d'extraction systématique décrite par la figure 2.15 sur la version figée *PDB release #1* de 2004. Des complexes obtenus, seuls ceux ayant au moins un partenaire isolé existant dans la *Protein Data Bank* ont été retenus. La procédure d'extraction fournit un jeu de 102 complexes décrit dans les tableaux 2.1 et 2.2.

Tab. 2.1: 22 complexes non-liés/non-liés issus de l'extraction systématique

Code PDB du complexe	Chaînes des partenaires dans le complexe		Code PDB du premier partenaire	Chaîne dans le premier partenaire	Code PDB du second partenaire	Chaîne dans le second partenaire
1s6v	A	B	1kok	A	1irw	- -
1ml0	A	D	1mkf	A	1dol	- -
1fin	A	B	1pf8	A	1vin	- -
1ugh	E	I	1akz	- -	2ugi	A
1b2s	A	D	1bnj	A	1a19	A
1ghq	A	B	1c3d	- -	1ly2	A
1f6m	A	C	1cl0	A	1keb	A
1e6e	A	B	1e1n	A	1cje	A

TAB. 2.1: 22 complexes non-liés/non-liés issus de l'extraction systématique (suite)

Code PDB du complexe	Chaînes des partenaires dans le complexe		Code PDB du premier partenaire	Chaîne dans le premier partenaire	Code PDB du second partenaire	Chaîne dans le second partenaire
1u7f	A	B	1khx	A	1dd1	A
1dkf	A	B	1lbd	- -	1xap	A
1nbf	A	C	1nb8	A	1f9j	A
1kac	A	B	1nob	A	1f5w	A
1exb	A	E	1qrq	A	1qdv	A
1noc	A	B	1r35	A	1q23	A
1we3	A	O	1srv	A	1wnr	A
7cei	A	B	1unk	A	1m08	A
1pxv	A	C	1x9y	A	1nyc	A
1p2k	A	I	1xuk	- -	2hex	A
1jk9	A	B	1yaz	A	1qup	A
1n95	A	B	1ft1	A	1fpp	B
1kgy	A	E	1nuk	A	1iko	P
1t6b	X	Y	1acc	- -	1shu	X

TAB. 2.2: 80 complexes non-liés/liés issus de l'extraction systématique

Code PDB du complexe	Chaînes des partenaires dans le complexe		Code PDB du partenaire non-lié	Chaîne dans le partenaire non-lié
1s3s	A	G	1e32	A
1k9o	E	I	1ane	- -
1dtd	A	B	1aye	- -
1ava	A	C	1bg9	- -
1us7	A	B	1bgq	- -
1i1r	A	B	1bqu	A
1sg1	A	X	1btg	A
1oxb	A	B	1c03	A
3ygs	C	P	1cy5	A
1xb2	A	B	1d2e	A
1c4z	A	D	1d5f	A
1f93	A	E	1dcp	A
1qe1	A	B	1dlo	A
1oc0	A	B	1dvm	A
1f1b	A	B	1ekx	A
1f02	I	T	1f00	I
1t0f	A	C	1f1z	A
1f80	A	D	1f7t	A

TAB. 2.2: 80 complexes non-liés/liés issus de l'extraction systématique (suite)

Code PDB du complexe	Chaînes des partenaires dans le complexe		Code PDB du partenaire non-lié	Code PDB du second partenaire
1dp5	A	B	1fq8	A
1h2s	A	B	1gue	A
1gl4	A	B	1h4u	A
1usu	A	B	1hk7	A
1ktd	A	B	1ieb	A
1d2z	A	B	1ik7	A
1j7v	L	R	1inr	- -
1jtd	A	B	1jwz	A
1l2w	A	I	1jya	A
1jzd	A	C	1jzo	A
1hx1	A	B	1kaz	- -
1lqv	A	C	1l8j	A
1k3z	A	D	1my5	A
1ujw	A	B	1nqe	A
1ta3	A	B	1om0	A
1ory	A	B	1orj	A
1dhk	A	B	1ose	- -
1x79	A	B	1oxz	A
1pdk	A	B	1qpp	A
1ewy	A	C	1que	- -
1nvu	Q	S	1rvd	A
1uea	A	B	1sln	- -
1sv0	A	C	1sv4	A
1t6g	A	C	1t6e	X
1kzy	A	C	1uol	A
1tf0	A	B	1uor	- -
1jlt	A	B	1vpi	- -
1m9f	A	C	1w8v	A
1xtg	A	B	1xtf	A
1qty	V	X	2vpf	A
1e44	A	B	3eip	A
1bgx	H	T	5ktq	A
1ma9	A	B	1s22	A
1vg9	A	B	1vg8	A
1a4y	A	B	2ang	A
1fqk	A	B	1fqi	A
1tco	A	C	1qpl	A
1m4u	A	L	1bmp	- -
1nf5	A	B	1fgx	A
1lw6	E	I	1ciq	A

TAB. 2.2: 80 complexes non-liés/liés issus de l'extraction systématique (suite)

Code PDB du complexe	Chaînes des partenaires dans le complexe		Code PDB du partenaire non-lié	Code PDB du second partenaire
1d4v	A	B	1dg6	A
1eer	A	B	1ern	A
1lzw	A	B	1k6k	A
1svx	A	B	1r6z	P
1kfu	L	S	1nx3	A
1jiw	I	P	1akl	- -
1f3v	A	B	1ca4	A
1ct0	E	I	1ds3	I
1dn1	A	B	1ez3	A
1wpl	A	K	1jg5	A
1npe	A	B	1klo	- -
1ib1	A	E	1kuy	A
1gpw	A	B	1kxj	A
1ofu	A	X	1oft	A
1qav	A	B	1qau	A
1rke	A	B	1qkr	A
1r4q	A	B	1qnu	A
1tue	A	B	1qqh	A
1gfw	A	B	1qx3	A
1y01	A	B	1sdl	A
1xdt	R	T	1tox	A
1rj9	A	B	2ng1	- -

2.4.3 Génération de complexes non-natifs

Pour chacun des 102 complexes du jeu précédent, nous avons pris les partenaires séparés et utilisé le programme *DOCK* pour générer des conformations non-natives. Nous avons réalisé un échantillonnage systématique avec un pas angulaire de 10° sur toute la gamme d'angles, de façon à obtenir 18.10^6 conformations différentes pour chacun des complexes. Pour la suite, nous avons sélectionné aléatoirement à chaque fois 200 complexes parmi ces 18.10^6 conformations, pour en faire un jeu de complexes non natifs. À chaque sélection, nous avons vérifié qu'aucun complexe « presque » natif ne se trouvait dans le jeu de travail.

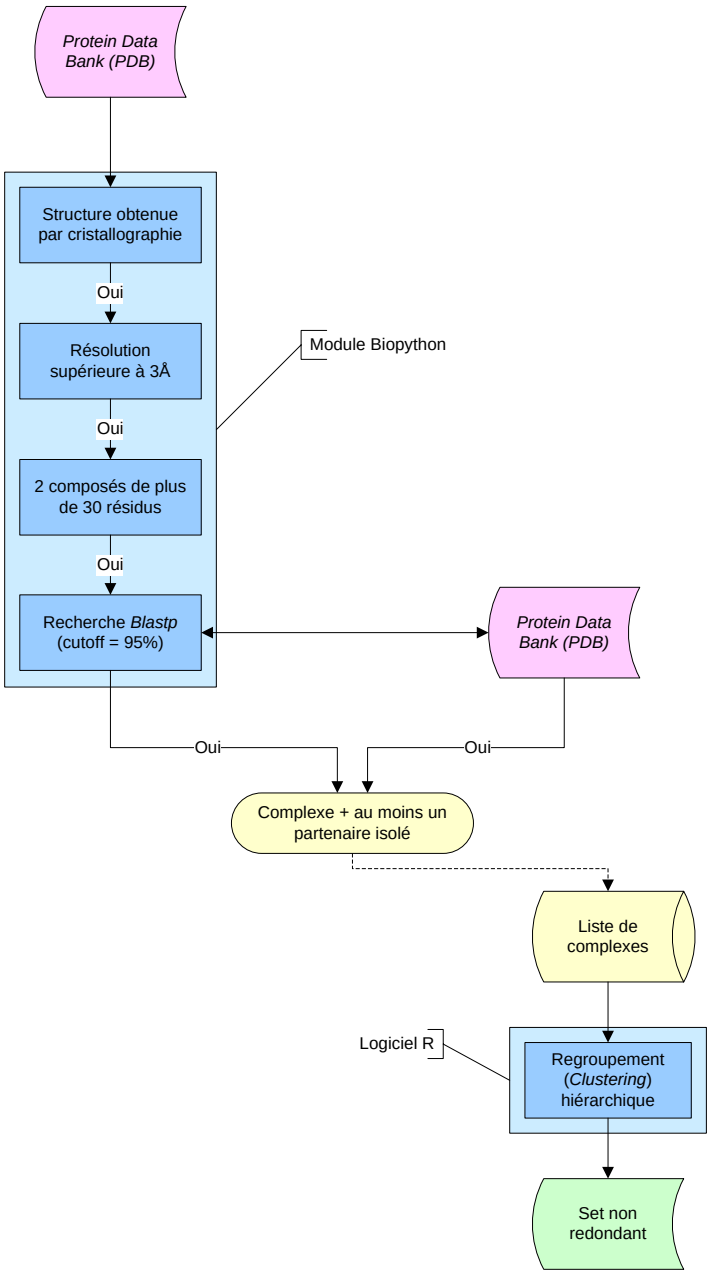


FIG. 2.15 – Procédure d'extraction des complexes binaires de la *Protein Data Bank*. *À l'aide du module* Biopython *[92], nous avons extrait les entrées correspondant à des structures cristallographiques de résolution supérieure à 3 Å et qui contiennent deux chaînes polypeptidiques longues de plus de 20 résidus. Ensuite, nous avons cherché avec* Blastp *[3] les entrées contenant les partenaires isolés et conservé ceux ayant plus de 95% d'identité de séquence avec la chaîne rencontrée dans le complexe. Une procédure de groupement (*clustering*) hiérarchique, effectuée à l'aide du logiciel* R *[182], permet enfin d'obtenir un jeu non redondant de 102 complexes (22 non-liés/non-liés et 80 non-liés/liés).*

2.5 L'apprentissage

Les valeurs des paramètres présentés précédemment (paragraphe 2.3.1.4 page 41) ont été mesurées sur les 102 complexes natifs et sur les complexes non-natifs du jeu d'apprentissage. Ces valeurs ont ensuite été utilisées en entrée de procédures d'apprentissage (*Machine Learning*) pour optimiser des fonctions de score de façon à ce qu'elles discriminent le mieux possible entre complexes natifs et non-natifs.

2.5.1 La fonction logistique

2.5.1.1 Présentation

Pour sélectionner les paramètres d'apprentissage, nous avons choisi d'utiliser un premier modèle d'apprentissage très simple : la fonction logistique. La fonction logistique est aussi appelée perceptron, c'est un réseau de neurones sans couche cachée. Ce modèle a déjà été utilisé par nos collaborateurs dans le cadre d'études de la structure des protéines [155, 156]. Il présente l'avantage, en plus d'être très simple, de permettre de connaître l'influence relative directe de chacun des paramètres, ce qui aide à l'interprétation.

Au départ, cette étude impliquait l'utilisation de 461 variables. Pour des raisons de convergence, l'échantillon d'apprentissage étant relativement petit, nous avons utilisé les 84 variables présentées au paragraphe 2.3.1.4 page 41.

L'entrée standard d'une fonction logistique est un vecteur $X[X_i]$ des descripteurs correspondant ici aux 84 variables présentées précédemment. La probabilité d'observer un complexe natif est évaluée par :

$$P(X) = 1/\left[1 + exp\left(-w_o - \sum_i w_i X_i\right)\right] \tag{2.23}$$

où w_i est le poids du descripteur à la position i du vecteur d'entrée et w_0 un poids initial global.

Le vecteur de poids $W[w_i]$ est estimé par maximum de vraisemblance sur le jeu d'apprentissage à l'aide du modèle linéaire généralisé (GLM) du logiciel *R* [182]. On obtient une fonction de score pour laquelle on connaît le poids relatif de chacun des paramètres.

2.5.1.2 Règle de décision et évaluation de la prédiction

Le taux d'erreur associé à la prédiction est défini comme :

$$R = P(Pred = Biol/Obs = Decoy) \cdot\ P(Obs = Decoy) + P(Pred = Decoy/Obs = Biol) \cdot\ P(Obs = Biol)$$

où *Pred* et *Obs* représentent, respectivement, la prédiction et la réalité ; *Biol* et *Decoy*, un complexe natif et un complexe non-natif. Ce taux d'erreur correspond à la probabilité de prédire à la fois un complexe natif quand un non-natif est observé, et de prédire un complexe non-natif quand un complexe natif est observé. La précision de la prédiction Q est égale à $1 - R$.

Étant donnée une fonction logistique et son vecteur de poids associé, la règle de décision consiste à tester si la sortie $P(X)$ de la fonction logistique est supérieure ou inférieure à un seuil Pth. Si $P(X) > Pth$, la prédiction est *Biol* ; sinon, elle est *Decoy*. Après avoir estimé les poids du modèle logistique, nous avons déterminé le seuil en minimisant le taux d'erreur R. Cela a été réalisé en découpant l'intervalle $[0, 1]$ en segments de 0,01, puis en calculant la valeur de R à chaque étape, et en gardant la valeur de la tranche pour laquelle on obtenait un R minimal.

2.5.1.3 Validation croisée

Pour permettre la validation croisée des procédures de prédiction, on découpe le jeu de départ en 10 sous-ensembles sélectionnés aléatoirement. Pour chacun de ces sous-ensembles, les 9 autres sous-ensembles sont utilisés comme jeu d'apprentissage, alors que le sous-ensemble restant est utilisé pour la validation.

2.5.2 *ROc based GEnetic learneR (ROGER)*

En étroite collaboration avec Jérôme Azé du LRI à Orsay, nous avons aussi utilisé un algorithme génétique fondé sur l'étude de la courbe de ROC. Cet algorithme, appelé *ROGER (ROc based GEnetic learneR)* [188, 200, 201] utilise le critère de ROC (*Receiver Operating Characteristics*) qui est connu comme bon critère pour l'évaluation et la comparaison de classificateurs [142, 180].

2.5.2.1 Description

Les courbes de ROC, issues du traitement du signal, et rendues populaires par l'analyse de données médicales, représentent la relation entre le taux de vrais positifs et le taux de faux positifs. Le but des algorithmes d'apprentissage peut alors être envisagé comme un problème d'optimisation multi-objectifs : maximiser le taux de vrais positifs tout en minimisant le taux de faux positifs. Le cas idéal correspond donc au point (0,1) (voir figure 2.16), avec 100% d'exemples qui sont des vrais positifs. Ainsi, l'aire sous la courbe (notée AUC pour*Area Under the Curve*) peut être vue comme une mesure globale de l'efficacité de l'apprentissage.

L'optimisation de l'aire sous la courbe (AUC) est un problème NP-complet, qui a été traité de différentes façons. Le programme *ROGER* traite le problème en utilisation des stratégies d'évolutions dans un algorithme génétique efficace pour l'optimisation numérique [6].

2.5.2.2 Algorithme

ROGER explore l'espace des hypothèses continues, projetant l'espace des exemples dans l'espace des réels $\mathbb{R}$. Le jeu de données, noté E, est composé de n exemples $(x_i, y_i), i = 1 \dots n$, où $x_i \in X$ indique la description du i-ème exemple ($X \subset \mathbb{R}^d, d$ étant le nombre de caractéristiques) et y_i indique l'étiquette correspondante ($y_i = \pm 1$).

Dans sa première version, *ROGER* explore l'espace des hypothèses linéaires sur $\mathbb{R}$. À chaque génotype $w = (w_i, \dots, w_d) \in \mathbb{R}^d$ est associé une hypothèse (phénotype) h_w définie sur X comme :

$$h_w(x) =< w, x > \tag{2.24}$$

La fonction d'adaptation (*fitness*) $F(w)$ associée au génotype w est définie comme la fraction de paires d'exemples (positifs, négatifs) qui sont classés correctement selon h_w :

$$F(w) = Pr(h_w(x_i) > h_w(x_j) | y_i > y_j) \tag{2.25}$$

Si on se réfère à la courbe de ROC, on voit que cela permet de se rapprocher du cas idéal, donc cela maximise l'aire sous la courbe.

Certains types d'hypothèses non-linéaires peuvent aussi être considérés en doublant l'espace de recherche. En particulier, à un individu $z = (w_1, \dots, w_d, c_1, \dots, c_d) \in \mathbb{R}^{2d}$ peut être associé à

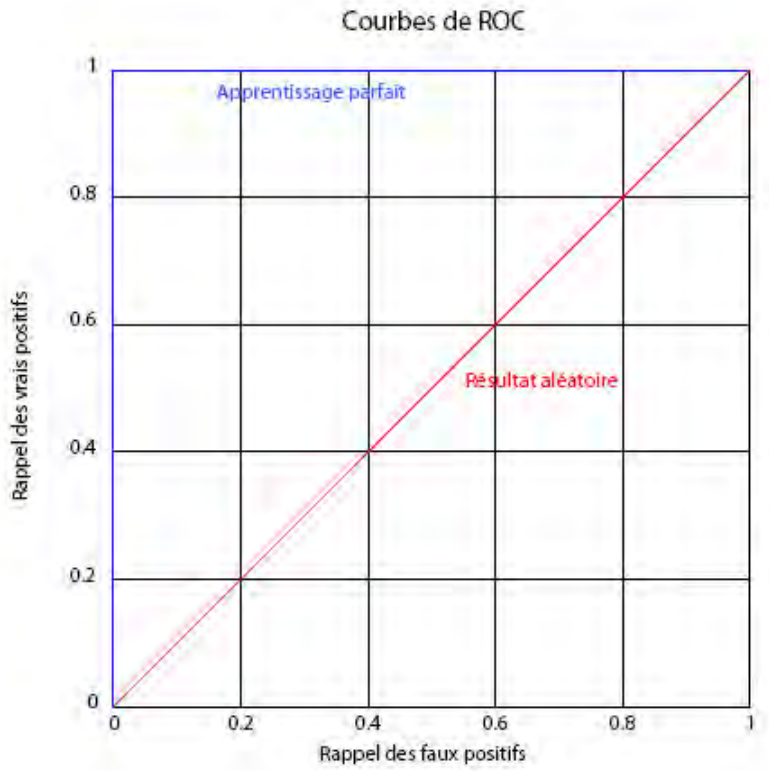


FIG. 2.16 – Exemple de courbe de ROC (*Receiver Operating Characteristics*). *En bleu, la courbe idéale, avec 100% de vrais positifs (AUC=1) et en rouge, la diagonale, qui correspond à un résultat aléatoire (AUC=0,5).*

une hypothèse h_x définie comme :

$$h_x(x = (x^1, \ldots, x^d)) = \sum_{j=1}^{d} w_i \times |x^j - c_j| \tag{2.26}$$

La fonction d'adaptation est alors calculée comme précédemment.

Dans ces deux cas, l'optimisation de F est obtenue par une stratégie d'évolution, utilisant la mutation auto-adaptative et le crossover (figure 2.17).

2.5.3 Séparateurs à Vaste Marge (*SVM*)

2.5.3.1 Présentation

Avec un ensemble de vecteurs d'apprentissage étiquetés positivement et négativement, l'algorithme d'apprentissage par séparateurs à vaste marge (*SVM* ou *Support Vector Machines*) apprend un classificateur qui peut ensuite être utilisé pour étiqueter des exemples test non étiquetés [57, 198]. Les exemples en entrée de l'apprentissage $\{y_1, \ldots, y_n\}$ sont projetés dans l'espace multidimensionnel des caractéristiques et l'algorithme recherche un hyperplan qui sépare les exemples positifs des exemples négatifs, avec la marge la plus grande possible, c'est-à-dire la distance au point le plus proche (voir figure 2.18).

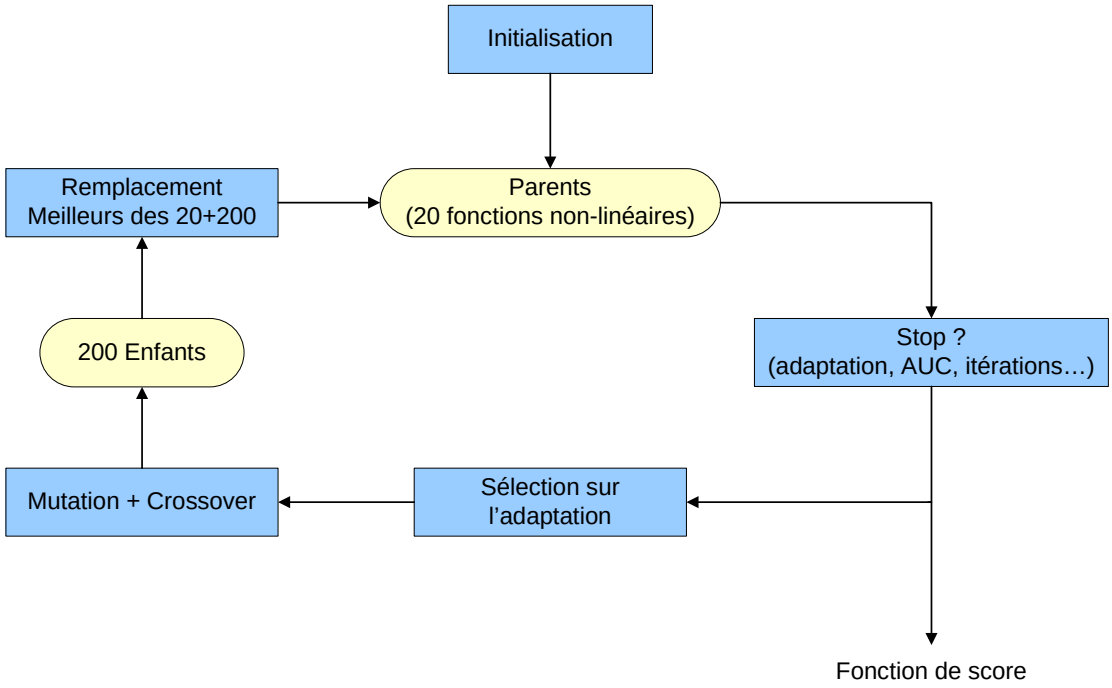


FIG. 2.17 – Principe de l'algorithme génétique de ROGER (ROc based GEnetic learneR). *Après une étape d'initialisation, on dispose de 20 fonctions non-linéaires, les parents, dont l'adaptation va être calculée puis comparée à une condition d'arrêt. Ensuite, il y a sélection sur l'adaptation, mutation et crossover et on obtient 200 fonctions, les enfants. Parmi les 200 enfants et les 20 parents, on sélectionne les 20 meilleurs qui deviennent les parents à l'itération suivante.*

Si le jeu d'apprentissage n'est pas séparable linéairement, les séparateurs à vaste marge trouvent un hyperplan qui réalise un compromis entre bonne classification et large marge.

2.5.3.2 Noyaux

Au lieu de projeter explicitement les objets dans un espace des caractéristiques multidimensionnel H, les SVM travaillent implicitement dans l'espace des caractéristiques en calculant uniquement le noyau correspondant $K(\vec{x}, \vec{y})$ (aussi appelé *kernel*), entre deux objets x et y, défini par :

$$K(\vec{x}, \vec{y}) = \Phi(\vec{x}).\Phi(\vec{y}) \tag{2.27}$$

où Φ est la projection dans l'espace des caractéristiques H.

Les logiciels de SVM comprennent en général les noyaux suivants :

– le noyau linéaire :

$$K(\vec{x}, \vec{y}) = \vec{x}.\vec{y} \tag{2.28}$$

– le noyau polynomial :

$$K(\vec{x}, \vec{y}) = (\vec{x}.\vec{y} + 1)^d \tag{2.29}$$

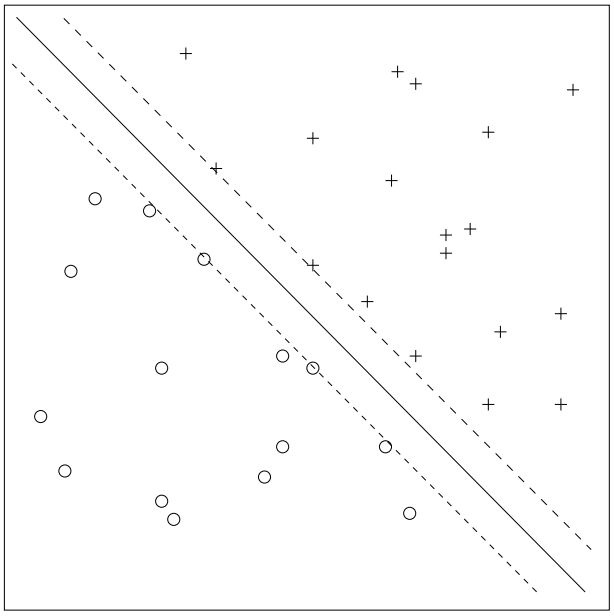


FIG. 2.18 – Schéma de séparation par un hyperplan (SVM). *Les croix et les cercles représentent respectivement les exemples positifs et négatifs. Les lignes pointillées montrent les vecteurs support.*

– le noyau gaussien (ou *RBF* pour *Radial Basic Function*) :

$$K(\vec{x}, \vec{y}) = exp\left(\frac{\|\vec{x} - \vec{y}\|}{\sigma}\right)^2 \tag{2.30}$$

où σ est la largeur du noyau (des petites valeurs de σ rendent la limite de décision moins contraignante).

2.5.3.3 Calculs

Pour projeter les objets dans l'espace multidimensionnel des caractéristiques, nous avons testé les différents noyaux précédents. Pour les noyaux polynomial et gaussien, nous avons testé plusieurs valeurs de d et de σ. Le paramètre C, qui contrôle le compromis entre erreurs d'apprentissage et marge, a été gardé à sa valeur par défaut :

$$C = \frac{N}{\Sigma K\left(\vec{x_i}, \vec{x_i}\right))} \tag{2.31}$$

où N est la taille du jeu d'apprentissage. Des essais d'optimisations à l'aide de la fonction *tune* de *libSVM* ont été réalisés, sans permettre de parvenir à de meilleurs résultats.

Les calculs ont été effectués à l'aide des logiciels *libSVM* [36] et *SVMTorch* [47].

2.5.4 Traitement des données manquantes

La taille de l'échantillon d'apprentissage ainsi que la définition des paramètres font que les mesures en entrée de la procédure d'apprentissage contiennent de nombreuses valeurs manquantes. À l'heure actuelle, plusieurs méthodes sont disponibles pour traiter les cas d'entrées qui contiennent des valeurs manquantes. Ce sont principalement :

- enlever les entrées ayant des valeurs manquantes ;
- remplacer les valeurs manquantes par la moyenne/médiane sur une classe d'entrée (les complexes natifs ou non-natifs) ou sur l'ensemble du jeu de données ;
- inférer les valeurs manquantes à l'aide d'une méthode de régression ou d'une autre méthode itérative.

Dans cette étude, il ne nous était pas possible de mettre de côté les entrées ayant des valeurs manquantes, car toutes les entrées ont des valeurs manquantes. Au vu des courbes de répartitions (voir paragraphe 3.1.1 page 60) et de différents essais d'apprentissage, nous avons décidé de remplacer les valeurs manquantes par la médiane du paramètre observée sur les complexes natifs. La répartition n'étant pas gaussienne, la médiane est en effet moins sensible aux variations que la moyenne, et l'observation de la mesure sur les complexes natifs uniquement permet d'obtenir un apprentissage plus stable.

Chapitre 3

Résultats et Discussion

Sommaire

3.1 Un modèle simple mais utile

Dans cette partie, seront présentées les mesures effectuées pour tester le modèle de représentation de la structure protéique au moyen du diagramme de Voronoï, et le comparer aux études précédentes. Ces mesures ont été réalisées sur les 102 structures de complexes extraits de la PDB (voir paragraphe 2.4.2 page 48).

3.1.1 Mesures

3.1.1.1 Surfaces des interfaces et compositions en acides aminés

Surfaces Les surfaces des interfaces obtenues (voir figure 3.1) sur le jeu de complexes sont conformes aux données de la littérature [144]. Pour un complexe donné, la surface telle que mesurée par notre méthode est donc proche de celle calculée par la méthode de Lee et Richards. Cela montre en particulier que le fait de plonger la protéine dans les sphères de solvant dont la taille est très différente de celle des molécules d'eau n'affecte pas l'aire de la surface.

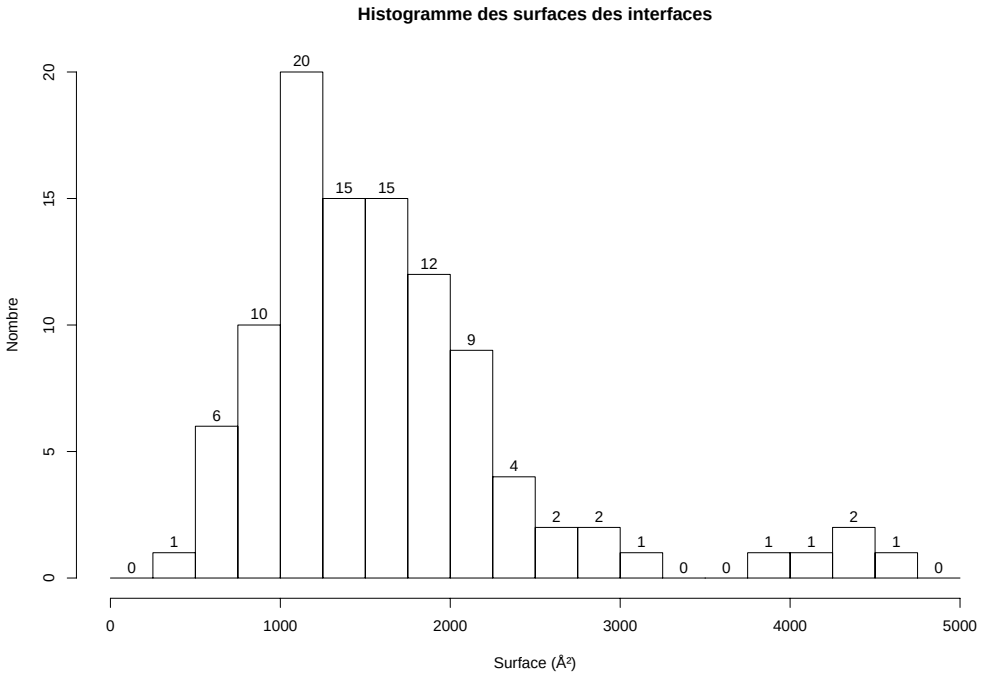


FIG. 3.1 – Graphe de répartition des surfaces des interfaces.

Composition Pour chaque acide aminé, on peut observer les différences de composition en acides aminés à l'interface (figure 3.2). On peut observer que notre méthode donne plus de « petits » acides aminés et moins de « gros » acides aminés à l'interface (sauf pour la phénylalanine). On pouvait s'attendre à ce résultat étant donné la méthode de comptage. On peut cependant noter que les différences dans les pourcentages ne sont pas uniquement corrélées à la taille du résidu. La méthode de mesure utilisée par Lo Conte et collaborateurs ne prend en compte que les résidus ayant des atomes directement impliqués dans l'interface. Notre définition du voisinage permet de mesurer également l'influence des résidus situés immédiatement sous la surface en raison de la définition du voisinage (voir figure 2.9).

Le fait que les différences entre les différences de pourcentage ne soient pas uniquement corrélées à la taille du résidu pourrait signifier que cette « sous-couche » joue un rôle important dans l'interaction.

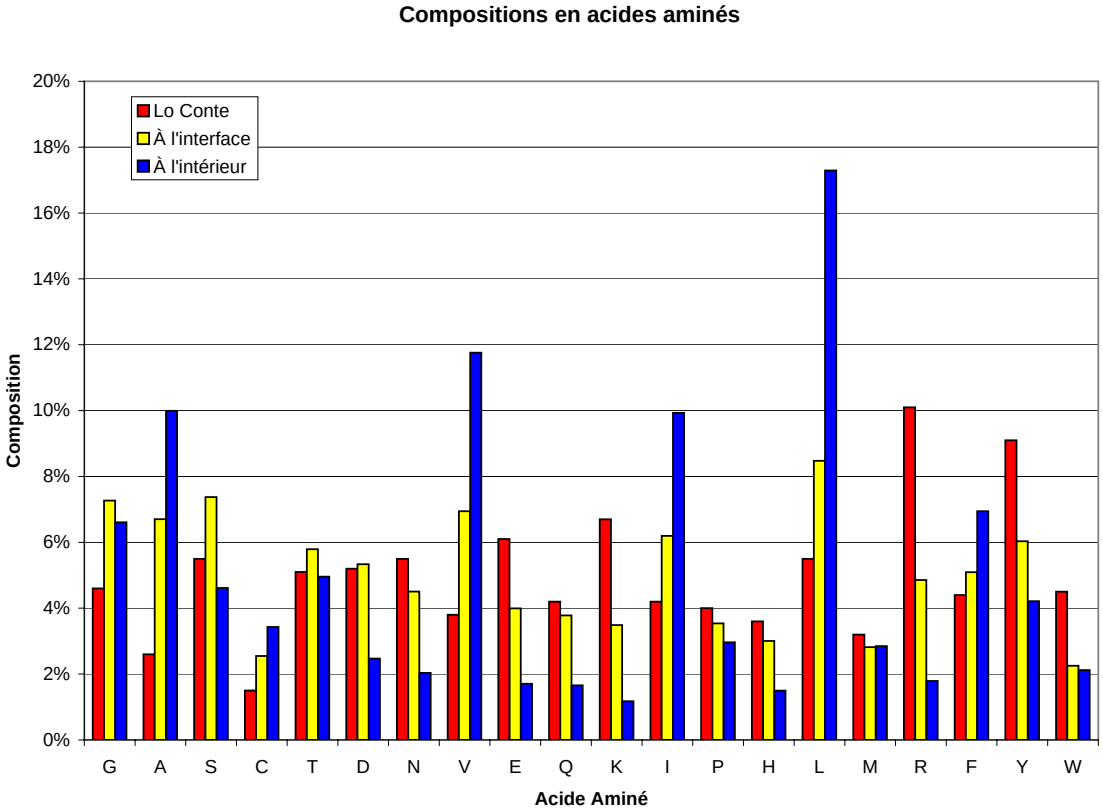


FIG. 3.2 – Composition en acides aminés. *Les acides aminés sont classés par volume de Pontius croissant [177].*

3.1.1.2 Affinité entre acides aminés à l'interface

Distances À l'interface entre les différents partenaires, les distances entre acides aminés voisins au sens de Voronoï ont été calculées. À partir de ces mesures, on peut obtenir les distances moyennes entre résidus à l'interface, et en tracer les répartitions (figures 3.3, 3.4 et 3.5).

Ces répartitions, bien que non gaussiennes, peuvent êtres ajustées sur des distributions statistiques classiques (fonction de Pearson par exemple). Il pourrait être intéressant dans l'avenir d'utiliser ces répartitions par exemple au travers de potentiels. Cependant, c'est actuellement impossible, car pour une majorité de types de paires, les données sont insuffisantes.

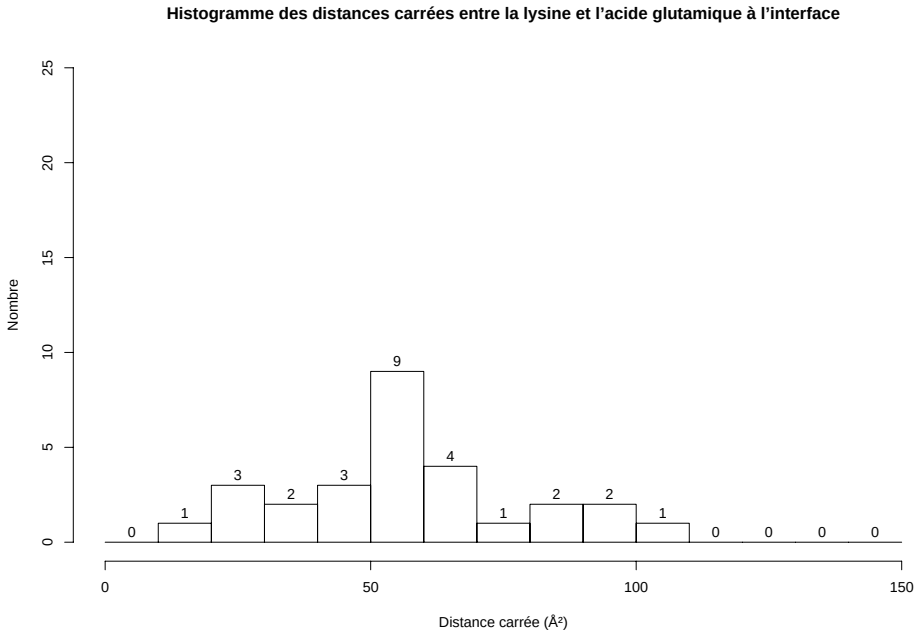


FIG. 3.3 – Graphe de répartition des distances entre la lysine et l'acide glutamique à l'interface.

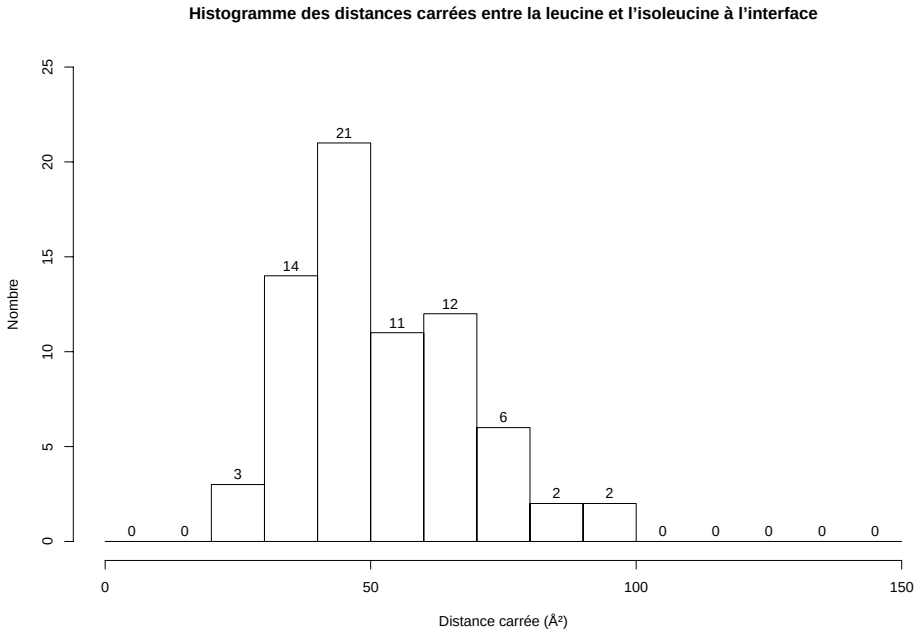


FIG. 3.4 – Graphe de répartition des distances entre la leucine et l'isoleucine à l'interface.

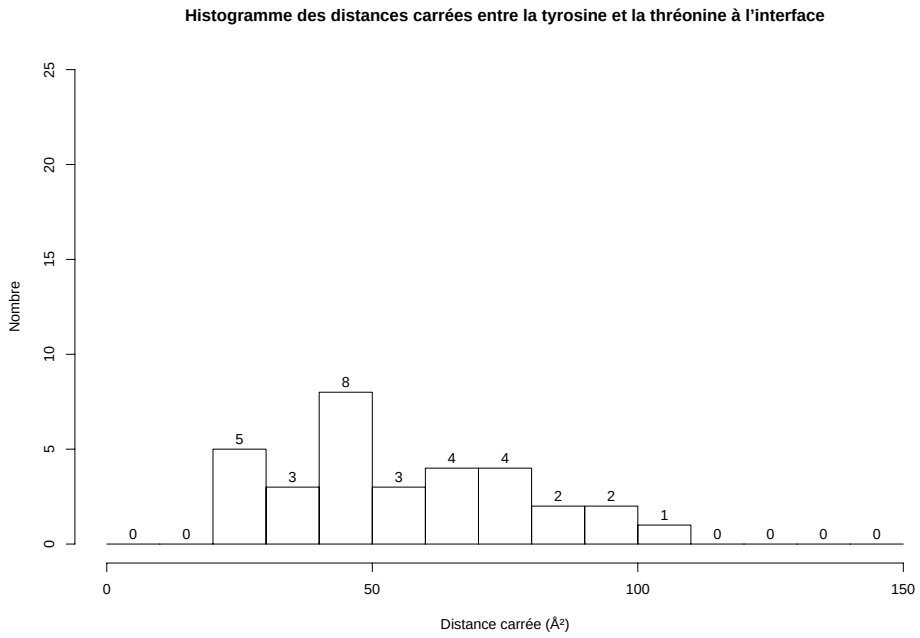


FIG. 3.5 – Graphe de répartition des distances entre la tyrosine et la thréonine à l'interface.

3.1.1.3 Mesure de l'empilement

Les volumes moyens des cellules des acides aminés à l'interface entre deux chaînes et à l'intérieur d'une même chaîne ont été calculés et comparés entre eux, et comparés aux volumes des résidus obtenus par J. Pontius [177] (calculés par un Voronoï atomique). Ce type de mesure permet d'observer la compacité de l'empilement des résidus dans les protéines [76, 207].

Il apparaît immédiatement sur l'histogramme présenté en figure 3.6, que les volumes que nous calculons pour les « petits » acides aminés sont très surévalués, et ceux des « gros » très sous-évalués. Nous avons cependant pu montrer qu'ils sont très bien corrélés [178]. Ce phénomène est essentiellement dû au fait que nous utilisons un diagramme non pondéré.

On remarque aussi que les volumes moyens des cellules obtenus à l'interface entre deux chaînes sont supérieurs à ceux obtenus à l'intérieur d'une même chaîne et cette différence est significative. À l'interface entre deux chaînes, l'assemblage des chaînes latérales est moins compact. Cette différence de compacité est compensée par la présence de molécules d'eau que nous ne prenons pas en compte dans ces mesures. J. Janin et ses collaborateurs ont d'ailleurs montré que le nombre de molécules d'eau est caractéristique du type d'interface [189].

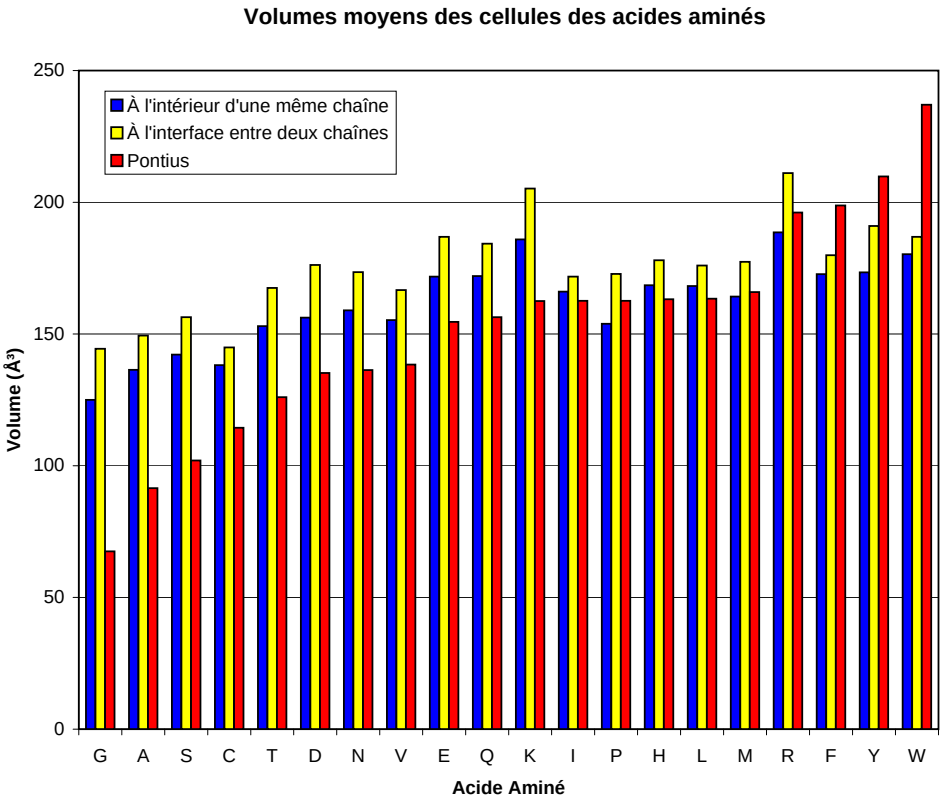


FIG. 3.6 – Volumes des cellules de Voronoï des différents acides aminés. *En bleu, les volumes à l'intérieur d'une chaîne ; en jaune, ceux à l'interface et en rouge, les volumes de l'étude de J. Pontius [177].*

3.1.2 Diagramme de Laguerre

3.1.2.1 Motivation

Comme indiqué précédemment, les volumes calculés dans notre modèle sont plus grands que ceux de Pontius pour les petits acides aminés et plus petits pour les gros acides aminés. La raison est que la face qui sépare deux acides aminés est toujours située à mi-distance entre les deux centres de gravité, quelle que soit la nature des acides aminés. Le diagramme de Voronoï ne dépend donc de la nature des acides aminés que par le fait que la distribution de leurs centres de gravité dépend de leur encombrement stérique. Afin de mieux prendre en compte cet aspect, nous avons essayé de construire le diagramme de Laguerre (paragraphe 2.2.1.4 page 31), en attribuant à chaque acide aminé certain rayon de manière à ce que le volume de sa cellule soit en moyenne égal à son volume de Pontius.

3.1.2.2 Problèmes

Avec cette construction, nous avons rencontré un certain nombre de problèmes qui nous ont fait préférer le diagramme de Voronoï :

1. Les écart-types des volumes sont très importants. Certaines cellules, comme celle de la glycine par exemple, disparaissent même complètement dans certaines conditions (figure

3.7), et il arrive fréquemment que le centre géométrique de la chaîne latérale ne soit pas situé dans la cellule. Même si ceci n'est pas gênant d'un point de vue mathématique, cela rend plus difficile l'interprétation biologique.

2. Le calcul est plus lourd et le serait encore plus si on associait un rayon en fonction de l'environnement du résidu, pour éviter le problème précédent.

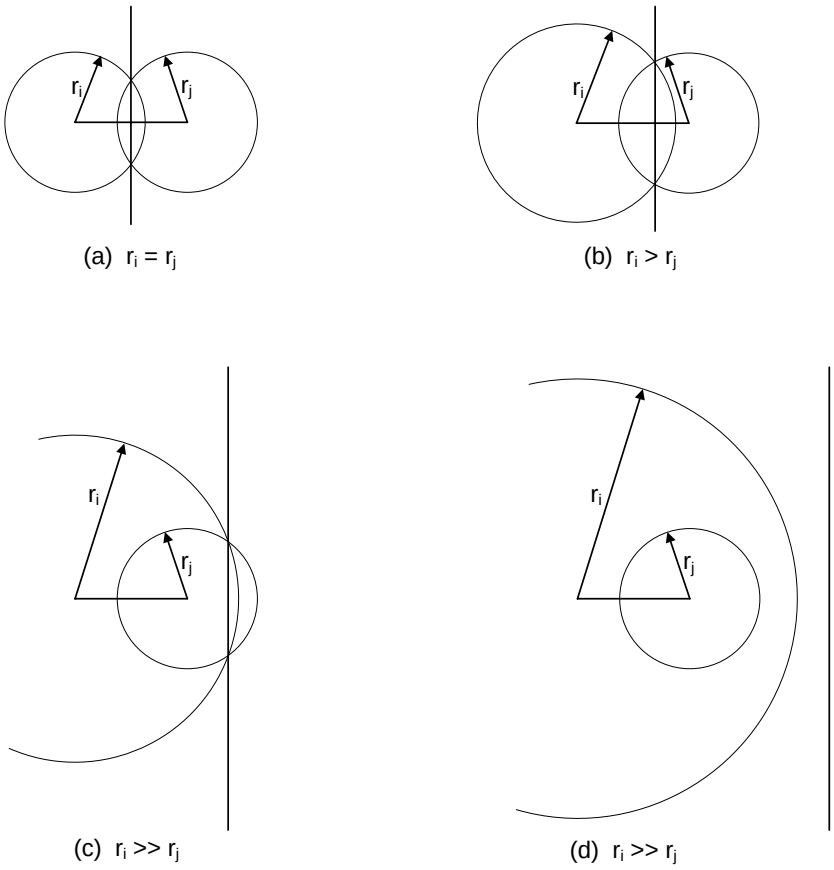


FIG. 3.7 – Position des faces pour un diagramme de Laguerre. *(a) Les deux acides aminés ont le même rayon, le plan qui correspond à un coté d'une face de cellule coupe le segment formé par les deux centres de gravité en son milieu. (b) Lorsque les deux rayons sont différents, le plan est décalé vers l'acide aminé de rayon le plus faible. (c) Ce plan peut ainsi ne pas couper le segment formé par les deux centres de gravité. Dans ce cas, le centroïde n'est pas dans sa cellule. (d) Et, dans les cas extrêmes, la sphère la plus petite n'intersecte pas sa région de Laguerre. Il peut même arriver, avec plus de deux centroïdes, que la région de Laguerre associée « n'existe pas ».*

3.1.3 Choix du centroïde : test du Cα

Pour vérifier que le choix des centres géométriques des chaînes latérales par rapport aux carbones α était pertinent, nous avons également construit le diagramme de Voronoï à partir des carbones α.

Les diagrammes obtenus présentent des différences importantes (figure 3.8) ; en particulier, les surfaces exposées au solvant par les petits acides aminés et les acides aminés hydrophobes sont plus importantes.

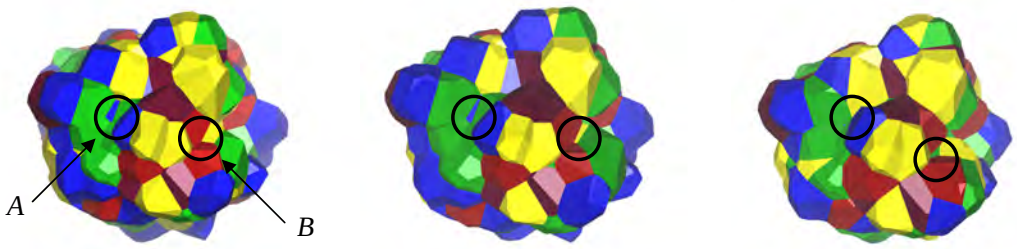


FIG. 3.8 – Comparatif des diagrammes Voronoï / Laguerre / Voronoï Cα pour la thrombine extraite du complexe 1BTH. *À gauche, le diagramme de Voronoï, au centre le diagramme de Laguerre et à droite, le diagramme de Voronoï Cα. Dans la zone A, on observe, pour le diagramme de Voronoï obtenu à partir des Cα, la « disparition » d'un résidu polaire exposé et dans la zone B, l' « exposition » d'un résidu hydrophobe entièrement enfoui lorsque le calcul est effectué à partir des centres de gravité.*

Les volumes moyens des cellules sont significativement différents de ceux obtenus précédemment et sont nettement moins bien corrélés aux volumes de Pontius.

De plus, la forme des cellules calculées à partir des carbones α correspond moins bien à la forme de la chaîne latérale (figure 3.9).

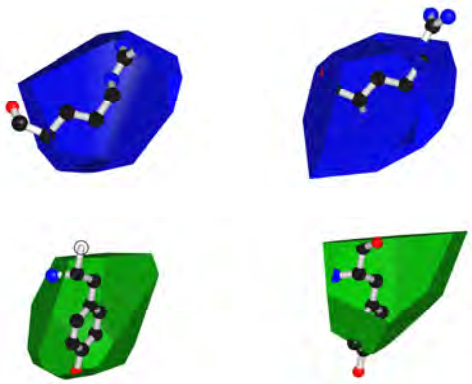


FIG. 3.9 – Cellules de Voronoï de l'arginine 4 et la tyrosine 14 du complexe 1BTH. *À gauche, les cellules calculées à partir des centres de gravité, et à droite, celles calculées à partir des carbones α. On peut observer que la chaîne latérale « sort » de la cellule dans le cas d'une construction à partir des carbones α.*

3.1.4 Premiers essais de classification

Méthode Avant de réaliser l'extraction des complexes sur l'ensemble de la *PDB* (paragraphe 2.4.2 page 48), nous avons testé si les descripteurs que nous proposions étaient ou non discriminants. Pour chacun des complexes décrits dans l'article de L. Lo Conte [144], nous avons calculé

les paramètres présentés au paragraphe 2.3.1.4 et utilisés ceux-ci en entrée d'un apprentissage par fonction logistique (paragraphe 2.5.1) à l'aide des logiciels *R* [182] et *WEKA* [81]. Après avoir calculé la fonction logistique pour l'ensemble des complexes, nous avons effectué une validation croisée en découpant le jeu en 10.

Résultats Nous avons fixé un score de 0 pour les complexes non-natifs et un score de 1 pour les complexes natifs. En réinjectant chacun des complexes du jeu d'apprentissage et en évaluant son score, on obtient le graphe présenté à la figure 3.10.

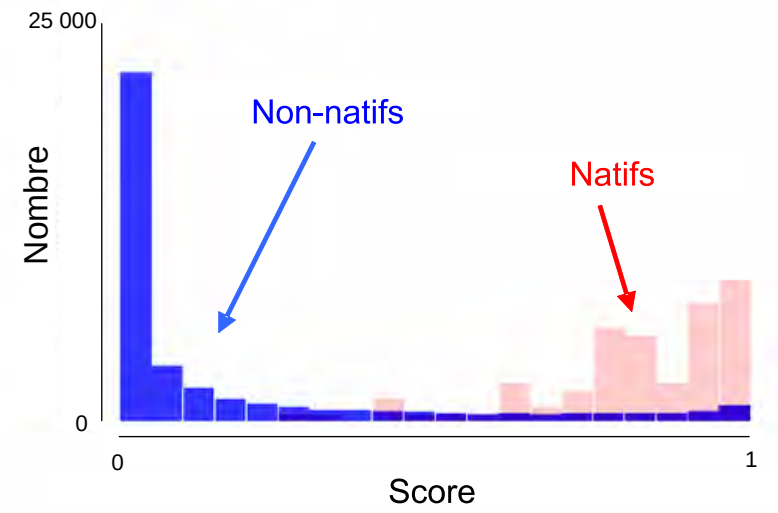


FIG. 3.10 – Graphes des scores obtenus par la fonction logistique.*En rouge, les complexes natifs et en bleu, les complexes non natifs.*

Nous avons également testé l'écart à la moyenne. Sur le même jeu de données qu'au paragraphe précédent, nous avons comparé les courbes de ROC obtenue par la fonction logistique avec celles obtenues par un score qui serait défini par l'écart à la moyenne des paramètres :

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x_i})^2}{\sigma_i} \tag{3.1}$$

On obtient les courbes présentées dans la figure 3.11.

Pour déterminer si tous les types paramètres étaient importants, nous avons aussi effectué des « tests emboîtés » : l'apprentissage avec la fonction logistique a été calculé en enlevant chaque groupe de paramètres. Les tests ont montré une meilleure performance en utilisant l'ensemble des paramètres.

Conclusion Les résultats montrent très clairement que :

- l'écart à la moyenne donne des résultats quasi aléatoires ;
- la fonction logistique donne des résultats encourageants qui montrent que les paramètres sont discriminants. Cependant, la classification obtenue par cette méthode est peu sélective. En particulier, on obtient beaucoup de faux positifs quelque soit le seuil choisi.

Dans la suite, nous avons donc utilisé ces mêmes paramètres, mais en utilisant des méthodes d'apprentissage plus puissantes : l'algorithme ROGER (paragraphe 2.5.2 page 54) et les SVM (paragraphe 2.5.3 page 55).

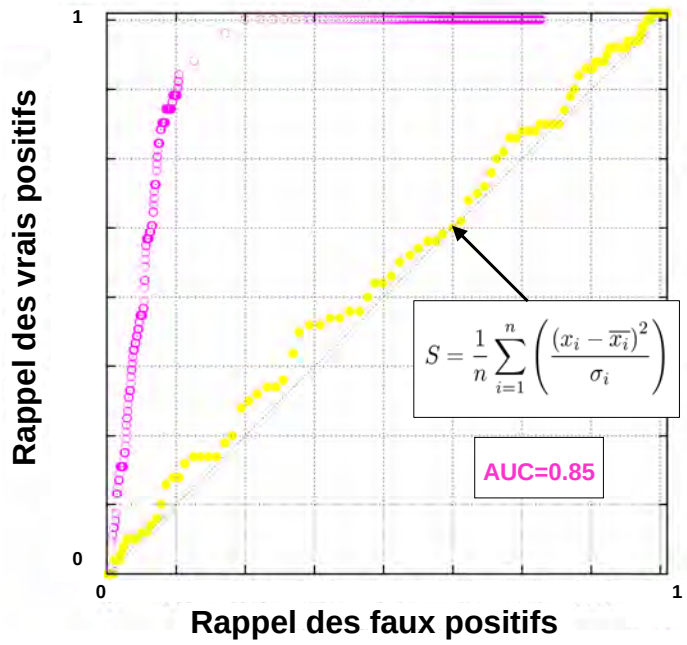


FIG. 3.11 – Courbes de ROC obtenues pour la fonction logistique et l'écart à la moyenne. *En magenta, la courbe de ROC pour la fonction logistique et en jaune, la courbe de ROC pour l'écart à la moyenne qui correspond à un résultat aléatoire.*

3.2 Un modèle en accord avec la physique du problème

3.2.1 Introduction

Pour comparer le modèle de structure que nous utilisons à ceux précédemment mis au point, nous avons cherché le spectre d'énergie associé à la formation d'un complexe et nous l'avons comparé au modèle de l'énergie aléatoire [59]. Ce modèle des systèmes désordonnés simple, utilise le fait que les corrélations entre les niveaux d'énergie sont dans ce cas négligeables. Il met en évidence une transition de phase et une phase de basse température qui est complètement gelée.

C'est sur le modèle de l'étude de J. Janin publié en 1996 [101] que nous avons réalisé cette étude. Comme précédemment, on peut assimiler le score de docking à une énergie pour étudier d'autres paramètres biophysiques de la reconnaissance protéine-protéine.

Pour chaque complexe protéique, on peut calculer un spectre d'énergie à partir du modèle d'amarrage utilisant *DOCK* et la fonction de score construite à partir de la tessellation de Voronoï et l'apprentissage par le logiciel *ROGER*. L'étude de ce spectre dans le cadre du modèle de l'énergie aléatoire permet de déduire, pour chaque complexe, une température de transition vitreuse et une température critique montrant que le complexe à l'état natif est prédominant à 300 K et qu'un modèle énergétique à deux états est inadapté pour la représentation de l'association entre protéines.

3.2.2 Article : *A docking analysis of the statistical physics of protein-protein recognition*

INSTITUTE OF PHYSICS PUBLISHING PHYSICAL BIOLOGY
Phys. Biol. **2** (2005) S17–S23 doi:10.1088/1478-3975/2/2/S02

A docking analysis of the statistical physics of protein–protein recognition

Julie Bernauer[1], Anne Poupon[1], Jérôme Azé[2] and Joël Janin[3]

[1] Yeast Structural Genomics Laboratory, IBBMC UMR CNRS 8619, Bâtiment 430, Université Paris-Sud, 91405-Orsay, France
[2] Laboratoire de Recherche en Informatique, Bâtiment 490, Université Paris-Sud, 91405-Orsay, France
[3] Laboratoire d'Enzymologie et Biochimie Structurales, UPR 9063 CNRS, 91198-Gif-sur-Yvette, France

E-mail: janin@lebs.cnrs-gif.fr

Received 1 February 2005
Accepted for publication 12 April 2005
Published 13 May 2005
Online at stacks.iop.org/PhysBio/2/S17

Abstract
We describe protein–protein recognition within the frame of the random energy model of statistical physics. We simulate, by docking the component proteins, the process of association of two proteins that form a complex. We obtain the energy spectrum of a set of protein–protein complexes of known three-dimensional structure by performing docking in random orientations and scoring the models thus generated. We use a coarse protein representation where each amino acid residue is replaced by its Voronoï cell, and derive a scoring function by applying the evolutionary learning program ROGER to a set of parameters measured on that representation. Taking the scores of the docking models to be interaction energies, we obtain energy spectra for the complexes and fit them to a Gaussian distribution, from which we derive physical parameters such as a glass transition temperature and a specificity transition temperature.

List of abbreviations

AUC	area under a ROC curve
CAPRI	Critical Assessment of Predicted Interactions
PDB	Protein Data Bank
REM	random energy model
ROC	receiver operating characteristics
ROGER	ROc based GEnetic learneR algorithm

1. Introduction

The specific recognition of a biological macromolecule by another is a fundamental process in all fields of biology, and structural biologists strive to give it a physical–chemical basis at the atomic level. In the case of DNA, the double helix and the complementary pairing of the nucleotide bases are an elegant solution that Watson and Crick discovered 50 years ago, but for proteins, we have no general answer of this kind. Proteins are designed to interact specifically with all sorts of objects, from metal ions to small organic molecules to proteins, DNA or RNA, each with a different mode of recognition. Protein–protein recognition is the most diverse of these processes, and the one that occupies the center of the stage nowadays, thanks to the recent accumulation of data from genome sequencing and genome-wide genetic or biochemical experiments [1–3].

Computer simulations are one of many ways to approach specific recognition. Docking algorithms take two molecules defined by their atomic coordinates and search for favorable ways to put them together. For proteins, the earliest algorithm is that of Wodak and Janin [4, 5]. It docks together two proteins represented by a set of rigidly linked residue centroids with only six degrees of freedom. Many other docking procedures, reviewed in [6–8], have been developed in recent years to perform the same task. The principal application of the docking procedures is to predict the structure of protein–protein complexes from that of the components. Their capacity to do so is currently being evaluated in blind predictions by the CAPRI (Critical Assessment of Predicted Interactions) community-wide experiment [9]. The results suggest that, although the procedures efficiently explore the degrees of freedom of the system, scoring the solutions is still a difficult task: the native solution corresponding to the experimental structure is often lost among false positives. Developing scoring functions that correctly represent the physics of

recognition and filter out the false positives is therefore an active field of study. This can be done in an empirical way by examining features of protein–protein complexes deposited in the Protein Data Bank (PDB) [10]. Our particular approach employs a coarse protein model in which we replace the atomic description by a set of Voronoï polyhedra built around each amino acid residue. We measure a set of 84 parameters on the Voronoï representation of 79 protein–protein complexes and on models generated by docking in all orientations the components of these complexes with program DOCK [11]. We then employ an evolutionary learning program called ROGER [12, 13] to derive a scoring function from these data.

In this paper, we use DOCK and the ROGER scoring function for a purpose other than structure prediction: to generate interaction energy spectra and apply the random energy model (REM) of statistical physics [14]. REM has had a number of applications in protein studies, including specific recognition [15–17]. We obtain spectra by docking proteins in random orientations and taking the score of the docking models as the interaction energy. We show that, although the score does not efficiently discriminate the native from the non-native models, its distribution follows the random energy model, allowing parameters such as the glass transition temperature and the specificity transition temperature to be determined.

2. Methods and results

2.1. The Voronoï tessellation as a descriptor of protein–protein interaction

Given a finite set of nodal points p_i in space, the Voronoï cell of node p_i is the region of space in which all points are closer to p_i than to any other node. Voronoï cells are convex polyhedra that may be non-bonded. Each point of space belongs to either one Voronoï cell or a common face between cells: Voronoï cells form a tessellation. For computational efficiency, the Voronoï tessellation is best constructed using its dual, the Delaunay triangulation, as implemented in the Computational Geometric Algorithm Library [18]. This incremental randomized construction ensures the optimal time complexity, which for the three-dimensional problem is $O(n^2)$.

A protein structure may be described by a set of Voronoï cells, the nodes being protein atoms [19], or as we do here, the geometric centers of the residue side chains plus the Cα atom [20]. Nodes representing solvent molecules must be added in order to prevent surface atoms or residues from having non-bonded Voronoï cells. Following [20], we place spheres of diameter 6.5 Å on a water-like lattice, so that solvent cells have a volume similar to the average residue Voronoï cell. This representation has been shown to yield valuable results on the molecular packing and other structural properties of proteins [20–22]. Its application to protein–protein complexes allows the following definitions:

- two residues are neighbors if their Voronoï cells share a common face;
- a residue belongs to the protein interior if all its neighbors are residues of the same protein;
- a residue belongs to the protein surface if one or more of its neighbors are solvent;
- a residue belongs to the protein–protein interface if one or more of its neighbors belong to the other protein;
- an interface residue belongs to the core of the interface if none of its neighbors is solvent;
- the cell faces shared by residues of both proteins constitute the interface.

2.2. Systematically sampling docking models

On the basis on previous studies [23, 24], we selected 79 PDB entries that contain protein–protein complexes of known x-ray structure and illustrate specific recognition: 1a2k 1acb 1agr 1ahw 1aip 1ak4 1ao7 1atn 1avw 1avz 1bql 1bth 1bvk 1cbw 1cgi 1dan 1dee 1dfj 1dhk 1dkg 1dvf 1efu 1eo8 1fbi 1fc2 1fin 1fle 1fq1 1fss 1gg2 1gla 1got 1gua 1hez1 1hez2 1hia 1hwg 1iai 1igc 1jhl 1kb5 1mah 1mda 1mhh 1mkw 1mlc 1nca 1nfd 1nmb 1nsn 1osp 1ppf 1qfu 1seb 1spb 1stf 1tab 1tbq 1tco 1tgs 1toc 1tx4 1udi 1ugh 1vfb 1wej 1wq1 1ycs 2btf 2jel 2kai 2pcc 2sic 2trc 2vir 3hfl 3hfm 3hhr 4htc.

We take each of the 79 complexes apart and apply the rigid-body docking program DOCK [11] to the component proteins. DOCK uses five angles and a distance to represent the translational/rotational degrees of freedom of one component relative to the other. For each combination of the five angular parameters, DOCK uses the Wodak–Janin algorithm [5] to determine the distance that brings the two proteins in contact, and builds what we call here a docking model. The native state being the x-ray structure, docking in any other orientation generates a non-native model. We cover all possible modes of rigid-body interaction by sampling the five angles in 10° steps over their whole respective range, which yields 18×10^6 docking models for each complex. Working sets of models where the components interact over different parts of their surface and in different orientations are then randomly selected among those for further evaluation.

2.3. Scoring with the ROGER algorithm

The function that we use to score docking models contains 84 parameters derived from the Voronoï tessellation. The parameters refer to the interface and belong to six classes:

P1. the interface area (1 parameter);
P2. the number of core interface residues (1 parameter);
P3. their Voronoï volumes (20 parameters);
P4. the frequency of each residue type at the interface (20 parameters);
P5. the frequency of the pairs of residues in contact (21 parameters);
P6. the distance between the centroids of core interface residues in contact (21 parameters).

When estimating parameters of classes P5 and P6, the 20 amino acid types are binned in six physical–chemical groups in order to reduce the number of parameters: (ILMV), (FYW), (HKR), (DE), (NQ), (AGSTCP). The scoring function is written as

$$f(x) = \Sigma_i w_i |x_i - c_i|. \quad (1)$$

The weights w_i and central values c_i were derived by the ROGER evolutionary learning program [12, 13]. ROGER (ROc based GEnetic learneR) is designed to maximize the area under a ROC curve (AUC). Receiver operating characteristics (ROC) curves, issued from signal processing, represent the trade-off between true and false positive rates when interpreting hypotheses. The ideal hypothesis, which generates no false positive and 100% of the true positives, has a ROC curve that is a step function and an AUC of 1. A random selection yielding true and false positives in equivalent numbers has an AUC of 0.5. The goal of a learning algorithm being to maximize the rate of true positives while minimizing that of false positives, the AUC can be viewed as a global measure of the learning efficiency. Its optimization constitutes a NP-complete problem and the ROGER algorithm uses evolution strategies to solve it.

2.4. Implementation

The training set contained values for the 84 interface parameters defined above measured on the 79 complexes and on 8400 non-native models of these complexes generated with DOCK. We performed systematic docking in angular steps of 10°, clustered the results with the geometric procedure of [11], and randomly selected 8400 clusters represented by their average position. Not all interface parameters could be measured on every model of the training set. For instance, an amino acid type might not be present at a given interface, leading to missing values of parameters in classes P3, P5 and P6. As ROGER does not handle missing data, they were replaced by the median value of the corresponding attribute in either the native or the non-native members of the training set.

We ran ROGER on this set with nonlinear hypotheses in a (20 + 200) evolution strategy where 20 parents are selected deterministically from the 20 parents of the previous generation plus 200 off-springs, using self-adaptative mutation and uniform crossover with crossover rate 60%. Twenty-one independent runs yielded 21 scoring functions taking the best of each run. The results were assessed by performing a ten-fold stratified cross-validation that led to 210 scoring functions. To score a docking model, all 210 functions were evaluated and a two-step bagging was applied from leafs to root of the tree associated to the ten-fold cross-validation: first by taking the median value of the 21 independent runs in each fold, then by taking the median of the 10 values.

2.5. Energy–entropy relationships in the random energy model

We take the score attributed to a docking model to be an estimate of the energy E of the interaction between the two-component proteins in that particular state, and analyze the distribution of E as in [17]. We draw an energy spectrum by counting states with energies between E and $E + \mathrm{d}E$. If there are $m(E)$ such states, the entropy is

$$S(E) = k_B \ln m(E). \tag{2}$$

The native state has an energy E_0 which we take to be 0 for convenience, and it is unique, so that $S(E_0) = 0$. It is separated by an energy gap Δ from the non-native state of lowest energy. The total number of states (including the native) is

$$N = 1 + \int_{\Delta}^{\infty} m(E)\,\mathrm{d}E. \tag{3}$$

At thermodynamic equilibrium and temperature T, all states coexist and their relative abundance $n(E)$ follows Boltzmann's law:

$$n(E) = m(E)\,\exp\left(-\frac{E}{k_B T}\right). \tag{4}$$

We write the partition function Z:

$$Z = 1 + r = 1 + \int_{\Delta}^{\infty} n(E)\,\mathrm{d}E. \tag{5}$$

The native state contributes 1 to Z, the non-native, r. As specific recognition implies $r \ll 1$, the gap Δ should be large relative to the thermal energy $k_B T$. The condition $r = 1/2$ defines a temperature T_S that was called the specificity transition temperature [17]: below T_S, the native state is dominant, above T_S, non-native states take over. T_S can be obtained by drawing the tangent to the curve from the origin. The tangent at $E = \Delta$ defines another characteristic temperature of the system, its critical temperature T_c, also called the glass transition temperature in spin glass studies [14]. Below T_c, the only non-native states that compete with the native are those with energy near $E = \Delta$. Above T_c, many states of higher energy are populated. T_S and T_c are easily calculated by fitting an analytical expression to the energy spectrum. The random energy model assumes a Gaussian distribution [14].

2.6. The energy spectrum of a protein–protein complex

We took PDB entry 1eo8, a complex between the flu virus hemagglutininin and the Fab fragment of a monoclonal antibody [25], and generated docking models of that complex by running the program DOCK on its components in 10° angular steps. We then selected 10 000 models at random and evaluated their ROGER score. Figure 1 shows a histogram of the distribution of the score values. ROGER scores are designed to be 0 for a perfect solution and 1 for a random solution. The native state has the best score (0.258) and is separated by a gap from the non-native models, which have scores in the range 0.3–1.5. The mean value of the non-native scores is 1.00, the expected value for random solutions, but the distribution is somewhat asymmetric with fewer models scoring above 1 than below and a median of 1.02. Nevertheless, figure 1 shows that a Gaussian fits the left half (low score region) of the distribution to a good approximation.

Accepting that the ROGER score represents the energy of the interaction between the flu hemagglutinin and the antibody, figure 1 is the energy spectrum complex of that system. We may calculate entropies by equation (2) and draw an entropy–energy curve. In figure 2, the left part of that curve is fitted to the parabola equivalent to the Gaussian in figure 1. The curve deviates from the parabola for scores above 1, but this part of the curve contains only high-energy states. Equation (4) indicates they are not populated except at very

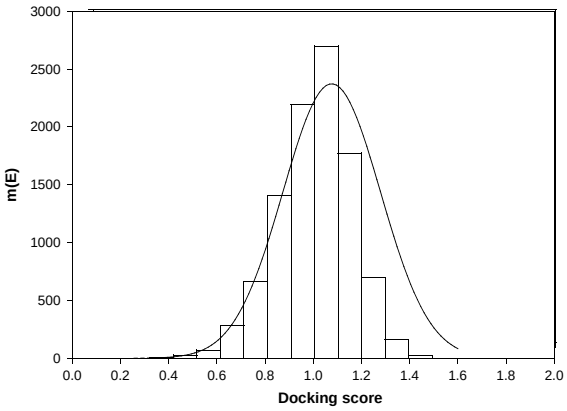


Figure 1. The energy spectrum of the flu hemagglutininin/Fab complex 1eo8. The scores of 10 000 randomly selected docking solutions are represented as a histogram. The native state has a score of 0.258. The full line is a Gaussian fit of the left half of the histogram ($R^2 = 0.994$).

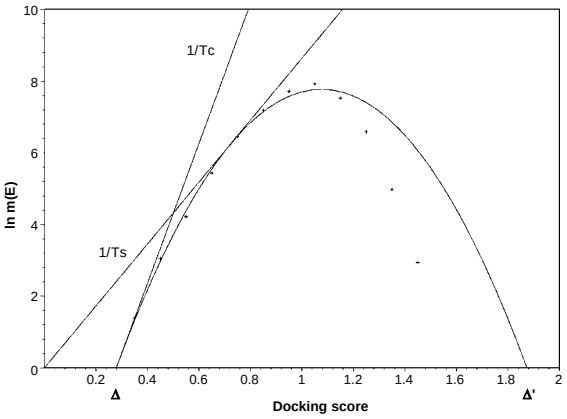


Figure 2. Entropy–energy curve for complex 1eo8. The points are from the histogram in figure 1. The full line is the parabola that corresponds to the Gaussian curve in figure 1. It intersects the horizontal axis at $E = \Delta = 0.28$ and $\Delta' = 1.87$. The tangent line at $E = \Delta$ has a slope $1/T_c$, the line drawn from the origin, a slope $1/T_S$. In the REM of protein–protein recognition [17], $T_c = 0.116$ is the glass transition temperature and $T_S = 0.05$ is the specificity transition temperature.

high temperatures, and they can be ignored. The parabola intersects the horizontal line at two points with scores Δ and Δ', and the width of the spectrum is $\Delta' - \Delta$. Δ is the energy gap between the native state and the first non-native docking model taking the distribution to be Gaussian and the native state to have the best possible score $E = 0$ instead of its actual value of 0.258. Under these assumptions, the tangent to the parabola drawn from the origin determines the specificity transition temperature T_S, which is a factor of 2.3 larger than the glass transition temperature T_c measured from the slope of the tangent at $E = \Delta$.

We went on to analyze the other complexes in the same way. On each, we ran DOCK in 10° angular steps

Table 1. Physical parameters derived from the energy spectrum of protein–protein complexes.

Parameter[a]	Mean value	SD
Rank of native	157	309
Score of native	0.52	0.15
Mean non-native score	0.88	0.12
Gap energy Δ[b]	0.29	0.08
Center of parabola	0.97	0.27
Spectrum width $\Delta' - \Delta$	1.36	0.57
T_S/T_c	2.5	0.4

[a] Data on 70 protein–protein complexes in the training list (1efu, 1fc2, 1nmb, 1qfu, 2btf, 2kai, 3hfl, 3hhr and 4htc were omitted for technical reasons). The mean value and standard deviation (SD) of the parameters were calculated on 4000 docking models of each of the complexes.
[b] Δ, Δ', T_S and T_c were obtained by fitting a parabola to the entropy–energy curve and assuming that the native state had a score of 0.

and evaluated the ROGER score of 4000 randomly selected docking models. Seven complexes behaved like 1eo8: their native states had scores lower than the best non-native, with a positive energy gap. In the other complexes, the energy gap was negative and, on average, the native ranked 157th of 4000 models. Histograms were made and entropy–energy curves drawn after rescaling $m(E)$ in order to have $N = 10\,000$ as for 1eo8. The left three quarters of each curve were fitted to a parabola. The fit was good (R^2 values above 0.98) except for one complex. The physical parameters listed in table 1 are average values derived from the parabolic fits. Although the mean non-native score is less than 1, the parabolas are centered near the random value of 1, and their width is in the range 1–3. Because the specificity transition temperature T_S cannot be calculated when the gap is negative, values cited in table 1 assume the native score to be 0 as in figure 1, and we cite the dimensionless ratio T_S/T_c instead of T_S and T_c which are in arbitrary units like the ROGER score. The ratio is in the range 1.7–3.4, with a mean of 2.5 and a relatively small standard deviation.

3. Discussion

3.1. The ROGER score

We derived a score function for docking models by running the evolutionary learning program ROGER on parameters measured on the Voronoï representation of 79 protein–protein complexes. This score function did poorly in discriminating between the native structure of the complexes and models obtained by docking their components in random orientations: the native had the best score in only 10% of the test cases. Thus, the ROGER score is not suitable for prediction at the present stage. Nevertheless, we could take it to represent an interaction energy to be used in the frame of REM. Most elements of the score have physical counterparts that contribute to the free enthalpy of association ΔG_a of protein–protein complexes, and therefore to their stability in solution. For instance, the interface area (parameter P1) is linearly

related to the free enthalpy contribution of the hydrophobic effect [26–28]. The residue volumes (parameter class P3) describe the atomic packing, which determines the Van der Waals interaction energy. The amino acid composition (class P4) distinguishes the core of protein–protein interfaces from the rest of the protein surface. It expresses propensities for interfaces that correlate with the solubility of the amino acids in water and their desolvation free enthalpy [29–31]. Last, the pair parameters of classes P5 and P6 reflect the role of the Van der Waals and electrostatic interactions at the interface [32, 33].

These contributions to ΔG_a are all on different scales, and the task of the learning algorithm is to weight them properly. The poor discrimination achieved by the score may have several origins. First, ROGER may have failed to determine optimal weights due to the small number of native states in the training set. Second, all atomic details are blurred in the Voronoï representation. Third, our choice of parameters incompletely represents the physics of the system. Our study assumes that proteins associate as rigid bodies and ignores the entropy and energy cost of conformation changes. Conformation changes occur in several of the complexes of our list, and they could in principle be incorporated in the analysis. However, the changes observed when two proteins associate are highly variable in nature and amplitude from one system to another [23, 34], and a much larger sample would be needed for a learning algorithm to perform correctly.

3.2. Simulating bimolecular association

We took the distribution of the ROGER scores in docking models to be an energy spectrum and the quantity in equation (2) to be an entropy. This is justified if we consider that docking simulates the random collision of protein molecules diffusing in solution. A collision creates a short-lived encounter pair that may or may not proceed to form a stable complex depending on which regions of the two protein surfaces are in contact and on their relative orientation at the time of the collision [35–40]. The native state represents encounter pairs that do convert to the complex; the non-native models are encounter pairs that dissociate quickly. Observed rates of bimolecular association in systems like antibody–antigen or enzyme–inhibitor complexes indicate that, in the absence of long-range electrostatic steering, one collision in 10^3–10^6 is productive [35, 40–42]. This suggests that the native state is in competition with 10^3–10^6 non-natives. We retained 10 000 non-native models of the flu hemagglutinin/antibody complex, a number in the range 10^3–10^6. In DOCK, four of the angular parameters are latitudes and longitudes that locate the region of the surface of each of the two protein molecules in contact after docking; the fifth is a twist rotation about the line of centers [4, 5]. $N = 10\,000$ is the number of docking models generated by sampling each of the five angles in steps of approximately 36°. Thus, the value of S that we obtain with equation (2) is the entropy cost of orienting and locating with the accuracy of a 36° angular step, the regions of the two molecular surfaces that the collision brings in contact. This step corresponds to a 4 Å displacement on the surface of a protein molecule of mean radius 20 Å, a plausible range for the local interactions that discriminate between the native and non-native models [43], and for the dimensions of the 'binding funnel' that leads to the stable complex [44–46].

3.3. Specific versus non-specific interaction

ROGER scores are on a scale where 0 is for a perfect solution, 1 for random. On that scale, the energy spectrum that we obtain for non-native docking models of the flu hemagglutinin/antibody complex extends from 0.3 to 1.5 (figure 1), and its parabolic fit has a width $\Delta' - \Delta = 1.6$. Other complexes yield similar values. The Δ' intersect corresponds to docking models where the two proteins barely touch and their energy of interaction must be close to zero. The Δ intersect corresponds to the best non-native docking model, a model that has extensive contacts, but misses the local interactions that determine the specificity. The molecular packing of protein crystals provides examples of extensive non-specific modes of protein–protein interaction. Whereas most crystal packing interfaces are small [48], a minority of the pairwise interfaces that occur in protein crystals is comparable in size to the interfaces of the complexes we study here. These large crystal packing interfaces differ from specific interfaces in their physical–chemical properties. On average, they are less hydrophobic, less tightly packed and contain fewer hydrogen bonds [49]. Moreover, their amino acid composition is similar to the rest of the protein surface and different from the core of the specific interfaces. The parameters that contribute to the ROGER score take some of these properties into account. Nevertheless, the distinction between crystal packing and specific interfaces is non-trivial [49–52].

We cannot measure free enthalpies for crystal packing interactions, but we may assume that their range covers a few tens of kcal mol^{-1}, the standard state value of the association-free enthalpies being typically 10–20 kcal mol^{-1} for specific complexes. Thus, a unit of the ROGER score may be worth 10–20 kcal mol^{-1}. On that scale, the mean value of Δ in table 1 is equivalent to 3–6 kcal mol^{-1}, 5–10 times the thermal energy at 300 K. For 1eo8, this yields a value of T_S in the range 600–1200 K, insuring that the native complex is the dominant species at 300 K. However, T_c is 2.5 times lower than T_S, and the glass transition temperature is probably below ambient. In that case, the non-native modes of association that compete with the native state are not those with the lowest energy Δ, but a large subset of states belonging to the left half of the energy spectrum. Then, the commonly accepted two-state model where all non-native states have the same energy is inappropriate and REM is a better description of the equilibrium properties of the system.

4. Conclusion and outlook

The Voronoï representation is well adapted to an analysis of protein–protein interaction at a coarse level and useful when simulating random collisions. The score that we derived from it ignores the detailed atomic structure. It could not account for

specificity, but it met the requirements of the Random Energy Model and enabled us to represent events that occur in solution both *in vitro* and *in vivo* and implicate non-native modes of protein–protein association. These modes contribute to the physical process of recognition even in cases where only the native state is biologically relevant.

Acknowledgments

J Janin acknowledges financial support from the EIDIPP program of Action Concertée Incitative IMPBio.

Glossary

Native state. The x-ray structure of a protein–protein complex taken from the PDB.

Docking model. Model of a protein–protein complex obtained by docking its components in the computer.

Non-native model. All models of a complex obtained by docking are non-native unless the components are positioned and oriented as in the native state.

Tessellation. A way of dividing space into adjacent cells that do not overlap; the tessellation used here was invented by the French/Russian mathematician Voronoï in 1908.

References

[1] Marcotte E M, Pellegrini M, Ng H L, Rice D W, Yeates T O and Eisenberg D 1999 Detecting protein function and protein–protein interactions from genome sequences *Science* **285** 751–13

[2] Janin J and Wodak S J 2003 Protein modules and protein–protein interaction: towards a global view *Adv. Prot. Chem.* **61** 1–8

[3] Janin J and Séraphin B 2003 Genome-wide studies of protein–protein interaction *Curr. Opin. Struct. Biol.* **13** 383–8

[4] Wodak S and Janin J 1978 Computer analysis of protein–protein interactions *J. Mol. Biol.* **124** 323

[5] Janin J and Wodak S J 1985 Reaction pathway for the quaternary structure change in hemoglobin *Biopolymers* **24** 509–26

[6] Halperin I, Ma B, Wolfson H and Nussinov R 2002 Principles of docking: an overview of search algorithms and a guide to scoring functions *Proteins* **47** 409–43

[7] Wodak S J and Janin J 2003 The structural basis of macromolecular recognition *Adv. Prot. Chem.* **61** 9–73

[8] Smith G R and Sternberg M J 2002 Prediction of protein–protein interactions by docking methods *Curr. Opin. Struct. Biol.* **12** 28–85

[9] Janin J, Henrick K, Moult J, Eyck L T, Sternberg M J, Vajda S, Vakser I and Wodak S J 2003 CAPRI: a critical assessment of predicted interactions *Proteins* **52** 2–9

[10] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucl. Acid Res.* **28** 235–42

[11] Cherfils J, Duquerroy S and Janin J 1991 Protein–protein recognition analyzed by docking simulation *Proteins* **11** 271–80

[12] Roche M *et al* 2004 *ROC Analysis in Artificial Intelligence, 1st Int. Workshop, ROCAI-2004 (Valencia, Spain, 22 August 2004)* ed J Hernandez-Orallo, C Ferri, N Lachiche and P A Flach pp 81–8

[13] Sebag M, Azé J and Lucas N 2004 *ROC-Based Evolutionary Learning: Application to Medical Data Mining Proc. 6th Int. Conf. on Artificial Evolution, EA 2003 (Marseille, France)* pp 384–96

[14] Derrida B 1981 Random energy model: an exactly solvable model of disordered systems *Phys. Rev.* B **24** 2613–26

[15] Onuchic J N, Luthey-Schulten Z and Wolynes P G 1997 Theory of protein folding: the energy landscape perspective *Ann. Rev. Phys. Chem.* **48** 545–600

[16] Lancet D, Sadovsky E and Seidemann E 1993 Probability model for molecular recognition in biological receptor repertoires: significance to the olfactory system *Proc. Natl Acad. Sci. USA* **90** 3715–9

[17] Janin J 1996 Quantifying biological specificity: the statistical mechanics of molecular recognition *Proteins* **25** 438–45

[18] Boissonnat J D *et al* 1999 *Symp. on Computational Geometry* p 421

[19] Richards F M 1974 The interpretation of protein structures: total volume, group volume distributions and packing density *J. Mol. Biol.* **82** 1–14

[20] Soyer A, Chomilier J, Mornon J P, Jullien R and Sadoc J F 2000 Voronoi tessellation reveals the condensed matter character of folded proteins *Phys. Rev. Lett.* **85** 3532–5

[21] Angelov B, Sadoc J F, Jullien R, Soyer A, Mornon J P and Chomilier J 2002 Nonatomic solvent-driven Voronoi tessellation of proteins: an open tool to analyze protein folds *Proteins* **49** 446–56

[22] Poupon A 2004 Voronoi and Voronoi-related tessellations in studies of protein structure and interaction *Curr. Opin. Struct. Biol.* **14** 233–41

[23] Lo Conte L, Chothia C and Janin J 1999 The atomic structure of protein–protein recognition sites *J. Mol. Biol.* **285** 2177–98

[24] Chakrabarti P and Janin J 2002 Dissecting protein–protein recognition sites *Proteins* **47** 334–43

[25] Fleury D, Daniels R S, Skehel J J, Knossow M and Bizebard T 2000 Structural evidence for recognition of a single epitope by two distinct antibodies *Proteins* **40** 572–8

[26] Chothia C and Janin J 1975 Principles of protein–protein recognition *Nature* **256** 705–8

[27] Sharp K A, Nicholls A, Friedman R and Honig B 1991 Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models *Biochemistry* **30** 9686–97

[28] Young L, Jernigan R L and Covell D G 1994 *Protein Sci.* **3** 717–29

[29] Jones S and Thornton J M 1997 Principles of protein–protein interactions *Proc. Natl Acad. Sci. USA* **93** 13–20

[30] Tsai C J, Lin S L, Wolfson H J and Nussinov R 1997 Studies of protein–protein interfaces: a statistical analysis of the hydrophobic effect *Protein Sci.* **6** 53–64

[31] Zhang C, Vasmatzis G, Cornette J L and DeLisi C 1997 Determination of atomic desolvation energies from the structures of crystallized proteins *J. Mol. Biol.* **267** 707–26

[32] Miyazawa S and Jernigan R L 1996 Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading *J. Mol. Biol.* **256** 623–44

[33] Ofran Y and Rost B 2003 Analysing six types of protein–protein interfaces *J. Mol. Biol.* **325** 377–87

[34] Betts M J and Sternberg M J 1999 An analysis of conformational changes on protein–protein association: implications for predictive docking *Protein Eng.* **12** 271–83

[35] Northrup S H and Erickson H P 1992 Kinetics of protein–protein association explained by Brownian

dynamics computer simulation *Proc. Natl Acad. Sci. USA* **89** 3338–42
[36] Janin J 1997 The kinetics of protein–protein recognition *Proteins* **28** 153–61
[37] Gabdoulline R R and Wade R C 2001 Protein–protein association: investigation of factors influencing association rates by Brownian dynamics simulations *J. Mol. Biol.* **306** 1139–55
[38] Gabdoulline R R and Wade R C 2002 Biomolecular diffusional association *Curr. Opin. Struct. Biol.* **12** 204–13
[39] Selzer T and Schreiber G 2001 New insights into the mechanism of protein–protein association *Proteins* **45** 190–8
[40] Schreiber G 2002 Kinetic studies of protein–protein interactions *Curr. Opin. Struct. Biol.* **12** 41–7
[41] Camacho C J, Kimura S R, DeLisi C and Vajda S 2000 Kinetics of desolvation-mediated protein–protein binding *Biophys. J.* **78** 1094–105
[42] Schlossauer M and Baker D 2004 Realistic protein–protein association rates from a simple diffusional model neglecting long-range interactions, free energy barriers and landscape ruggedness *Protein Sci.* **13** 1660–9
[43] Camacho C J and Vajda S 2001 Protein docking along smooth association pathways. *Proc. Natl Acad. Sci. USA* **98** 10636–41
[44] Camacho C J and Vajda S 2002 Protein–protein association kinetics and protein docking *Curr. Opin. Struct. Biol.* **12** 36–40
[45] Kumar S, Ma B, Tsai C J, Sinha N and Nussinov R 2000 Folding and binding cascades: dynamic landscapes and population shifts *Protein Sci.* **9** 10–9
[46] Zhang C, Chen J and DeLisi C 1999 Protein–protein recognition: exploring the energy funnels near the binding sites *Proteins* **34** 255–67
[47] Rajamani D, Thiel S, Vajda S and Camacho C J 2004 Anchor residues in protein–protein interactions *Proc. Natl Acad. Sci. USA* **101** 11287–92
[48] Rodier F and Janin J 1995 Protein–protein interaction at crystal contacts *Proteins* **23** 580–7
[49] Bahadur R P, Chakrabarti P, Rodier F and Janin J 2004 A dissection of specific and non-specific protein–protein interfaces *J. Mol. Biol.* **336** 943–55
[50] Jones S and Thornton J M 1997 Analysis of protein–protein interaction sites using surface patches *J. Mol. Biol.* **272** 121–32
[51] Ponstingl H, Henrick K and Thornton J M 2000 Discriminating between homodimeric and monomeric proteins in the crystalline state *Proteins* **41** 47–57
[52] Mintseris J and Weng Z 2003 Atomic contact vectors in protein–protein recognition *Proteins* **53** 629–39

3.3 Application aux cibles de *CAPRI (Critical Assessment of PRediction of Interactions)*

3.3.1 Introduction

Afin de tester la performance de la fonction de score sur des cas d'étude réalistes, n'appartenant pas à l'échantillon d'apprentissage, nous avons travaillé sur les cibles proposées dans le cadre de l'expérience *CAPRI* (voir paragraphe 1.2.3.4 page 18).

Deux approches ont été envisagées :

1. D'abord, aucune information biologique n'est prise en compte et l'exploration est réalisée simplement à l'aide du programme *DOCK* ;
2. Ensuite, l'exploration est « guidée » par l'information biologique. La fonction de score utilisée ayant été calibrée avec *DOCK*, elle est simplement appliquée aux résultats obtenus par *HADDOCK*.

Dans la plupart des cas, après reclassement, une des meilleures solutions se trouve dans les dix premiers résultats. La fonction de score que nous avons obtenue permet donc de reclasser *a posteriori* les résultats obtenus par un algorithme d'amarrage. De façon évidente, plus le jeu de complexes à reclasser est de bonne qualité, meilleur est le résultat, la prise en compte de l'information biologique lors de l'exploration (*HADDOCK*) permettant d'obtenir de bien meilleurs résultats. De plus, même si un affinement atomique est bien sûr requis à l'issue de ces étapes d'amarrage, celui-ci pourra aussi être envisagé dans le cadre d'une représentation de type Voronoï.

3.3.2 Article : *A new protein-protein docking scoring function based on interface residue properties*

A new protein-protein docking scoring function based on interface residue properties

[1]Bernauer, J. [2]Azé, J. [3]Janin, J. [1]Poupon, A.

[1]Yeast Structural Genomics – IBBMC UMR 8619 – Bâtiment 430 – Université Paris-Sud – 91405 ORSAY –- FRANCE

[2]Equipe de Bioinformatique – LRI UMR 8623– Bâtiment 430 – Université Paris-Sud – 91405 ORSAY –- FRANCE

[3]LEBS – UPR 9063 – Bâtiment 34 – CNRS – 91198 GIF-SUR-YVETTE –- FRANCE

Yeast Structural Genomics

IBBMC UMR 8619 – Bâtiment 430 – Université Paris-Sud – 91405 ORSAY

FRANCE

☎ (+33) 1 69 15 31 57

Correspondence should be addressed to:

julie.bernauer@ibbmc.u-psud.fr

ABSTRACT

A protein-protein docking procedure traditionally consists in two successive tasks: a search algorithm generates a large number of candidate solutions, and then a scoring function is used to rank them. To address the second step, we developed a scoring function based on a Voronoi tessellation of the protein three-dimensional structure. We showed that the Voronoi representation may be used to describe in a simplified but useful manner, the geometric and physico-chemical complementarities of two molecular surfaces. We measured a set of parameters on native protein-protein complexes and on decoys, and used them as attributes in several statistical learning procedures: a logistic function, Support Vector Machines (SVM), and a genetic algorithm. For the later, we used ROGER, a genetic algorithm designed to optimize the area under the ROC curve. To further test the scores derived with ROGER, we ranked models generated by two different docking algorithms on targets of a blind prediction experiment, improving in almost all cases the rank of native-like solutions.

Proc Virt Conf Genom and Bioinf PLOS (0):00-00
Print ISSN 1547-383X
Online ISSN 1547-7320

www.virtualgenomics.org

CATEGORY

- Protein Structural Analysis and Prediction

Keywords: protein-protein interaction, docking, machine learning, Voronoi tessellation, genetic algorithm

1. INTRODUCTION

The three-dimensional structure of a protein-protein complex is a crucial information for its functional study. Whereas structural genomics projects have considerably increased our knowledge of individual protein structures, protein-protein complexes are still beyond the reach of high-throughput methods. Reliable, fast and automatic docking algorithms that can assemble individual proteins into complexes are therefore of great value.

A docking procedure comprises two tasks, generally consecutive and largely independent. The first is a rigid-body search over the rotational/translational degrees of freedom. It generates a large number of candidate solutions where the two partners contact each other in many different orientations, avoiding steric clashes. Then, the best solutions are selected by evaluating a score. Scoring functions express the geometric complementarity of the two molecular surfaces in contact, and also the strength of the interaction, based on the physico-chemical characteristics of the amino acids in contact with each other. Most procedures handle the geometric and physical-chemical criteria separately.

The formation of a complex often induces changes in the structure of both partners. These changes may concern only the side chain conformation of amino acid residues at the interface, or they may imply motions of the protein backbone, the nature and amplitude of which remains very difficult to predict. In some cases, an amino acid substitution in one of the partners can cause gross changes in the complex even though it hardly affects the structure of the protein itself [11]. This makes "unbound" docking predictions much more difficult than "bound" ones: bound docking starts from protein structures taken from the complex and ignores these motions, whereas unbound docking uses the structures of the free partners and must take the physico-chemistry into account.

One major barrier for the study of macromolecular assemblies is is their sheer complexity. The problem cannot be solved without simplifications, and even with those, the best performing algorithms consume great amounts of CPU time. Biological information is often available for a given protein-protein structure, requiring “human” post-processing [18]. Yet, postgenomic

functional studies need a procedure that can search for plausible protein-protein complexes in a complete genome with thousands of genes. This requires the procedure to be: (i) very reliable, which means that the score of the best solution is high only if the two proteins do form a complex, and that the best solution is close to the native complex; and (ii) very fast, since the inspection of a whole genome requires the modeling of many hundreds of thousands potential complexes.

In this work we use the Voronoi tessellation as a descriptor of the protein structure [20]. The tessellation, which is based on amino acid residues rather than individual atoms has been shown to yield valuable mathematical results on the molecular packing [25], molecular recognition [3] and other structural properties of proteins. We show here that its use in the description of protein-protein interfaces leads to valuable scoring functions for docking algorithms.

To generate a scoring function, we choose a limited number of parameters that can be measured on the Voronoi tessellation of a complex. These parameters are measured in a set of native-like protein-protein complexes and in a set of decoys, and used as attributes in statistical learning methods. Three different types of learning algorithms are compared: a logistic function, SVM and a genetic algorithm called ROGER. ROGER gives the best results by far, and we use its score to re-rank models of protein-protein complexes generated by two different docking algorithms on targets of CAPRI (Critical Assessment of PRedicted Interactions), a blind prediction experiment designed to test docking procedures [13]. The results show that the ROGER score improves the ranking of native-like solutions and suggest that it will be of great value in early steps of a fully automated docking procedure.

2. MATERIALS AND METHODS

2.1 The Voronoi Construction

Voronoi diagrams, also known as Dirichlet or Thiessen tessellations, have been used in many fields of sciences. Given a set of points in space (centroids), the Voronoi tessellation divides the space into Voronoi cells centered on each point. A Voronoi cell includes all points of space that are closer to the cell centroid than to any other centroid, and it is the smallest polyhedron defined by bisecting planes between its centroid and all others. The Delaunay tessellation is obtained by tracing the vertices joining centroids, which have a common face in their Voronoi cells. The two tessellations are dual from each other. They are uniquely related, and for efficiency, it is

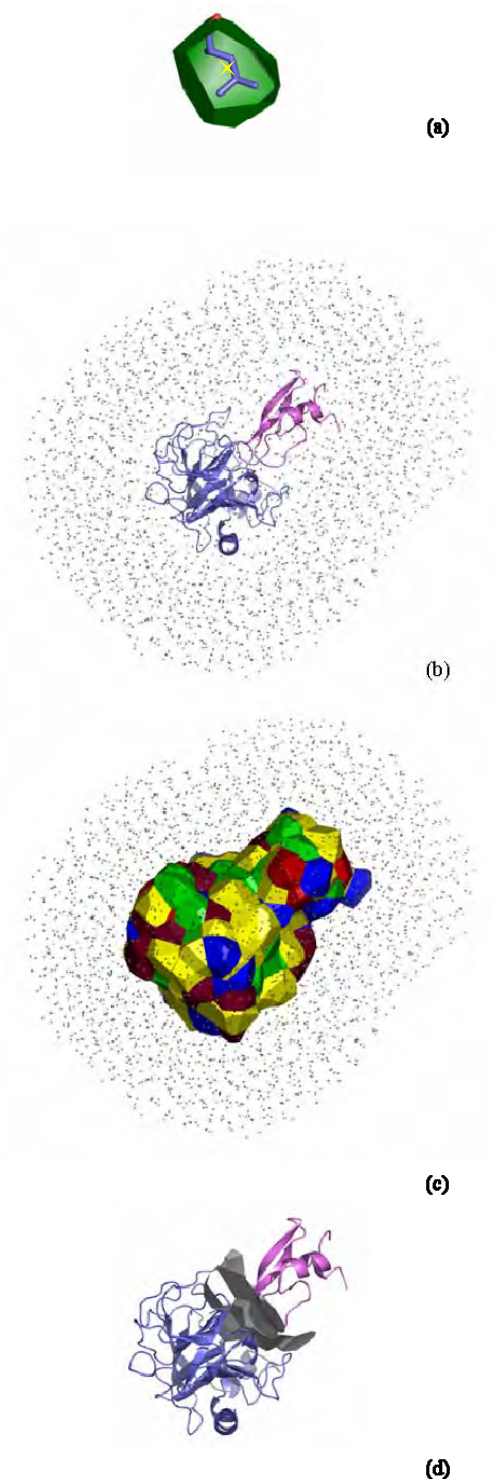


Figure 1: Voronoi description of protein-protein interfaces. (a) The Voronoi cell of a leucine residue in complex 1p2k. The cross marks the centroid used for constructing the cell (b) Complex 1p2k; the two protein chains are in blue and pink, and solvent is drawn as dots (c) Voronoi envelope of complex 1p2k (d) Facets in gray are shared by Voronoi polyhedra of residues of the two proteins, representing the interface of the complex

best to compute the Delaunay tessellation first and derive Voronoi cells from it.

We performed the computation by using as centroids the center of mass of amino acid side chains, including Cα (Figure 1). Our procedure uses the CGAL (Computational Geometric Algorithms Library), which implements an incremental randomized algorithm [4] of optimal $O(n^2)$ complexity, n being the number of centroids. Solvent was modeled around the protein to prevent surface residues from having unbound Voronoi cells (Figure 1). We placed spheres of radius 6.5 Å on a water-like lattice [25] to give solvent cells a volume similar to the average residue Voronoi cell.

The tessellation of a protein-protein complex leads to the following definitions:

- two residues are neighbors if their Voronoi cells share a common face.
- a residue belongs to the protein interior if all its neighbors are residues of the same protein.
- a residue belongs to the protein surface if one or more of its neighbors is solvent.
- a residue belongs to the protein-protein interface if one or more of its neighbors belongs to the other protein.
- an interface residue belongs to the core of the interface if none of its neighbors is solvent
- the cell facets shared by residues of both proteins constitute the interface.

2.2 Training Set

The training set consists in two subsets: positive examples, complexes of known 3D structure; and negative examples generated from the positive examples using a docking procedure.

2.2.1 Complexes of Known 3D Structures

The set of models on which we trained statistical learning methods was computed from the 2004 release #1 of the Protein Data Bank (PDB) [2] (Figure 2). First, we used a BioPython module [12] to extract entries reporting X-ray structures with resolution higher than 3Å and with two polypeptide chains longer than 20 residues, putative partners in a complex. Then we searched the PDB with Blastp [1] for entries containing the free partners, and retained those having more than 95% identity. Hierarchical clustering by R software [21] then gave a non-redundant set of 102 complexes (22 Unbound/Unbound and 80 Bound/Unbound) described in Table 1.

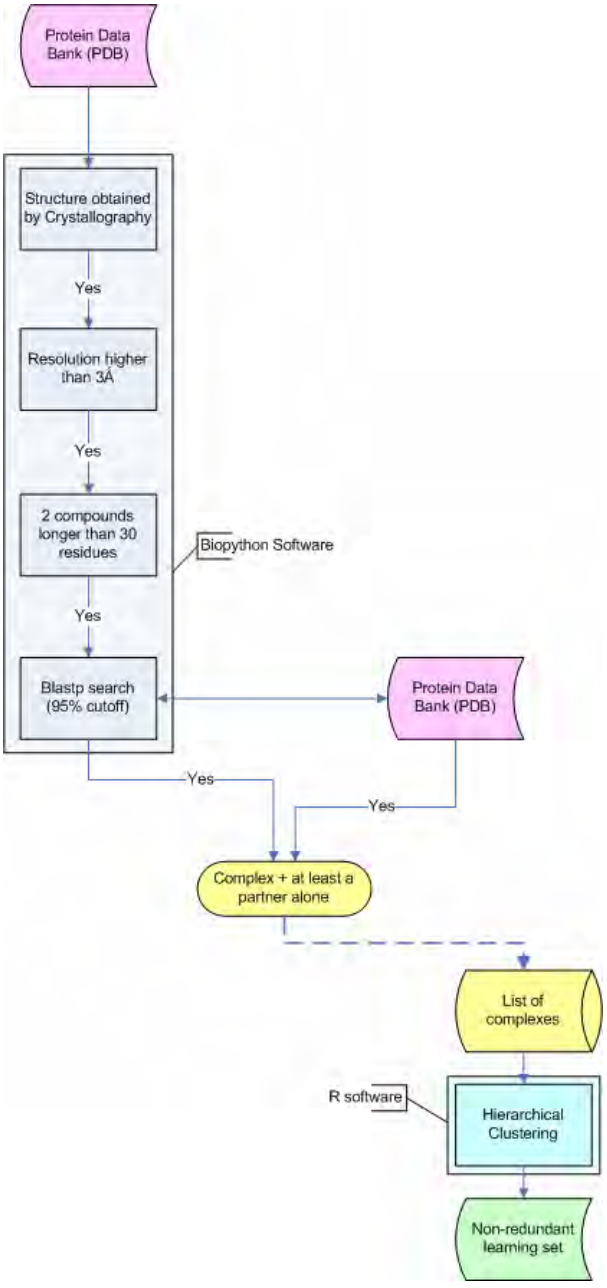


Figure 2 : Procedure for extracting the training set from the PDB

Table 1 : Learning set

a) Unbound/Unbound

PDB code of complex	Chains of partners in the complex		PDB code of first partner	Chain in first partner	PDB code of second partner	Chain in second partner
1s6v	A	B	1kok	A	1irw	- -
1ml0	A	D	1mkf	A	1dol	- -
1fin	A	B	1pf8	A	1vin	- -
1ugh	E	I	1akz	- -	2ugi	A
1b2s	A	D	1bnj	A	1a19	A
1ghq	A	B	1c3d	- -	1ly2	A
1f6m	A	C	1cl0	A	1keb	A
1e6e	A	B	1e1n	A	1cje	A
1u7f	A	B	1khx	A	1dd1	A
1dkf	A	B	1lbd	- -	1xap	A
1nbf	A	C	1nb8	A	1f9j	A
1kac	A	B	1nob	A	1f5w	A
1exb	A	E	1qrq	A	1qdv	A
1noc	A	B	1r35	A	1q23	A
1we3	A	O	1srv	A	1wnr	A
7cei	A	B	1unk	A	1m08	A
1pxv	A	C	1x9y	A	1nyc	A
1p2k	A	I	1xuk	- -	2hex	A
1jk9	A	B	1yaz	A	1qup	A
1n95	A	B	1ft1	A	1fpp	B
1kgy	A	E	1nuk	A	1iko	P
1t6b	X	Y	1acc	- -	1shu	X
1ktd	A	B	1ieb	A		
1d2z	A	B	1ik7	A		
1j7v	L	R	1inr	- -		
1jtd	A	B	1jwz	A		
1l2w	A	I	1jya	A		
1jzd	A	C	1jzo	A		
1hx1	A	B	1kaz	- -		
1lqv	A	C	1l8j	A		
1k3z	A	D	1my5	A		
1ujw	A	B	1nqe	A		
1ta3	A	B	1om0	A		
1ory	A	B	1orj	A		
1dhk	A	B	1ose	- -		
1x79	A	B	1oxz	A		
1pdk	A	B	1qpp	A		
1ewy	A	C	1que	- -		
1nvu	Q	S	1rvd	A		
1uea	A	B	1sln	- -		
1sv0	A	C	1sv4	A		
1t6g	A	C	1t6e	X		
1kzy	A	C	1uol	A		
1tf0	A	B	1uor	- -		
1jlt	A	B	1vpi	- -		
1m9f	A	C	1w8v	A		
1xtg	A	B	1xtf	A		
1qty	V	X	2vpf	A		
1e44	A	B	3eip	A		

b) Unbound/Bound

PDB code of complex	Chains of partners in the complex		PDB code of unbound partner	Chain in unbound partner
1s3s	A	G	1e32	A
1k9o	E	I	1ane	- -
1dtd	A	B	1aye	- -
1ava	A	C	1bg9	- -
1us7	A	B	1bgq	- -
1i1r	A	B	1bqu	A
1sg1	A	X	1btg	A
1oxb	A	B	1c03	A
3ygs	C	P	1cy5	A
1xb2	A	B	1d2e	A
1c4z	A	D	1d5f	A
1f93	A	E	1dcp	A
1qe1	A	B	1dlo	A
1oc0	A	B	1dvm	A
1f1b	A	B	1ekx	A
1f02	I	T	1f00	I
1t0f	A	C	1f1z	A
1f80	A	D	1f7t	A
1dp5	A	B	1fq8	A
1h2s	A	B	1gue	A
1gl4	A	B	1h4u	A
1usu	A	B	1hk7	A
1bgx	H	T	5ktq	A
1ma9	A	B	1s22	A
1vg9	A	B	1vg8	A
1a4y	A	B	2ang	A
1fqk	A	B	1fqi	A
1tco	A	C	1qpl	A
1m4u	A	L	1bmp	- -
1nf5	A	B	1fgx	A
1lw6	E	I	1ciq	A
1d4v	A	B	1dg6	A
1eer	A	B	1ern	A
1lzw	A	B	1k6k	A
1svx	A	B	1r6z	P
1kfu	L	S	1nx3	A
1jiw	I	P	1akl	- -
1f3v	A	B	1ca4	A
1ct0	E	I	1ds3	I
1dn1	A	B	1ez3	A
1wpl	A	K	1jg5	A
1npe	A	B	1klo	- -
1ib1	A	E	1kuy	A
1gpw	A	B	1kxj	A
1ofu	A	X	1oft	A
1qav	A	B	1qau	A
1rke	A	B	1qkr	A
1r4q	A	B	1qnu	A
1tue	A	B	1qqh	A
1gfw	A	B	1qx3	A
1y01	A	B	1sdl	A
1xdt	R	T	1tox	A
1rj9	A	B	2ng1	- -

2.2.2 Decoys

The decoys were generated by applying a docking algorithm to each complex in Table 1. The starting proteins structures were those of the free partners when available, or taken from the complex for the bound partner of unbound/bound complexes.

To generate non-native solutions (decoys), we used the program DOCK [6, 14]. DOCK explores the solution space using five angles and a distance as rigid-body parameters. The program was run in grid mode sampling the whole range of the five angles in 10° steps; 200 solutions were randomly chosen among the 18.10^6 thus generated. We checked that they were all non-native, that is, far from the X-ray structure.

2.3 Training Attributes

We chose properties of interface residues and residue pairs as training attributes. The number of parameters that may be used in training is limited by the size of the training set. With only 102 native complexes in the set, attributes could be attached to each of the 20 amino acid residue types, but not to each pair. To define pair attributes, we grouped residue types in six categories: hydrophobic *H* (ILFMV), aromatic ϕ (FYW), positively charged + (HKR), negatively charged - (DE), polar *P* (NQ) and small *S* (AGSTCP). The final set of attributes includes 84 parameters in five classes:

P1 The Voronoi interface area, measured by summing the areas of the cell faces constituting this interface (1 parameter)

P2 the total number of core interface residues (1 parameter)

P3 the number fraction of each type of core interface residues (20 parameters)

P4 the mean volume of the Voronoi cells for the core interface residues of each type (20 parameters)

P5 the number fraction of pairs of each category (21 parameters)

P6 the mean centroid-to-centroid distance in pairs of each category (21 parameters)

2.4 Statistical Learning Methods

The values of the 84 parameters were measured on the 102 native complexes and on the decoys of the training set. These values were then used to train statistical learning procedures that optimize score functions to best discriminate between the native and the decoy models.

2.4.1 Logistic Function

A logistic function is a linear combination of the parameters with weights optimized to yield a value close to 1 on the native models and 0 on the decoys. The vector of weights $W[w_i]$ was estimated on the learning set by the maximum likelihood method using the general linear model (GLM) of the R software [21].

2.4.2 SVM: Support Vector Machines

Support vector machines aim to divide a high-dimensional space into regions containing only positive examples or only decoys, each represented by a point with the values of the different parameters as coordinates. The functions defining these regions, which can be of different types, are called kernels. We tested support vector machines with linear, polynomial and radial basic function (RBF) kernels. Computations were carried out using SVMTorch [7].

2.4.3 ROGER: a ROc based GEnetic learner

The ROC (Receiver Operating Characteristics) procedure is often used evaluate learning procedures by cross-validation on examples taken from the learning set. Plotting the proportion of true positives against the proportion of false positives yields a ROC curve, and the area under that curve is the ROC criterion.

ROGER[24] uses a genetic algorithm to find a family of functions that optimizes the ROC criterion. In our implementation, the function optimized was the sum of the weighted and centered parameters:

$$f(x) = \sum_i \omega_i \left| x_i - c_i \right|$$

Thus, two values are determined for each attribute: a central value c_i and a weight ω_i. We applied a ten-fold cross-validation procedure by forming 10 groups of models, each excluding 10% of the training set, and repeated the training procedure 21 times for each set. This was required by the heuristic nature of the genetic algorithm, which can end in a local minimum. Thus, the learning procedure generated 210 functions. For a given complex, the 210 functions were evaluated, and the median value wais retained as score.

2.4.4 Missing Data in Learning

Interfaces in the learning set contained on average 18 core residues per partner protein. Thus, the least abundant residue types and some of the category pairs were absent

from many members of the set, leading to a null value of the amino acid frequency (parameter class 3) and to missing values of the volumes (class 4) and of the pair parameters in classes 5 and 6. Because learning methods including the logistic functions and ROGER, generally cannot handle missing data, we replaced each missing value by either:

- The mean value on the whole set
- The median value on the whole set
- The mean value in the category (native or decoy) to which the example belong
- The median value in the category

Missing values also had to be replaced when scoring a model generated by docking in the test phase. Here again, many combinations are possible: replacing by the mean or median values of the whole learning set, of the whole test set, by category or not.

In this study, we replaced missing values by their median on the whole learning set during both the learning procedure and the test phase. We tested other alternatives and found that they performed less well, possibly because our parameters have non-Gaussian distributions.

3. RESULTS AND DISCUSSION

3.1 Performance of the Learning Procedures

The ROC curve was evaluated on the training set for four different scores: the sum of the mean square deviation of the attributes from their mean values, the logistic function, the classification made with SVMs, and the scoring function obtained with ROGER. A perfect selection (100% true positives and no false positives) should make the area under the ROC curve AUC equal to 1; a random selection yielding true positives and false positives in equivalent numbers should have an AUC of 0.5.

Taking the sum of the mean square deviations as a score yielded an AUC close to 0.5, indicating that individual parameters discriminate very poorly between the native and the decoys, and that they must be properly weighted and combined, which is what a learning procedure aims to. With the logistic function, the area under the ROC curve increased to 0.85, indicating a better discrimination.

ROGER and the SVMs did much better, achieving AUC of respectively 0.98 and 0.99. More important perhaps, the initial slope of the ROC curve was very steep in both cases, implying that ROGER and the SVMs had very few false positives among their best scoring solutions. Thus, both learning procedures were successful, and we retained the ROGER score for further studies as the SVMs only give a binary classification, ill-suited to our problem of “finding a needle in a hay stack”. As docking generates hundreds of thousands of non native solutions along with only a few near-native ones, a binary classification would still yield many false positives. In contrast, the ROGER score could easily be combined with other functions if needed.

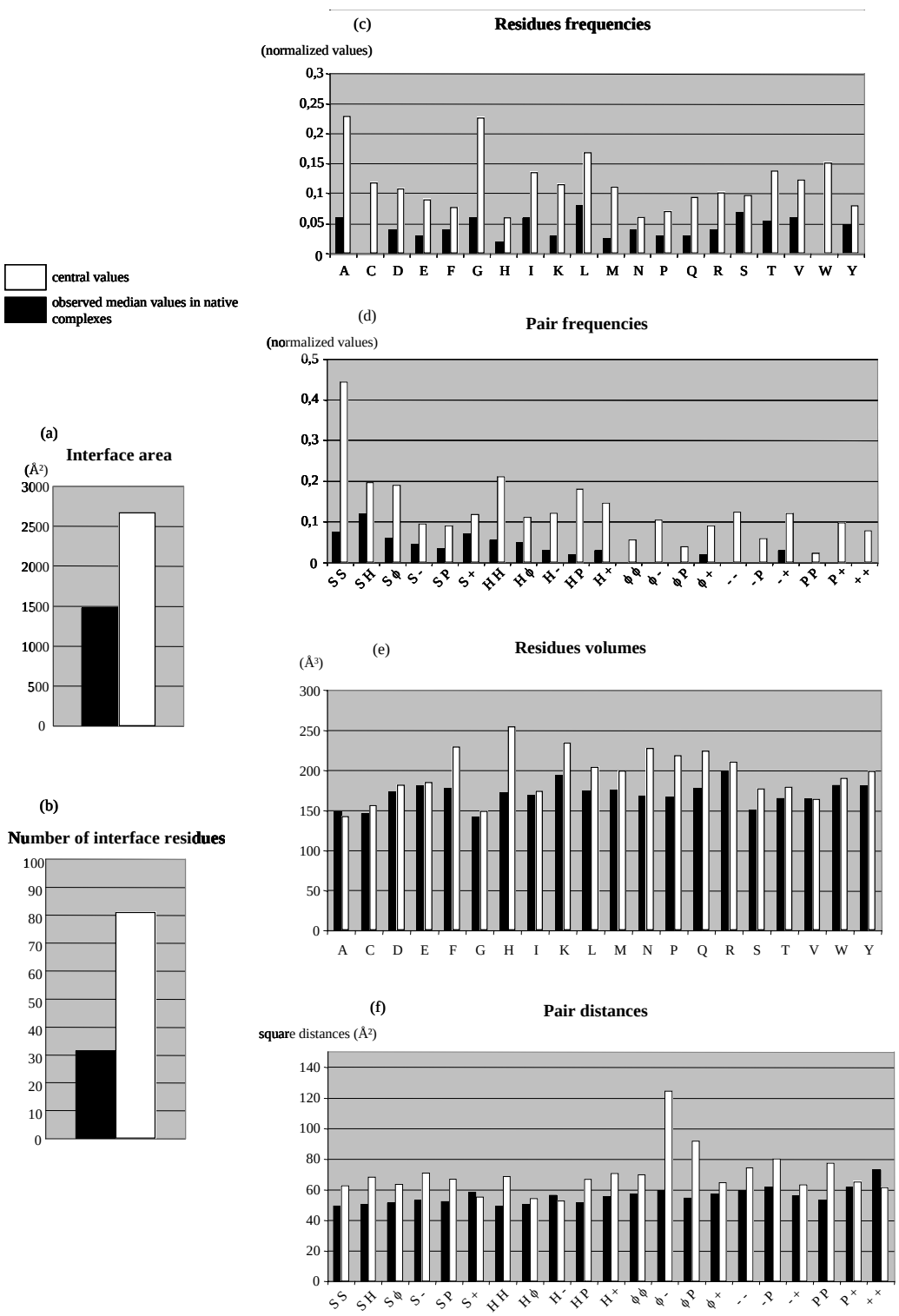


Figure 3: Bar diagrams on parameters

(a) P1 Voronoi interface area (Å²) (b) P2 number of interface residues

(c) P3 Residues frequencies (normalized) (d) P5 Pair frequencies (normalized)

(e) P4 Residues Voronoi volumes (Å^3) (f) P6 Pair square distances

(b)

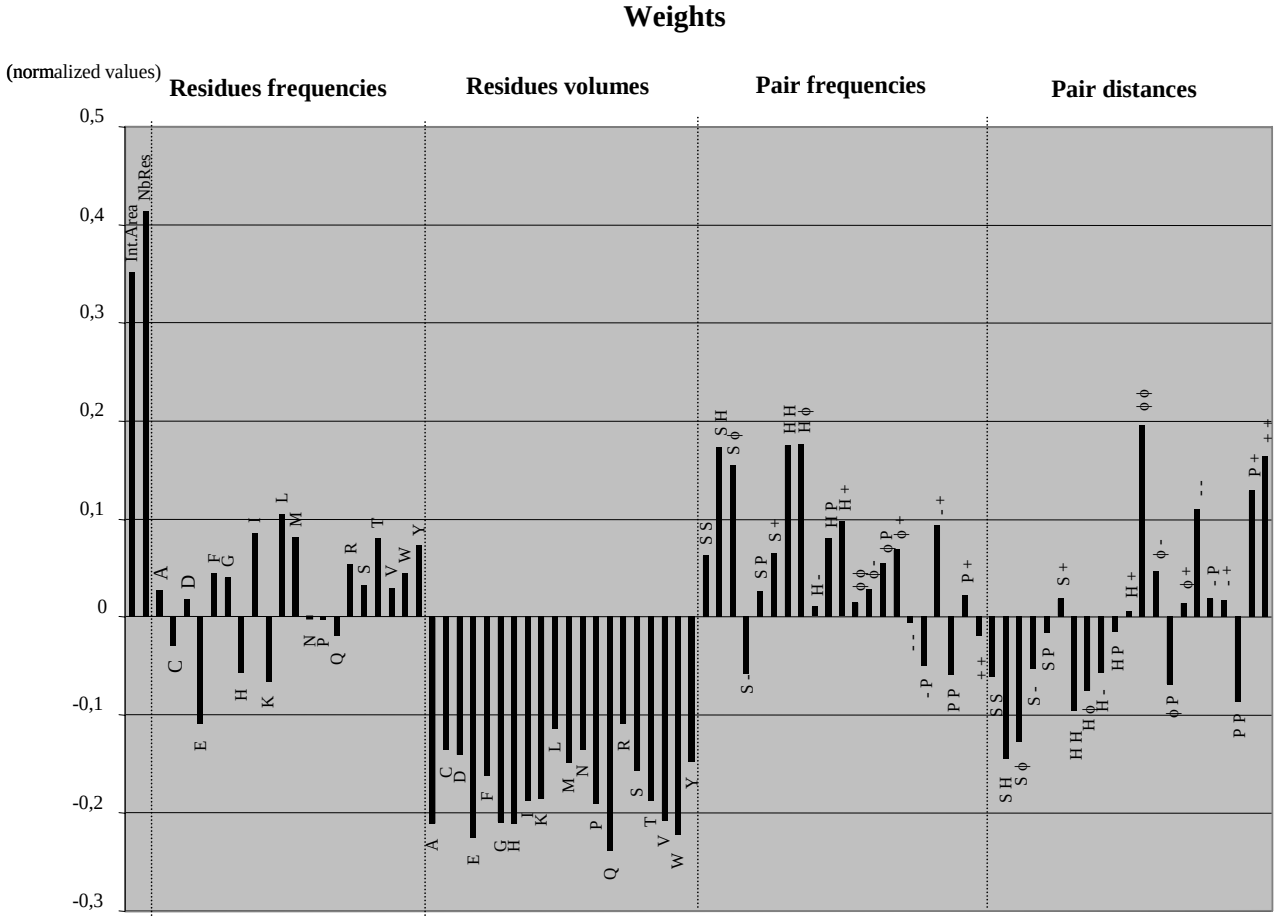


Figure 4: Bar diagrams on weights

3.2 Weight and Central Value of the Attributes in the ROGER scoring functions

The ROGER score of a docking model is the median value of 210 functions generated in separate trainings as described under Methods. The role of an attribute in the function depends on both the absolute value and the sign of its weight.

Because the algorithm is designed to minimize the score of native-like solutions, the most significant attributes are those having large weights and/or central values that are very different from the median value. The greater the difference between the central and median value, the more influence a parameter has.

Figure 3 and 4 show that the largest positive weights are for the interface area and number of interface residues, two attributes that measure the size of the interfaces. They are highly correlated and both have central values that are much larger than the median values observed in native complexes. Thus, large interfaces are preferred. Residue and pair frequencies also have central values that are larger than the median. The weights are small and of both signs for residue frequencies, which govern the amino acid composition of the interface. They are larger and mostly positive for pair frequencies, especially for pairs involving hydrophobic residues, implying that the score favors hydrophobic contacts.

In contrast, the central value is close to the median for most of the residue volumes and pair distances. The residue volumes and the pair distances involving hydrophobic residues have large negative weights, and therefore, they are important attributes. A tight packing of

hydrophobic residues at the interface, which makes their volume and the distance to neighbors less than the average, is particularly favored by the ROGER score.

3.3 Results on the Targets of CAPRI Rounds 3 to 6

We tested the scoring functions on models of the targets of CAPRI Rounds 3-6 generated by two docking programs: DOCK [6, 14] run as in the learning set, and HADDOCK[9].

Models generated by the two programs were grouped into classes depending on the fraction Fnat of the residue-residue contacts in the native structure that were present in the model, and on the fraction Fint of the interface residues that were correctly predicted. As described in the legend of Table 2, classes 1 to 4 were defined by ranges of Fnat; class 5 had Fnat =0 (no native contact), but Fint >0 (some interface residues correctly predicted); class 6 was for incorrect models.

As all classes were not represented for each target, Table 2 cites the "best class" that was present in a given data set. Taking target 11 as an example, the best model in the data set generated by HADDOCK had Fnat=0.54. Thus, the best class in that data set was 2 (0.5<Fnat<0.75).

3.3.1 Re-ranking HADDOCK Solutions

Models for five CAPRI targets generated by HADDOCK [9] were kindly communicated by Dr. A. Bonvin. Their ROGER score was computed and the result of their re-ranking is given in Table 2.

With Target 11, a model of the best class (class 2), re-ranked 4, and one of the next best class (class 3) was 1st. The top 50 ROGER scores included 4 models of class 2, and 44 models of class 3. Thus very few if any of the top 50 were false positives. With Targets 13 and 14, the data set contained class 1 models that were ranked 1st or 2d by ROGER, although the top 50 contained more false positives than for Target 11. The results were less satisfactory on Targets 12 and 15, no model of the best class present in these data sets scoring in the top 50.

3.3.2 Re-ranking DOCK Solutions

We ran a complete grid search of the five angles in steps of 10 degrees on ten of the CAPRI targets of Rounds 3-6. The models generated by DOCK were clustered, the average position in each cluster was retained, and its ROGER score was evaluated.

Table 2 shows that the DOCK data sets were generally poorer than for HADDOCK. None contained models of class 1 or 2. The best models had Fnat<0.5 (class 3 or 4), or even Fnat=0 (class 5) in the case of Target 15. Nevertheless, the re-ranking with ROGER performed correctly. In all but one data set, the top 50 included models of the best class that was present in the set. The 1st rank was a model of the best class (class 4) in the case of Target 16, and a model of the second best class in four other cases.

Table 2: Summary of rescoring results on Dock and Haddock data sets

Program	Target	Number of solutions scored	Best class appearing in set	rank1	rank2	rank3	Number of hits
HADDOCK	11	200	2	4	1	9	4
	12	200	1	79	78	49	0
	13	200	1	1	7	5	1
	14	200	1	2	1	3	10
	15	200	3	57	1	20	0
DOCK	9	3446	4	4	1	3	10
	11	32	4	11	3	2	3
	12	139	3	27	14	20	4
	13	2263	4	84	7	1	0
	14	2460	4	5	352	1	2
	15	9	5	9	1	8	1
	16	165	4	1	2	5	2
	17	72	4	39	1	2	1
	18	972	3	24	1	4	1
	19	1979	4	22	25	1	2

Definitions of the classes:

Class 1: Fnat > 0.75
Class 2: 0.5 < Fnat ≤ 0.75
Class 3: 0.25 < Fnat ≤ 0.5
Class 4: 0 < Fnat ≤ 0.25
Class 5: Fint > 0 and Fnat = 0
Class 6: Fint = 0 and Fnat = 0

Fnat: fraction of contacts in the native structure that are present at the docking model
Fint: fraction of correctly predicted interface residues

rank1, rank2, rank3: rank1 is the best ROGER rank achieved by a model that belong to the best class present in the set; rank2 is the best rank achieved by of model of the next best class, and so on for rank3.
Number of hits: number of models with a ROGER rank less than 50 that belong to the best class present in the set.

3.3.3 Comparison with other docking programs

Recent experiment in protein-protein docking such as CAPRI (Critical Assessment of PRedicted Interactions) has shown improvements in docking procedures now able to take certain flexibility into account. Lately scoring functions appeared to play a greater role in rescoring the solutions [18], especially during a refinement stage.

They usually rely on shape complementarity supplemented by additional energy terms such as van der Waals Coulomb or desolvation in different and ingenious ways [5, 8, 10, 15, 16, 19, 23, 26, 28, 29]. Taking into account residue conservation and biological information from literature has also been added to the process with varying results [17, 27].

Lately novel energy scoring functions involving statistical methods have made a breakthrough, first in predicting protein structure in the CASP experiment [22] but now also for the protein-protein docking problem with very promising first attempts [8, 30].

Our work falls within this scope of statistical method but affords both the measure of geometric criteria and protein structure knowledge described by previous energy-based methods.

4. CONCLUSION

The residue-based Voronoi tessellation provides a convenient low-resolution description of protein structure and protein-protein interfaces. We built a set of parameters derived from that description and a data set of native or decoys models obtained by docking, and used these sets to train the ROGER statistical learning procedure. It returned a score that we tested on docking models of CAPRI targets generated by two different docking programs.

For most targets, a best or second best class solution was found in the top 10 ranking solutions, and in more than half of the cases the top ranking solution belonged to the best or second best class. We may thus hope that further refinement of the parameters and of the scoring function will have a class 1 or 2 solution as top ranking in all cases. The quick and efficient selection of docking models based on whole residues will be an asset in whole-genome studies. On the other hand, these models remain crude and we cannot expect to have very accurate solutions unless go back to an atomic model and an appropriate scoring function.

Last, a study that deals only with the scoring function is obviously dependent on the quality of the original data set. Further refinement of the method will require the implementation of a new exploration method that could also be based on Voronoi tessellation

5. ACKNOWLEDGEMENTS

We thank Dr A. Bonvin for kindly communicating data sets of CAPRI models, and the EIDIPP program of Action Concertée Incitative IMPBio for financial support.

6. REFERENCES

[1] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403-10.

[2] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. 2000. The Protein Data Bank. Nucleic Acids Res. 28(1):235-42.

[3] Bernauer, J.; Poupon, A.; Azé, J.; Janin, J. 2005. A docking analysis of the statistical physics of protein-protein recognition. Physical Biology. 2(2):S17.

[4] Boissonnat, J. D.; Devilliers, O.; Pion, S.; Teillaud, M.; Yvinec, M. 2002. Triangulations in CGAL. Comput. Geom. Theory Appl. 225-19.

[5] Carter, P.; Lesk, V. I.; Islam, S. A.; Sternberg, M. J. 2005. Protein-protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI. Proteins. 60(2):281-8.

[6] Cherfils, J.; Duquerroy, S.; Janin, J. 1991. Protein-protein recognition analyzed by docking simulation. Proteins. 11(4):271-80.

[7] Collobert, R.; Bengio, S. 2001. SVMTorch: Support Vector Machines for Large-Scale Regression Problems. Journal of Machine Learning Research. 1143-160.

[8] Daily, M. D.; Masica, D.; Sivasubramanian, A.; Somarouthu, S.; Gray, J. J. 2005. CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. Proteins. 60(2):181-6.

[9] Dominguez, C.; Boelens, R.; Bonvin, A. M. 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc. 125(7):1731-7.

[10] Fernandez-Recio, J.; Abagyan, R.; Totrov, M. 2005. Improving CAPRI predictions: optimized desolvation for rigid-body docking. Proteins. 60(2):308-13.

[11] Graille, M.; Zhou, C. Z.; Receveur-Brechot, V.; Collinet, B.; Declerck, N.; van Tilbeurgh, H. 2005. Activation of the LicT transcriptional antiterminator involves a domain swing/lock mechanism provoking massive structural changes. J Biol Chem. 280(15):14780-9.

[12] Hamelryck, T.; Manderick, B. 2003. PDB file parser and structure class implemented in Python. Bioinformatics. 19(17):2308-10.

[13] Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J. 2003. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins. 52(1):2-9.

[14] Janin, J.; Wodak, S. J. 1985. Reaction pathway for the quaternary structure change in hemoglobin. Biopolymers. 24(3):509-26.

[15] Law, D.; Hotchko, M.; Ten Eyck, L. 2005. Progress in computation and amide hydrogen exchange for prediction of protein-protein complexes. Proteins. 60(2):302-7.

[16] Lee, K.; Sim, J.; Lee, J. 2005. Study of protein-protein interaction using conformational space annealing. Proteins. 60(2):257-62.

[17] Ma, X. H.; Li, C. H.; Shen, L. Z.; Gong, X. Q.; Chen, W. Z.; Wang, C. X. 2005. Biologically enhanced sampling geometric docking and backbone flexibility treatment with multiconformational superposition. Proteins. 60(2):319-23.

[18] Mendez, R.; Leplae, R.; Lensink, M. F.; Wodak, S. J. 2005. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. Proteins. 60(2):150-69.

[19] Mustard, D.; Ritchie, D. W. 2005. Docking essential dynamics eigenstructures. Proteins. 60(2):269-74.

[20] Poupon, A. 2004. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. Curr Opin Struct Biol. 14(2):233-41.

[21] R Development Core Team, R: A language and environment for statistical computing, 2005.

[22] Rohl, C. A.; Strauss, C. E.; Chivian, D.; Baker, D. 2004. Modeling structurally variable regions in homologous proteins with rosetta. Proteins. 55(3):656-77.

[23] Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. 2005. Geometry-based flexible and symmetric protein docking. Proteins. 60(2):224-31.

[24] Sebag, M.; Azé, J.; Lucas, N. 2004. ROC-Based Evolutionary Learning: Application to Medical Data Mining. Lecture Notes in Computer Science.

[25] Soyer, A.; Chomilier, J.; Mornon, J. P.; Jullien, R.; Sadoc, J. F. 2000. Voronoi tessellation reveals the condensed matter character of folded proteins. Phys Rev Lett. 85(16):3532-5.

[26] Terashi, G.; Takeda-Shitaka, M.; Takaya, D.; Komatsu, K.; Umeyama, H. 2005. Searching for protein-protein interaction sites and docking by the methods of molecular dynamics, grid scoring, and the pairwise interaction potential of amino acid residues. Proteins. 60(2):289-95.

[27] Tress, M.; de Juan, D.; Grana, O.; Gomez, M. J.; Gomez-Puertas, P.; Gonzalez, J. M.; Lopez, G.; Valencia, A. 2005. Scoring docking models with evolutionary information. Proteins. 60(2):275-80.

[28] van Dijk, A. D.; de Vries, S. J.; Dominguez, C.; Chen, H.; Zhou, H. X.; Bonvin, A. M. 2005. Data-driven docking: HADDOCK's adventures in CAPRI. Proteins. 60(2):232-8.

[29] Wiehe, K.; Pierce, B.; Mintseris, J.; Tong, W. W.; Anderson, R.; Chen, R.; Weng, Z. 2005. ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. Proteins. 60(2):207-13.

[30] Zhang, C.; Liu, S.; Zhou, Y. 2005. Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function. Proteins. 60(2):314-8.

3.4 Une discrimination entre dimères biologiques et dimères cristallographiques

3.4.1 Introduction

Lors de la résolution d'une structure de protéine par cristallographie, on obtient les coordonnées des atomes dans l'unité asymétrique pour un groupe d'espace donné (voir partie II page 101). L'unité asymétrique est la plus petite entité qui permet de reconstruire tout le cristal par opérations de symétrie [21]. On obtient ainsi l'empilement cristallin (voir figure 3.18).

Dans l'exemple de l'allantoïcase (figure 3.18), l'unité asymmétrique contient un homodimère alors que l'entité biologique est un homohexamère (figure 3.19). Cet exemple montre un des problèmes qui se posent en cristallographie : comment, à partir de la structure obtenue dans l'unité asymétrique, déterminer l'entité biologique. Les mesures à l'interface ne permettent pas de trancher si la surface d'interaction est importante. Seules les méthodes biochimiques permettent de trancher.

Le calcul des paramètres précédemment décrits pour l'amarrage sur une liste 310 dimères spécifiques (également dits biologiques car existants dans les conditions physiologiques) et non-spécifiques (n'existant que dans le cristal) montrent qu'ils sont, là encore, discriminants.

3.4.2 Méthodes et logiciels

Jeu d'apprentissage Le jeu d'apprentissage sur lequel nous avons travaillé est celui de l'étude de R. Bahadur et collaborateurs [9]. Il contient :

- 188 dimères « cristallographiques »dont :
  - 85 sans symétrie d'axe 2,
  - 103 avec symétrie d'axe 2 ;
- 122 dimères « biologiques ».

Ce jeu, vérifié manuellement a été longuement étudié [8, 9, 106, 176, 241].

Mesures et apprentissage Pour chacun de ces complexes, nous avons calculé les paramètres présentés au paragraphe 2.3.1.4 et effectué un apprentissage par séparateurs à vaste marge (paragraphe 2.5).

Pour éviter les problèmes de surapprentissage, nous avons séparé aléatoirement notre jeu en deux :

- l'apprentissage est effectué sur deux tiers des données ;
- la validation est réalisée sur le tiers restant.

L'apprentissage est effectué avec une 5-validation croisée et l'ensemble de la procédure (séparation du jeu, apprentissage et validation croisée) est répété 5 fois.

Les essais d'optimisation des paramètres C et γ (voir paragraphe 2.5) n'ayant pas permis d'obtenir de meilleures performances ne seront pas détaillés. Les valeurs des paramètres C et γ ont donc été prises par défaut. Les meilleures performances ont été obtenues avec un noyau gaussien.

Deux types d'apprentissage ont été effectués (figure 3.20) :

1. Une discrimination simple entre dimères « biologiques »et dimères « cristallographiques »et une discrimination simple entre dimères « cristallographiques »sans symétrie d'axe 2 et dimères « cristallographiques »avec symétrie d'axe 2 (deux étapes à deux classes) ;

[21] Pour plus d'informations, on pourra aussi se reporter à l'ouvrage de J. Drenth [62].

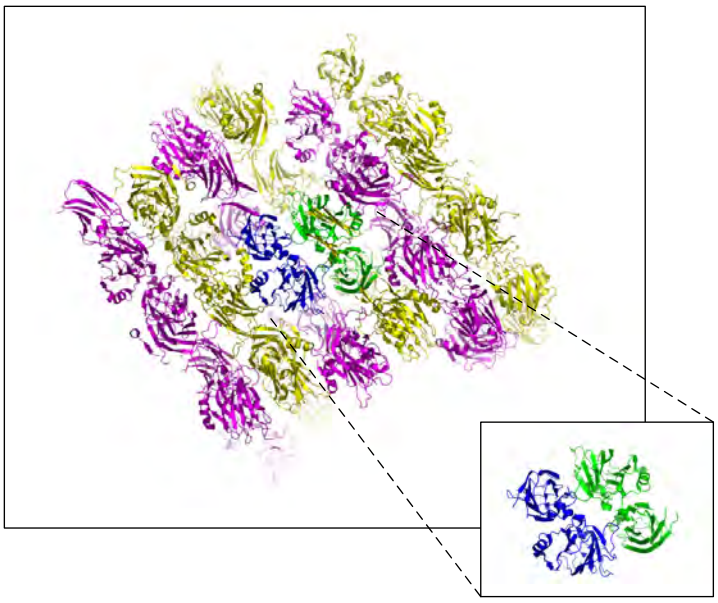


FIG. 3.18 – Unité asymétrique et empilement cristallin : exemple de l'allantoïcase de la levure [129]. *En bleu et vert, les deux monomères de l'unité asymétrique et en jaune et magenta, les monomères obtenus par application des opérations de symétrie.*

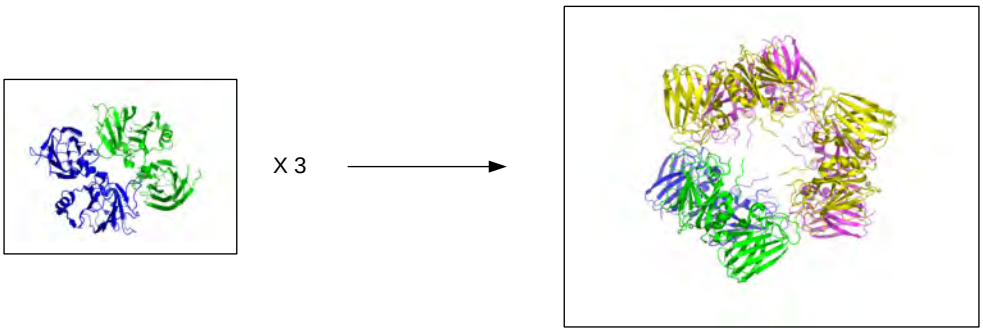


FIG. 3.19 – Dimère « cristallographique » et hexamère « biologique » de l'allantoïcase. *En bleu, le monomère ; en bleu et vert, le dimère de l'unité asymétrique et en jaune et magenta, les monomères obtenus par application des opérations de symétrie pour former l'homohexamère « biologique ».*

2. Un apprentissage en une étape, mais avec trois classes pour distinguer entre dimères « biologiques », dimères « cristallographiques »sans symétrie d'axe 2 et dimères « cristallographiques »avec symétrie d'axe 2.

Les calculs ont été réalisés à l'aide de la bibliothèque *libSVM* [36] et du logiciel *R* [182].

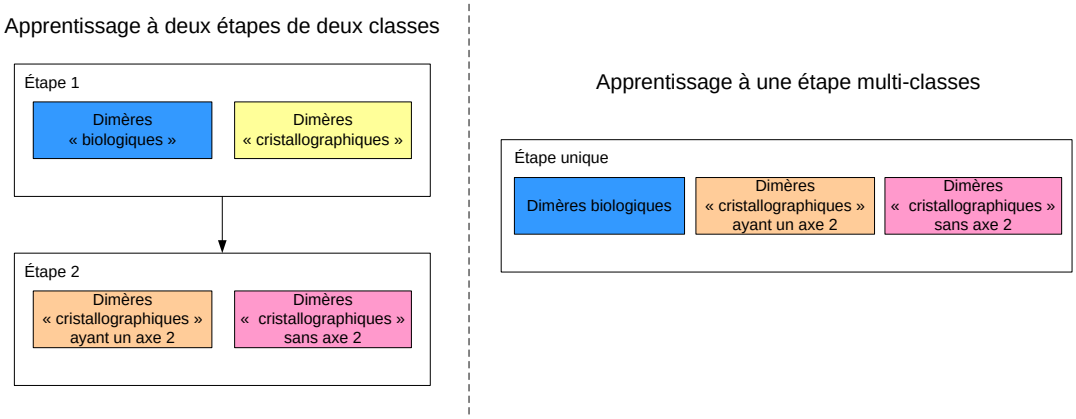


FIG. 3.20 – Schéma des différentes procédures d'apprentissage pour la discrimination entre complexes « biologiques » et « cristallographiques ». *Un apprentissage en deux étapes et un apprentissage en une étape ont été effectués.*

3.4.3 Résultats et discussion

Résultats Pour chaque type d'apprentissage, nous avons tracé et mesuré l'aire sous la courbe de ROC (figure 3.21).

Apprentissage en deux étapes, chacune contenant deux classes La première étape, de discrimination entre dimères « cristallographiques »et « biologiques », présente une aire sous la courbe moyenne de 0,89±0,03. La deuxième étape, de discrimination entre dimères « cristallographiques »sans symétrie d'axe 2 et avec symétrie d'axe 2, présente une aire sous la courbe moyenne de 0,71±0,05.

Apprentissage en une étape avec trois classes Dans cette procédure multiclasses, on obtient une aire sous la courbe moyenne de 0,85±0,03.

Comparaison avec les classifications existantes Nous avons repris l'étude de classification de H. Zhu et collaborateurs [241] pour comparer les performances obtenues par notre méthode, dans laquelle les descripteurs sont issus de la construction de Voronoï avec ceux utilisés traditionnellement.

Sur le jeu de R. Bahadur, les résultats sont plus performants avec les descripteurs que nous avons utilisés, même si nous ne disposons pas des courbes de ROC pour l'étude de H. Zhu (voir tableau 3.4 page 96).

Nous avons effectué les mêmes calculs que précédemment sur le jeu de H. Zhu qui contient :
- 106 dimères cristallographiques ;
- 137 dimères biologiques dont :

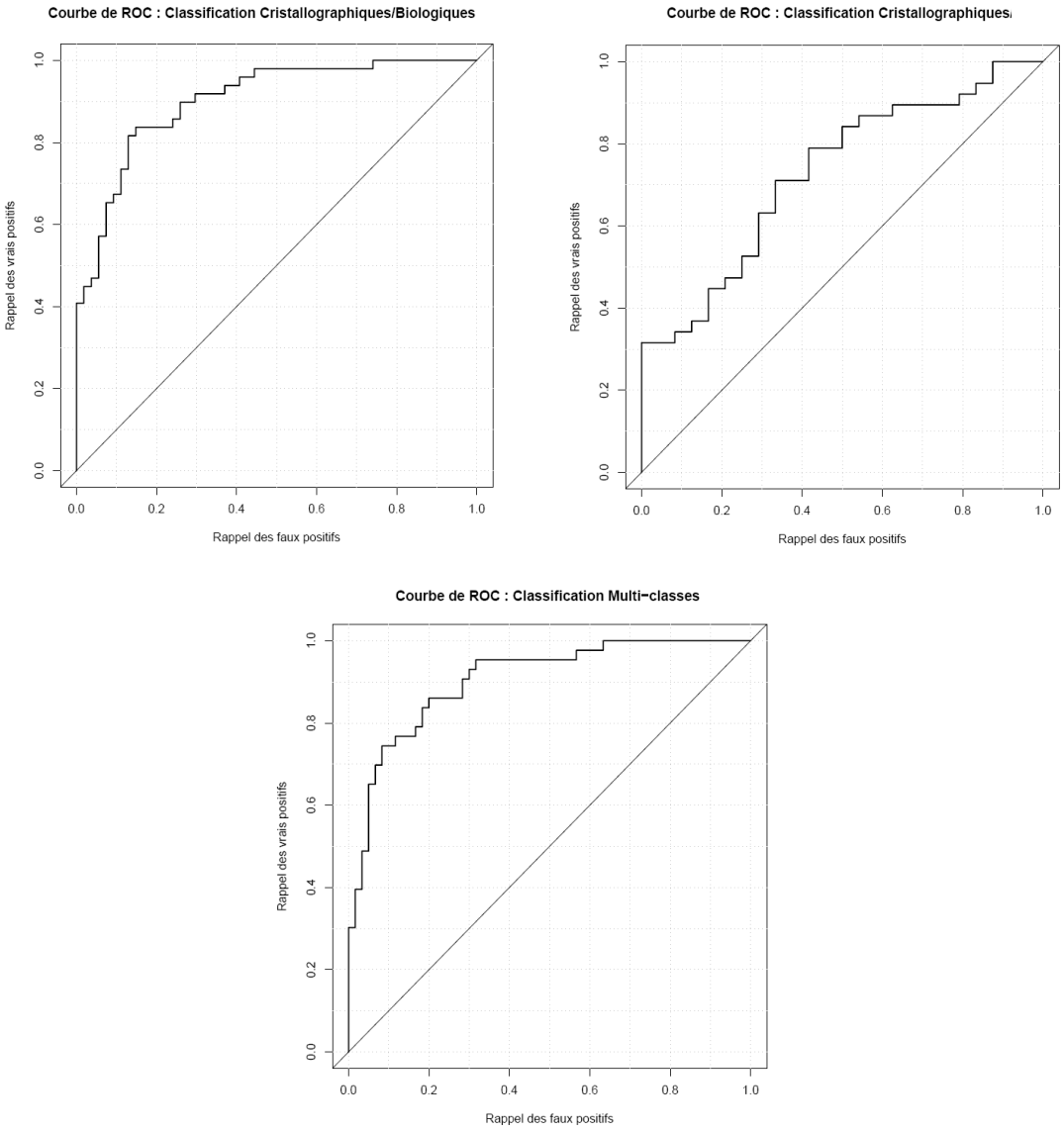


FIG. 3.21 – Courbes de *ROC* pour chacune des procédures d'apprentissage. *Ici ne sont représentées que les courbes obtenues sur les jeux tests pour une des quinze exécutions du calcul.*

– 75 obligées (*obligates*) ;
– 62 non-obligées (*non-obligates*).

Sur les trois jeux décrits dans ce travail (122 complexes extraits de la PDB, jeu de R. Bahadur, et jeu de H. Zhu), nous avons tracé les répartitions des surfaces sous forme de boîtes à moustaches. On obtient les résultats présentés à la figure 3.22.

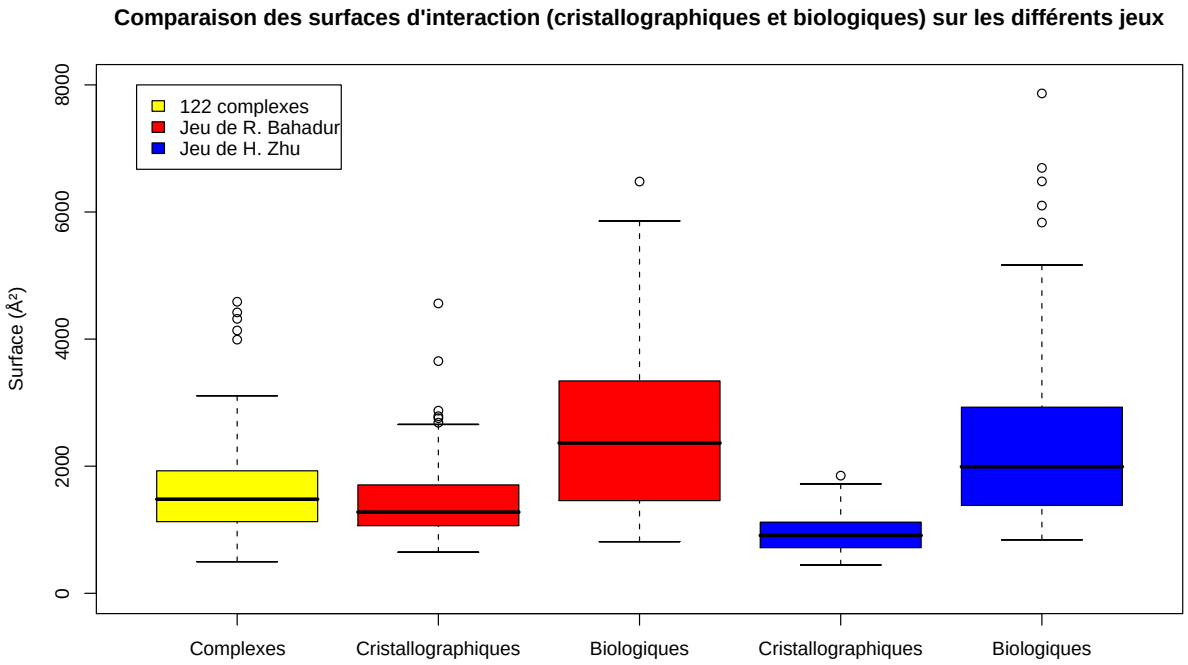


FIG. 3.22 – Répartition des surfaces pour les interactions biologiques et cristallines sur les trois jeux de complexes étudiés.

On observe que les complexes que nous avons extraits automatiquement ont des surfaces intermédiaires entre celles des dimères cristallins et celles des dimères biologiques. Ce résultat n'est pas surprenant car les complexes sont moins stables que les homodimères, mais plus stables que les dimères cristallins. Les interactions cristallines ont des surfaces plus variables sur le jeu de R. Bahadur que sur celui de H. Zhu, ce qui peut expliquer les chutes de performances de la méthode de H. Zhu sur le set de R. Bahadur [241].

Nous avons ensuite tracé les répartitions des surfaces sous forme de boîtes à moustaches pour chaque jeu pris séparément. Pour les deux jeux, on obtient des répartitions de surfaces conformes à celle présentées dans les articles. La surface de Voronoï est donc en accord avec les surfaces calculées selon la méthode de Lee et Richards.

Pour le jeu de R. Bahadur (figure 3.23), on voit que la différence entre dimères « cristallographique »ayant un axe 2 et monomères sans axe 2 est faible. La surface, qui est un descripteur important dans l'apprentissage ne pourra, seule, permettre de différencier ces types d'interactions : c'est ce qui rend cette discrimination plus difficile.

Pour le jeu de H. Zhu (figure 3.24), on voit que la différence de surface entre interactions non-obligées et obligées est importante, ces deux catégories seront donc plus aisées à différencier

avec une procédure d'apprentissage attribuant un poids important à la surface que les monomères de l'étude de R. Bahadur.

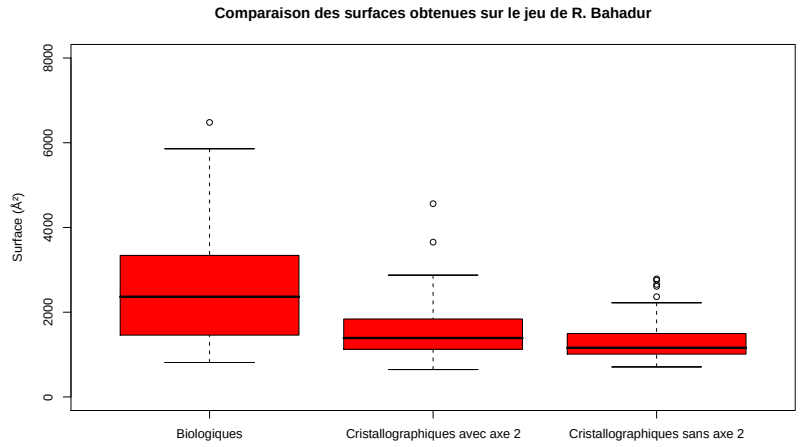


FIG. 3.23 – Répartition des surfaces sur le jeu de R. Bahadur

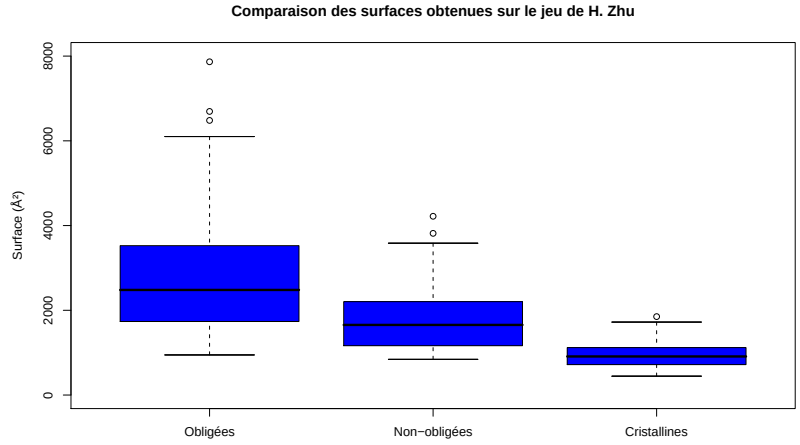


FIG. 3.24 – Répartition des surfaces sur le jeu de H. Zhu

On observe donc que la surface, qui est l'un des descripteurs les plus importants de l'apprentissage de H. Zhu ne peut à elle seule permettre de discriminer entre les différents types d'interactions.

Résultats Pour l'apprentissage en deux étapes :

- la première étape, de discrimination entre interactions cristallographiques et interactions biologiques, présente une aire sous la courbe de 0,91±0,03 ;
- la deuxième étape, de discrimination entre interactions obligées et interactions non-obligées, présente une aire sous la courbe de 0,77±0,08.

Pour l'apprentissage multiclasses, l'aire sous la courbe est de 0,87±0,06 (voir figure 3.25)

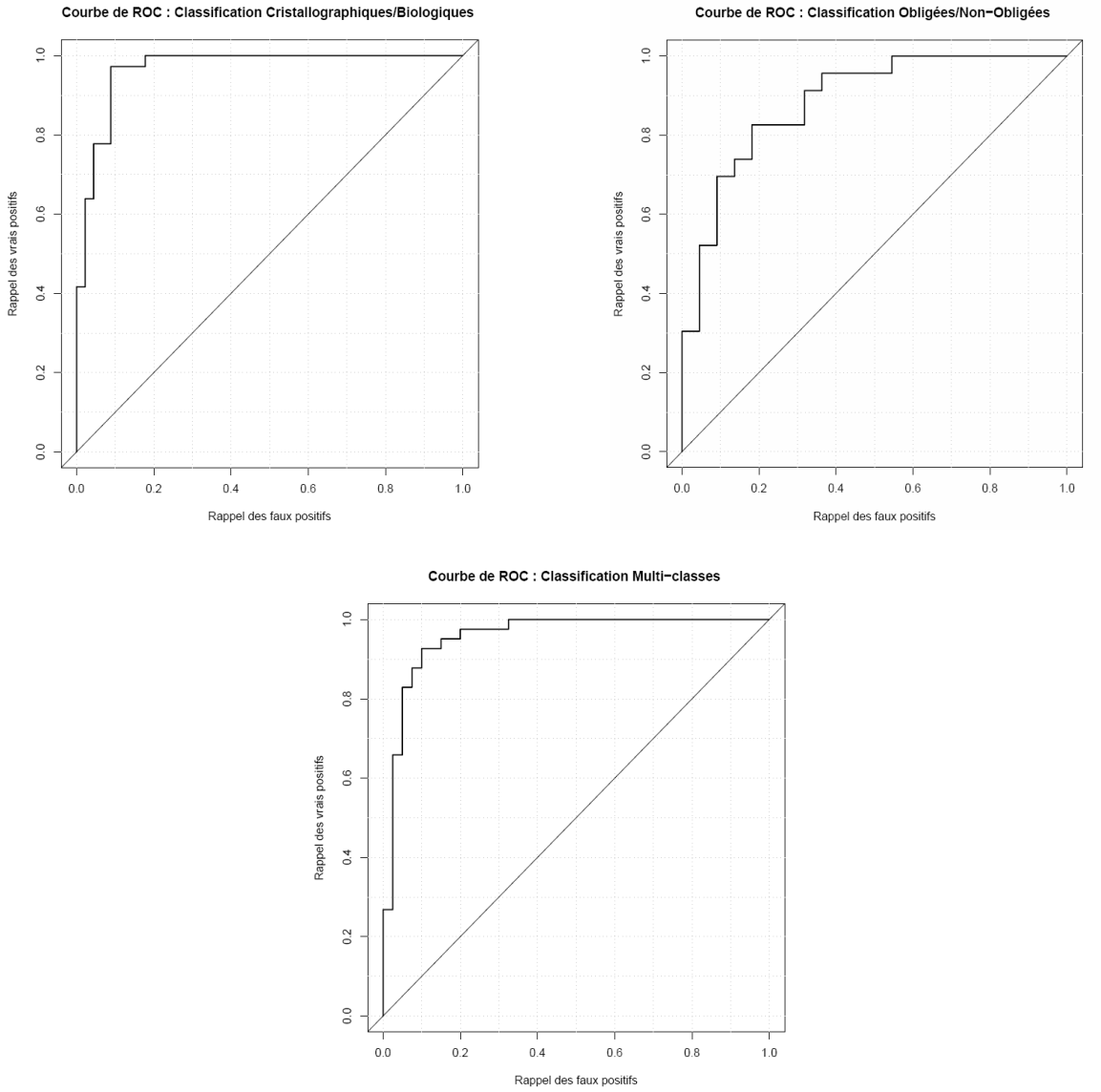


FIG. 3.25 – Courbes de *ROC* pour chacune des procédures d'apprentissage sur le jeu de H. Zhu. *Ici ne sont représentées que les courbes obtenues sur les jeux tests pour une des quinze exécutions du calcul.*

Comparaison des précisions entre les deux méthodes La précision correspond au nombre de bonnes prédictions sur le nombre total de prédictions. Sur le jeu de R. Bahadur, on obtient les resultats du tableau 3.4 et pour le jeu de H. Zhu, on obtient les résultats du tableau 3.5.

TAB. 3.4: Comparaison des précisions entre les deux méthodes pour le jeu de R. Bahadur

Apprentissage	Notre Méthode	Méthode de H. Zhu
Dimères « cristallographiques »/ Dimères « biologiques »	0,84±0,02	0,80
Dimères « cristallographiques » avec axe 2 / sans axe 2	0,70±0,04	indisponible
Multiclasses	0,85±0,03	indisponible

TAB. 3.5: Comparaison des précisions entre les deux méthodes pour le jeu de H. Zhu

Apprentissage	Notre Méthode	Méthode de H. Zhu
Cristallines / Obligées	0,86±0,04	0,94±0,01
Obligées / Non-obligées	0,78±0,05	0,75±0,03
Multiclasses	0,75±0,08	0.81±0,01

Nous sommes donc en mesure de prédire correctement si un dimère est cristallographique ou biologique dans 84% des cas.

Conclusion Dans les deux cas, on obtient un apprentissage quasiment aussi performant avec les descripteurs issus de la construction de Voronoï. Il est d'ailleurs plus performant avec un remplacement des valeurs manquantes par classe.

On observe aussi que les classificateurs issus de la construction de Voronoï permettent de compléter efficacement l'information apportée par la surface. Il serait intéressant de compléter cette étude en intégrant dans les paramètres d'apprentissage issus de la construction de Voronoï la fraction de résidus non-polaires rencontrée.

Chapitre 4

Conclusion de la première partie

Dans cette étude, nous avons utilisé la tessellation de Voronoï pour modéliser la structure des protéines. Nous avons utilisé un certain nombre de paramètres, calculés sur ce modèle, pour construire une fonction de score utilisable dans les procédures d'amarrage. Cette étude a permis de montrer que :

- ce type de diagramme est un bon modèle de l'empilement des résidus dans la structure. Même simplifié, ce modèle permet de retrouver les mesures de compacité précédemment observées. De plus, son implémentation efficace permet de nombreuses mesures ;
- on peut, avec ce modèle, déterminer des affinités entre types de résidus et des compositions aux interfaces en accord avec les paramètres physico-chimiques et les études précédentes. Les résultats ainsi obtenus montrent que la surface d'interaction, bien qu'importante, et la proportion d'atomes enfouis ne sont pas les seuls critères permettant de décrire les interfaces entres protéines ;
- les séparateurs à vaste marge (*SVM*) et l'algorithme génétique *ROGER*, même avec un traitement des valeurs manquantes simpliste, nous ont permis de discriminer entre différents types d'assemblages protéiques.

Nous avons aussi pu mettre au point un protocole d'extraction automatique et exhaustif des jeux de données de complexes sur la *Protein Data Bank* qui permet d'utiliser des procédures d'apprentissage pour le problème de l'amarrage protéine-protéine.

Toutefois, un certain nombre d'études complémentaires seront nécessaires pour obtenir des fonctions qui classent la meilleure solution en premier.

Tout d'abord, un meilleur traitement des valeurs manquantes est nécéssaire. En particulier, l'utilisation des dernières techniques d'imputation des données manquantes [174] ayant donné de bons résultats sur les données d'expression issues des puces à ADN [111] pourrait permettre d'obtenir de meilleurs résultats.

L'utilisation de l'algorithme *ROGER* qui maximise l'aire sous la courbe de ROC, a donné de bons résultats, mais quelques améliorations simples pourraient y être apportées. Au lieu de maximiser l'aire sous la courbe de *ROC*, il serait possible de maximiser le ratio entre proportion de vrais positifs et proportion de faux positifs. Ainsi, la pente de la courbe à l'origine serait plus importante, permettant d'obtenir de très bonnes solutions dans les tout premiers résultats obtenus, ce qui est le but recherché.

Une étude complémentaire des descripteurs et des poids obtenus par l'algorithme de *ROGER* sur les structures issues des différentes études de mutations sur une même structure serait intéressante. Par exemple, on pourrait utiliser les structures obtenues dans les différentes études

structurales de mutagenèse de Roy Mariuzza [71, 134, 209]. Ainsi, on pourrait déterminer plus précisément l'influence de chaque type de résidu et de paramètre à l'interface sur la fonction de score, cela permettant, soit d'affiner les paramètres déjà utilisés, soit d'en intégrer de nouveaux.

Ensuite, il pourrait être intéressant, dans le cadre de l'amarrage protéine-protéine, de travailler sur des complexes « presque-natifs ». Par « presque-native », nous entendons la meilleure conformation qui puisse être obtenue par la procédure géométrique qui calcule les conformations. En effet, c'est celle -ci que nous devons « sortir » de l'ensemble des conformations, et non pas la conformation native.

Enfin, en utilisant la procédure de Voronoï et les méthodes mises au point au niveau atomique, il serait possible d'affiner la description atomique des interfaces en fonction des différents types de complexes [152], une fois que l'on a obtenu une ou plusieurs conformations proches de la solution.

Grâce à cette méthode, plusieurs applications et améliorations d'applications devraient voir le jour prochainement.

D'abord, la fonction de score pour l'amarrage et la fonction de discrimination entre interactions « biologiques » et « cristallographiques » seront améliorées et rendues accessibles.

La fonction de score mise au point pourra être introduite comme module du programme *HADDOCK* comme le souhaiterait A. Bonvin, pour améliorer le score au fur et à mesure des étapes d'affinement et de clustering. L'utilisation d'informations biologiques fiables telles que les données de trace évolutionnaire [140, 184, 183] combinée à notre fonction de score statistique laisse envisager de bons résultats.

À plus long terme, la méthode mise au point pourrait être utilisée pour le docking flexible en adaptant un algorithme géométrique de docking existant tel que celui développé par Y. Wang [223]. Les premiers tests sur les résultats de tri par la fonction de score sur cet algorithme sont satisfaisants et une adaptation de celui-ci pour prendre en compte des liaisons de type rotule à certains endroits de la protéine sont envisageables.

Il est aussi envisageable d'utiliser la même méthode pour accélérer les méthodes de recherche dans la modélisation de structures *ab initio*. En effet, la description des protéines par une tessellation de Voronoï se prête bien à l'étude par méthodes d'apprentissage et cette modélisation, bien que simplifiée, garde l'information importante sur l'empilement.

Enfin, ce type de méthode basé sur des tessellations de Voronoï et associé aux travaux sur les sommes de Minkowski (qui permettent d'autres types de pavages d'objets dans l'espace) pourrait être appliqué au recalage des assemblages de structures de protéines dans des enveloppes de densité électronique provenant d'expériences de microscopie électronique ou de diffraction des rayons X aux petits angles.

Deuxième partie

Un exemple d'étude structurale d'une protéine tétramérique : la thymidylate synthase X

Chapitre 5

Introduction

Sommaire

5.1 Résolution d'une structure de protéine par cristallographie

À l'heure actuelle, deux méthodes expérimentales sont principalement utilisées pour déterminer la structure tridimensionnelle des protéines : la résonance magnétique nucléaire (RMN) et la cristallographie aux rayons X. La cristallographie aux rayons X est une technique de diffraction dans laquelle l'image représente le réseau des atomes dans un cristal, permettant de déterminer la géométrie du réseau et des entités qui le composent. Ce sont les électrons, entourant chaque atome, qui interagissent avec les photons X dans cette technique et non les noyaux atomiques.

La première structure de protéine résolue fut celle la myoglobine, déterminée en 1957 par Max Perutz et John Cowdery Kendrew, ce qui, entre autres, leur valut le prix Nobel de Chimie en 1962. Depuis, plus de 18 000 structures de protéines ont été résolues ainsi et déposées dans la *Protein Data Bank*.

5.1.1 Du gène au cristal

Lorsqu'on dispose de la partie d'un gène codant pour une protéine (cadre ouvert de lecture ou *ORF Open Reading Frame*) que l'on souhaite étudier, de nombreuses étapes expérimentales sont nécessaires avant d'obtenir un cristal (figure 5.1). Il faut en effet réussir à produire la protéine en

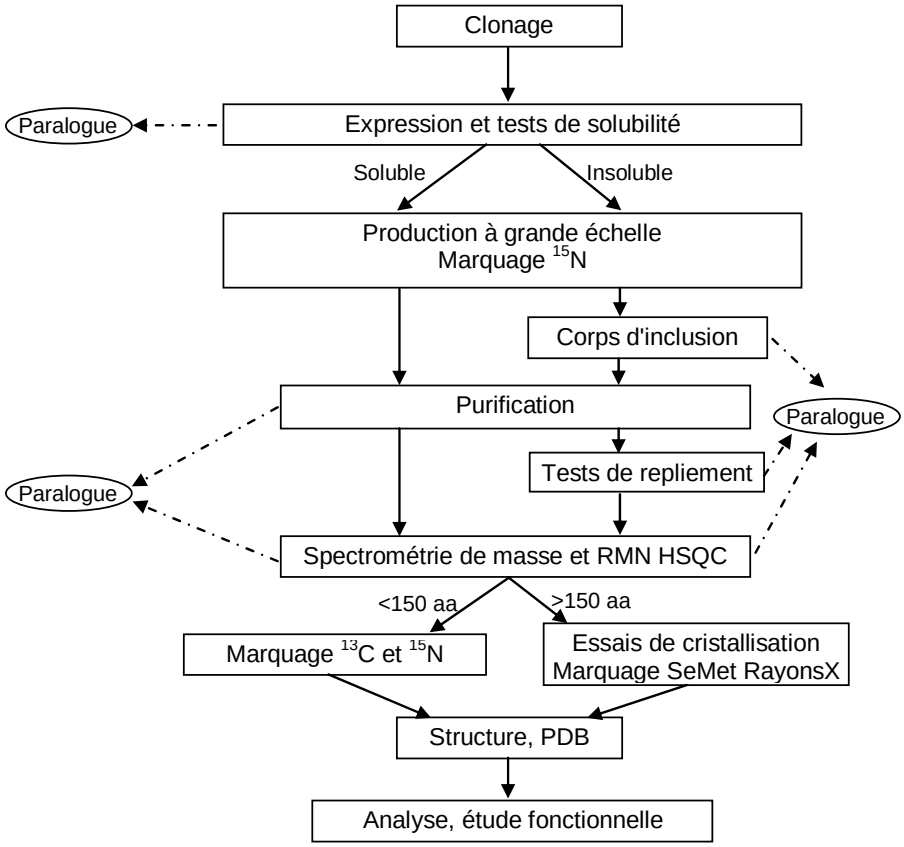


FIG. 5.1 – Du gène à la structure protéique : procédure expérimentale dans le projet de génomique structurale de la levure. *Après une étape de clonage, les protéines sont exprimées dans* E.coli. *Ensuite, le procédé dépend de la solubilité de la protéine mais les étapes principales restent la production à grande échelle puis la purification ainsi qu'une étape de spectrométrie de masse permettant de savoir si on dispose de la bonne protéine et une expérience de RMN (HSQC) pour savoir si celle-ci est repliée. Enfin, en fonction de la technique de résolution utilisée, on réalise un marquage, ou pour la cristallographie, on lance des essais de cristallisation. Il est à noter qu'à chaque étape, on peut utiliser un paralogue de la protéine d'intérêt si ce paralogue se prête mieux au procédé.*

quantité importante tout en ayant une solution pure. À chaque étape, il y a de nombreux risques d'échec. Par exemple, dans le projet pilote de génomique structurale de la levure, sur 288 cibles (gènes) sélectionnées, 259 ont pu être clonées, 215 exprimées, 131 solubilisées, 83 purifiées, 25 seulement cristallisées et 15 structures ont pu être résolues en deux ans. On peut voir dans cet exemple qu'avoir une solution de protéine pure est une condition nécessaire, mais pas suffisante. En effet, la cristallisation est une étape limitante de ce procédé. Comme il n'existe pas de moyen théorique de déterminer les conditions de cristallisation d'une protéine, de nombreux essais sont nécessaires : c'est une procédure longue et incertaine. De plus, une fois le cristal obtenu, sa qualité doit être telle qu'il doit diffracter et fournir une quantité d'informations suffisante pour pouvoir résoudre la structure à une résolution acceptable (inférieure à 3 Å).

Le cristal obtenu (figure 5.2) est ensuite cryogénisé (congelé rapidement) dans de l'azote liquide. Cette congélation, en plus de réduire l'agitation thermique à l'intérieur du cristal, permet

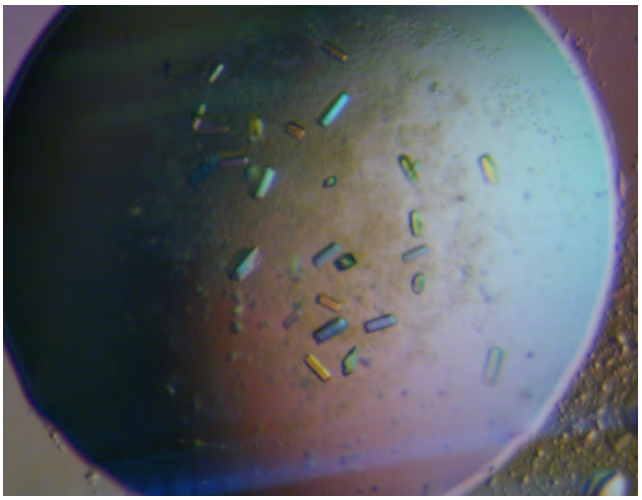


FIG. 5.2 – Photographie de cristaux de thymidylate synthase en lumière polarisée.

une meilleure résistance aux rayons X, de façon à pouvoir prendre une mesure sur un cristal non détérioré.

5.1.2 Diffraction

Le cristal est soumis à des rayons X, obtenus au laboratoire par un générateur à anode tournante, ou par un synchrotron. On obtient des images de diffraction (figure 5.3) constituées de tâches que l'on va combiner et utiliser pour construire une carte de densité électronique de la molécule qui a été cristallisée. Celle-ci est ensuite reconstruite à l'intérieur de la densité.

5.1.2.1 Mesure de l'intensité diffractée et problème des phases

Pour connaître le réseau cristallin et la structure de la protéine, on doit déterminer le signal caractérisé par une amplitude et une phase (voir paragraphe 6.3). À partir des images de diffraction, on dispose de l'intensité du signal. Cette intensité est proportionnelle au carré de l'amplitude du signal. On ne dispose cependant pas de la phase du signal qui est nécessaire à la résolution.

En cristallographie chimique, la phase φ peut être obtenue *ab initio* par des méthodes itératives, mais en raison de nombreux paramètres dont, en particulier, la complexité des molécules,

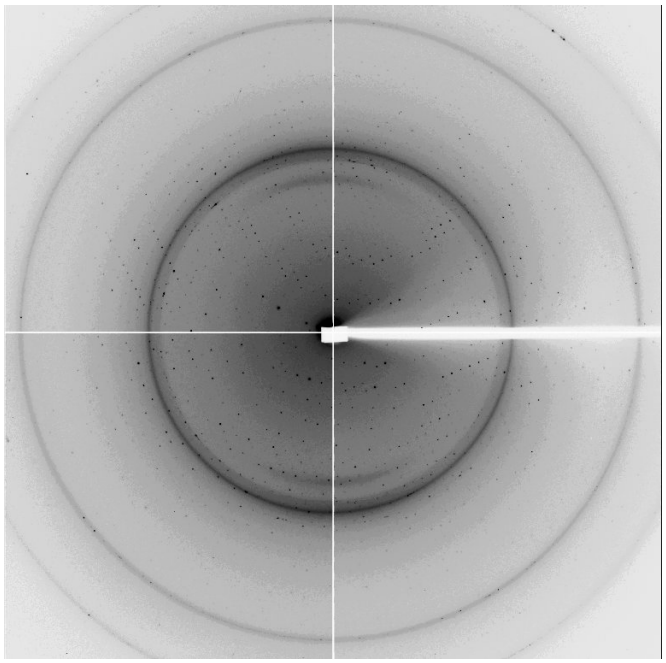


FIG. 5.3 – Image de diffraction d'un cristal de thymidylate synthase

cela n'est pas possible en biocristallographie. Ce problème est connu sous le nom de *problème des phases*.

5.1.2.2 Méthodes de détermination des phases

Pour résoudre ce problème, différentes méthodes ont été mises au point en biocristallographie. Les plus utilisées sont :

- le remplacement moléculaire (MR) ;
- le remplacement isomorphe multiple ou simple (MIR ou SIR) ;
- la dispersion anomale (MAD ou SAD).

Le remplacement moléculaire consiste à utiliser la structure d'une protéine homologue à celle étudiée, d'en calculer la diffraction théorique et d'utiliser les phases obtenues. En injectant ces phases dans le calcul, par méthode itérative, on peut parvenir à obtenir les phases correspondant à la protéine que l'on étudie. De nombreux programmes permettent de réaliser cette manipulation [119, 147, 162, 218]. C'est cette méthode que j'ai utilisée pour résoudre la structure de la thymidylate synthase X.

Pour le remplacement isomorphe multiple, on place des atomes lourds (Mercure, Platine, Uranium...) dans la protéine (par trempage par exemple) et l'image de diffraction du cristal correspondant permettra d'obtenir les phases. En effet, les atomes lourds ayant beaucoup plus d'électrons, leur signal sera plus important et contrasté, et on pourra en déterminer les phases [217]. On reviendra ensuite aux phases du signal correspondant à la protéine par des méthodes itératives.

Lorsque l'énergie du photon X est très proche du seuil d'absorption d'un atome, il y a résonance. Il y a donc des changements dans la phase et l'amplitude du faisceau diffusé, c'est la dispersion anomale. On utilise des atomes naturellement présents dans la protéine (Soufre) ou que l'on peut introduire dans les méthionines (Sélénium) car leur seuil d'absorption correspond

aux longueurs d'ondes obtenues sur les lignes synchrotron. Il suffit de mesurer le signal de diffraction à trois longueurs d'onde différentes bien choisies, plus ou moins proche du seuil d'absorption de l'atome, et les différences d'intensité vont permettre de déterminer les phases recherchées.

5.1.3 Reconstruction et affinement

Une fois le signal de diffraction analysé, on dispose d'une enveloppe qui correspond à la densité électronique de la protéine. Il s'agit alors d'effectuer ce qu'on appelle la reconstruction, à savoir replacer les atomes de la protéine dans la densité correspondante. On vérifie ensuite que la structure obtenue est en accord avec les propriétés physico-chimiques de la molécule en réalisant un affinement, c'est-à-dire une minimisation d'énergie en accord avec les valeurs expérimentales. Ces deux étapes sont réalisées de nombreuses fois l'une après l'autre de façon à pouvoir reconstruire le plus complètement possible la molécule.

L'accord de la structure obtenue avec les données expérimentales est mesuré par un paramètre statistique appelé R_{free} [27] (voir paragraphe 6.2.6).

5.2 La thymidylate synthase X : une cible antibactérienne potentielle

5.2.1 ADN et synthèse des pyrimidines

5.2.1.1 Structure de l'ADN

La structure en double hélice de l'ADN a été découverte en 1953 par James Watson et Francis Crick [225], ce qui leur a valu, avec Maurice Wilkins, d'obtenir le prix Nobel de Médecine en 1962.

La molécule d'ADN est formée de deux brins : chaque brin est constitué par un enchaînement de nucléotides (figure 5.4).

Chaque nucléotide est constitué d'un sucre, d'un phosphate et de l'une des cinq bases : Adénine (A), Thymine (T), Uracile (U), Cytosine (C) et Guanine (G). On trouve les quatre bases ATGC dans l'ADN et AUGC dans l'ARN [22]. On peut classer ces bases en deux groupes en raison de leur structure chimique : les purines (A et G) et les pyrimidines (T, U et C) (figure 5.5).

Dans la double hélice d'ADN, les deux brins s'associent sur la base de nombreuses interactions : l'effet hydrophobe et les liaisons π pour la superposition des bases, mais aussi, en ce qui concerne la spécificité des associations entre brins, à travers de nombreuses liaisons hydrogène. En effet, il existe des interactions de paires complémentaires A-T (ou A-U) et C-G (figure 5.4). Plus une molécule d'ADN contient de paires complémentaires, plus forte est l'interaction entre les deux brins.

5.2.1.2 Synthèse des pyrimidines : le cas particulier de la thymine

Il existe deux types de synthèse des pyrimidines, la synthèse *de novo* ou la synthèse à partir de précurseurs comme la thymidine. Dans le cas de la synthèse *de novo*, contrairement aux purines,

[22] L'uracile est rarement présent dans l'ADN sauf dans les cas de la dégradation de la cytosine, ou dans l'ADN de certains virus.

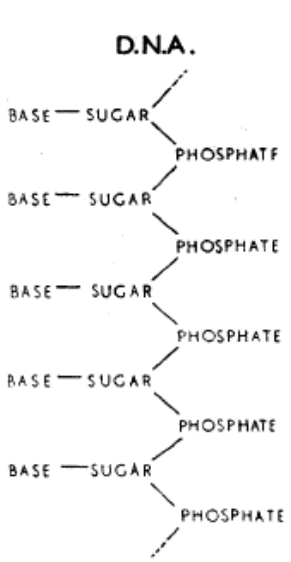


Fig. 1. Chemical formula of a single chain of deoxyribonucleic acid

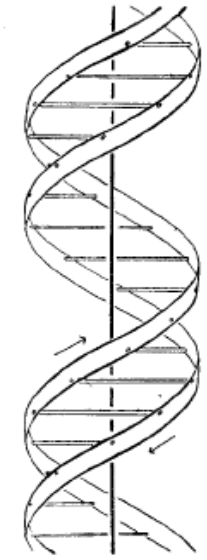


Fig. 2. This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

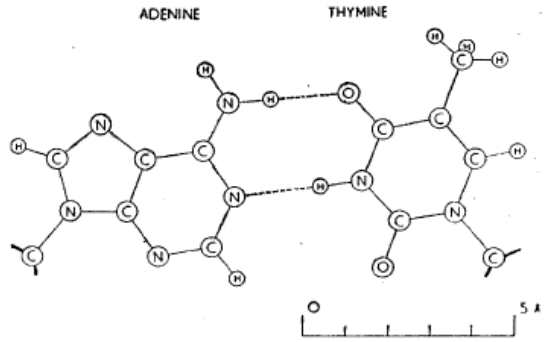


Fig. 4. Pairing of adenine and thymine. Hydrogen bonds are shown dotted. One carbon atom of each sugar is shown

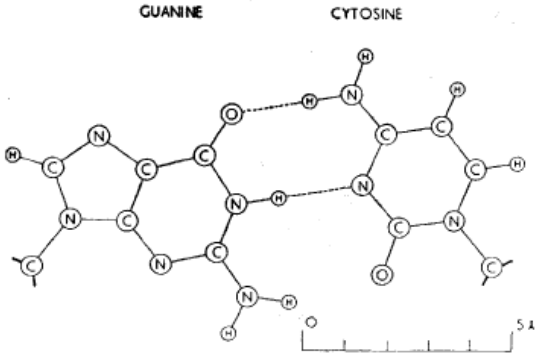


Fig. 5. Pairing of guanine and cytosine. Hydrogen bonds are shown dotted. One carbon atom of each sugar is shown

FIG. 5.4 – Figures originales de l'article de James Watson et Francis Crick, décrivant la structure en double hélice de la molécule d'ADN et les interactions entre paires complémentaires [224].

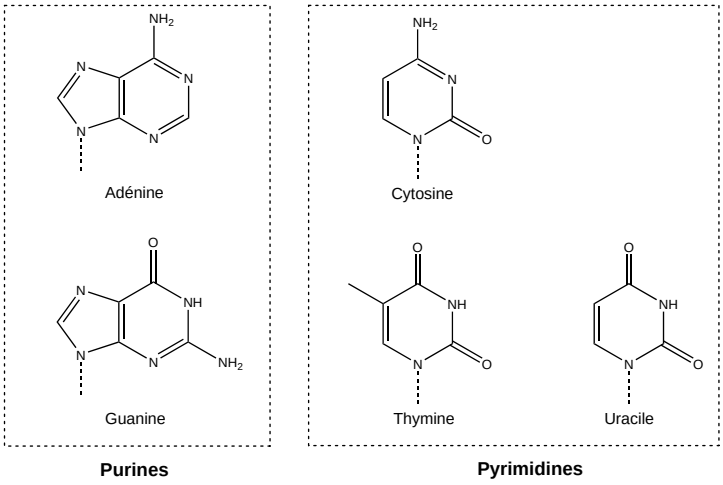


FIG. 5.5 – Les bases : purines et pyrimidines

les pyrimidines sont assemblées avant d'être attachées au 5-phosphoribosyl-1-pyrophosphate (PRPP).

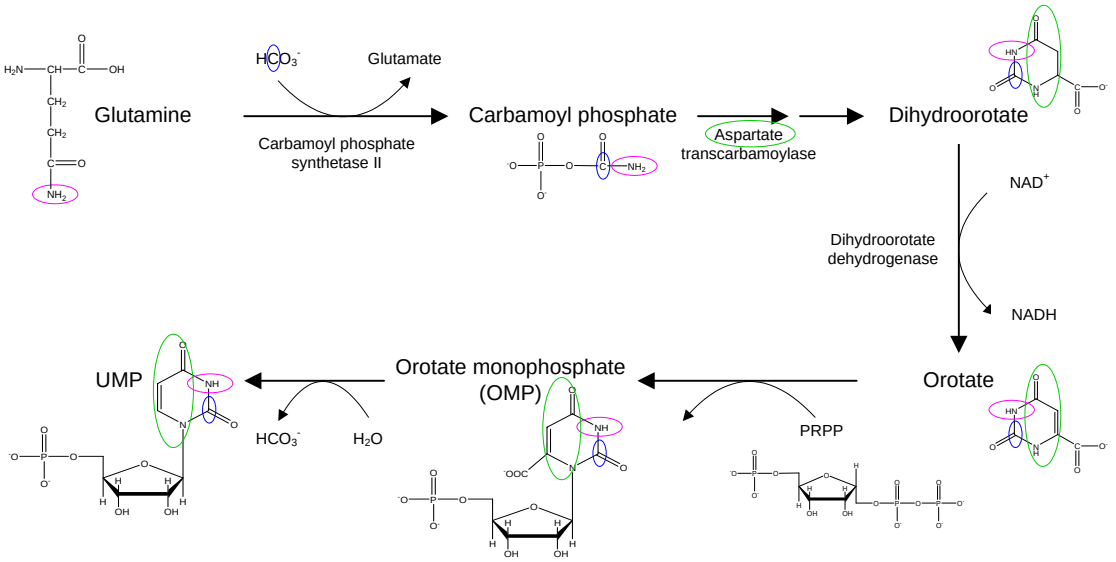


FIG. 5.6 – La synthèse des pyrimidines. *De façon simplifiée, la première étape est la formation de carbamoyl phosphate par la carbamoyl phosphate synthetase II. La deuxième étape importante est la création d'acide carbamoyl aspartique par l'aspartate transcarbamoylase. Ensuite, une fois l'acide orotique formé à partir de l'acide carbamoyl aspartique, il est combiné au PRPP pour former l'uridine-3'-phosphate (UMP). Le dUMP est ensuite phosphorylé 2 fois pour obtenir le dUTP.*

La thymidine, que l'on retrouve uniquement dans les molécules d'ADN, est synthétisée à partir d'uridine. En effet, le dTMP est formé à partir de dUMP par une enzyme qui réalise le transfert de méthyle. Chez les eucaryotes supérieurs, cette enzyme est la thymidylate synthase A, mais chez de nombreux autres organismes, dont de nombreux hyperthermophiles, le gène codant pour cette enzyme n'est pas présent.

C'est à partir de cette constatation que Hannu Myllykallio a recherché *in silico* une thymidylate synthase alternative qui a mené à la découverte de la thymidylate synthase X [126, 159]. C'est cette enzyme qui est responsable du transfert de méthyle chez les organismes ne possédant pas ThyA. C'est pourquoi, ThyX est une cible antibiotique de choix, son inhibition permettant d'empêcher la synthèse d'ADN chez la plupart des bactéries sans être toxique pour l'hôte qui lui, ne possède pas ThyX.

5.2.2 Un mécanisme controversé

5.2.2.1 Un problème d'oxydoréduction complexe

Tout comme ThyA, ThyX catalyse le transfert de méthyle du méthylènetétrahydrofolate vers le dUMP pour former le dTMP. Mais contrairement à ThyA, ThyX est une flavoprotéine, elle utilise le FAD qui joue un rôle d'oxydoréduction important pendant la réaction. ThyX nécessite aussi un autre substrat, en plus du méthylènetétrahydrofolate et du dUMP, le NAPDH (voir figure

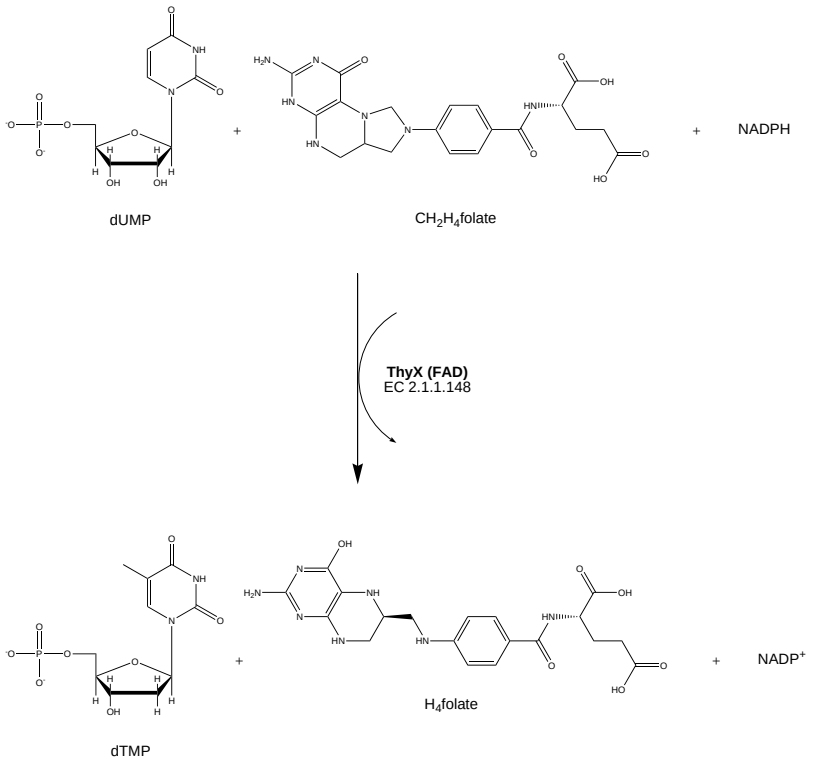


FIG. 5.7 – Réaction catalysée par la protéine ThyX

5.7). Durant la réaction, il y a donc : réduction du FAD, oxydation du NADPH, déprotonation du dUMP et trans-méthylation du CH_2H_4folate vers le dUMP.

5.2.2.2 Fixation simultanée ou successive ?

Deux mécanismes réactionnels ont été proposés pour cette enzyme. Le premier suggère un mécanisme sequentiel de type *ping-pong* dans lequel il y aurait formation d'une enzyme méthylée, intermédiaire de la réaction [89], mais le FdUMP ne semble pas former d'intermédiaire covalent avec les enzymes ThyX [125].

Nos collaborateurs Sébastien Graziani, Ursula Liebl et Hannu Myllykallio, proposent eux un mécanisme légèrement différent [88], en accord avec le modèle présenté par Agrawal et ses collaborateurs [2], où le complexe stable ThyX-dUMP joue un rôle primordial tout au long de la catalyse (figure 5.8).

Dans une optique de traitement antibiotique, cela peut-être important : en effet, les modes d'inhibition par des analogues du dUMP ou du CH_2H_4folate peuvent être différents et la synthèse d'un inhibiteur sélectif peut dépendre de la simultanéité de la fixation.

5.2.3 Vers une meilleure compréhension du mécanisme

Durant notre étude, deux autres structures de thymidylates synthases X (*T.maritima* et *M.tuberculosis*) ont été résolues. Nous nous sommes attachés à la résolution de la structure de la thymidylate synthase X du Chlorella Virus-1 de *Paramecium bursaria* en collaboration avec

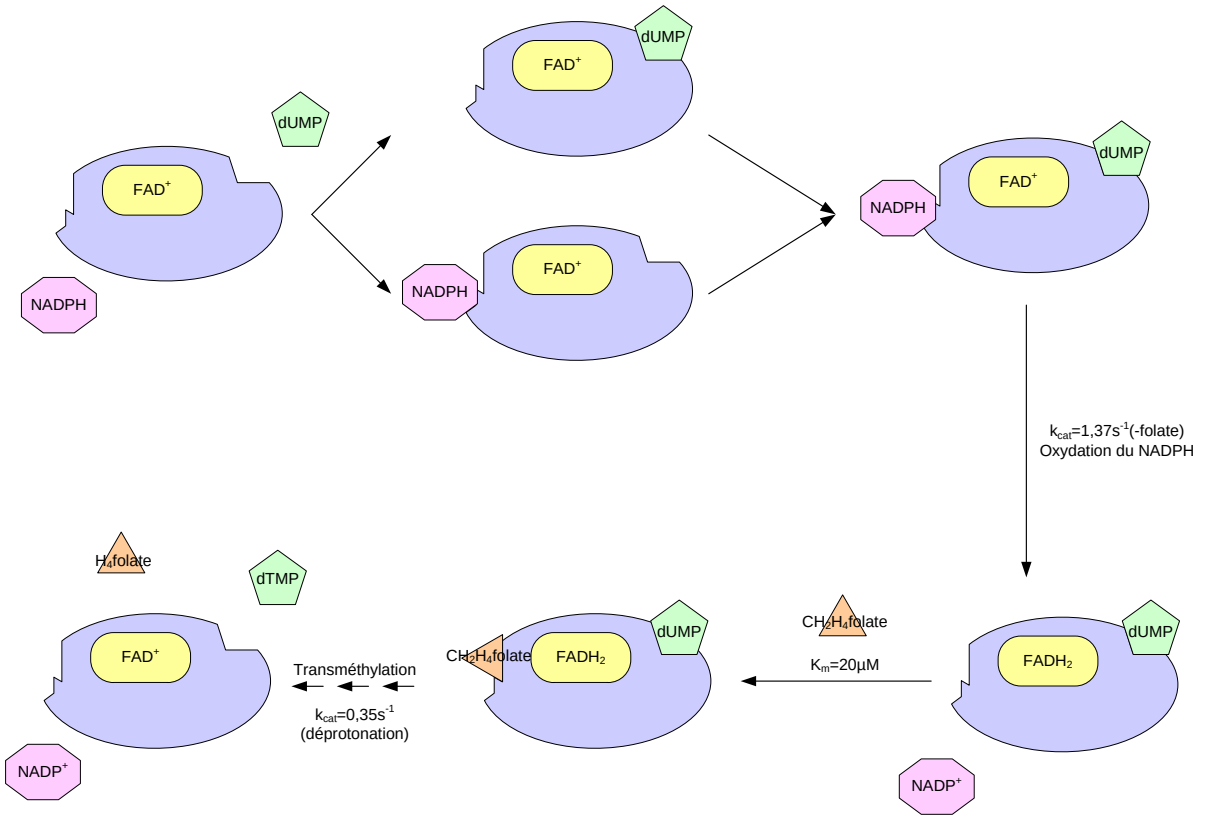


FIG. 5.8 – Schéma de la réaction catalysée par ThyX

Sébastien Graziani, Ursula Liebl et Hannu Myllykallio de l'Institut de Génétique et Microbiologie à Orsay, et James L. van Etten de l'Université du Nebraska, qui a découvert ce virus.

La résolution de la structure a été faite en lien très étroit avec une étude biochimique de cette protéine dont de nombreux mutants ont été réalisés par nos collaborateurs. Malheureusement, les tests de cristallisation et/ou diffraction avec différents inhibiteurs de ce complexe protéique ont échoué, limitant la portée de notre étude.

Chapitre 6

Étude structurale de la thymidylate synthase X du Chlorella Virus-1 de *Paramecium bursaria*

Sommaire

6.1 Introduction

À l'aide des données structurales et biochimiques, nous nous sommes intéressés dans ce travail au mécanisme catalytique d'une thymidylate synthase X virale : la thymydilate synthase X du Chlorella Virus-1 de *Paramecium bursaria*. L'objectif du travail présenté ici est de comprendre en détail le mécanisme enzymatique de cette protéine très active.

Pour cela, nous avons tout d'abord identifié biochimiquement plusieurs résidus impliqués dans la fixation du substrat et/ou dans la catalyse de ThyX. Ensuite, nous avons résolu la structure de cette protéine tétramérique qui nous a permis de comprendre d'un point de vue structural le rôle de ces résidus et nous a aidé dans l'interprétation des données fonctionnelles.

Nous observons que la conformation des résidus du site actif est différente de celles reportées dans les structures précédemment résolues, un réarrangement du site actif ayant certainement lieu pendant la catalyse. L'étude cinétique détaillée permet de mettre en évidence que la protéine ThyX étudiée utilise un mécanisme de réaction séquentiel et c'est la raison pour laquelle, nous pensons que le mécanisme catalytique implique la formation de complexes ternaires.

6.2 Article : *Catalytic mechanism and structure of a viral flavin-dependent thymidylate synthase ThyX*

CATALYTIC MECHANISM AND STRUCTURE OF VIRAL FLAVIN-DEPENDENT THYMIDYLATE SYNTHASE ThyX

Sébastien Graziani [1#], Julie Bernauer[3#], Stéphane Skouloubris[2], Marc Graille[3], Cong-Zhao Zhou[3§], Christophe Marchand[3], Paulette Decottignies[3], Herman van Tilbeurgh[3], Hannu Myllykallio[2], and Ursula Liebl[1]

[1]Laboratory of Optics and Biosciences, INSERM U696 - CNRS UMR 7645, Ecole polytechnique, 91128 Palaiseau, France ; [2]INSERM AVENIR group, Université Paris Sud, Institut de Génétique et Microbiologie –CNRS UMR 8621, 91405 Orsay, France ; [3]Institut de Biochimie et de Biophysique Moléculaire et Cellulaire (CNRS-UMR 8619), Université Paris Sud, 91405 Orsay, France ; [§] Present address: School of Life Science, University of Science and Technology of China, Hefei Anhui, 230027, PR China

Running Title: Viral Thymidylate Synthase ThyX

Address correspondence to: Ursula Liebl, Laboratory of Optics and Biosciences, Ecole polytechnique, 91128 Palaiseau, France, Tel +33169334740; Fax +33169333017; E-Mail: ursula.liebl@polytechnique.fr
Hannu Myllykallio, Institut de Génétique et Microbiologie, Université Paris Sud, 91405 Orsay, France, Tel +33169158170; Fax + 33169157808; E-Mail: hannu.myllykallio@igmors.u-psud.fr
[#]These authors contributed equally to this work.

Using biochemical and structural analyses we have investigated the catalytic mechanism of the recently discovered flavin-dependent thymidylate synthase ThyX from *Paramecium bursaria* chlorella virus-1 (PBCV-1). Site-directed mutagenesis experiments have identified several residues implicated in either NADPH oxidation or deprotonation activity of PBCV-1 ThyX. Chemical modification by DEPC and mass spectroscopic analyses identified a histidine residue (H53), crucial for NADPH oxidation and located in the vicinity of the redox active N5 atom of the FAD ring system. Moreover, we observed that the conformation of active site key residues of PBCV-1 ThyX differs from earlier reported ThyX structures, suggesting structural changes during catalysis. Steady-state kinetic analyses continue to support a sequential reaction mechanism where ThyX catalysis proceeds via formation of distinct ternary complexes without formation of a methyl-enzyme intermediate.

All cellular organisms need thymidylate (deoxythymidine 5'-monophosphate or dTMP) for the replication of their chromosomes, as dTMP is required for the biosynthesis of deoxythymidine 5'-triphosphate (dTTP), a building block of DNA. Cells can produce thymidylate either *de novo* from deoxyuridine 5'-monophosphate (dUMP) or incorporate thymidine using thymidine kinase (Tdk). The *de novo* pathway of dTMP synthesis requires a specific enzyme, thymidylate synthase, which methylates dUMP at position 5 of the pyrimidine ring. Two structurally and mechanistically distinct classes of thymidylate synthases exist. The well-studied ThyA proteins (EC 2.1.1.45) catalyse the reductive methylation reaction of dUMP, with methylenetetrahydrofolate (CH_2H_4folate) serving as one-carbon donor and as source of reductive power [reviewed in (1)].

On the other hand, the recently discovered ThyX (E.C. 2.1.1.148) family of thymidylate synthases contains flavin adenine dinucleotide (FAD) (2) that is tightly bound by a novel fold (3). FAD mediates hydride transfer from reduced nicotinamide (NADPH) during catalysis (4-6). Consequently, in the reaction catalysed by ThyX, CH_2H_4folate serves only as carbon donor, leading to the prediction that H_4folate (and not H_2folate as is the case for ThyA) is produced (2). This prediction has recently been confirmed by identifying H_4folate as a reaction product of *Chlamydia trachomatis* ThyX using high pressure liquid chromatography (7).

The catalytic reaction of thymidylate synthase ThyA is a sequential ordered

Abbreviations used: DEPC, diethylpyro-carbonate; MALDI-TOF MS, matrix assisted laser desorption ionization-time of flight mass spectrometry; r.m.s.d., root-mean-square distance

mechanism in which dUMP binding is followed by the entry of CH_2H_4folate and subsequent ternary complex formation with dUMP and CH_2H_4folate simultaneously bound to the enzyme (8,9). This was demonstrated by thorough steady-state kinetic measurements using varying concentrations of these two substrates of the ThyA reaction. Moreover, using F-dUMP in the reaction mixtures, this covalent ternary complex can readily be trapped for ThyA proteins (10). Although ThyX catalysis is of considerable interest for detecting and designing new anti-microbial compounds (2), our understanding of the reaction mechanism of this enzyme is still incomplete. Several models propose that the catalytic cascade of different ThyX proteins starts with oxidation of NADPH (4-6), a reaction step that does not occur during ThyA catalysis. We have proposed earlier that *Paramecium bursaria* chlorella virus-1 (PBCV-1) ThyX uses a sequential reaction mechanism with the formation of a ternary complex of CH_2H_4folate and dUMP bound to the enzyme (6). This proposal is compatible with a structural model where dUMP and CH_2H_4folate are simultaneously docked at the active site of ThyX from *Thermotoga maritima* (4). However, a ping-pong mechanism involving the formation of a methyl-enzyme as reaction intermediate has been proposed for *Chlamydia trachomatis* ThyX proteins (7). F-dUMP does not seem to form a covalent intermediate with ThyX enzymes (11), complicating analysis of the ThyX reaction. Consequently, it is currently unknown whether this discrepancy results from experimental differences or rather indicates that ThyX proteins from viral and cellular sources might use different reaction mechanisms.

The goal of this work is to obtain detailed insight into the enzymatic mechanism of the highly active PBCV-1 ThyX protein using a combination of biochemical and structural approaches. We therefore identified and further investigated several residues required for substrate binding and/or catalysis of PBCV-1 ThyX. Moreover, we report the crystal structure of the PBCV-1 ThyX tetramer that provides a structural basis for the interpretation of the obtained functional data. We observed that the conformation of active site key residues of PBCV-1 ThyX is different from earlier reported ThyX structures, suggesting that the active site undergoes structural changes during catalysis. As our detailed steady-state kinetic analyses further indicate that ThyX uses a sequential reaction mechanism for catalysis, we continue to favour a scenario where ThyX catalysis proceeds via formation of distinct ternary complexes.

EXPERIMENTAL PROCEDURES

Bacterial strains - The bacterial strains used in this study are *Escherichia coli* BL21 (*F⁻ ompT hsdSB* (r_B^- m_B^-) *gal dcm*; Novagen), and the thymidine-auxotroph DH5α [*ΔthyA*::Erm (12)]. *E. coli* strains were grown at 37°C in Luria Bertani or in M9 (Difco) minimal medium (3 g/L Na_2HPO_4, 1.5 g/L KH_2PO_4, 0.25 g/L NH_4Cl and 0.15 g/L NaCl) supplemented with 2 mM $MgSO_4$, 0.1 mM $CaCl_2$ and 0.3 % glycerol. One hundred µg/ml ampicillin and 1 mM IPTG (isopropyl-beta-D-thiogalactopyranoside) were added for plasmid maintenance or protein induction, respectively. Complementation tests were performed as described previously (11).

Molecular genetic techniques and construction of plasmids - The pVEX plasmid containing the *thyX* gene from PBCV-1 is a pGEX-2T derivative that has been described previously (6). Site-directed mutations were introduced using the Quik Change mutagenesis kit (Stratagene) with pVEX (T7tag-a674r6His) as template using the following primer couples (all sequences are indicated in 5' to 3' direction): (CCACAAGCATGGTCAATC) and (CATAGGCGGTGTTCGTAACC) for H53Q, (CCACAAGAAGTGGTCAATC) and (CATAGGCGGTGTTCTTCACC) for H53K (CGCCAGCGGAGCTTCCACTTC) and (CGAGTCCAAGAAGCGGTCGCC) for H79Q, (CGCAAGCGGAGCTTCCACTTC) and (GTCCAAGAAGCGTTCGCCTCG) for H79K, (CAGGCGTACGCATCTGTGATG) and (CGTACGCCTGGGAAAATTCCT) for R90A, (GGATCCAGTACATCGAACTGC) and (CCCTAACCTAGGTCATGTAGC) for H177Q, (GGATCAAGTACATCGAACTGC) and (CCCTAACCTAGTTCATGTAGC) for H177K, (CGAACTGG CGACTTCAAACGG) and (GCTTGACCGCTGAAGTTTGCC) for R182A. The E190G substitution was obtained as described earlier (6).

Protein expression and purification - The PBCV-1 wild-type and mutant ThyX proteins were expressed in either *E. coli* DH5α (Δ*thyA*)

or BL21 at 37°C in 800 ml LB medium containing 100 µg/ml ampicillin. Protein expression was induced by adding 1 mM IPTG to early exponential phase cultures (O.D.600 ~ 0.5) for 3 hours. His-tagged proteins were purified from cell-free extracts by gravity-flow chromatography on Ni-NTA agarose (Qiagen) and gel filtration using an S-200 column (Amersham). Eluted proteins were stored at -80°C in 50 mM N-(2 hydroxethyl)piperazine-N'-2-ethanesulfonic acid (HEPES), pH 7.0, supplemented with 10% glycerol. Protein samples were analyzed on 12.5% SDS PAGE and were more than 95% pure. Control experiments with non-complementing mutants established that *E. coli* thymidylate synthase ThyA does not bind to the column under these conditions.

ThyX activity measurements - Tritium release assays for measuring PBCV-1 ThyX activity *in vitro* were performed essentially as described earlier for the *Helicobacter pylori* ThyX protein (11) with 37.5 pmoles of enzyme in 25 µL reaction mixture. Reactions were terminated after 3.5 min incubation. Kinetic parameters were determined using non-linear regression using the software package GRAPHPAD. Typical reactions contained 10 mM $MgCl_2$, 0.5 mM NADPH, 60 µM FAD, 500µM CH_2H_4folate, 12.5µM dUMP, 10% glycerol in 50 mM HEPES, pH 7.5. The specific activity of the [5-^{3}H]-dUMP stock (Moravek) used in the experiments was 13.6 Ci/mmol. Activities of mutant proteins analysed were compared with those of wild-type protein obtained in parallel experiments.

In double titration experiments, the concentrations of dUMP and CH_2H_4folate were varied from 3 to 50 µM and from 3 to 100 µM, respectively; CH_2H_4folate and NADPH were varied from 3 to 100 µM and from 100 to 300 µM respectively, and the concentrations of dUMP and NADPH were varied from 3 to 100 µM and from 6.25 to 200 µM respectively. Where indicated, deprotonation assays were also performed using enzyme treated with diethyl-pyrocarbonate (see below).

NADPH oxidation activity of ThyX proteins was measured using a total volume of 100 µl. Reaction components were 50 µM dUMP, 10% glycerol, 1 mM $MgCl_2$, 400 µM NADPH, 10 µM FAD. Control experiments established that the addition of 50 µM of CH_2H_4folate substantially inhibited NADHP oxidation activity. Activity was monitored by net *decrease* of absorbance at 340 nm using a CARY 50 spectrophotometer (Varian). An extinction coefficient of 6400 cm^{-1} at 340 nm (ε_{340}) was used to quantify absorption changes. Note that this assay is different from the spectrophotometric assay established for ThyA proteins that catalyze the oxidation of H_4folate to H_2folate resulting in net *increase* in absorption at 340 nm.

Chemical modification of PBCV-1 ThyX with diethyl pyrocarbonate (DEPC) and reversal of reaction with hydroxylamine - DEPC has been used to analyze the functional role of histidine residues in a number of proteins, although under certain conditions non-specific reactions with serine and threonine residues can occur. In the presence of DEPC, histidine residues yield an *N*-carbethoxy-histidyl derivative that is reversible upon addition of hydroxylamine (NH_2OH). In the experiments shown, DEPC was freshly diluted with absolute ethanol before each use. Its concentration was determined by reaction with imidazole as described (13). Modification of PBCV-1 ThyX was performed at 25°C for 20 min in 980 µL final volume containing 15 µM PBCV-1 ThyX wild type protein (15 nmoles), 50 mM potassium phosphate buffer pH 7.0 and 250 µM DEPC (the final concentration of ethanol never exceeded 3% (v/v)). A control experiment was performed under the same conditions without DEPC. 54 µL of 1 M hydroxylamine hydrochloride (adjusted to pH 7.0) were added to 490 µL of DEPC-treated PBCV-1 ThyX WT and control solutions, and the reaction was carried out at 25°C for 20 min. DEPC and hydroxylamine were removed by size exclusion chromatography using a 5 mL Sephadex G-25 desalting column (Amersham Bioscience) equilibrated in 50 mM potassium phosphate buffer pH 7.0. Samples were routinely concentrated to 80 µL in a Microcon YM10 concentrator (Amicon) before proteolytic digestion and mass spectrometry.

Proteolytic digestion and MALDI-TOF mass spectrometry - 1 µL of trypsin (1mg/mL) was added to 15 µL (~1 nmole) of concentrated control protein and DEPC-inactivated ThyX before and after hydroxylamine treatment. The digestion was carried out for 5 h at 37°C in 50 mM potassium phosphate buffer (pH 7.0) in a

final volume of 25 µL. Peptide mass fingerprints were recorded in reflector positive-ion mode (accelerating voltage 20 kV, grid voltage 73%, guide wire 0.002%, delay 200 ns) on a Voyager DE-STR MALDI-TOF mass spectrometer (PerSeptive-Applied Biosystems) equipped with a 337 nm nitrogen laser using a close external calibration covering the range 750-4000 Da. 1 µL of peptide solutions was mixed with 3 µL of 50% acetonitrile, 0.3% trifluoroacetic acid and 6 µL of saturated solution of µ-cyano-4-hydroxycinnamic acid in 30% acetonitrile, 0.3% trifluoroacetic acid. 1.3 µL of this premix was then deposited onto the sample plate and allowed to dry at room temperature.

Chemicals - CH_2H_4folate was a generous gift of Dr. Moser (Eprova). FAD, dUMP, DEPC and hydroxylamine were purchased from Sigma.

Crystallization and structure determination - Protein samples were stored in 0.1 M Tris-HCl (pH 8.5). Protein crystals were grown at 18°C from a 1:1 µL mixture of 10 mg/ml protein solution with 10% PEG 400, 0.1 M $MgCl_2$, 18% isopropanol and 5% PEG MME 550. Crystals were flash-cooled and diffraction data were collected at 100K on the ID14-EH2 beam line (ESRF, Grenoble, France). These data were processed using XDS package (14). The crystals belong to the $P2_12_12$ space group with predicted two molecules per asymmetric unit and a solvent content of 51%. Molecular replacement was done using the program Molrep (15) and coordinates of the previously solved crystal structure of TSCP (TM0449) from *Thermotoga maritima* (PDB code 1kq4, (3)). In the resulting model, the two monomers of ThyX had an R_{factor} of 58.8% for data in the 20 Å to 4 Å range. However, by applying symmetry operations, the biologically active tetramer previously observed for TM0449 is reconstituted (3), indicating that this model is correct. This was further confirmed by automated refinement and rebuilding of the model using the program ARP/wARP (16), which led to automatic construction of 83% of the model and significant improvement of the R_{free}. The structure was refined using the program Refmac (17) and manually rebuilt with Turbo Frodo (18). As some regions were found to adopt different conformations, NCS restraints were not used during refinement. The final model contains residues 1-35, 39-89, 124-216 for chain A and 1-88, 125-215 for chain B. In addition, two FAD and 247 water molecules have been built. All these residues are well defined in the 2Fo-Fc electron density map and fall within the allowed region of the Ramachandran plot, as defined by Procheck (19). Statistics for data collection and refinement are reported in supplemental table 1.

RESULTS

Structure of PBCV-1 ThyX - The crystal structure of the PBCV-1 ThyX protein complexed to its FAD cofactor was solved by molecular replacement using the *Thermotoga maritima* TM0449 structure (hereafter TmThyX) as starting model and refined to 2.3 Å resolution (data acquisition and refinement statistics are presented in supplemental table 1). As shown in Figure 1A, the PBCV-1 ThyX monomer adopts the same hammerhead shark shaped structure as Tm and *Mycobacterium tuberculosis* (Mtb) ThyX with approximate dimensions 30×35×70 $Å^3$ (3,20). The monomer is made of a central α/β domain composed of a curved four-stranded anti-parallel β-sheet (β1, β2, β4, and β3) and six helices (α1, α2, α3, α6, α7 and α8) packed against the same face of the sheet. Two additional long α helices (α4 and α5) form a distinct domain on top of the core. The r.m.s.d. value between PBCV-1 and Tm or Mtb ThyX monomers is 1.5 Å or 1.95 Å, respectively (values calculated for 160-170 Cα atoms). The differences between these structures reside mainly in small variations in the orientation of the helices that are not in direct contact with the β–sheet of the central core domain (α1, α2, α8, α4 and α5).

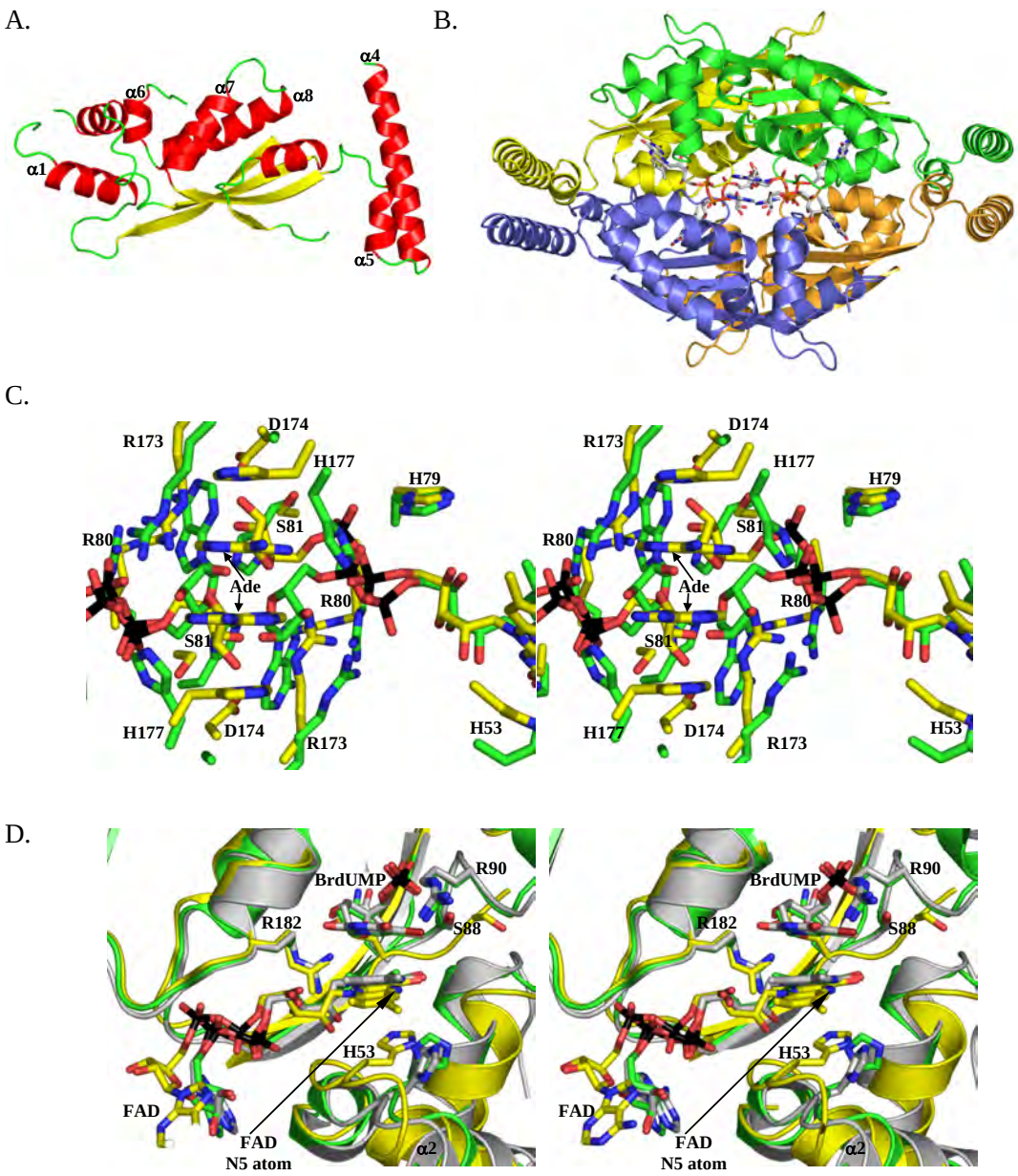


Figure 1: PBCV-1 ThyX structure.
A. Ribbon representation of the PBCV-1 ThyX monomer.
B. Ribbon representation of the ThyX homotetramer, each monomer is coloured differently. The four bound FAD molecules are shown as sticks. The monomer highlighted in green is related to the representation in Fig. 1A by a 90° rotation along the x axis.
C. Stereoview representation of the comparison of the ThyX FAD binding mode. PBCV-1 and Mtb ThyX are coloured yellow and green, respectively. For clarity, only PBCV-1 ThyX numbering is indicated. As Tm ThyX FAD binding mode is closely similar to that of Mtb ThyX, it has been omitted. Phosphorous atoms from FAD are coloured black.
D. Stereoview representation of the superposition of the active sites from PBCV-1 (yellow), Mtb (green) and Tm ThyX (grey). For clarity, only PBCV-1 ThyX numbering is indicated. Phosphorous atoms from FAD are coloured black.

Although, only two PBCV-1 ThyX monomers (chains A and B) are present in the asymmetric unit (r.m.s.d. between these two chains is 0.27 Å for all main chain atoms), a homotetramer similar to the one previously described for Tm and Mtb ThyX can be obtained by applying the space group symmetry operations (Fig. 1B, r.m.s.d. between the tetrameric forms of Tm and PBCV-1 ThyX is 3.81 Å for main chain atoms only) (3,20). The homotetramer has approximate dimensions of 50×60×85 $Å^3$. As previously observed, the tetramer is mostly formed by stacking of helices from the core domains as well as by pairwise interaction of the long helices α4 and α5 that are detached from the core domain.

FAD binding mode - The purified PBCV-1 ThyX protein is characterized by a yellow colour, indicating that the oxidized form of the flavine cofactor remains tightly bound to this viral protein during all the purification steps (3,20). Similarly to other structures of ThyX proteins, the FAD cofactors lie in large clefts at the interface between the four monomers and adopt an elongated conformation. The FAD molecules are deeply buried since only 15% from each FAD molecule surface remains solvent accessible. This accessible region corresponds to the isoalloxazine (flavin) ring of the FMN moiety where the redox chemistry takes place. On the opposite side of the FAD cofactor, the ribityl and AMP parts are strongly fixed onto the protein and fully buried. Surprisingly, superposition of the *Thermotoga martima* (Tm) and *Mycobacterium tuberculosis* (Mtb) structures of flavin - dependent thymidylate synthase ThyX onto the PBCV-1 ThyX-FAD complex shows that the FAD adenosine ring adopts a different conformation in the PBCV-1 enzyme (Figs. 1C, D). In PBCV-1 ThyX, two adenosine moieties bound to two distinct monomers are stacked together and sandwiched in between two histidine rings (H177 from chain A and B'), thus forming a four layered ring stack. In Tm and Mtb ThyX, the corresponding adenosines point away from each other and lie on the side chain from the residue directly following the RHR signature: H98 (Mtb ThyX) and I81 (Tm ThyX). In PBCV-1 ThyX, the pocket that is equivalent to the Mtb and Tm ThyX AMP binding site is blocked by the side chains from H83 from monomer A and E58, T171, R173 and D174 from monomer B. The phosphoribityl binding mode exhibits a high degree of similarity with previously described structures and hence will not be described here (see supplementary table 1). The tricyclic isoalloxazine carries the reactive moiety of FAD. In the absence of bound dUMP, the isoalloxazine ring in PBCV-1 ThyX is solvent accessible on its *si*-face while the *re*-face packs onto the H53 side chain, involving this residue in hydrophobic packing interactions. In the Mtb and Tm ThyX structures, the ring of the corresponding histidine and the isoalloxazine ring are not stacked but bound in a perpendicular direction at the edge of this ring. Comparison of the residues homologous to H53 in these three enzymes shows that in PBCV-1 ThyX, helix α2, that precedes the loop bearing H53, has slipped along the helical axis by about 2.7 Å, resulting in a more buried position for H53 (Fig. 1D).

In addition, the isoalloxazine pyrimidine ring makes six hydrogen bonds with residues originating from three different monomers (A, A' and B'). The interaction of the $N_{\eta}1$ atom from the invariant R78 residue, (corresponding to the first arginine residue of the RHR sequence motif, characteristic of ThyX proteins), with the O_2 atom of the FAD molecule is of interest from a catalytic point of view. The presence of a largely conserved positively charged residue near this part of the FAD was suggested to be functionally relevant either in modulating the redox potential of the co-factor or in stabilising the anionic form of the reduced flavin (21). The five remaining hydrogen bonds are made by the S55 hydroxyl group (monomer B') with FAD N_1, R78 (monomer A) and the E86 amide group (monomer A') with FAD O_2 as well as the Q85 $O\varepsilon_2$ and the E86 carbonyl oxygen (both from monomer A') with FAD N_3.

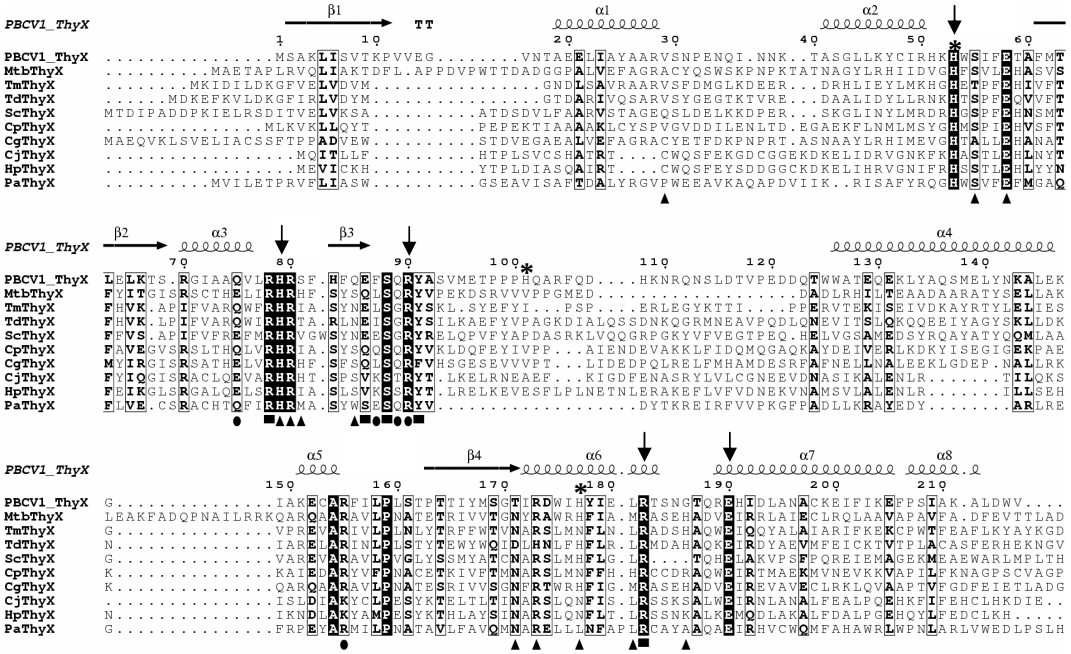


Figure 2: Structure-based sequence alignment of ThyX proteins. PBCV-1 ThyX secondary structure elements are indicated above the sequence. ThyX positions involved in *T. maritima* ThyX structure, in FAD or dUMP binding or both are indicated under the sequence, by triangles, circles or squares, respectively. Arrows indicate residues investigated in this work using site-directed mutagenesis. Asterisks refer to PBCV-1 ThyX histidine residues modified by DEPC. Strictly conserved residues are in white on a black background. Partially conserved amino acids are boxed. This figure was generated using the ESPript program (25). PBCV-1: *Paramecium bursaria* chlorella virus-1; Mtb: *M. tuberculosis*; Tm: *T. maritima*; Td: *Treponema denticola*; Sc: *Streptomyces coelicolor*; Cp: *Clostridium perfringens*; Cg: *Corynebacterium glutamicum*; Cj: *Campylobacter jejuni*; Hp: *H. pylori*; Pa: *Pyrobaculum aerophilum.*

Identification of functionally important ThyX residues using a structure-based sequence alignment - In order to identify functionally important ThyX residues, we first performed a structure-based sequence alignment of a diverse set of ThyX proteins (Fig. 2). A number of conserved ThyX residues have been previously characterized using site-directed mutagenesis approaches (7,11,20). For instance, in *Helicobacter pylori* ThyX, mutation of the histidine residue equivalent to PBCV-1 ThyX H53 results in no detectable activity although the protein is folded, as it is still able to bind FAD (11). In parallel, S88 (all residue numbering used hereafter refers to PBCV-1 ThyX) has been proposed to act as the nucleophile in the catalytic reaction, although in none of the reported ThyX structures this residue is optimally located for a nucleophilic attack to occur at position 6 of dUMP (3,20).

Amino acid residues involved in the binding of FAD or dUMP in ThyX have been identified by crystallographic analyses of ternary complexes (indicated by circles and squares in Fig. 2, (3,20) but little is known about the residues implicated in NADPH or CH_2H_4folate binding.

In this work we analyzed the role of several conserved residues in ThyX catalysis: H53, H79, R90, H177, R182 and E190 (indicated by arrows in Fig. 2). Except for H53, the role of these residues in ThyX catalysis and/or substrate binding has not been investigated in detail prior to this study.

Mutational analysis of conserved residues of PBCV-1 ThyX - Several site-directed mutants of the above residues were constructed and their roles in ThyX catalysis were studied

Substitution	Complementation	Oxidation activity	Deprotonation activity kcat [min^{-1}] / (kcat/Km) [$min^{-1}\mu M^{-1}$]		
			dUMP	CH_2H_4folate	NADPH
Wild type	**++++**	**100%**	**15.2 / (0.43)**	**2.6 / (0.11)**	**2.3 / (0.09)**
H53Q	**- (protein insoluble)**	**Nd**	-	-	-
H53K	**- (protein insoluble)**	**Nd**	-	-	-
H79Q	**++**	**94%**	**8.4 / (0.19)**	**1.47 / (0.05)**	**3.2 / (0.09)**
H79K	-	**Nd**	-	-	-
R90A	-	**56%**	-	-	-
H177Q	**++**	**31%**	**4.6 / (0.09)**	**1.15 / (0.05)**	**4.3 / (0.045)**
H177K	-	**9.5%**	-	-	-
R182A	-	**20%**	-	-	-
E190G	-	-	-	-	-

Table 1: Biochemical analyses of the PBCV-1 ThyX mutant proteins. For the oxidation test 100% corresponds to the level of activity measured for wild type protein in the presence of 200 µM NADPH, 1 mM $MgCl_2$, 10% glycerol, 50 µM dUMP, 10 µM FAD together with 0.17 µM enzyme. CH_2H_4folate was found to inhibit oxidation activity and was not included in reaction mixtures for oxidation tests. Deprotonation activities were measured as described in Materials and Methods. Nd, Not determined; -, Not detected (less than 3% of the value observed for the wildtype protein).

(Table 1). Complementation tests indicated that, with the exception of the H79Q and H177Q substitutions, all mutants analyzed did not confer thymidine-independent growth to an *E. coli* strain lacking thymidylate synthase ThyA, thus emphasizing their importance for ThyX activity. The H53Q and H53K mutant strains did not produce soluble protein for biochemical analyses, and the lack of complementation could simply result from improper folding of the protein. All other mutant proteins produced comparable amounts of soluble protein after induction (data not shown). Therefore, the lack of functional complementation suggests that these residues play important roles in ThyX catalysis and/or substrate binding.

To investigate the biochemical basis for the observed loss of complementation activity, the soluble PBCV-1 mutant proteins were expressed and purified. As expected, the complementing mutant proteins, H79Q and H177Q, were still active in NADPH oxidation and deprotonation assays (Table 1). However, both purified histidine mutants were instable and their k_{cat}/K_m values changed simultaneously for all three substrates. This could be explained by the role played by both residues in FAD binding. First, the H79 side chain packs against the FAD ribityl part and is also hydrogen bonded to the phosphate group of the AMP moiety (Fig. 1C). Second, as mentioned above, the H177 side chain is involved in binding of the FAD adenine base in a four-layer stacking. The R90A mutant has lost 44% of its oxidation activity and shows no measurable deprotonation activity. On the basis of the *Thermotoga maritima* ThyX structural data, the region containing R90 participates in dUMP binding (3). Together with our earlier demonstration that dUMP binding is necessary for NADPH oxidation by PBCV-1 ThyX (see supplemental figure 1) this provides a feasible explanation for its decreased oxidation activity. The E190G substitution was capable of binding FAD at wild type level, but nevertheless lacked detectable oxidation and deprotonation. On the

other hand, the R182A mutant did not copurify with oxidized FAD, but was able to oxidize NADPH in the presence of 400 µM FAD. For the corresponding residue in Mtb ThyX a role in substrate positioning has been proposed to enable hydride transfer during ThyX catalysis (20). We have shown earlier qualitatively that CH_2H_4folate inhibits NADPH oxidation activity (6). For the R182A mutant protein, we now determined an IC50 value of 48 µM for CH_2H_4folate whereas the analogous value for the wildtype was 16 µM (Fig. 3), indicating that the R182A mutant may affect CH_2H_4folate binding. These results suggest for the first time participation of R182 in NADPH and/or CH_2H_4folate binding.

ThyX activity is inhibited by DEPC treatment - As we failed to purify the H53Q and H53K mutant proteins in soluble form, we used chemical modification by diethylpyrocarbonate (DEPC) to further investigate the proposed catalytic role(s) of histidine residues (11). For these experiments, the deprotonation activity of PBCV-1 ThyX wild type protein was measured before and after DEPC treatment. The DEPC treatment decreased ThyX deprotonation activity at least 20-fold and was partially reversible with hydroxylamine (NH_2OH) treatment (see below), indicating that the DEPC-modified histidine residues are implicated in the ThyX catalytic mechanism. More detailed titration experiments with the DEPC-treated enzyme in the presence of varied substrates were performed to further investigate the effect of the chemical modification (Fig. 4A-C). We found that when dUMP or CH_2H_4folate were titrated with the chemically modified enzyme some activity was still detectable, whereas when NADPH titration was performed, no activity was detected. A possibility to explain this observation is that NADPH, when added after CH_2H_4folate, binds poorly to the chemically modified enzyme. Due to interfering absorbance at 340 nm, oxidation tests could not be performed with DEPC-treated ThyX, but fluorescence detection measurements at 460 nm (after 340 nm excitation) in the presence of 100 µM DEPC did not reveal oxidation activity (data not shown).

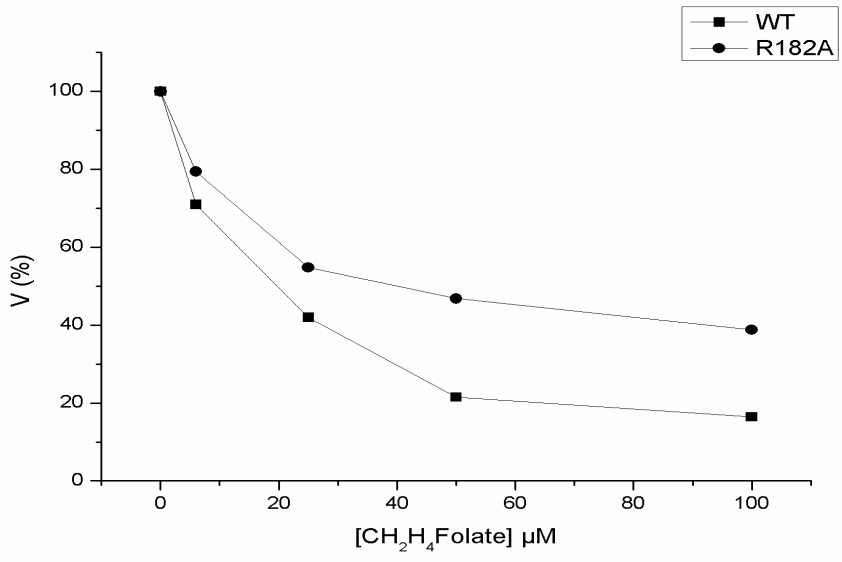


Figure 3: IC 50 values for CH_2H_4folate, determined for WT PBCV-1 ThyX and the R182A mutant.

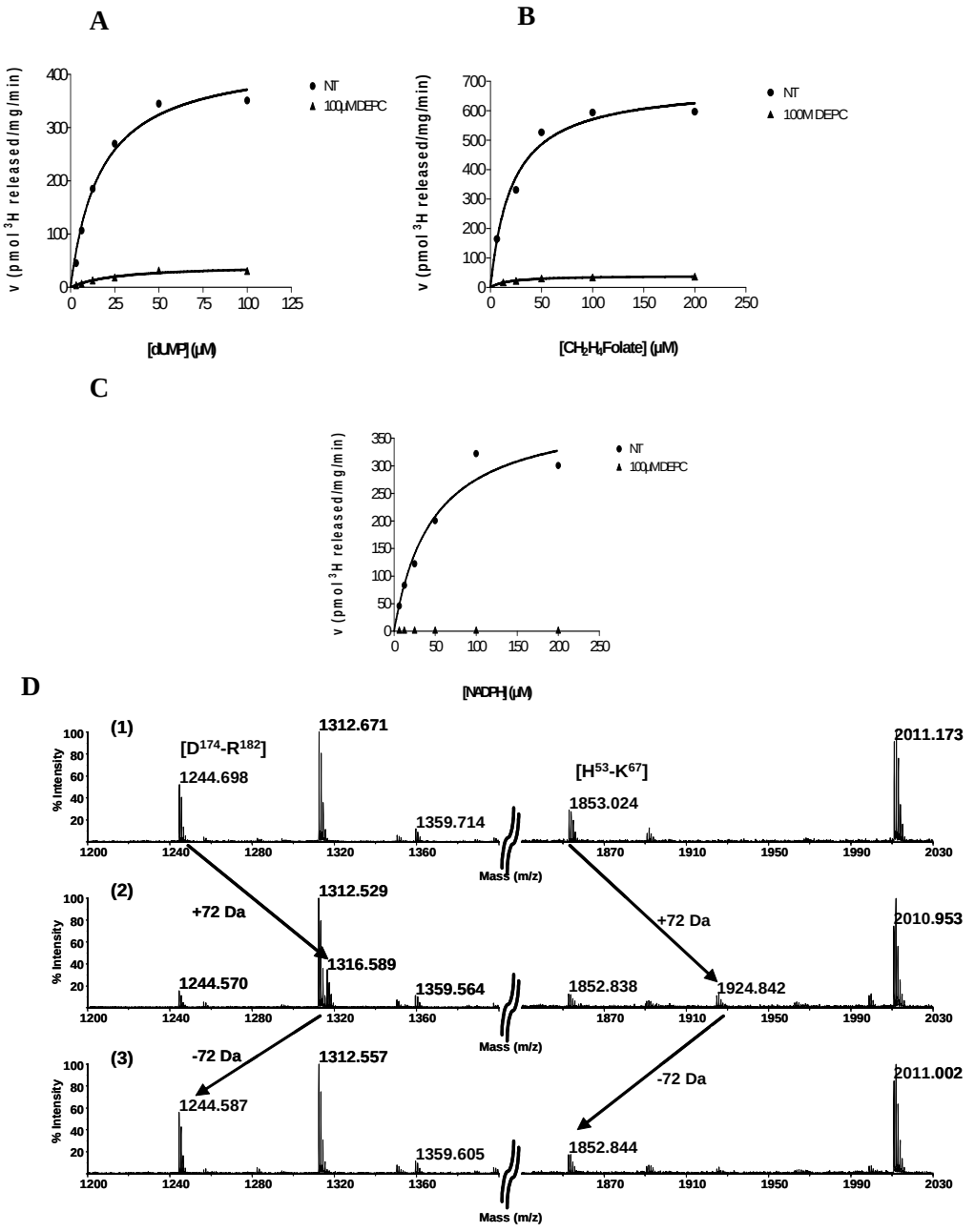


Figure 4: Saturation kinetics of PBCV-1 ThyX with and without DEPC treatment. Kinetics was determined for dUMP (**A**), CH_2H_4folate (**B**) and NADPH (**C**) in the presence and absence of 100µM DEPC. Enzyme activity was determined at various concentrations of the indicated substrates, with all other substrates at fixed saturating concentrations as described in "Experimental Procedures". **Panel D** corresponds to the partial MALDI mass spectra (after trypsic digestion) of *untreated* ThyX (spectrum 1), of *ThyX after DEPC treatment* (spectrum 2) and of *ThyX after DEPC treatment followed by NH_2OH reversal* (spectrum 3).

MALDI-TOF mass spectrometry after tryptic cleavage was performed in order to identify the modified residue(s). As this derivatization is quite unstable, depending on the conditions (optimum pH 6.0-7.0) (22), all experiments were performed in phosphate buffer pH 7.0 as quickly as possible, although these conditions were not optimal for trypsin proteolysis and mass spectrometry (no desalting prior to data acquisition). Peptide mass fingerprints clearly showed that three peptides, namely [D174-R182], [H53-K67] (Fig. 4D) and [Y94-R104] (data not shown), exhibited a 72 Da mass increase after DEPC treatment, in agreement with the derivatization of a histidine residue into carbethoxyhistidine. All of the "fingerprints" contain only one histidine residue, respectively in positions 53, 101 and 177 (modified histidines are indicated in Fig. 2). DEPC has been reported to occasionally react with other nucleophilic groups (principally hydroxyl groups, e. g. tyrosine). It has been reported that the effect of DEPC on histidine and tyrosine can be reversed by mild treatment with hydroxylamine, but the reversion is much more rapid in the case of the DEPC-derivatized histidine (13). Thus, in order to unambiguously assign the mass shift to the specific derivatization of His residues, we have investigated the effect of hydroxylamine on DEPC-treated PBCV-1 ThyX. Figure 4D (spectrum 3) shows that the tryptic profile of DEPC-treated PBCV-1 ThyX rapidly returns to the pattern of the control sample after treatment with hydroxylamine. In addition, the mass spectrum of the ThyX protein treated with only hydroxylamine was identical to the one of a control sample (data not shown), indicating that no modification artefact was induced. We can thus conclude that three histidine residues have been modified: H53, H101 and H177. Of these residues, H101 is not conserved (Fig. 2) and the H177Q substituted protein is still functional (Table 1). Therefore, our chemical modification experiments provide further support for a crucial role of H53 in ThyX oxidation activity.

Two-substrate kinetics - To further investigate the kinetic mechanism of PBCV-1 ThyX, we systematically measured deprotonation activity for all possible combinations of the three ThyX substrates, NADPH, dUMP and CH_2H_4folate, at several fixed concentrations of one substrate while varying the concentration of the second substrate (Fig. 5A-C). When plotted in double reciprocal plots, these data allow to distinguish between sequential and ping-pong reaction mechanisms. For all substrate couples, our data unambiguously demonstrated a set of several converging lines, indicative of a sequential kinetic mechanism, different from what has been reported for *C. trachomatis* ThyX (7). These findings are compatible with our earlier proposal indicating that ThyX catalysis proceeds via formation of ternary complexes of at least two substrates bound to the enzyme at the same time. Considering that CH_2H_4folate inhibits NADPH binding/oxidation in a competitive way, the binding sites for these two substrates are expected to overlap. Consequently, formation of a quaternary complex where all three substrates are simultaneously bound to the enzyme seems unlikely.

Figure 5

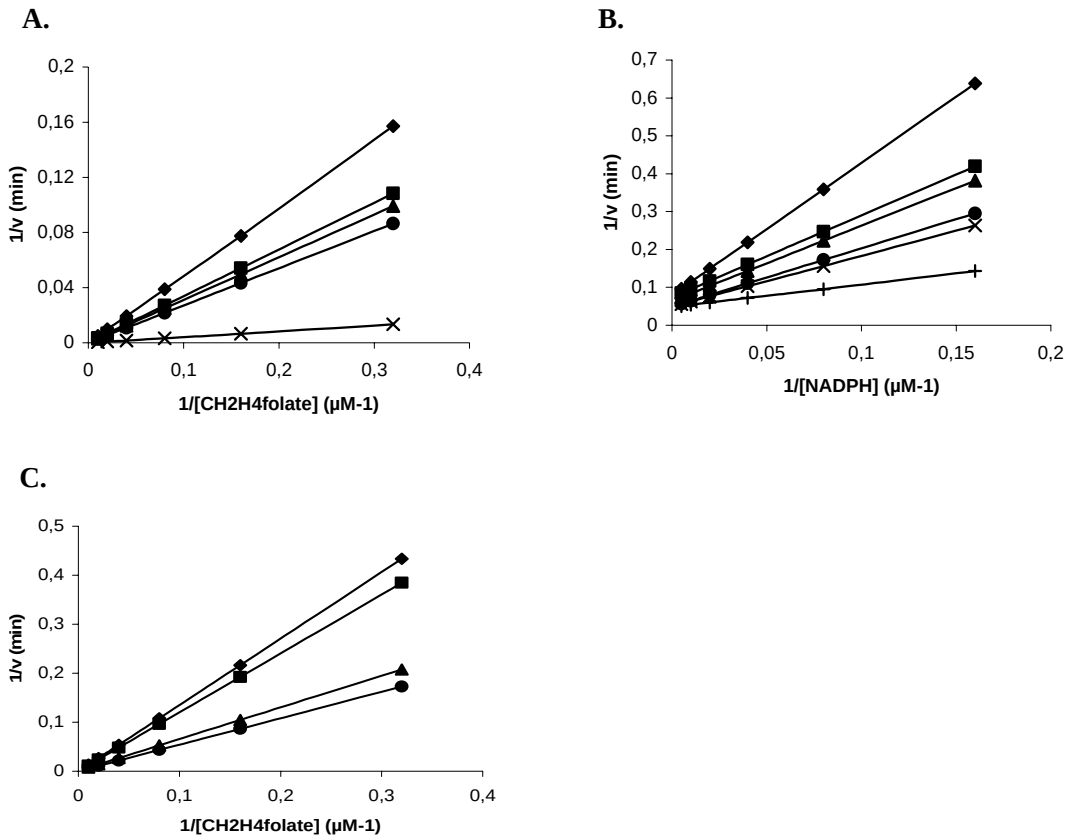


Figure 5: Steady state kinetics of PBCV-1 ThyX catalysis. Lineweaver-Burk transformation plot corresponding to A) Titration of CH_2H_4folate with increasing concentrations of dUMP (♦) 50µM, (■) 25µM, (▲) 12.5µM, (x) 6.25µM, and (+) 3.125µM. Activity was monitored as release of ^3H from [^3H-5]-dUMP, resulting in formation of $^3H_20$. B) Titration of NADPH using increasing concentrations of dUMP. (♦) 100µM, (■) 50µM, (▲) 25µM, (x) 12.5µM, (*) 6.25µM, and (•) 3.125µM. C) Titration of CH_2H_4folate with increasing concentrations of NADPH: (♦) 300µM, (■) 200µM, (▲) 150µM and (•) 100µM. Values used for linear conversion where obtained by non-linear regression.

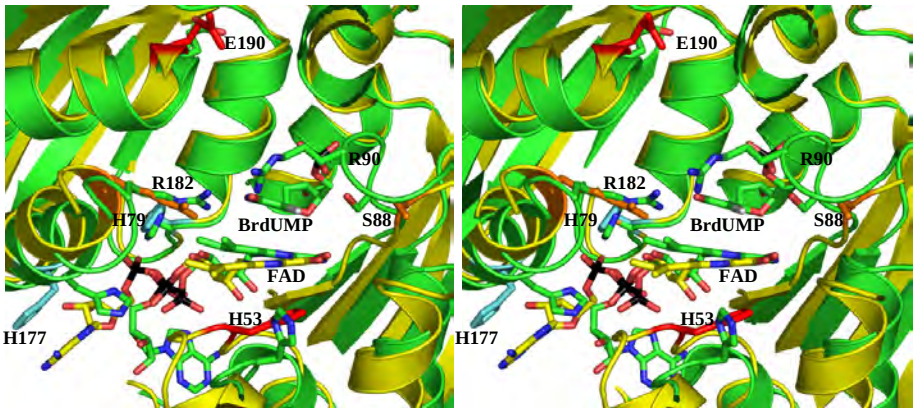


Figure 6: Mapping of the ThyX residues involved in catalysis. PBCV-1 and Mtb ThyX are coloured yellow and green, respectively. PBCV-1 ThyX residues whose mutation completely affects NAD(P)H oxidation or methyl transfer are represented in red and orange, respectively. The two positions H79 and H177 for which substitution by lysine or glutamine has either dramatic or no effect on ThyX activity are coloured in blue. The bromine atom from BrdUMP is coloured in grey. For clarity, only PBCV-1 ThyX numbering is indicated. Phosphorous atoms from FAD and BrdUMP are coloured in black.

DISCUSSION

To date no structural information is available on the binding modes of ThyX with the substrates NADPH or tetrahydrofolate and so far our extensive efforts to obtain quality diffracting crystals of PBCV-1 ThyX complexed to its different substrates did not provide exploitable information.

Since our kinetic experiments are indicative for simultaneous binding of dUMP and NAD(P)H, we attempted to model the ternary complex with NAD(P)H using the complexes of Tm and Mtb ThyX enzymes bound to FAD and dUMP (3,20) to position dUMP into the active site of PBCV-1 ThyX. As shown in Figure 2, all residues involved in dUMP binding (squares and circles) are conserved between PBCV-1, Tm and Mtb ThyX. Only minor side chain adjustments of residues E86 and R155 are needed to accommodate dUMP in the PBCV-1 ThyX active site in the same configuration as for the Tm and Mtb enzymes. In these complexes the uracil base stacks onto the central ring of the FAD isoalloxazine moiety and the ribose phosphate is clamped between helix α3 and the substrate binding loop (residues 86-97). In our PBCV-1 ThyX-FAD complex, this loop is mostly disordered and the hydroxyl group of the putative S88 catalytic nucleophile (11) is 8 to 9Å away from dUMP. In the Tm ThyX apo-structure this loop is equally unstructured while becoming folded upon interaction with dUMP. Hence, upon substrate binding to PBCV-1 ThyX, this loop could become ordered and bring the S88 hydroxyl group in closer contact to the substrates.

In order to fit NADPH into the PBCV-1 ThyX FAD complex we have searched within the Protein Data Bank for all the non-redundant structures describing ternary protein complexes with NAD(P)H and FAD or FMN. We identified 16 of these complexes that could be grouped into three different binding clusters. In the first cluster (grouping nine complexes, best exemplified by glutathione reductase (23), the nicotinamide ring from the NAD(P)H stacks onto the central isoalloxazine ring in a way that is very similar to the binding mode of dUMP complexed with FAD-ThyX. In the second binding mode (1 complex: 2,4-dienoyl-CoA reductase (24), the plane of the nicotinamide ring lies perpendicular to the FAD isoalloxazine pyrimidine ring and occupies the same position as the ribose in the dUMP-ThyX complex. In complexes of the third cluster (grouping 6 cases) FAD (or FMN) is more than 10Å away from NAD(P)H. None of these modes is fully compatible with our experimental data. However, our biochemical data show that NAD(P)H binding and/or

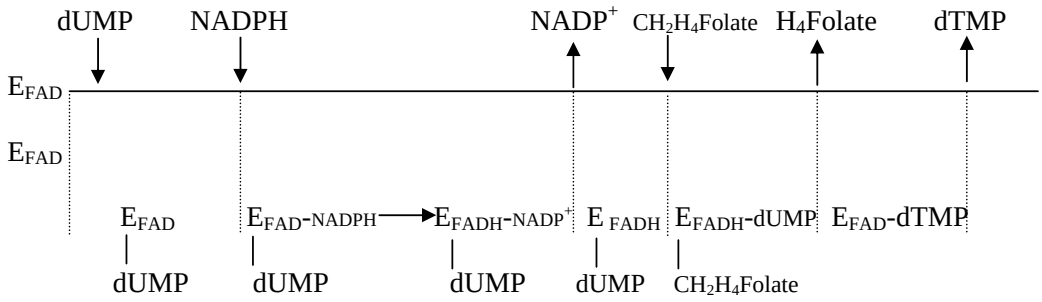


Figure 7: Cleland plot for the proposed sequential mechanism of ThyX proteins. Note that the order in which products are released from the active site has not been investigated in detail.

oxidation is substantially increased by the presence of dUMP, suggesting that dUMP binding re-orients NADPH at the active site to optimize electron transfer. Structural data with bound NADPH analogs (with or without dUMP) will be required to understand the NADPH binding mode of ThyX proteins.

Our mutagenesis results have identified residues that are important for NAD(P)H oxidation (Table 1). These residues have been mapped in Figure 6. Apart from H53, mutation of additional ThyX residues provokes moderate (R90A and R182A) to drastic effects (E190G) on NAD(P)H oxidation activity. These three purified mutant proteins still bind FAD, indicating that the tetramer was assembled correctly. In consequence, the loss of NAD(P)H oxidation might either reflect a decreased affinity for NAD(P)H and/or dUMP (the latter being necessary for PBCV-1 ThyX NAD(P)H oxidase activity) or a deleterious effect on catalytic activity. The side chains of the strictly conserved arginine residues at positions 90 and 182 are hydrogen bonded to the dUMP uracil base. Their substitution by alanine should therefore reduce ThyX affinity for dUMP. The role of E190 is more complex. In the structures of Tm and Mtb ThyX, its carboxylate is hydrogen bonding with S86 (Mtb ThyX) or R105 (Tm ThyX). It is only poorly solvent accessible in the ternary ThyX-FAD-dUMP complex and is unlikely to be directly interacting with NAD(P)H. The complete loss of activity of the E190G is probably due to a local destabilization of the active site. Despite their partial NADPH oxidation activity no *in vivo* complementation is observed with the R90A and R182A mutants, due to their lack of protonation activity. From our structural model and in agreement with the proposal of Agrawal *et al.* (4), we conclude that R90 and R182 are positioned in a way that they could be involved both in CH_2H_4folate and dUMP binding or methyl transfer to dUMP, explaining why they fail to complement *in vivo*.

The PBCV-1 ThyX H53Q and H53K mutant strains did not produce soluble protein for biochemical analyses, and the lack of complementation could simply result from improper folding of the protein or be due to the more buried position of the histidine side chain (Fig. 1D). Site-directed mutagenesis and activity measurements on the corresponding histidine residue in *Helicobacter pylori* ThyX (H48; (11)) as well as chemical modification studies on PBCV-1 ThyX (Fig. 4D) indicate that this histidine side chain is essential for catalysis. In Tm and Mtb ThyX, it is in close vicinity of the redox active N5 atom of the FAD molecule suggesting that it could act as proton donor and/or acceptor as observed in most flavin-dependent enzymes (21). In the actual PBCV-1 ThyX structure, H53 is not ideally positioned to exert this role. However, H53 reacts readily with DEPC, indicating that this residue is accessible and undergoes structural alterations. In agreement with the helix $\alpha 2$ sliding observed between PBCV-1 and Tm or MtB ThyX structures (Fig. 1D), a possible conformational change, possibly induced by substrate binding, might bring the

H53 Nε2 atom within a more favourable position to abstract (or receive) a proton from the isoalloxazine N5.

Our site-directed mutagenesis, chemical modification and structural data have provided new insight into the active site of ThyX proteins. Whereas our kinetic data continue to support a sequential reaction mechanism for ThyX proteins (Fig. 7), an earlier study has proposed a ping-pong mechanism implicating the formation of a methyl enzyme intermediate during turnover of *C. trachomatis* ThyX (7). It is of note that we have been unable to detect the proposed covalent intermediate using either radioactively marked $^{14}CH_2H_4$folate or mass spectrometric analyses for either PBCV-1 or C. trachomatis ThyX proteins (data not shown). The reasons for these experimental discrepancies are currently unknown. However, double reciprocal plots apparently suggesting a ping-pong mechanism have also been measured for Lactobacillus casei ThyA protein (8), although all other kinetic data supported a sequential mechanism. Moreover, the order of substrate binding of ThyA proteins has been reported to depend on buffer composition used in the experiments (1). Additional experimental evidence is required to elucidate the role of the proposed ternary complexes of PBCV-1 ThyX in catalysis.

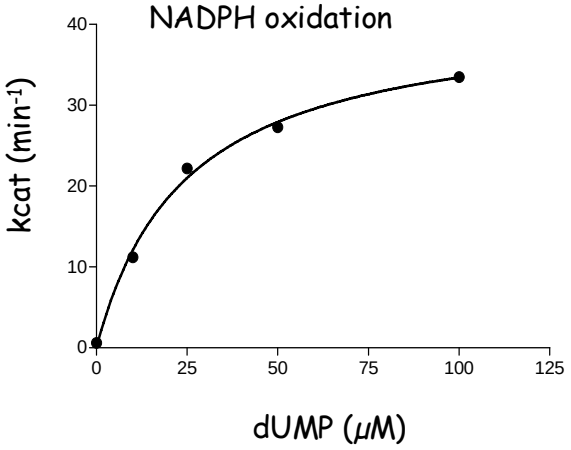
Acknowledgements – We thank James van Etten and Yuannan Xia for the pVEX plasmid and Yap Boum for his help with the NADPH oxidation essays. This research was supported by a grant from the Programme Microbiologie Fondamentale to UL and HM. HM and SS gratefully acknowledge support from the INSERM (Bioavenir and Jeune Chercheur Programs) and the Fondation Bettencourt Schuller.

REFERENCES

1. Carreras, C. W., and Santi, D. V. (1995) *Annu Rev Biochem* **64**, 721-762
2. Myllykallio, H., Lipowski, G., Leduc, D., Filee, J., Forterre, P., and Liebl, U. (2002) *Science* **297**(5578), 105-107
3. Mathews, A. M., Deacon, J. M., Canaves, D., McMullan, S. A., Lesley, S., Agarwalla, and Kuhn, P. (2003) *Structure (Camb)* **11**, 677-690
4. Agrawal, N., Lesley, S. A., Kuhn, P., and Kohen, A. (2004) *Biochemistry* **43**(32), 10295-10301
5. Gattis, S. G., and Palfey, B. A. (2005) *J Am Chem Soc* **127**(3), 832-833
6. Graziani, S., Xia, Y., Gurnon, J. R., Van Etten, J. L., Leduc, D., Skouloubris, S., Myllykallio, H., and Liebl, U. (2004) *J Biol Chem* **279**(52), 54340-54347
7. Griffin, J., Roshick, C., Iliffe-Lee, E., and McClarty, G. (2005) *J Biol Chem* **280**(7), 5456-5467
8. Daron, H. H., and Aull, J. L. (1978) *J Biol Chem* **253**(3), 940-945
9. Lorenson, M. Y., Maley, G. F., and Maley, F. (1967) *J Biol Chem* **242**(14), 3332-3344
10. Pogolotti, A. L., Jr., Ivanetich, K. M., Sommer, H., and Santi, D. V. (1976) *Biochem Biophys Res Commun* **70**(3), 972-978
11. Leduc, D., Graziani, S., Lipowski, G., Marchand, C., Le Marechal, P., Liebl, U., and Myllykallio, H. (2004) *Proc Natl Acad Sci U S A* **101**(19), 7252-7257
12. Demarre, G., Guerout, A. M., Matsumoto-Mashimo, C., Rowe-Magnus, D. A., Marliere, P., and Mazel, D. (2005) *Res Microbiol* **156**(2), 245-255
13. Miles, E. W. (1977) *Methods Enzymol* **47**, 431-442
14. Kabsch, W. (1993) *Journal of Applied Crystallography* **26**, 795-800
15. Vagin, A., and Teplyakov, A. (1997) *J. Appl. Cryst.* **30**, 1022-1025
16. Perrakis, A., Morris, R., and Lamzin, V. S. (1999) *Nat Struct Biol* **6**(5), 458-463
17. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) *Acta Crystallogr D Biol Crystallogr* **53**(Pt 3), 240-255
18. Roussel, A., and Cambillau, C. (1989) Silicon Graphics Geometry Partner Directory. In., Silicon Graphics, Mountain View, CA
19. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) *Journal of Applied Crystallography* **26**, 283-291
20. Sampathkumar, P., Turley, S., Ulmer, J. E., Rhie, H. G., Sibley, C. H., and Hol, W. G. (2005) *J Mol Biol* **352**(5), 1091-1104
21. Fraaije, M. W., and Mattevi, A. (2000) *Trends Biochem Sci* **25**(3), 126-132
22. Lemaire, M., Schmitter, J.-M., Issakidis, E., Miginiac-Maslow, M., Gadal, P., and Decottignies, P. (1994) *J. Biol. Chem.* **269**, 27291-27296
23. Karplus, P. A., and Schulz, G. E. (1989) *J Mol Biol* **210**(1), 163-180
24. Hubbard, P. A., Liang, X., Schulz, H., and Kim, J. J. (2003) *J Biol Chem* **278**(39), 37553-37560
25. Gouet, P., Courcelle, E., Stuart, D. I., and Metoz, F. (1999) *Bioinformatics* **15**(4), 305-308

SUPPLEMENTARY MATERIALS

Supplemental Figure 1: dUMP dependency of NADPH oxidation



Supplemental Table 1 : Data collection statistics

Data collection statistics		
	Space group	$P2_12_12$
	Wavelength	0.933 Å
	Unit cell parameters	a= 69.264 Å b= 76.991 Å c= 93.437 Å α= 90° β= 90° γ= 90°
	Resolution	20.0-2.3 Å
	Number of reflections	109258
	Number of unique reflections	22844
	Multiplicity	4.8
	R_{sym} [1] (%)	9.1(32.7)
	Completeness (%)	99.2(97.4)
	I/σ (I)	14.1(5.9)
Refinement statistics		
	Reflections (working/test)	21351/1121
	R/R_{free} (%) [2]	18.6/24.2
	Non-hydrogens atoms	3200
	R.m.s.d. bonds (Å)	0.024
	R.m.s.d. angles (deg)	2.146
	Mean B-factor (Å^2)	43.050
Ramachandran plot		
	Most favored	91.1%
	Allowed	8.9%

Values in parentheses are for the highest resolution shell.

[1] $R_{sym} = \sum_h \sum_i \left| I_{hi} - \langle I_h \rangle \right| / \sum_h \sum_i I_{hi}$, where I_{hi} is the ith observation of the reflection h, while $\langle I_h \rangle$ is the mean intensity of reflection h.

[2] $R_{factor} = \sum \left\| |F_O| - |F_C| \right\| / |F_O|$. R_{free} was calculated with a small fraction (5%) of randomly selected reflections.

6.3 Annexe : diffraction et problème des phases

Pour que le signal de diffraction soit détectable, il faut que de nombreux photons soient diffractés de la même façon par les électrons correspondants dans des molécules arrangées de façon périodique.

Le signal diffracté est relié au réseau cristallin par le facteur de structure qui s'écrit comme suit :

$$F_{hkl} = \sum_{j=1}^{N} q_j f_j e^{-B_j \frac{\sin^2\theta}{\lambda^2}} e^{2i\pi(hx_j+ky_j+lz_j)} \tag{6.1}$$

avec :

q_j : facteur d'occupation de l'atome j ;

f_j : facteur de diffusion atomique de l'atome j ;

B_j : facteur d'agitation thermique de l'atome j ;

θ : angle de diffraction ;

λ : longueur d'onde ;

h, k, l : indices de réflexion ;

x, y, z : coordonnées de l'atome ;

N : nombre d'atomes.

La densité electronique est obtenue à partir du facteur de structure :

$$\rho(x,y,z) = \frac{1}{V} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} |F| e^{i\varphi_{hkl}} e^{-2i\pi(hx+ky+lz)} \tag{6.2}$$

avec :

V : volume de la maille cristalline.

L'intensité du signal mesuré en chaque tache est proportionnelle au carré du facteur de structure. On a :

$$I \propto |F|^2 \tag{6.3}$$

Pour pouvoir déterminer comment est constitué le cristal et donc connaître la densité électronique, il faut connaître la phase φ, or celle-ci n'est pas accessible par la mesure de l'intensité.

En biocristallographie, on ne peut obtenir la phase *ab initio* par des méthodes itératives. Ce problème est connu sous le nom de *problème des phases*.

Chapitre 7

Conclusion de la deuxième partie

Les études de mutagenèse dirigée de nos collaborateurs au laboratoire d'Optique et de Biosciences et à l'Institut de Génétique et Microbiologie ont permis d'identifier plusieurs résidus impliqués soit dans l'oxydation du NAPDH, soit dans la déprotonation du dUMP. Quant aux études complémentaires de modification chimique couplée à la spectrométrie de masse réalisées par nos collaborateurs du laboratoire, elles ont permis d'identifier un résidu histine (H53) ayant un rôle crucial dans l'oxydation du NADPH. L'étude structurale montre que ce résidu est de plus situé au voisinage de l'azote N5, redox actif, de l'isoalloxazine du FAD.

D'un point de vue structural, nous avons aussi observé que la conformation des résidus du site actif de ThyX PBCV-1 est différente de celles observées dans les structures précedemment résolues, un réarrangement du site actif devant avoir lieu pendant la catalyse. Les analyses cinétiques confortent le mécanisme séquentiel proposé, dans lequel le processus de catalyse met en jeu des complexes ternaires distincts sans formation d'un intermédiaire méthyl-enzyme.

Bien sûr, d'autres preuves expérimentales de la formation de ces complexes ternaires sont nécessaires pour trancher. De plus, il serait souhaitable à l'avenir de résoudre les structures de complexes de cette protéine avec des analogues des ligands naturels pour pouvoir mieux caractériser les variations structurales ayant lieu lors de la catalyse.

Conclusion générale

Au cours de ma thèse, j'ai travaillé à la modélisation de la structure des protéines par des diagrammes de Voronoï. Cette modélisation m'a permis de définir et calculer des descripteurs qui se sont révélés discriminants dans deux problèmes : la prédiction de la structure des complexes protéine-protéine, et la discrimination entre dimères biologiques et dimères cristallographiques.

Dans les deux cas, nous avons utilisé des méthodes d'apprentissage statistique qui ont permis de calculer des fonctions de score. Ces fonctions, bien que ne permettant pas encore d'atteindre le but que nous nous étions fixé, permettent d'espérer arriver rapidement à la solution.

En couplant cette fonction de score à une procédure rapide et efficace de génération des conformations possibles, il sera alors possible de cribler très rapidement un grand nombre de complexes, permettant des études fonctionnelles à l'échelle génomique.

L'étude structurale de la protéine ThyX m'a permis de mieux comprendre ce que représentent les coordonnées atomiques présentes dans les fichiers de la *PDB*.

Enfin, de nombreuses étapes de ce travail ont été faites en collaboration avec des chercheurs de différentes disciplines : apprentissage, géométrie algorithmique, statistiques, biochimie et biologie structurale. Ces aperçus de différents domaines de la recherche ont été pour moi très enrichissants.

Bibliographie

[1] L. Adamian, R. Jackups, Jr, T. Binkowski, and J. Liang. Higher-order interhelical spatial interactions in membrane proteins. *J Mol Biol*, 327(1) :251–72, 2003.

[2] N. Agrawal, S. Lesley, P. Kuhn, and A. Kohen. Mechanistic studies of a flavin-dependent thymidylate synthase. *Biochemistry*, 43(32) :10295–301, 2004.

[3] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3) :403–10, 1990.

[4] B. Angelov, J. Sadoc, R. Jullien, A. Soyer, J. Mornon, and J. Chomilier. Nonatomic solvent-driven Voronoi tessellation of proteins : an open tool to analyze protein folds. *Proteins*, 49(4) :446–56, 2002.

[5] J. Austin, K. Rodgers, and T. Spiro. Protein structure from ultraviolet resonance Raman spectroscopy. *Methods Enzymol*, 226 :374–96, 1993.

[6] T. Back. *Evolutionary Algorithms in Theory and Practice : Evolution Strategies, Evolutionary Programming, Genetic Algorithms*. Oxford University Press, 1996.

[7] D. J. Bacon and W. F. Anderson. A Fast Algorithm for Rendering Space-Filling Molecule Pictures. *J. Mol. Graphics*, 6 :219–220, 1988.

[8] R. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53(3) :708–19, 2003.

[9] R. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4) :943–55, 2004.

[10] D. Bakowies and W. Van Gunsteren. Water in protein cavities : A procedure to identify internal water and exchange pathways and application to fatty acid-binding protein. *Proteins*, 47(4) :534–45, 2002.

[11] E. Ben-Zeev, A. Berchanski, A. Heifetz, B. Shapira, and M. Eisenstein. Prediction of the unknown : inspiring experience with the CAPRI experiment. *Proteins*, 52(1) :41–6, 2003.

[12] E. Ben-Zeev and M. Eisenstein. Weighted geometric docking : incorporating external information in the rotation-translation scan. *Proteins*, 52(1) :24–7, 2003.

[13] A. Berchanski, D. Segal, and M. Eisenstein. Modeling oligomers with Cn or Dn symmetry : application to CAPRI target 10. *Proteins*, 60(2) :202–6, 2005.

[14] H. Berman, T. Bhat, P. Bourne, Z. Feng, G. Gilliland, H. Weissig, and J. Westbrook. The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*, 7 Suppl :957–9, 2000.

[15] H. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, 10(12) :980, 2003.

[16] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28 :235–242, 2000.

[17] M. Betts and R. Russell. *Bioinformatics for Geneticists*, chapter Amino acid properties and consequences of substitutions. John Wiley & Sons, 2003.

[18] T. Binkowski, L. Adamian, and J. Liang. Inferring functional relationships of proteins from local sequence and spatial surface patterns. *J Mol Biol*, 332(2) :505–26, 2003.

[19] D. Blow, C. Wright, D. Kukla, A. Ruhlmann, W. Steigemann, and R. Huber. A model for the association of bovine pancreatic trypsin inhibitor with chymotrypsin and trypsin. *J Mol Biol*, 69(1) :137–44, 1972.

[20] J. Boissonnat and M. Yvinec. *Géométrie Algorithmique*. 1995.

[21] J.-D. Boissonnat, F. Cazals, F. Da, O. Devillers, S. Pion, F. Rebufat, M. Teillaud, and M. Yvinec. Programming with CGAL : The example of triangulations. In *Symposium on Computational Geometry*, pages 421–422, 1999.

[22] D. Bostick and I. Vaisman. A new topological method to measure protein structure similarity. *Biochem Biophys Res Commun*, 304(2) :320–5, 2003.

[23] L. Boulu, G. Crippen, H. Barton, H. Kwon, and M. Marletta. Voronoi binding site model of a polycyclic aromatic hydrocarbon binding protein. *J Med Chem*, 33(2) :771–5, 1990.

[24] P. Bourne. CASP and CAFASP experiments and their findings. *Methods Biochem Anal*, 44 :501–7, 2003.

[25] M. Bradley and G. Crippen. Voronoi modeling : the binding of triazines and pyrimidines to L. casei dihydrofolate reductase. *J Med Chem*, 36(21) :3171–7, 1993.

[26] P. Bradley, L. Malmstrom, B. Qian, J. Schonbrun, D. Chivian, D. Kim, J. Meiler, K. Misura, and D. Baker. Free modeling with Rosetta in CASP6. *Proteins*, 61 Suppl 7 :128–34, 2005.

[27] A. Brunger. Assessment of phase accuracy by cross validation : the free R value. Methods and applications. *Acta Crystallogr D Biol Crystallogr*, 49(Pt 1) :24–36, 1993.

[28] A. Brunger, P. Adams, G. Clore, W. DeLano, P. Gros, R. Grosse-Kunstleve, J. Jiang, J. Kuszewski, M. Nilges, N. Pannu, R. Read, L. Rice, T. Simonson, and G. Warren. Crystallography & NMR system : A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, 54 (Pt 5) :905–21, 1998.

[29] Brunger A.T. *X-PLOR, a System for Crystallography and NMR*. Yale University, New Haven, yale university press edition, 1992.

[30] N. Calimet, M. Schaefer, and T. Simonson. Protein molecular dynamics with the generalized Born/ACE solvent model. *Proteins*, 45(2) :144–58, 2001.

[31] C. Camacho. Modeling side-chains using molecular dynamics improve recognition of binding region in CAPRI targets. *Proteins*, 60(2) :245–51, 2005.

[32] C. Carter, Jr, B. LeFebvre, S. Cammer, A. Tropsha, and M. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J Mol Biol*, 311(4) :625–38, 2001.

[33] P. Carter, V. Lesk, S. Islam, and M. Sternberg. Protein-protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI. *Proteins*, 60(2) :281–8, 2005.

[34] F. Cazals, F. Proust, and J. Janin. Revisiting the description of Protein-Protein interfaces. Part II : Experimental study. *J. Mol. Biol. (submitted)*, 2006.

[35] S. Chakravarty, A. Bhinge, and R. Varadarajan. A procedure for detection and quantitation of cavity volumes proteins. Application to measure the strength of the hydrophobic driving force in protein folding. *J Biol Chem*, 277(35) :31345–53, 2002.

[36] C.-C. Chang and C.-J. Lin. Libsvm : a library for support vector machines (version 2.31), 2001.

[37] R. Chen, L. Li, and Z. Weng. ZDOCK : an initial-stage protein-docking algorithm. *Proteins*, 52(1) :80–7, 2003.

[38] R. Chen, J. Mintseris, J. Janin, and Z. Weng. A protein-protein docking benchmark. *Proteins*, 52(1) :88–91, 2003.

[39] R. Chen, W. Tong, J. Mintseris, L. Li, and Z. Weng. ZDOCK predictions for the CAPRI challenge. *Proteins*, 52(1) :68–73, 2003.

[40] R. Chen and Z. Weng. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, 47(3) :281–94, 2002.

[41] R. Chen and Z. Weng. A novel shape complementarity scoring function for protein-protein docking. *Proteins*, 51(3) :397–408, 2003.

[42] J. Cherfils, S. Duquerroy, and J. Janin. Protein-protein recognition analyzed by docking simulation. *Proteins*, 11(4) :271–80, 1991.

[43] C. Chothia. Hydrophobic bonding and accessible surface area in proteins. *Nature*, 248(446) :338–9, 1974.

[44] C. Chothia. Structural invariants in protein folding. *Nature*, 254(5498) :304–8, 1975.

[45] C. Chothia and M. Gerstein. Protein evolution. How far can sequences diverge ? *Nature*, 385(6617) :579, 581, 1997.

[46] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256(5520) :705–8, 1975.

[47] R. Collobert and S. Bengio. SVMTorch : Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1 :143–160, 2001.

[48] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, and J. Mornon. Comparison of three algorithms for the assignment of secondary structure in proteins : the advantages of a consensus assignment. *Protein Eng*, 6(4) :377–82, 1993.

[49] S. Comeau and C. Camacho. Predicting oligomeric assemblies : N-mers a primer. *J Struct Biol*, 150(3) :233–44, 2005.

[50] S. Comeau, S. Vajda, and C. Camacho. Performance of the first protein docking server ClusPro in CAPRI rounds 3-5. *Proteins*, 60(2) :239–44, 2005.

[51] M. Connolly. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612) :709–13, 1983.

[52] M. Connolly. Molecular surface triangulation. *J Appl Crystallogr*, 18 :499–505, 1985.

[53] M. Connolly. Shape complementarity at the hemoglobin alpha 1 beta 1 subunit interface. *Biopolymers*, 25(7) :1229–47, 1986.

[54] M. Connolly. Molecular interstitial skeleton. *Computers & Chemistry*, 15(1) :37–45, 1991.

[55] D. Cozzetto, A. Di Matteo, and A. Tramontano. Ten years of predictions ... and counting. *FEBS J*, 272(4) :881–2, 2005.

[56] G. Crippen. Voronoi binding site models. *NIDA Res Monogr*, 112 :7–20, 1991.

[57] N. Cristiani and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Mar. 2000.

[58] M. Daily, D. Masica, A. Sivasubramanian, S. Somarouthu, and J. Gray. CAPRI rounds 3-5 reveal promising successes and future challenges for RosettaDock. *Proteins*, 60(2) :181–6, 2005.

[59] B. Derrida. Random-energy model : An exactly solvable model of disordered systems. *Phys. Rev.*, B(24) :2613–2626, 1981.

[60] O. Devillers. The Delaunay hierarchy. *Internat. J. Found. Comput. Sci.*, 13 :163–180, 2002.

[61] C. Dominguez, R. Boelens, and A. Bonvin. HADDOCK : a protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc*, 125(7) :1731–7, 2003.

[62] J. Drenth. *Principles of X-Ray Crystallography.* Springer-Verlag, 1999.

[63] F. Dupuis, J. Sadoc, R. Jullien, B. Angelov, and J. Mornon. Voro3D : 3D Voronoi tessellations applied to protein structures. *Bioinformatics*, 21(8) :1715–6, 2005.

[64] S. Dutta and H. Berman. Large macromolecular complexes in the Protein Data Bank : a status report. *Structure (Camb)*, 13(3) :381–8, 2005.

[65] H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Pac Symp Biocomput*, pages 272–87, 1996.

[66] H. Edelsbrunner, M. Facello, and J. Liang. On the definition and the construction of pockets in macromolecules. *Discr Appl Math*, 88 :83–102, 1998.

[67] H. Edelsbrunner and P. Koehl. The weighted-volume derivative of a space-filling diagram. *Proc Natl Acad Sci U S A*, 100(5) :2203–8, 2003.

[68] F. Eisenhaber, P. Lijnzaad, C. Sander, P. Argos, and M. Scharf. The Double Cubic Lattice Method : Efficient Approaches to Numerical Integration of Surface Area and Volume and to Dot Surface Contouring of Molecular Assemblies. *J. Comp. Chem.*, 16(3) :273–284, 1995.

[69] M. Eisenstein. Introducing a 4th dimension to protein-protein docking. *Structure (Camb)*, 12(12) :2095–6, 2004.

[70] J. Fernandez-Recio, R. Abagyan, and M. Totrov. Improving CAPRI predictions : optimized desolvation for rigid-body docking. *Proteins*, 60(2) :308–13, 2005.

[71] B. Fields, F. Goldbaum, W. Dall'Acqua, E. Malchiodi, A. Cauerhff, F. Schwarz, X. Ysern, R. Poljak, and R. Mariuzza. Hydrogen bonding and solvent structure in an antigen-antibody interface. Crystal structures and thermodynamic characterization of three Fv mutants complexed with lysozyme. *Biochemistry*, 35(48) :15494–503, 1996.

[72] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230) :245–6, 1989.

[73] J. Finney. Volume occupation, environment and accessibility in proteins. The problem of the protein surface. *J Mol Biol*, 96(4) :721–32, 1975.

[74] D. Fischer, S. Lin, H. Wolfson, and R. Nussinov. A geometry-based suite of molecular docking processes. *J Mol Biol*, 248(2) :459–77, 1995.

[75] D. Fischer, R. Norel, H. Wolfson, and R. Nussinov. Surface motifs by a computer vision technique : searches, detection, and implications for protein-ligand recognition. *Proteins*, 16(3) :278–92, 1993.

[76] P. Fleming and F. Richards. Protein packing : dependence on protein size, secondary structure and amino acid composition. *J Mol Biol*, 299(2) :487–98, 2000.

[77] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *The Computer Journal*, 7(2) :149–154, 1964.

[78] D. Frishman and P. Argos. The future of protein secondary structure prediction accuracy. *Fold Des*, 2(3) :159–62, 1997.

[79] H. Gan, A. Tropsha, and T. Schlick. Lattice protein folding with two and four-body statistical potentials. *Proteins*, 43(2) :161–74, 2001.

[80] E. Gardiner, P. Willett, and P. Artymiuk. GAPDOCK : a Genetic Algorithm Approach to Protein Docking in CAPRI round 1. *Proteins*, 52(1) :10–4, 2003.

[81] S. Garner. Weka : The waikato environment for knowledge analysis. 1995.

[82] A. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, C. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868) :141–7, 2002.

[83] B. Gellatly and J. Finney. Calculation of protein volumes : an alternative to the Voronoi procedure. *J Mol Biol*, 161(2) :305–22, 1982.

[84] M. Gerstein and C. Chothia. Packing at the protein-water interface. *Proc Natl Acad Sci U S A*, 93(19) :10167–72, 1996.

[85] M. Gerstein, J. Tsai, and M. Levitt. The volume of atoms on the protein surface : calculated from simulation, using Voronoi polyhedra. *J Mol Biol*, 249(5) :955–66, 1995.

[86] A. Goede, R. Preissner, and C. Frömmel. Voronoi cell : New method for allocation of space among atoms : Elimination of avoidable errors in calculation of atomic volume and density. *J Comp Chem*, 18(9) :1113–1123, 1997.

[87] J. Gray, S. Moughon, T. Kortemme, O. Schueler-Furman, K. Misura, A. Morozov, and D. Baker. Protein-protein docking predictions for the CAPRI experiment. *Proteins*, 52(1) :118–22, 2003.

[88] S. Graziani, Y. Xia, J. Gurnon, J. Van Etten, D. Leduc, S. Skouloubris, H. Myllykallio, and U. Liebl. Functional analysis of FAD-dependent thymidylate synthase ThyX from Paramecium bursaria Chlorella virus-1. *J Biol Chem*, 279(52) :54340–7, 2004.

[89] J. Griffin, C. Roshick, E. Iliffe-Lee, and G. McClarty. Catalytic mechanism of Chlamydia trachomatis flavin-dependent thymidylate synthase. *J Biol Chem*, 280(7) :5456–67, 2005.

[90] D. Halaby and J. Mornon. The immunoglobulin superfamily : an insight on its tissular, species, and functional diversity. *J Mol Evol*, 46(4) :389–400, 1998.

[91] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking : An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4) :409–43, 2002.

[92] T. Hamelryck and B. Manderick. PDB file parser and structure class implemented in Python. *Bioinformatics*, 19(17) :2308–10, 2003.

[93] Y. Harpaz, M. Gerstein, and C. Chothia. Volume changes on protein folding. *Structure*, 2(7) :641–9, 1994.

[94] A. Heifetz, E. Katchalski-Katzir, and M. Eisenstein. Electrostatics in protein-protein docking. *Protein Sci*, 11(3) :571–87, 2002.

[95] R. Herrmann. Theory of hydrophobic bonding. II. Correlation of hydrocarbon solubility in water with solvent cavity surface area. *J Phys Chem*, 76(19) :2754–2759, 1972.

[96] Y. Ho, A. Gruhler, A. Heilbut, G. Bader, L. Moore, S. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. Willems, H. Sassi, P. Nielsen, K. Rasmussen, J. Andersen, L. Johansen, L. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. Sorensen, J. Matthiesen, R. Hendrickson, F. Gleeson, T. Pawson, M. Moran, D. Durocher, M. Mann, C. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868) :180–3, 2002.

[97] G. Huang. High-throughput DNA sequencing : a genomic data manufacturing process. *DNA Seq*, 10(3) :149–53, 1999.

[98] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *J. Mol. Graphics*, 14 :33–38, 1996.

[99] Y. Inbar, D. Schneidman-Duhovny, I. Halperin, A. Oron, R. Nussinov, and H. Wolfson. Approaching the CAPRI challenge with an efficient geometry-based docking. *Proteins*, 60(2) :217–23, 2005.

[100] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein-protein interaction map of the budding yeast : A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A*, 97(3) :1143–7, 2000.

[101] J. Janin. Quantifying biological specificity : the statistical mechanics of molecular recognition. *Proteins*, 25(4) :438–45, 1996.

[102] J. Janin. Assessing predictions of protein-protein interaction : the CAPRI experiment. *Protein Sci*, 14(2) :278–83, 2005.

[103] J. Janin. Sailing the route from Gaeta, Italy, to CAPRI. *Proteins*, 60(2) :149, 2005.

[104] J. Janin. The targets of CAPRI rounds 3-5. *Proteins*, 60(2) :170–5, 2005.

[105] J. Janin, K. Henrick, J. Moult, L. Eyck, M. Sternberg, S. Vajda, I. Vakser, and S. Wodak. CAPRI : a Critical Assessment of PRedicted Interactions. *Proteins*, 52(1) :2–9, 2003.

[106] J. Janin, F. Rodier, P. Chakrabarti, and R. Bahadur. Molecular recognition in the Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, to appear, 2006.

[107] J. Janin and B. Seraphin. Genome-wide studies of protein-protein interaction. *Curr Opin Struct Biol*, 13(3) :383–8, 2003.

[108] J. Janin and S. Wodak. Reaction pathway for the quaternary structure change in hemoglobin. *Biopolymers*, 24(3) :509–26, 1985.

[109] J. Janin and S. Wodak. Protein modules and protein-protein interaction. Introduction. *Adv Protein Chem*, 61 :1–8, 2002.

[110] F. Jiang and S. Kim. "Soft docking" : matching of molecular surface cubes. *J Mol Biol*, 219(1) :79–102, 1991.

[111] R. Jornsten, H. Wang, W. Welsh, and M. Ouyang. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21(22) :4155–61, 2005.

[112] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta. Crystal. A*, 32 :922–923, 1976.

[113] W. Kabsch. A discussion of the solution for the best rotation to related two sets of vectors. *Acta. Crystal. A*, 34 :827–828, 1978.

[114] W. Kabsch and C. Sander. Dictionary of protein secondary structure : pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12) :2577–637, 1983.

[115] S. Karlin, Z. Zhu, and F. Baud. Atom density in protein structures. *Proc Natl Acad Sci U S A*, 96(22) :12500–5, 1999.

[116] S. Karlin, M. Zuker, and L. Brocchieri. Measuring residue associations in protein structures. Possible implications fro protein folding. *J. Mol. Biol.*, 264 :121–136, 1994.

[117] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. Friesem, C. Aflalo, and I. Vakser. Molecular surface recognition : determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci U S A*, 89(6) :2195–9, 1992.

[118] J. Kendrew, R. Dickerson, B. Standberg, R. Hart, D. Davies, and D. Phillips. Structure of myoglobin. A three dimensional Fourier synthesis at 2Å resolution. *Nature*, 185 :422–427, 1960.

[119] C. Kissinger, D. Gehlhaar, and D. Fogel. Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr D Biol Crystallogr*, 55 (Pt 2) :484–91, 1999.

[120] N. Kobayashi, T. Yamato, and N. Go. Mechanical property of a TIM-barrel protein. *Proteins*, 28(1) :109–16, 1997.

[121] K. Komatsu, Y. Kurihara, M. Iwadate, M. Takeda-Shitaka, and H. Umeyama. Evaluation of the third solvent clusters fitting procedure for the prediction of protein-protein interactions based on the results at the CAPRI blind docking study. *Proteins*, 52(1) :15–8, 2003.

[122] B. Krishnamoorthy and A. Tropsha. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12) :1540–8, 2003.

[123] A. Kryshtafovych, C. Venclovas, K. Fidelis, and J. Moult. Progress over the first decade of CASP experiments. *Proteins*, 61 Suppl 7 :225–36, 2005.

[124] D. Law, L. Ten Eyck, O. Katzenelson, I. Tsigelny, V. Roberts, M. Pique, and J. Mitchell. Finding needles in haystacks : Reranking DOT results by using shape complementarity, cluster analysis, and biological information. *Proteins*, 52(1) :33–40, 2003.

[125] D. Leduc, S. Graziani, G. Lipowski, C. Marchand, P. Le Marechal, U. Liebl, and H. Myllykallio. Functional evidence for active site location of tetrameric thymidylate synthase X at the interphase of three monomers. *Proc Natl Acad Sci U S A*, 101(19) :7252–7, 2004.

[126] D. Leduc, S. Graziani, L. Meslet-Cladiere, A. Sodolescu, U. Liebl, and H. Myllykallio. Two distinct pathways for thymidylate (dTMP) synthesis in (hyper)thermophilic Bacteria and Archaea. *Biochem Soc Trans*, 32(Pt 2) :231–5, 2004.

[127] B. Lee and F. M. Richards. The interpretation of protein structures : estimation of static accessibility. *J. Mol. Biol.*, 55(3) :379–400, 1971.

[128] R. Lee and G. Rose. Molecular recognition. I. Automatic identification of topographic surface features. *Biopolymers*, 24(8) :1613–27, 1985.

[129] N. Leulliot, S. Quevillon-Cheruel, I. Sorel, M. Graille, P. Meyer, D. Liger, K. Blondeau, J. Janin, and H. van Tilbeurgh. Crystal structure of yeast allantoicase reveals a repeated jelly roll motif. *J Biol Chem*, 279(22) :23447–52, 2004.

[130] C. Levinthal, S. Wodak, P. Kahn, and A. Dadivanian. Hemoglobin interaction in sickle cell fibers. I : Theoretical approaches to the molecular contacts. *Proc Natl Acad Sci U S A*, 72(4) :1330–4, 1975.

[131] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol*, 104(1) :59–107, 1976.

[132] L. Li, R. Chen, and Z. Weng. RDOCK : refinement of rigid-body protein docking predictions. *Proteins*, 53(3) :693–707, 2003.

[133] X. Li, C. Hu, and J. Liang. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. *Proteins*, 53(4) :792–805, 2003.

[134] Y. Li, M. Urrutia, S. Smith-Gill, and R. Mariuzza. Dissection of binding interactions in the complex between the anti-lysozyme antibody HyHEL-63 and its antigen. *Biochemistry*, 42(1) :11–22, 2003.

[135] J. Liang and K. Dill. Are proteins well-packed ? *Biophys J*, 81(2) :751–66, 2001.

[136] J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules : I. Molecular area and volume through alpha shape. *Proteins*, 33(1) :1–17, 1998.

[137] J. Liang, H. Edelsbrunner, P. Fu, P. Sudhakar, and S. Subramaniam. Analytical shape computation of macromolecules : II. Inaccessible cavities in proteins. *Proteins*, 33(1) :18–29, 1998.

[138] J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities : measurement of binding site geometry and implications for ligand design. *Protein Sci*, 7(9) :1884–97, 1998.

[139] J. Liang and S. Subramaniam. Computation of molecular electrostatics with boundary element methods. *Biophys J*, 73(4) :1830–41, 1997.

[140] O. Lichtarge, H. Bourne, and F. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, 257(2) :342–58, 1996.

[141] S. Lin, R. Nussinov, D. Fischer, and H. Wolfson. Molecular surface representations by sparse critical points. *Proteins*, 18(1) :94–101, 1994.

[142] C. X. Ling, J. Huang, and H. Zhang. AUC : A better measure than accuracy in comparing learning algorithms. In Y. Xiang and B. Chaib-draa, editors, *Advances in Artificial Intelligence, 16th Conference of the Canadian Society for Computational Studies of Intelligence, AI 2003, Halifax, Canada, June 11-13, 2003, Proceedings*, volume 2671 of *Lecture Notes in Computer Science*, pages 329–341. Springer, 2003.

[143] J. Linge, S. O'Donoghue, and M. Nilges. Automated assignment of ambiguous nuclear overhauser effects with ARIA. *Methods Enzymol*, 339 :71–90, 2001.

[144] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5) :2177–98, 1999.

[145] J. Mandell, V. Roberts, M. Pique, V. Kotlovyi, J. Mitchell, E. Nelson, I. Tsigelny, and L. Ten Eyck. Protein docking using continuum electrostatics and geometric fit. *Protein Eng*, 14(2) :105–13, 2001.

[146] B. McConkey, V. Sobolev, and M. Edelman. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics*, 18(10) :1365–73, 2002.

[147] A. McCoy, R. Grosse-Kunstleve, L. Storoni, and R. Read. Likelihood-enhanced fast translation functions. *Acta Crystallogr D Biol Crystallogr*, 61(Pt 4) :458–64, 2005.

[148] R. Mendez, R. Leplae, L. De Maria, and S. Wodak. Assessment of blind predictions of protein-protein interactions : current status of docking methods. *Proteins*, 52(1) :51–67, 2003.

[149] R. Mendez, R. Leplae, M. Lensink, and S. Wodak. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60(2) :150–69, 2005.

[150] E. Meng, B. Shoichet, and I. Kuntz. Automated docking with grid-based energy evaluation. *J Comp Chem*, 13 :505–524, 1992.

[151] I. Mihalek, I. Res, and O. Lichtarge. A structure and evolution-guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins. *Bioinformatics*, 22(2) :149–56, 2006.

[152] J. Mintseris and Z. Weng. Atomic contact vectors in protein-protein recognition. *Proteins*, 53(3) :629–39, 2003.

[153] J. Mitchell, R. Kerr, and L. Ten Eyck. Rapid atomic density methods for molecular shape characterization. *J Mol Graph Model*, 19(3-4) :325–30, 388–90, 2001.

[154] J. Moult, K. Fidelis, B. Rost, T. Hubbard, and A. Tramontano. Critical assessment of methods of protein structure prediction (CASP)–round 6. *Proteins*, 61 Suppl 7 :3–7, 2005.

[155] M. Mucchielli-Giorgi, S. Hazout, and P. Tuffery. PredAcc : prediction of solvent accessibility. *Bioinformatics*, 15(2) :176–7, 1999.

[156] M. Mucchielli-Giorgi, S. Hazout, and P. Tuffery. Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins*, 46(3) :243–9, 2002.

[157] P. Munson and R. Singh. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. *Protein Sci*, 6(7) :1467–81, 1997.

[158] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia. SCOP : a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4) :536–40, 1995.

[159] H. Myllykallio, G. Lipowski, D. Leduc, J. Filee, P. Forterre, and U. Liebl. An alternative flavin-dependent mechanism for thymidylate synthesis. *Science*, 297(5578) :105–7, 2002.

[160] K. Nadassy, I. Tomas-Oliveira, I. Alberts, J. Janin, and S. Wodak. Standard atomic volumes in double-stranded DNA and packing in protein–DNA interfaces. *Nucleic Acids Res*, 29(16) :3362–76, 2001.

[161] K. Nadassy, S. Wodak, and J. Janin. Structural features of protein-nucleic acid recognition sites. *Biochemistry*, 38(7) :1999–2017, 1999.

[162] J. Navaza. Implementation of molecular replacement in AMoRe. *Acta Crystallogr D Biol Crystallogr*, 57(Pt 10) :1367–72, 2001.

[163] R. Norel, D. Fischer, H. Wolfson, and R. Nussinov. Molecular surface recognition by a computer vision-based technique. *Protein Eng*, 7(1) :39–46, 1994.

[164] R. Norel, S. Lin, H. Wolfson, and R. Nussinov. Molecular surface complementarity at protein-protein interfaces : the critical role played by surface normals at well placed, sparse, points in docking. *J Mol Biol*, 252(2) :263–73, 1995.

[165] R. Norel, D. Petrey, H. Wolfson, and R. Nussinov. Examination of shape complementarity in docking of unbound proteins. *Proteins*, 36(3) :307–17, 1999.

[166] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu. *Spatial Tessellations : Concepts and Applications of Voronoi Diagrams*. Wiley series in probability and statistics. John Wiley & Sons, 2000.

[167] C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton. CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8) :1093–108, 1997.

[168] E. Paci and M. Marchi. Intrinsic compressibility and volume compression in solvated proteins by molecular dynamics simulation at high pressure. *Proc Natl Acad Sci U S A*, 93(21) :11609–14, 1996.

[169] P. Pancoska and T. Keiderling. Systematic comparison of statistical analyses of electronic and vibrational circular dichroism for secondary structure prediction of selected proteins. *Biochemistry*, 30(28) :6885–95, 1991.

[170] L. Pauling and R. Corey. The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A*, 37(5) :251–6, 1951.

[171] L. Pauling, R. Corey, and H. Branson. The structure of proteins ; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, 37(4) :205–11, 1951.

[172] K. Peters, J. Fauck, and C. Frommel. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J Mol Biol*, 256(1) :201–13, 1996.

[173] E. Pettersen, T. Goddard, C. Huang, G. Couch, D. Greenblatt, E. Meng, and T. Ferrin. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.*, 25 :1605–1612, 2004.

[174] A. Pickles. *Encyclopedia of social measurement*, chapter Missing data, problems and solutions, pages 689–694. Elsevier, 2005.

[175] B. Pierce, W. Tong, and Z. Weng. M-ZDOCK : a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*, 21(8) :1472–8, 2005.

[176] H. Ponstingl, K. Henrick, and J. Thornton. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, 41(1) :47–57, 2000.

[177] J. Pontius, J. Richelle, and S. Wodak. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J Mol Biol*, 264(1) :121–36, 1996.

[178] A. Poupon. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr Opin Struct Biol*, 14(2) :233–41, 2004.

[179] R. Pribic, I. van Stokkum, D. Chapman, P. Haris, and M. Bloemendal. Protein secondary structure from Fourier transform infrared and/or circular dichroism spectra. *Anal Biochem*, 214(2) :366–78, 1993.

[180] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In J. W. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconson, USA, July 24-27, 1998*, pages 445–453. Morgan Kaufmann, 1998.

[181] M. Quillin and B. Matthews. Accurate calculation of the density of proteins. *Acta Crystallogr D Biol Crystallogr*, 56 (Pt 7) :791–4, 2000.

[182] R Development Core Team. R : A language and environment for statistical computing. 2004. ISBN 3-900051-00-3.

[183] I. Res and O. Lichtarge. Character and evolution of protein-protein interfaces. *Phys Biol*, 2(1-2) :S36–43, 2005.

[184] I. Res, I. Mihalek, and O. Lichtarge. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, 21(10) :2496–501, 2005.

[185] F. Richards. The interpretation of protein structures : total volume, group volume distributions and packing density. *J Mol Biol*, 82(1) :1–14, 1974.

[186] F. Richards. Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol*, 115 :440–64, 1985.

[187] D. Ritchie. Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins*, 52(1) :98–106, 2003.

[188] M. Roche, A. J., Y. Kodratoff, and M. Sebag. Learning interestingness measures in terminology extraction. a ROC-based approach. In J. Hernández-Orallo, C. Ferri, N. Lachiche, and P. A. Flach, editors, *ROC Analysis in Artificial Intelligence, 1st International Workshop, ROCAI-2004, Valencia, Spain, August 22, 2004*, ROCAI, pages 81–88, 2004.

[189] F. Rodier, R. Bahadur, P. Chakrabarti, and J. Janin. Hydration of protein-protein interfaces. *Proteins*, 60(1) :36–45, 2005.

[190] G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716) :834–8, 1985.

[191] B. Rost, C. Sander, and R. Schneider. Redefining the goals of protein secondary structure prediction. *J Mol Biol*, 235(1) :13–26, 1994.

[192] A. Roussel and C. Cambillau. TURBO-FRODO. In Silicon Graphics Committee, editor, *Silicon Graphics Geometry Partners Directory*, pages 77–78, Silicon Graphics, Mountain View, CA, 1989.

[193] B. Sandak, R. Nussinov, and H. Wolfson. A method for biomolecular structural recognition and docking allowing conformational flexibility. *J Comput Biol*, 5(4) :631–54, 1998.

[194] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1) :56–68, 1991.

[195] M. Schaefer, C. Bartels, F. Leclerc, and M. Karplus. Effective atom volumes for implicit solvent models : comparison between Voronoi volumes and minimum fluctuation volumes. *J Comput Chem*, 22(15) :1857–1879, 2001.

[196] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. Wolfson. Geometry-based flexible and symmetric protein docking. *Proteins*, 60(2) :224–31, 2005.

[197] D. Schneidman-Duhovny, Y. Inbar, V. Polak, M. Shatsky, I. Halperin, H. Benyamini, A. Barzilai, O. Dror, N. Haspel, R. Nussinov, and H. Wolfson. Taking geometry to its edge : fast unbound rigid (and hinge-bent) docking. *Proteins*, 52(1) :107–12, 2003.

[198] B. Schoölhopf. Support vector learning, 1997.

[199] O. Schueler-Furman, C. Wang, and D. Baker. Progress in protein-protein docking : atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins*, 60(2) :187–94, 2005.

[200] M. Sebag, J. Azé, and N. Lucas. Impact studies and sensitivity analysis in medical data mining with roc-based genetic learning. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA*, pages 637–640. IEEE Computer Society, 2003.

[201] M. Sebag, J. Azé, and N. Lucas. Roc-based evolutionary learning : Application to medical data mining. In P. Liardet, P. Collet, C. Fonlupt, E. Lutton, and M. Schoenauer, editors, *Artificial Evolution, 6th International Conference, Evolution Artificielle, EA 2003, Marseilles, France, October 27-30, 2003*, volume 2936 of *Lecture Notes in Computer Science*, pages 384–396. Springer, 2004.

[202] B. Shoichet, I. Kuntz, and D. Bodian. Molecular docking using shape descriptors. *J Comp Chem*, 13 :380–397, 1992.

[203] A. Shrake and J. Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol*, 79(2) :351–71, 1973.

[204] R. Singh, A. Tropsha, and I. Vaisman. Delaunay tessellation of proteins : four body nearest-neighbor propensities of amino acid residues. *J Comput Biol*, 3(2) :213–21, 1996.

[205] G. Smith and M. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr Opin Struct Biol*, 12(1) :28–35, 2002.

[206] G. Smith and M. Sternberg. Evaluation of the 3D-Dock protein docking suite in rounds 1 and 2 of the CAPRI blind trial. *Proteins*, 52(1) :74–9, 2003.

[207] A. Soyer, J. Chomilier, J. Mornon, R. Jullien, and J. Sadoc. Voronoi tessellation reveals the condensed matter character of folded proteins. *Phys Rev Lett*, 85(16) :3532–5, 2000.

[208] M. Stolowitz. Chemical protein sequencing and amino acid analysis. *Curr Opin Biotechnol*, 4(1) :9–13, 1993.

[209] E. Sundberg, M. Urrutia, B. Braden, J. Isern, D. Tsuchiya, B. Fields, E. Malchiodi, J. Tormo, F. Schwarz, and R. Mariuzza. Estimation of the hydrophobic effect in an antigen-antibody protein-protein interface. *Biochemistry*, 39(50) :15375–87, 2000.

[210] G. Terashi, M. Takeda-Shitaka, D. Takaya, K. Komatsu, and H. Umeyama. Searching for protein-protein interaction sites and docking by the methods of molecular dynamics, grid scoring, and the pairwise interaction potential of amino acid residues. *Proteins*, 60(2) :289–95, 2005.

[211] R. Thiele, R. Zimmer, and T. Lengauer. Protein threading by recursive dynamic programming. *J Mol Biol*, 290(3) :757–79, 1999.

[212] J. Tsai and M. Gerstein. Calculations of protein volumes : sensitivity analysis and parameter database. *Bioinformatics*, 18(7) :985–95, 2002.

[213] J. Tsai, R. Taylor, C. Chothia, and M. Gerstein. The packing density in proteins : standard radii and volumes. *J Mol Biol*, 290(1) :253–66, 1999.

[214] J. Tsai, N. Voss, and M. Gerstein. Determining the minimum number of types necessary to represent the sizes of protein atoms. *Bioinformatics*, 17(10) :949–56, 2001.

[215] P. Tuffery. XmMol : an X11 and motif program for macromolecular visualization and modeling. *J. Mol. Graphics*, 13 :67–72, 1995.

[216] P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770) :623–7, 2000.

[217] A. Vagin and A. Teplyakov. A translation-function approach for heavy-atom location in macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*, 54 (Pt 3) :400–2, 1998.

[218] A. Vagin and A. Teplyakov. An approach to multi-copy search in molecular replacement. *Acta Crystallogr D Biol Crystallogr*, 56 Pt 12 :1622–4, 2000.

[219] A. van Dijk, S. de Vries, C. Dominguez, H. Chen, H. Zhou, and A. Bonvin. Data-driven docking : HADDOCK's adventures in CAPRI. *Proteins*, 60(2) :232–8, 2005.

[220] G. Voronoï. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik*, 133 :97–178, 1908.

[221] H. Wako and T. Yamato. Novel method to detect a motif of local structures in different protein conformations. *Protein Eng*, 11(11) :981–90, 1998.

[222] H. Wang. Grid-search molecular accessible surface algorithm for solving the protein docking problem. *J Comp Chem*, 12 :746–750, 1991.

[223] Y. Wang, P. K. Agarwal, P. Brown, H. Edelsbrunner, and J. Rudolph. Coarse and reliable geometric alignment for protein docking. *Pac Symp Biocomp*, 2005.

[224] J. Watson and F. Crick. Genetic implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361) :964–967, 1953.

[225] J. Watson and F. Crick. Molecular structure of nucleic acid : a structure for deoxyribose Nucleic Acid. *Nature*, 171(4356) :737–738, 1953.

[226] L. Wernisch, M. Hunting, and S. Wodak. Identification of structural domains in proteins by a graph heuristic. *Proteins*, 35(3) :338–52, 1999.

[227] J. Westbrook and P. Fitzgerald. *Structural Bioinformatics*, chapter The PDB format, mmCif formats and other data formats., pages 161–179. John Wiley & Sons, 2003.

[228] J. Westbrook, N. Ito, H. Nakamura, K. Henrick, and H. Berman. PDBML : the representation of archival macromolecular structure data in XML. *Bioinformatics*, 21(7) :988–92, 2005.

[229] K. Wiehe, B. Pierce, J. Mintseris, W. Tong, R. Anderson, R. Chen, and Z. Weng. ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. *Proteins*, 60(2) :207–13, 2005.

[230] D. Wishart, B. Sykes, and F. Richards. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J Mol Biol*, 222(2) :311–33, 1991.

[231] S. Wodak and J. Janin. Computer analysis of protein-protein interaction. *J Mol Biol*, 124(2) :323–42, 1978.

[232] S. Wodak and J. Janin. Structural basis of macromolecular recognition. *Adv Protein Chem*, 61 :9–73, 2002.

[233] S. Wodak and R. Mendez. Prediction of protein-protein interactions : the CAPRI experiment, its evaluation and implications. *Curr Opin Struct Biol*, 14(2) :242–9, 2004.

[234] H. Wolfson and R. Nussinov. Geometrical docking algorithms. A practical approach. *Methods Mol Biol*, 143 :377–97, 2000.

[235] K.-D. Zachmann, W. Heiden, M. Schlenkrich, and J. Brickmann. Topological analysis of complex molecular surfaces. *J Comp Chem*, 13 :76–84, 1992.

[236] R. J. Zauhar. A solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J. of Comp.-Aided Mol. Design.*, 9(2) :149–159, 1995.

[237] C. Zhang, S. Liu, and Y. Zhou. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Sci*, 13(2) :391–9, 2004.

[238] C. Zhang, S. Liu, and Y. Zhou. Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function. *Proteins*, 60(2) :314–8, 2005.

[239] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J Med Chem*, 48(7) :2325–35, 2005.

[240] J. Zhang. Protein-length distributions for the three domains of life. *Trends Genet*, 16(3) :107–9, 2000.

[241] H. Zhu, F. Domingues, I. Sommer, and T. Lengauer. NOXclass : prediction of protein-protein interaction types. *BMC Bioinformatics*, 7(1) :27, 2006.

[242] R. Zimmer, M. Wohler, and R. Thiele. New scoring schemes for protein fold recognition based on Voronoi contacts. *Bioinformatics*, 14(3) :295–308, 1998.

Table des figures

Liste des tableaux

Résumé

La fonction d'une protéine est souvent subordonnée à l'interaction avec un certain nombre de partenaires. L'étude de la structure tridimensionnelle de ces complexes, qui ne peut souvent se faire expérimentalement, permettrait la compréhension de nombreux processus cellulaires.

Le travail présenté ici se compose de deux parties. La première traite de la mise en place d'une fonction de score pour l'amarrage protéine-protéine et la deuxième de l'étude cristallographique d'une protéine tétramérique qui est une cible antibiotique potentielle : la thymidylate synthase X de *Paramecium bursaria* Chlorella virus.

La modélisation des complexes protéine-protéine ou *docking* comporte deux étapes successives : d'abord, un grand nombre de conformations sont générées, puis une fonction de score est utilisée pour les classer. Cette fonction de score doit prendre en compte à la fois la complémentarité géométrique des deux molécules et les propriétés physico-chimiques des surfaces en interaction.

Nous nous sommes intéressés à la seconde étape à travers le développement d'une fonction de score rapide et fiable. Ceci est possible grâce à la tessellation de Voronoï de la structure tridimensionnelle des protéines. En effet, les tessellations de Voronoï ou de Laguerre se sont avérées être de bons modèles mathématiques de la structure des protéines. En particulier, cette formalisation permet de faire une bonne description de l'empilement et des propriétés structurales des résidus.

Cette modélisation rend compte l'empilement des résidus à l'interface entre deux protéines. Ainsi, il est possible de mesurer un ensemble de paramètres sur des complexes protéine-protéine dont la structure est connue expérimentalement et sur des complexes leurres générés artificiellement. Ces paramètres, sont la fréquence d'apparition des résidus ou des paires de résidus, les volumes des cellules de Voronoï, les distances entre les résidus en contact à l'interface, la surface de l'interface et le nombre de résidus à l'interface. Ils ont été utilisés en entrée de procédures d'apprentissage statistique. Grâce à ces procédures (apprentissage logistique, séparateurs à vaste marge (SVM) et algorithmes génétiques), on peut obtenir des fonctions de score efficaces, capables de séparer les leurres des structures réelles.

Dans un deuxième temps, j'ai déterminé expérimentalement la structure de la thymidylate synthase X, cible antibiotique de choix. La thymidylate synthase X est une flavoprotéine qui a été découverte récemment. Elle intervient dans la synthèse du dTMP chez la plupart des procaryotes mais n'existe pas chez les eucaryotes supérieurs. Cette protéine catalyse le transfert de methyle du tétrahydrofolate vers le dUMP grâce à son cofacteur le FAD et au NADPH qui intervient comme substrat.

La structure tridimensionnelle de l'homotétramère de la thymidylate synthase X en présence de son cofacteur, le FAD, a été résolue à 2.4 Å par remplacement moléculaire. Comme pour les structures de thymidylate synthase X de *Thermotoga maritima* et de *Mycobacterium tuberculosis* précédemment résolues, le monomère se compose d'un coeur de feuillets β et de deux hélices α à son extrémité. Le site actif se trouve à l'interface de trois monomères, la partie isoalloxazine

du FAD étant accessible au solvant et proche d'une longue boucle flexible. La fixation du FAD dans cette structure est légèrement différente de celles déjà observées par la conformation de la partie adénine.

Cette structure, associée aux études de mutagénèse dirigée de nos collaborateurs, a permis de mettre évidence des résidus jouant un rôle majeur lors de la catalyse.

Mots-clés: complexes protéines-protéines, interactions, tessellation de Voronoï, procédures d'apprentissage, thymidylate synthase X.

Abstract

The function of a protein is often subordinated to its interaction with one or many partners. Yet, the tridimensional structure study of this complexes, that can't be done experimentally, would permit the understanding of many cellular processes.

This work contains two parts. The first part concerns the setting up of a scoring function for protein-protein docking and the second part concerns the crystallographic structure study of a tetrameric protein : the *Paramecium Bursaria* Chlorella Virus thymidylate synthase X, a potential antibacterial target.

Docking of protein-protein complexes consists in two successive steps : first a large number of putative conformations are generated, then a scoring function is applied to rank them. This scoring function has to take into account both geometric complementarity of the two molecules and physico-chemical properties of surfaces in interaction.

We addressed the second step of this problem through the development of a quick and reliable scoring function. This was done using Voronoi tessellation of the tridimensional structure of the proteins. Voronoi or Laguerre tessellations were shown to be good mathematical models of protein structure. In particular, this formalization leads to a good description of structural properties of the residues.

This modeling illustrates the packing of the residues at the interface between two proteins. Thus, it is possible to measure a set of parameters, on protein-protein complexes whose structure is known, and on decoys. These parameters are frequencies of residues and pair frequencies of the residues at the interface, volumes of Voronoi cells, distances between residues at the interface, interface area and number of residues at the interface. They were used as input in statistical machine learning procedures (logistic learning, support vector machines (SVM) and genetic algorithms). These led to efficient scoring functions, able to separate native structures from decoys.

In the second part, I describe the experimental determination of thymidylate synthase X tridimensionnal structure, an interesting antibacterial target.

Thymidylate synthase X is a flavoprotein discovered recently. It plays a key role in the synthesis of dTMP in most of the prokaryotic organisms, but does not exist in superior eukaryotic organisms. This protein catalyses the methyl transfer from tetrahydrofolate to dUMP using FAD as a cofactor and NADPH as substrate.

The tridimensional structure of ThyX homotetramer with its cofactor, FAD, was solved at 2.4Å by molecular replacement. As shown in the *Thermotoga maritima* and *Mycobacterium tuberculosis* ThyX structures, the monomer contains a core of β sheets and two α helices at its extremity. The active site is at the interface between three monomers, the isoalloxazine part of FAD being accessible to the solvent and close to a long flexible loop. FAD binding in this structure is a little different from those already observed, especially its the adenine part.

This structure, in association with directed mutagenesis experiments made by our collaborators, revealed residues playing a key role during the catalysis.

Keywords: protein-protein complexes, interactions, Voronoi tessellation, machine learning procedures, thymidylate synthase X.