HAL
open science

Exploitation des algorithmes génétiques pour la prédiction de structure de complexe protéine-protéine

Thomas Bourquard

To cite this version:

Thomas Bourquard. Exploitation des algorithmes génétiques pour la prédiction de structure de complexe protéine-protéine. Intelligence artificielle [cs.AI]. Université Paris Sud - Paris XI, 2009. Français. NNT : . tel-00782396

HAL Id: tel-00782396

https://theses.hal.science/tel-00782396v1

Submitted on 30 Jan 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre :

Thèse

de

L'Université Paris–Sud 11

présentée en vue de l'obtention du

Doctorat de l'Université Paris–Sud 11

École Doctorale Innovation Thérapeutique

Par

Thomas BOURQUARD

Exploitation des algorithmes génétiques pour la prédiction de structures protéine-protéine

M.	Daniel Gautheret	Professeur	Président du jury
M.	Raphaël Guerois	Chargé de Recherche	Rapporteur
M.	David Ritchie	Professeur	Rapporteur
Mme	Alessandra Carbone	Professeur	Examinatrice
M.	Alain Guénoche	Directeur de recherche	Examinateur
Mme	Anne Poupon	Directeur de recherche	Directrice de thèse
M.	Jérôme Azé	Maître de Conférences	co-Directeur de thèse

Équipe de Bioinformatique, Laboratoire de Recherche en Informatique, U.M.R. CNRS 8623
Équipe de Génomique Structurale, Institut de Biochimie et de Biophysique Moléculaire et Cellulaire, U.M.R. CNRS 8619

Table des matières

Liste des figures 9

Liste des tableaux 11

Liste des algorithmes 11

I Introduction et état de l'art 13

1 Introduction 15

1.1 La structure des protéines . . . 15

1.1.1 Introduction . . . 15

1.1.2 La structure primaire . . . 15

1.1.3 La structure secondaire . . . 18

1.1.4 La structure tertiaire . . . 20

1.1.5 La structure quaternaire . . . 22

1.2 Les complexes protéine-protéine . . . 23

1.2.1 Nature de l'interaction protéine-protéine . . . 23

1.2.2 Détection expérimentale . . . 24

1.3 L'amarrage protéine-protéine . . . 27

1.3.1 Principe et partitionnement du problème . . . 27

1.3.2 Les algorithmes . . . 28

1.3.3 CAPRI : une expérience à ne pas manquer . . . 30

1.4 Le diagramme de Voronoï et les constructions dérivées . . . 31

1.4.1 Un peu de géométrie... . . . 31

1.4.2 Analyse de la structure des protéines . . . 33

1.4.3 Assemblages . . . 35

1.5 L'apprentissage supervisé de concepts et les algorithmes génétiques 37
1.5.1 Principe des algorithmes évolutionnistes 37
1.5.2 Application à la résolution de problèmes 38
1.5.3 Opérateurs génétiques 39
1.5.4 Avantages et inconvénients des algorithmes évolutionnistes 40

II Matériel et Méthodes **43**

2 Méthode et logiciel **45**
2.1 Le diagramme de Voronoï et ses constructions dérivées 45
2.1.1 Définitions et notations 45
2.1.2 Triangulation régulière et triangulation de Delaunay 47
2.1.3 Applications aux protéines 49
2.1.4 Paramètres pour l'apprentissage 50
2.1.5 évaluation de la qualité des modèles 53
2.2 Échantillon d'apprentissage 55
2.3 Algorithme d'amarrage 70
2.3.1 Choix des points d'amarrage 70
2.3.2 Générations des conformations candidates 72
2.4 Algorithme génétique et courbe de *ROC* 76
2.4.1 Principe de la méthode 76
2.4.2 Individus, algorithme et opérateurs d'évolution 77
2.4.3 Fonctions d'adaptation 79
2.4.4 Fonction de score 79
2.4.5 Opérateurs mutation et croisement 81
2.5 Partitionnement des données d'apprentissage 82
2.5.1 Méta-attributs 82
2.5.2 Algorithme de partitionnement 83
2.5.3 Validation sur les cibles CAPRI 87

III Résultats **89**

3 Résultats et discussion **91**

3.1 Comparaison des tessellations de Voronoï et Laguerre dans le contexte de l'amarrage protéine-protéine . . . 91
3.1.1 Construction du diagramme de Laguerre . . . 92
3.1.2 Fonctions de score . . . 93
3.1.3 Étude comparée des tesselations de Voronoï et Laguerre pour la modélisation de complexe protéine-protéine . . . 95
3.1.4 Résultats et perspectives . . . 103
3.2 Affinement de la méthode d'évaluation des conformations candidates par algorithme génétique . . . 103
3.2.1 Le cœur et la couronne . . . 104
3.2.2 Partitionnement des données d'apprentissage . . . 106
3.2.3 Résultats complémentaires . . . 116
3.2.4 Conclusions et perspectives . . . 117

IV Conclusion Générale **119**

Bibliographie **133**

Table des figures

1.1 Schéma de la liaison peptique formée entre un acide aspartique et une phényalanine. 16

1.2 Contraintes spatiales des liaisons peptidiques telle que décrite par Linus Pauling [113]. 16

1.3 Regroupement des acides aminés en fonction de leurs propriétés physico-chimiques. *Chaque couleur repésente un groupe.* 17

1.4 L'hélice α telle que représentée dans l'article original de Linus Pauling [114]. . . . 19

1.5 Brins β tels que représentés dans l'article original de Linus Pauling [114]. 20

1.6 Prédictions de structures obtenues à l'aire du logiciel Rosetta [16] lors de la session CASP6 Image originale en couverture du journal **PROTEINS** : Structure, Function and Bioinformatics, volume 61 du 26 septembre 2005. 22

1.7 Interactions transitoires et cycle des protéines G. La liaison du récepteur à son ligand provoque la libération du complexe $G_\alpha - G_\beta - G_\gamma$−GDP. Le GDP est échangé contre un GTP, et la sous-unité G_α−GTP se dissocie du reste du complexe. C'est ce sous-complexe qui va, à son tour, aller activer d'autres protéines dans la cellule, ce qui mènera finalement à la production par la cellule d'une réponse biologique adaptée. 24

1.8 Principe de la méthode de détection de complexes protéiques par double-hybride dans la levure. La protéine Gal4 est l'activateur naturel des différents gènes intervenant dans le métabolisme du galactose. Elle agit en se fixant au niveau des UASG *Upstream Activating Sequence GAL* régulant la transcription. Les deux protéines X et Y dont on veut tester l'interaction sont fusionnées, l'une au domaine de fixation Gal4 (DNA binding Domain), l'autre au domaine de Gal4 activant la transcription (Activation Domain). Si une interaction entre X et Y existe, il se forme alors un activateur de transcription DBD-X/Y-AD. Cet activateur peut se lier au niveau des séquences UASG permettant la transcription d'un gène rapporteur par l'ARN polymérase II. 25

1.9 Principe de la méthode de détection de complexes protéiques par TAP-tag dans la levure. 26

1.10 Le problème de l'amarrage : Comment associer les partenaires A et B ensemble ? Les conformations putatives sont-elles suceptibles d'exister *in vivo* ? 27

1.11 Rappel des différentes tessellations. 32

1.12 Comparaison des volumes obtenus pour une décomposition de Voronoï atomique non-pondérée, par la méthode B de Richard (vert), décomposition de Voronoï non pondérée à partir des centres géométriques (bleu) et pour une décomposition de Laguerre à partir des centres géométriques (bleue). 34

1.13 Valeurs des paramètres mesurés sur des structures natives et non-natives. Pour ces mesures, le diagramme de Voronoï construit à partir des centres géométriques des chaînes latérales a été utilisé. Les paramètres sont mesurés sur le cœur de l'interface. 36

1.14 Principe des algorithmes évolutionnistes. 38

2.1 Diagramme de Voronoï. La cellule de Voronoï d'un nœud (en gris) peut également être vue comme sa zone d'influence. 46

2.2 La triangulation de Delaunay. Cette triangulation est construite, soit à partir de la tessellation de Voronoï, soit, comme illustré ici, en construisant les triangles dont les sommets sont les nœuds, et dont les cercles circonscrits ne contiennent aucun autre sommet. 47

2.3 La tessellation de Laguerre (en rose) et son dual, la triangulation régulière (en bleu). Les cercles figurent les poids attribués aux nœuds. 48

2.4 Le centre géométrique de la chaîne latérale (C_α compris) varie très peu entre les conformations A et B de l'aspartate. 49

2.5 Exemples de cellules calculés, dans la protéine et pour les mêmes acides aminés, en utilisant le centre géométrique des chaînes latérales (à gauche) ou le C_α (à droite). 50

2.6 Solvatation du complexe 1BTH et fermeture de ses cellules de Voronoï. 51

2.7 Voisinage au sens de Voronoï. Les acides aminés les plus proches ne sont pas nécessairement voisins. Ici, les acides aminés 1 et 2 sont plus éloignés l'un de l'autre que les acides aminés 3 et 4, ils ne sont pourtant pas voisins au sens de Voronoï. 52

2.8 Description des critères d'évaluation des prédictions dans l'expérience CAPRI *Critical Assessment of PRediction Interactions*. 54

2.9 Points d'ancrages à la surface de la protéine 1BTH. C'est à partir de ces points d'ancrage que sont calculés les vecteurs normaux. 71

2.10 Principe de la méthode de génération des conformations candidates. Les vecteurs normaux aux différents points d'ancrage de la surface sont construits pour les deux partenaires à assembler. Pour un couple de vecteurs appartenant chacun à l'un des partenaire, on effectue une translation du partnaire L afin d'amener les extrémités des deux vecteurs l'une sur l'autre. Puis on effectue une rotation de L afin d'amener les deux vecteurs face-à-face. Enfin, on applique une rotation à L autour de l'axe défini par les deux vecteurs, une conformation candidate étant générée tous les 5°. 72

2.11 Constuction de conformations “presque-natives” du complexe 1a4y. La conformation native est tracée en vert, des conformations qui lui sont proches, générées par notre algorithme sont figurées dans les autres couleurs. 75

2.12 Principe de l’apprentissage par algorithme génétique. 76

2.13 Courbe de ROC. La courbe de ROC est la fonction qui donne le taux de vrais positifs en fonction du taux de faux positifs. Pour une fonction aléatoire, il y a en moyenne autant de vrais positifs que de faux positifs, l’aire sous la courbe (AUC) est alors de 0,5. Plus la fonction est efficace, et plus on se rapproche de l’axe des ordonnées. Ainsi dans la figure, la fonction qui donne la courbe A est plus efficace que celle qui donne la courbe B. Cette différence peut être mesurée par l’aire sous la courbe, qui est plus grande pour A que pour B, et qui vaut 1 dans le cas idéal. 80

3.1 Comparaison des diagrammes de Voronoï et de Laguerre pour un ensemble de points. Le diagramme de Voronoï est représenté en lignes fines et le diagramme de Laguerre en lignes plus épaisses. Pour le diagramme de Laguerre, on attribue à chaque point un poids, qui a été ici choisi en fonction de la taille de l’acide aminé représenté par le point (en code à une lettre). On peut constater que les cellules correspondant aux acides aminés volumineux, comme le tryptophane (W) ou l’arginine (R) sont plus grandes dans le diagramme de Laguerre. On peut voir également que la cellule correspondant à la glycine (G, en noir) disparaît totalement dans le diagramme de Laguerre. 93

3.2 Fronts de Pareto des apprentissages en 10−validation croisée. Cette courbe donne la précision en fonction du rappel. La précision correspond à Vrais Positifs / (Faux Positifs + Vrais Positifs), et le rappel à Vrais Positifs / (Vrais Positifs + Faux Négatifs). Pour chaque type de fonction ou méthode d’apprentissage (L : fonction linéaire, NL : fonction non linéaire du premier degré, SVM : Support vector machine) les cibles sont classées suivant la somme des rangs (sumRank), la somme des scores (sumScore) ou la médiane (Med). 94

3.3 Rang de sortie de la native. Des ensembles comprenant une structure native et 81 conformations non-natives générées à partir des partenaires, sont évalués par les fonctions apprises sur l’ensemble du jeu de données et l’ensemble de l’interface (apprentissage global), ou sur les partitions. Dans les deux cas l’apprentissage est réalisé en leave-one-out. L’histogramme donne le nombre de natives classées entre 1 et 5, 6 et 10, etc. 107

Liste des tableaux

2.1 Liste des complexes natifs utilisés dans l'apprentissage. À chaque complexe est associé sa classe telle que Mintseris le référence à savoir E désigne les complexes Enzyme-Substrat A pour Anticorps-Antigène AB pour les complexes Antigène-Anticorps liés et O pour *Others* désignant tous les complexes n'appartenant pas aux trois classes précédentes. Sont définies le nom des chaînes et un critère lié (Bound ou "B"), non-lié (Unbound ou "U"), voire modélisé lorsque la structure du complexe est connu mais la structure de l'un des deux partenaires est inconnue (Modelized ou "M"). 69

2.2 Densité moyenne de compaction pour les acides aminés définis par Galzitkaya *et al.*. 85

3.1 Rang de la structure native, paramètres évalués sur le cœur de l'interface. 105

3.2 Rang de la structure native, paramètres évalués sur l'ensemble de l'interface (cœur et couronne). 106

3.3 $TopN$ et rang correspondant (entre parenthèses) de la conformation native et de la conformation au moins acceptable la mieux classée, le classement est fait par ordre décroissant de $Top5$, $Top10$ ou $Top15$. Apprentissage sur les partitions, modèle Voronoï, paramètres calculés sur l'ensemble de l'interface. 116

3.4 $TopN$ et rang correspondant (entre parenthèses) de la conformation native et de la conformation au moins acceptable la mieux classé, le classement est fait par ordre décroissant de $Top5$, $Top10$ ou $Top15$. Apprentissage sur les partitions, modèle Laguerre, paramètres calculés sur l'ensemble de l'interface. 117

Liste des Algorithmes

1 Algorithme "enveloppeConvexe($prot$)" permettant de déterminer si un acide aminé se trouve dans une région convexe ou concave. Une implantation efficace permet d'arrêter les boucles sur r_a, r_b et r_c dès que r est retiré de l'ensemble *convexes*. . 70

2 Algorithme d'apprentissage($\mathcal{E}$, μ, λ, n_{max}, $best_fitness$). 78

3 Algorithme de partitionement (c, C δ). 84

Première partie

Introduction et état de l'art

Chapitre 1

Introduction

1.1 La structure des protéines

1.1.1 Introduction

Les protéines sont des macromolécules biologiques dont le rôle est primordial dans tous les processus du vivant, de la machinerie cellulaire au maintien des parois cellulaires, ou à la structure des tissus osseux. La fonction d'une protéine est entièrement dépendante de sa structure tridimensionnelle, qui peut être décrite à quatre niveaux de complexité et d'organisation : les structures primaire, secondaire, tertiaire et quaternaire. La structure tridimensionnelle d'une protéine, y compris ses interactions avec différents partenaires tels que des petites molécules ou d'autres macromolécules, peut être déterminée expérimentalement par cristallographie aux rayons X ou par RMN. Cependant, même si ces techniques ont connu des avancées fondamentales, notamment grâce aux projets de génomique structurale, cette détermination reste délicate voire impossible. De plus, le nombre considérable de protéines connues, et la combinatoire découlant des interactions entre elles, et avec d'autres molécules, rendent inenvisageable le recours systématique à l'expérimentation.

L'approche *in silico,* qui consiste à prédire, et non plus à déterminer, la structure tridimensionnelle des protéines et des complexes, peut permettre de répondre à cette problématique. La modélisation *ab initio* permet aujourd'hui de prédire, avec un succès de plus en plus grand, le repliement d'une protéine isolée à partir de sa séquence. D'autre part, les efforts d'une communauté grandissante de chercheurs, communauté dans laquelle nous nous plaçons, portent sur la prédiction de la conformation des complexes macromoléculaires à partir des structures, déterminées expérimentalement ou prédites, des partenaires isolés.

1.1.2 La structure primaire

Une protéine est un assemblage linéaire d'acides aminés, appelé séquence primaire, ou chaîne polypeptidique.

Un acide aminé est composé d'un groupement NH2, un groupement carboxylique COOH et d'une chaîne latérale, reliés par un atome de carbone appelé carbone C_α. La liaison peptidique résulte de la réaction entre la fonction carboxylique COOH d'un acide aminé R1 avec la fonction

NH2 d'un acide aminé R2, qui libère secondairement une molécule d'eau (voir Figure 1.1).

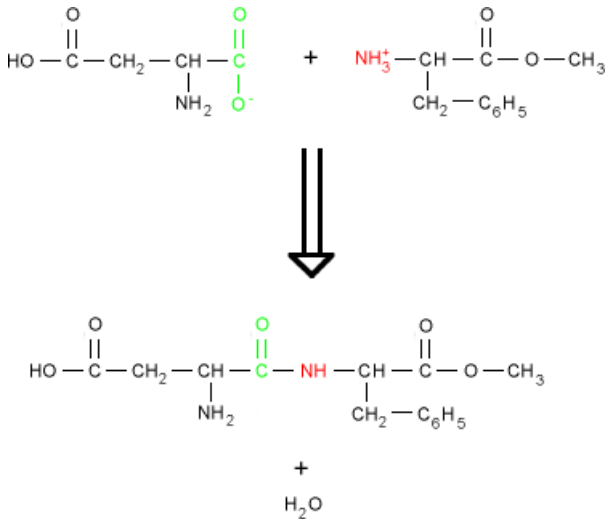

FIG. 1.1 – *Schéma de la liaison peptique formée entre un acide aspartique et une phényalanine.*

La nature de cette liaison impose par mésomérie 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 du second résidu sont coplanaires (voir Figure 1.2).

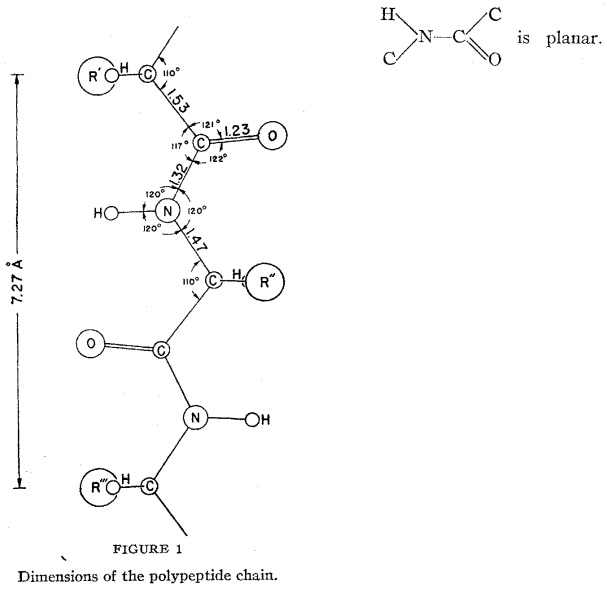

FIG. 1.2 – *Contraintes spatiales des liaisons peptidiques telle que décrite par Linus Pauling [113].*

Il existe 20 acides aminés principaux, se différenciant par leur chaîne latérale. En fonction des propriétés physico-chimiques de ces chaînes latérales, les acides aminés peuvent être regroupés en grands types : hydrophobes, aromatiques, polaires non chargés, chargés positivement, chargés négativement, et petits acides aminés (voir Figure 1.3).

Des modifications post-traductionnelles, c'est-à-dire survenant après l'assemblage des acides

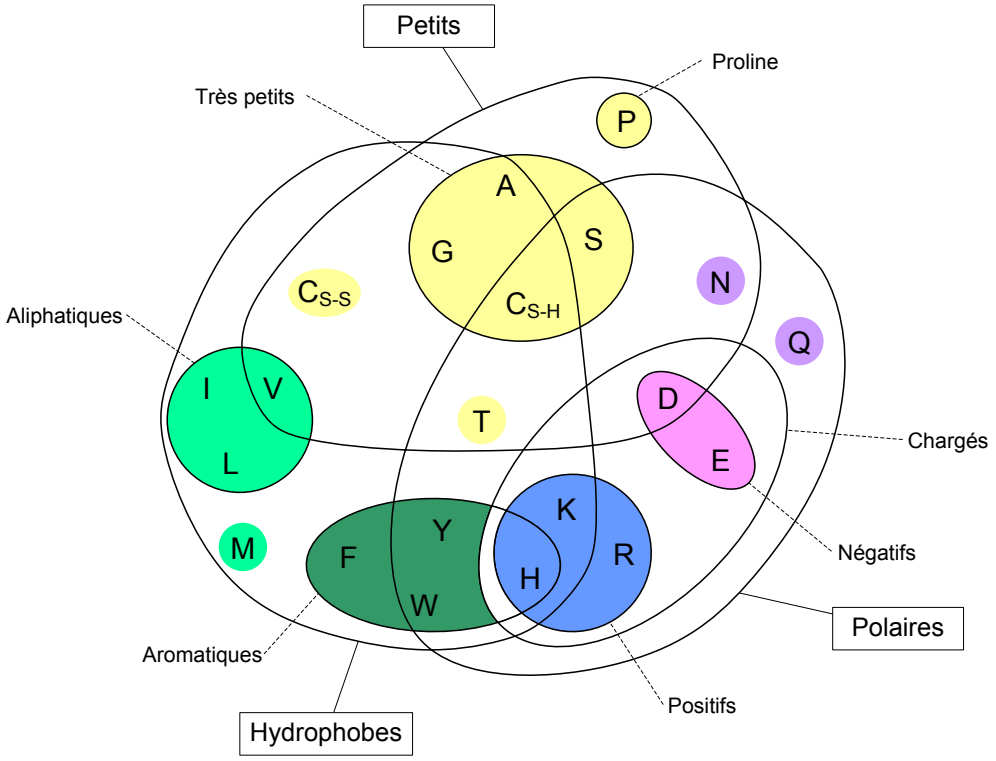

FIG. 1.3 – *Regroupement des acides aminés en fonction de leurs propriétés physico-chimiques.* *Chaque couleur repésente un groupe.*

aminés par le ribosome à partir de l'ARN messager, processus appelé traduction, peuvent affecter la séquence primaire. Ainsi certaines parties de la chaîne protéique, telles que les séquences d'adressage, peuvent être clivées. C'est essentiellement au niveau des chaînes latérales que les modifications sont observées. On peut par exemple citer la formation de ponts disulfures, la glycosilation ou l'acétylation.

Expérimentalement, la détermination de la structure primaire d'une protéine est le plus souvent réalisée de manière indirecte par le séquençage des nucléotides du gène qui code pour cette protéine, et de nombreuses techniques peuvent être utilisées pour le séquençage à haut débit. Si la taille de ces protéines est inférieure à 150 acides aminés, il est possible d'utiliser la spectrométrie de masse. Pour des longueurs de chaînes supérieures, elle peut être déterminée par micro séquençage et digestion enzymatique ou chimique en tronçons de longueur inférieure à 30 acides aminés.

1.1.3 La structure secondaire

La chaîne polypeptidique est capable de former des arrangements locaux périodiques, appelés structures secondaires. Ces structures particulières sont stabilisées par un ensemble de liaisons hydrogène entre les acides aminés proches dans la chaîne polypeptique. Les deux structures secondaires les plus courantes furent découvertes par Linus Pauling et Robert Carey : L'hélice α et le feuillet plissé β (voir Figures 1.5 et 1.4). Ces deux organisations, outre le fait de maximiser le nombre d'interactions non-covalentes entre résidus de la séquence, minimisent également les gênes stériques et les répulsions électrostatiques entre chaînes latérales.

La structure hélicoïdale est stabilisée par des liaisons hydrogènes entre les atomes impliqués dans une liaison peptidique et ceux distants de seulement 3.5 résidus en moyenne, présents en amont et en aval dans la chaîne polypeptidique (voir Figure 1.4).

À la différence de l'hélice α, le squelette de chaque brin d'un feuillet β possède une conformation plissée, stabilisée également par un ensemble de liaisons hydrogènes, qui ne sont pas colinéaires comme dans le cas d'une hélice α, mais orthogonales à la chaîne polypeptidique. Un brin β n'est donc stable qu'associé à au moins un autre brin β. Les brins de chaque feuillet peuvent être orientés tous dans le même sens, on parlera alors de feuillet parallèle, ou bien positionnés alternativement dans un sens et dans l'autre, on parlera alors de feuillet anti-parallèle (voir Figure 1.5).

Dans une protéine, en moyenne 50% des acides aminés sont impliqués dans l'un de ces deux types de structures secondaires. Les autres acides aminés forment des structures secondaires non régulières telles que les coudes ou les boucles. L'organisation tridimensionnelle et l'agencement de ces deux grands motifs structuraux ont conduit à classer les protéines en cinq catégories :

- Tout-α
- Tout-β
- α/β (en alternance)
- $\alpha + \beta$ (correspondant aux structures n'alternant pas de manière régulière mais contenant des hélices α et des feuilles β)
- Protéines composées de liaisons croisées ou domaines de réticulation, ne possédant pas ou peu d'éléments de structure secondaire régulière

Il est possible de déterminer expérimentalement la composition globale en structures secon-

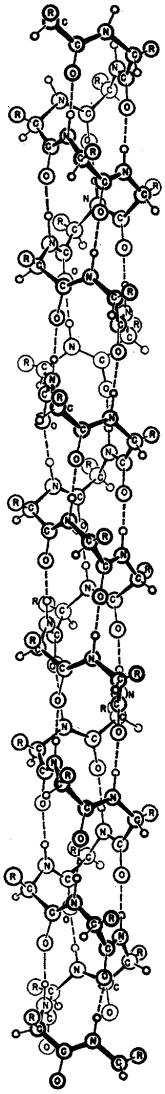

FIGURE 2
The helix with 3.7 residues per turn.

FIG. 1.4 – *L'hélice α telle que représentée dans l'article original de Linus Pauling [114].*

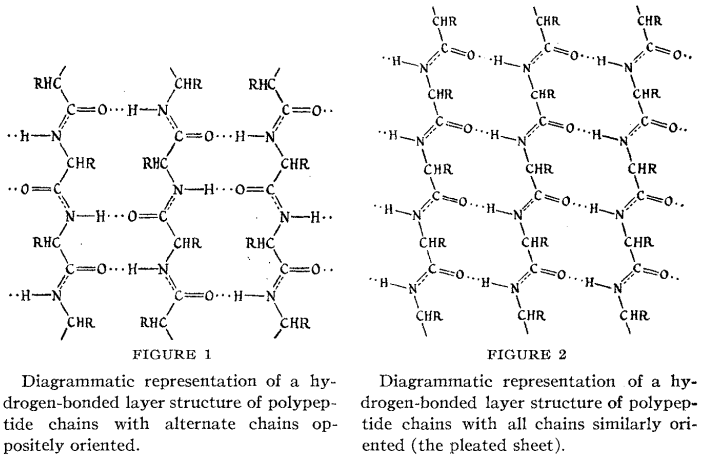

FIG. 1.5 – *Brins β tels que représentés dans l'article original de Linus Pauling [114].*

daires régulières d'une protéine en solution purifiée par différentes méthodes. On citera entre autres le dichroïsme circulaire vibrationnel, la spectroscopie infra-rouge, ou encore l'analyse des déplacements chimiques en RMN. Si ces méthodes permettent d'évaluer une composition globale en motifs structuraux, seule la résolution de la structure tridimensionnelle de la protéine permet de déterminer avec précision la position dans la séquence de ces différents motifs.

Un certain nombre de logiciels permet la prédiction de ces motifs répétés [44, 122]. L'expérience CASP[1] [13]) a permis de mettre en évidence que, dans plus de 75% des cas, ces prédictions étaient exactes. Ces logiciels ne s'appliquent cependant qu'aux protéines de petite taille (généralement moins de 150 acides aminés).

1.1.4 La structure tertiaire

La structure tertiaire décrit le repliement dans l'espace des éléments de structure secondaire d'une chaîne polypeptidique unique. Dans les conditions physiologiques, ce repliement est unique, spontané ou aidé dans sa maturation par d'autres protéines : les chaperonnes.

Les interactions servant à stabiliser la structure secondaire sont plus variées que celles qui stabilisent la structure secondaire :

- Interactions covalentes par l'intermédiaire de ponts disulfures.
- Interactions non covalentes entre les résidus constitutifs du cœur de la protéine par l'intermédiaire de liaisons hydrogène, électrostatiques ou de Van der Walls.
- Interactions avec des molécules de solvant [121], ou avec d'autres ions ou cofacteurs (hème, flavine, FAD, ...), notamment à la surface.

Si les petites protéines (moins de 200 acides aminés) se replient généralement autour d'un seul cœur, les protéines de plus grande taille peuvent constituer plusieurs ensembles, que l'on appellera alors domaines structuraux, reliés par des espaceurs qui peuvent ou non être flexibles.

Deux méthodes ont été particulièrement exploitées pour déterminer la structure tridimensionnelle des protéines : la spectroscopie par Résonance Magnétique Nucléaire (RMN) et la diffrac-

[1]CASP : **C**ritical **A**ssessment of Methods of Protein **S**tructure **P**rediction.

tion aux rayons X (cristallographie). Historiquement, la première structure résolue fut celle de la myoglobine par Max Perutz and J.C Kendrew [70]. Depuis, de nombreuses structures ont pu être déterminées et sont déposées dans la Protein Data Bank [4, 5]. Cette banque contenait, au 22 septembre 2009, plus de 51 000 structures référencées, dont 46 100 ont été déterminées par cristallographie, et 5 000 par RMN.

D'autres méthodes à basse résolution, comme la microscopie électronique ou le SAXS[2], peuvent également être utilisées, mais restent, à l'heure actuelle, moins efficaces.

Ces méthodes, bien que leurs performances aient été grandement améliorées par l'essor des projets de génomique structurale, restent sous la contrainte de conditions expérimentales restrictives. De l'expression à la purification des protéines, de leur concentration à l'obtention d'un cristal à la résolution et au phasage de ces structures, ce sont autant d'étapes synonymes de conditions d'arrêt pour un cristallographe. Les statistiques réalisées dans les projets de génomique structurale montrent que le taux de succès dans un temps raisonnable est d'environ 10%.

Quant à la spectroscopie RMN, même si la contrainte du cristal est supprimée, elle ne peut être appliquée que pour la détermination de structures de protéines de petite taille, généralement inférieure à 300 résidus, et pour un échantillon protéique à analyser pur à plus de 95%.

De nombreuses méthodes automatiques de prédiction de la structure tertiaire des protéines ont été développées.

Modélisation par homologie :

La modélisation par homologie exploite les structures précédemment résolues. Elle repose sur l'idée que deux séquences protéiques présentant une similitude au niveau de leurs séquences (idéalement un taux d'identité de séquence supérieur à 30%) adoptent des repliement similaires [23, 124].

Bien que le modèle ainsi obtenu ne soit pas exact, il rend compte des régions clés impliquées dans la reconnaissance et l'interaction avec d'autres partenaires ou des résidus présents dans le site actif, et permet une connaissance relativement précise du cœur de la protéine.

Modélisation par reconnaissance des repliements ou méthodes d'enfilage (threading) :

Cette méthode repose sur le fait que le repliement est mieux conservé par l'évolution que la séquence. De ce fait, même lorsque la comparaison des séquences ne permet pas de détecter une homologie, l'information tridimensionnelle, disponible pour l'une des protéine, peut permettre de retrouver les liens évolutifs. La méthode consiste à "enfiler" la séquence à modéliser sur une structure connue, puis à estimer la compatibilité entre la séquence et la structure. Cette méthode, relativement robuste, permet l'identification d'homologues distants au sein de familles de protéines, voire l'identification d'une fonction même lorsque le taux d'identité de séquence est relativement faible.

Modélisation *ab initio*

La modélisation *ab initio* vise à prédire la structure d'une protéine sur la seule connaissance de sa séquence. De nombreux modèles de calculs ont été développés, généralement basés sur l'optimisation et la minimisation d'une fonction d'énergie rendant compte de l'état de la protéine. La prédiction de ces structures est extrêmement gourmande en temps de calcul, mais les résultats

[2]SAXS : **Small Angle X-ray Scattering.**

sont en constante amélioration comme l'atteste l'expérience CASP [25, 73, 105]. Initié en 1994, ce test en aveugle permet d'évaluer la performance des méthodes de prédiction. Un des résultats d'une des sessions CASP est présenté sur la Figure 1.6.

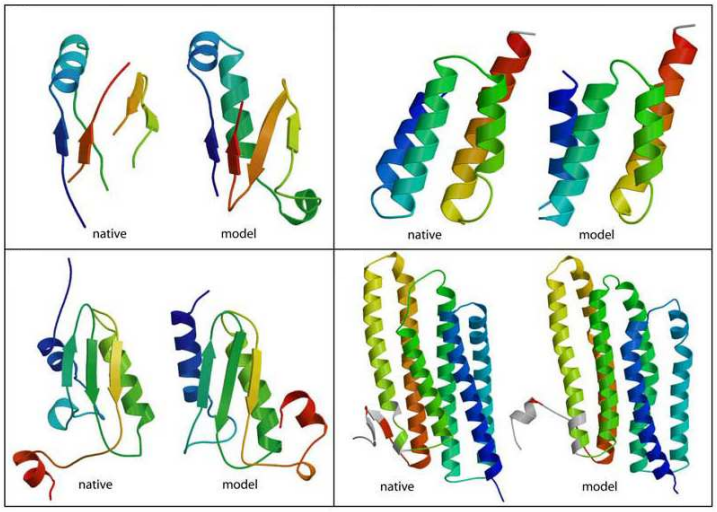

FIG. 1.6 – *Prédictions de structures obtenues à l'aire du logiciel Rosetta [16] lors de la session CASP6 Image originale en couverture du journal* **PROTEINS** *: Structure, Function and Bioinformatics, volume 61 du 26 septembre 2005.*

1.1.5 La structure quaternaire

Alors que certaines protéines sont constituées d'une chaîne polypeptidique unique, d'autres sont fonctionnelles uniquement sous forme d'oligomères, c'est-à-dire l'assemblage de plusieurs chaînes polypeptidiques, également appelées sous-unités. On parlera d'hétéromultimères ou d'homomultimères lorsque ces sous-unités sont respectivement différentes ou identiques entre elles. La stabilité de l'assemblage repose, comme celle de la structure tertiaire, sur un ensemble d'interactions à courte distance généralement non covalentes.

La structure quaternaire peut être obligatoire, c'est-à-dire que les sous-unités constitutives ne sont jamais trouvées isolées, cette structure quaternaire est alors appelée multimère biologique, puisqu'elle représente la forme active de la protéine. Certains assemblages sont, au contraire transitoires, et les partenaires peuvent être trouvés isolés. La plupart des protéines établissent de telles interactions transitoires avec d'autre protéines ou avec des acides nucléiques.

Plusieurs méthodes pour détecter la présence et la composition de ces assemblages ont été mises au point, notamment le double hybride dans la levure [58], ou le TAP-tag [46] et la spectroscopie de masse [55]. Ces méthodes permettent de connaître les différents partenaires d'une protéine avec un débit élevé, et d'obtenir ainsi une vue de l'interactome. Elles souffrent cependant d'un manque de précision relativement élevé.

L'agencement spatial des sous-unités peut être abordé par des méthodes à basse résolution, telles que la diffusion des rayons X ou des neutrons, ou encore la microscopie électronique. Ces méthodes permettent de connaître la forme de l'enveloppe de l'assemblage, et, à condition de connaître la structure tridimensionnelle des partenaires isolés, il est possible de reconstruire le complexe à l'intérieur de cette enveloppe.

Dans le cas où les structures des partenaires isolés sont connues, on peut également accéder à la structure de l'assemblage en déterminant, en particulier par mutagénèse dirigée, les acides aminés de chacun des partenaires qui participent à l'interaction.

Enfin, de même que les structures de protéines monomériques, les structures tridimensionnelles de complexes peuvent être déterminées expérimentalement par cristallographie ou par RMN. L'utilisation de la cristallographie implique cependant une stabilité suffisante de l'assemblage, ce qui exclut la majorité des complexes transitoires.

Historiquement, c'est en 1972 qu'a été réalisé le premier modèle de complexe protéine-protéine (trypsine, inhibiteur polypeptidique [9]). En 1978, le premier algorithme d'amarrage protéine-protéine a été publié par Wodak et Janin [146]. Depuis, ce domaine de recherche a pris un essor important, et différentes pistes on été explorées.

Les différentes méthodes d'amarrage traitent généralement le problème en deux étapes :

- une première phase de génération de conformations (applées pauses ou prédicats) entre les deux partenaires ;
- puis une seconde phase d'évaluation via une fonction de score, qui permet d'extraire des conformations proches de la structure biologique.

Ces méthodes s'appliquent à des protéines de rôle et de localisation cellulaires extrêmement variés. Les différentes méthodes d'amarrage seront présentées dans la section suivante.

1.2 Les complexes protéine-protéine

Les complexes protéine-protéine sont au cœur de la plupart des processus cellulaires. Par exemple, les voies de signalisation cellulaire, qui permettent l'intégration des signaux extra-cellulaires et la production d'une réponse biologique adaptée, reposent en grande partie sur l'association entre des protéine-kinases et leurs cibles.

D'autres mécanismes cellulaires essentiels, tels que la réplication de l'ADN, la traduction ou le transport cellulaire reposent également sur la formation de complexes protéiques. La détection et la détermination de l'organisation structurale de ces assemblages moléculaires constituent donc une étape majeure dans la compréhension du mécanisme et de la régulation de ces phénomènes.

1.2.1 Nature de l'interaction protéine-protéine

Comme évoqué précédemment, certaines interactions protéine-protéine sont obligatoires, à savoir que les protomères constitutifs de ces complexes (les sous-unités) ne sont pas stables pris indépendamment en solution, et leur fonction est inhérente à leur association.

Le répresseur *Arc*, impliqué dans la répression et la régulation de la transcription [11] est un homodimère obligatoire, c'est-à-dire obligatoirement constitué de deux sous-unités identiques. Les hétérodimères, constitués de l'assemblage de deux sous-unités différentes peuvent également être obligatoires.

C'est le cas par exemple des facteurs induits par l'hypoxie (HIF), qui sont obligatoirement constitués d'une sous-unité α et d'une sous-unité β, de tailles et de séquences différentes.

Au contraire, dans certains cas, les protomères sont stables à l'état isolé, mais également capables d'interagir. Le complexe est alors qualifié de transitoire. Par exemple, les protéines G, impliquées dans la transduction du signal dans la cellule, sont composées de trois sous-unités. Lorsque la protéine G, en absence de signal, est associée à un récepteur membranaire, les trois sous-unités forment un complexe. Lorsqu'un ligand vient se lier au récepteur, l'une des sous-unités est libérée, et agit sur des effecteurs cellulaires, notamment l'adenylate cyclase (voir Figure 1.7). L'association et la dissociation des différentes sous-unités sont tributaires d'une molécule de GTP ou de GDP.

Ces complexes sont des "interrupteurs cellulaires" permettant l'activation de toute une cascade de réactions menant à une réponse adaptée au signal [109].

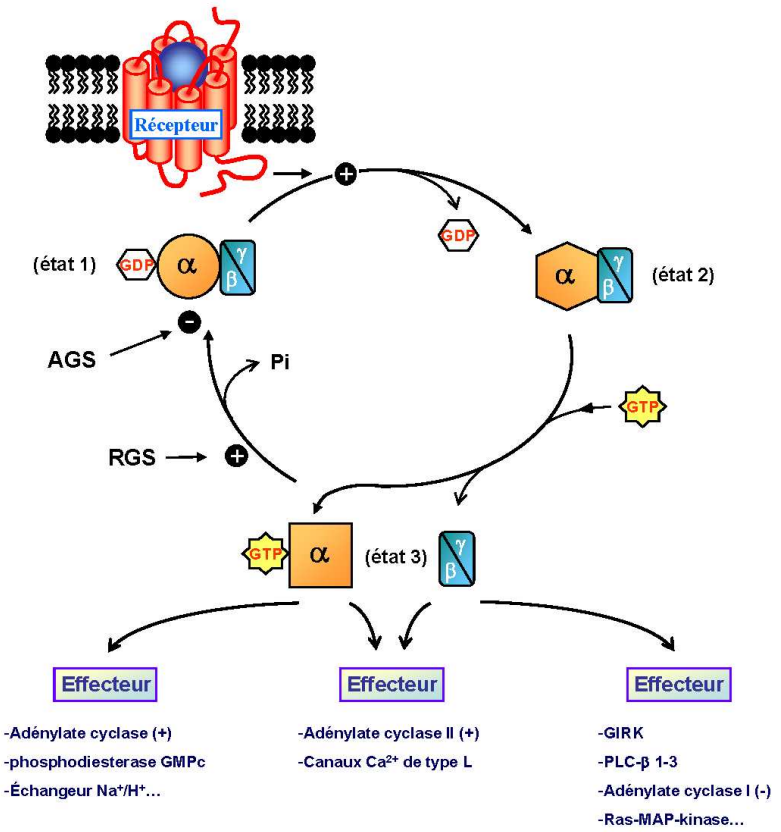

FIG. 1.7 – *Interactions transitoires et cycle des protéines G. La liaison du récepteur à son ligand provoque la libération du complexe $G_\alpha - G_\beta - G_\gamma$–GDP. Le GDP est échangé contre un GTP, et la sous-unité G_α–GTP se dissocie du reste du complexe. C'est ce sous-complexe qui va, à son tour, aller activer d'autres protéines dans la cellule, ce qui mènera finalement à la production par la cellule d'une réponse biologique adaptée.*

1.2.2 Détection expérimentale

De nombreuses méthodes permettent la détection des complexes présents dans la cellule, l'interactome. On peut citer par exemple le phage-display [115] ou la co-immunoprécipitation, qui

permettent une détection ponctuelle des interactions entre partenaires protéiques. Des méthodes permettant une analyse à grande échelle de l'interactome ont été developpées, les deux principales étant le double hybride dans la levure [37] et le TAP-tag [63].

Le double-hybride en levure

Cette méthode, très utilisée pour la détection de complexes à grande échelle, permet la détection indirecte de l'interaction, révélée par l'induction d'un gène rapporteur [37] (voir Figure 1.8).

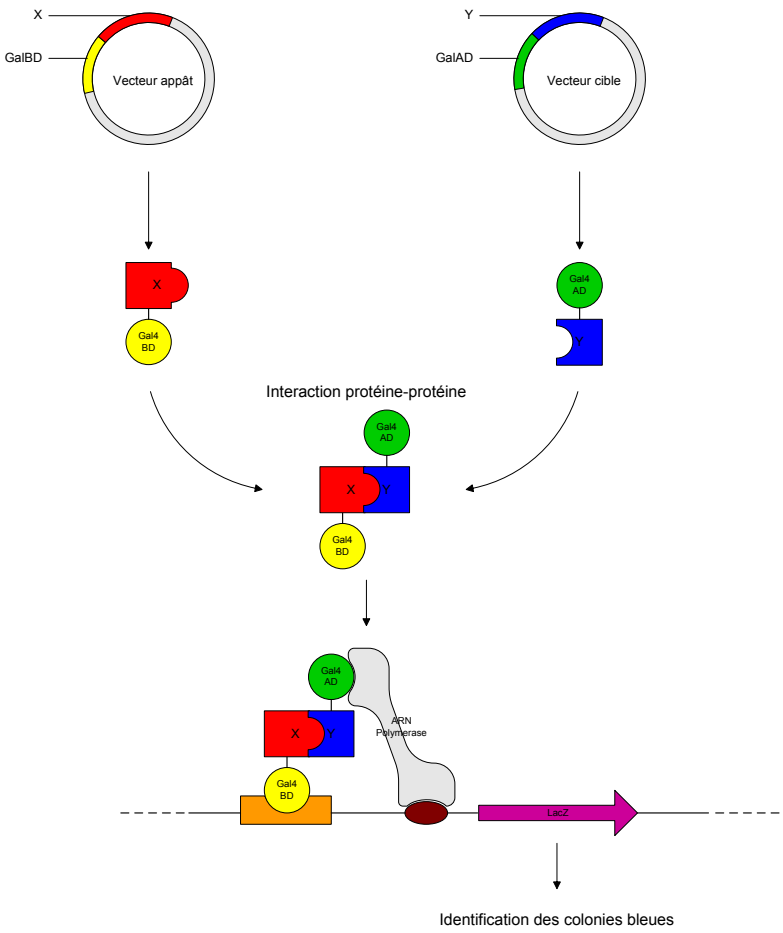

FIG. 1.8 – *Principe de la méthode de détection de complexes protéiques par double-hybride dans la levure. La protéine Gal4 est l'activateur naturel des différents gènes intervenant dans le métabolisme du galactose. Elle agit en se fixant au niveau des UASG* $Upstream\ Activating\ Sequence\ GAL$ *régulant la transcription. Les deux protéines X et Y dont on veut tester l'interaction sont fusionnées, l'une au domaine de fixation Gal4 (DNA binding Domain), l'autre au domaine de Gal4 activant la transcription (Activation Domain). Si une interaction entre X et Y existe, il se forme alors un activateur de transcription DBD-X/Y-AD. Cet activateur peut se lier au niveau des séquences UASG permettant la transcription d'un gène rapporteur par l'ARN polymérase II.*

Cette méthode présente l'inconvénient d'être relativement peu fiable, en effet les taux de faux positifs (induction du gène rapporteur alors que l'interaction n'existe pas) et de faux négatifs (interactions biologiques non détectées) sont relativement importants. Ces taux peuvent être

réduits en effectuant des détections complémentaires (inversion de la proie et de l'appât notamment), et en répétant les expériences. D'autre part, la phase d'analyse des résultats est déterminante, et des méthodes différentes peuvent conduire à des résultats très différents [72].

Plusieurs recherches systématiques de complexes par double-hybride ont été menées sur la levure [58, 139]. Le recouvrement entre ces différentes expériences est cependant relativement faible [58].

Le TAP-tag

Le principe du TAP-tag consiste à produire une protéine appât en fusion avec deux étiquettes de purification, généralement la calmoduline et la protéine A. La purification sur colonne d'affinité permet de retenir la protéine appât, qui est liée à ses partenaires cellulaires. Le complexe est ensuite élué (décroché de la colonne d'affinité), puis les composants sont séparés et identifiés par spectroscopie de masse (voir Figure 1.9).

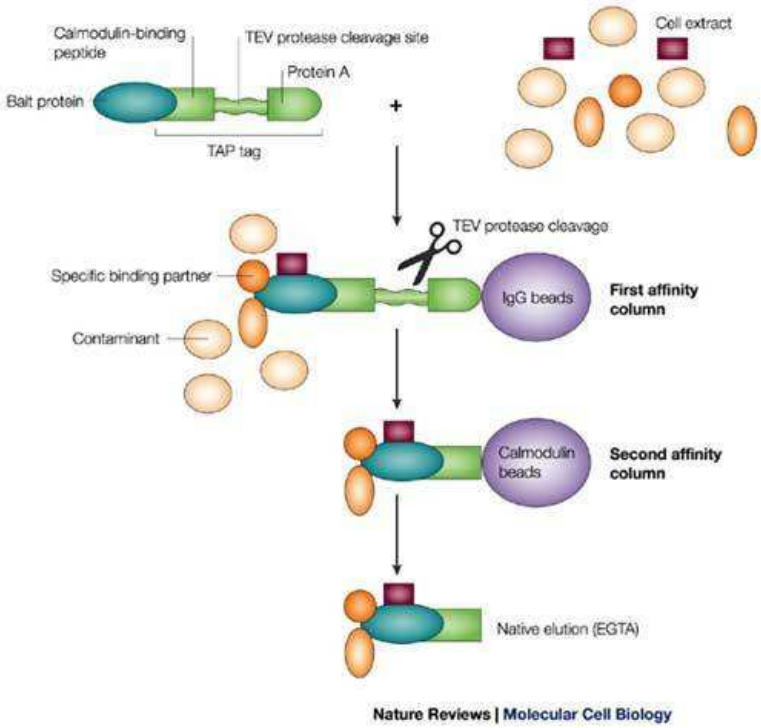

FIG. 1.9 – *Principe de la méthode de détection de complexes protéiques par TAP-tag dans la levure.*

Deux études menées sur la levure *S. cerevisiae* utilisant cette technique ont été menées [46, 55]. Ici encore, les recouvrements entre études menées à l'aide de la même méthode, sont relativement faibles. Le recouvrement avec les résultats obtenus en double-hybride sont encore plus faibles [58].

Ainsi, les méthodes expérimentales ne permettent d'accéder qu'à une petite partie de l'interactome. En effet, les problèmes techniques liés aux différentes méthodes, et la faible durée de vie de certains complexes, font que de nombreuses interactions échappent à la détection expérimentale. Ces études permettent cependant d'appréhender la taille du problème : elles ont permis de mettre en évidence plus de 12 000 complexes différents dans la levure.

1.3 L'amarrage protéine-protéine

1.3.1 Principe et partitionnement du problème

Le problème de l'amarrage protéine-protéine peut être formulé de la manière suivante : considérant deux partenaires protéiques dont la structure tridimensionnelle est connue, **quelle est la meilleure conformation pour l'assemblage de ces deux partenaires ?** Considérant cette conformation de complexe prédite, **celle-ci est elle suffisamment stable pour pouvoir exister** *in vivo ?*

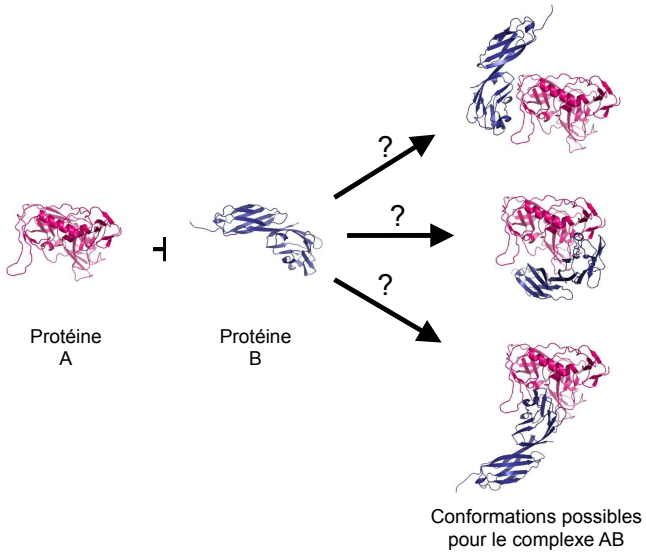

FIG. 1.10 – *Le problème de l'amarrage : Comment associer les partenaires A et B ensemble ? Les conformations putatives sont-elles suceptibles d'exister in vivo ?*

Les différentes procédures d'amarrage traitent généralement le problème en deux étapes : une première phase de génération, au cours de laquelle un grand nombre de conformations sont générées, puis une seconde phase au cours de laquelle ces différentes conformations sont évaluées afin d'extraire un petit nombre de conformations vraisemblables.

La génération des conformations candidates se fait généralement en "corps rigide", c'est-à-dire qu'aucune déformation n'est appliquée aux partenaires. La méthode consiste alors à échantilloner les différentes orientations possibles d'un partenaire par rapport à l'autre. Dans les procédures utilisant un système de grille, et une transformée de Fourier rapide, un échantillonage satisfaisant de l'espace des solutions conduit à générer environ 10^9 modes d'associations possibles pour des protéines de taille moyenne [24], dont seulement quelques dizaines sont proches de la structure biologique (la structure native). Cette étape est donc généralement très gourmande en temps de calcul, mais réduire la finesse de l'échantillonage réduit également la précision.

Il faut ensuite disposer de fonctions d'évaluation puissantes, permettant de retrouver ces conformations "presque-natives" parmi les conformations "non-natives".

Enfin, il faut noter que de nombreux réarrangements, essentiellement au niveau des chaînes latérales ont lieu au moment de la complexation. Il arrive, plus rarement que le squelette de l'un des partenaires, voire des deux, subisse des modifications.

Si les méthodes d'amarrage s'accommodent relativement bien des premiers, les modifications de squelettes ne sont, à l'heure actuelle, pas prises en compte. La structure d'un partenaire, lorsqu'il est isolé, sera appelée "non-liée" (*unbound*), et sa structure dans le complexe "liée" (*bound*).

Si la grande majorité des méthodes donnent de très bon résultats sur l'amarrage de deux protéines dans leur forme liée, notamment grâce à une complémentarité géométrique parfaite des deux partenaires, le problème de l'amarrage de deux protéines connues sous leur forme non-liée est beaucoup plus complexe. Cependant, dans l'optique d'une utilisation prédictive de ces méthodes, c'est bien sur le problème non-lié qui doit être résolu.

1.3.2 Les algorithmes

Le premier algorithme, fut mis au point par Shoshana Wodak et Joël Janin, d'après les travaux de Cyrus Levinthal [64, 146]. La phase d'exploration consiste à échantilloner les six degrés de liberté (5 rotations et une translation) d'une molécule par rapport à l'autre. Les conformations dans lesquelles les deux partenaires sont en contact sont évaluées suivant leur surface d'interaction approchée selon le modèle proposé par Levitt [80]. Cette fonction de score simple fut par la suite remplacée en 1991 par une fonction de minimisation d'énergie plus performante [144]. D'autres méthodes sont par la suite apparues :

- Méthodes basées sur la description de points critiques définis comme "trous et bosses" (*knobes and holes*) [24, 78, 149]. Ces algorithmes utilisent tous la complémentarité de forme comme déterminant de la validité du complexe généré (les solutions sélectionnées correspondant à celles établissant une concordance entre ces points critiques). Elles ont été améliorées par la suite, notamment en simplifiant la surface par l'application d'une grille [144] et par utilisation la transformation de Fourier rapide.
- Des méthodes de hachage géométrique ont permis d'étendre l'étude des méthodes de *knobs and holes* au modèle tel que l'avaient prédéfini Landam et Wolfson en 1988 [75]. Cette première version ne prenait en compte dans l'évaluation, que la complémentarité géométrique des deux partenaires. En 1993, les propriétés physico-chimiques des acides aminés en contact au niveau de l'interface ont été ajoutées. Ce modèle, extrêmement efficace pour l'amarrage de protéines sous forme liée reste cependant très sensible aux faibles variations de surface [40, 89]. Cette méthode fut améliorée par la suite grâce notamment à l'expérience CAPRI [39, 57, 110, 111, 112, 123, 125, 148].

Méthodes utilisant la transformée de Fourier rapide

Les méthodes utilisant la transformée de Fourier rapide [65] sont de loin les plus utilisées [2, 18, 21, 22, 69, 76, 103, 130, 145]. Le principe est d'appliquer une grille tridimensionnelle sur la protéine. À chaque point de la grille, on attribue alors un poids :

- négatif et élevé, si le point se situe à l'intérieur de l'un des partenaires protéiques.

– nul si le point est à l'extérieur.
– égal à 1 s'il est proche de la surface.

Les différentes conformations de l'assemblage de deux protéines A et B peuvent alors être évaluées en effectuant, à chaque point de la grille, le produit entre les valeurs attribuées à ce point avec A et avec B. Ainsi, si les protéines ne sont pas en contact, le produit sera nul en tout point, puisque la valeur sera toujours nulle pour au moins l'un des partenaires. Si les deux protéines se chevauchent, le produit total sera très élevé, puisque certains points se trouvent à la fois à l'intérieur de A et à l'intérieur de B. Enfin, si les deux protéines sont en contact, sans s'inter-pénétrer, le produit sera positif et peu élevé.

Ce principe a ensuite été étendu en incluant, dans les poids, des paramètres physico-chimiques, permettant par exemple de rendre compte d'interactions favorables entre les deux partenaires [3, 20, 93]. Dans certaines méthodes, des paramètres biologiques ont également été utilisés dans les poids, notamment pour rendre compte de la probabilité d'un acide aminé de se trouver à l'interface, basées par exemple sur des expériences de mutagenèse dirigée.

HADDOCK

HADDOCK (**H**igh **A**mbiguity **D**riven **DOCK**ing, [29]) est la méthode mise au point par Alexandre Bonvin et collaborateurs, et fait partie des méthodes d'amarrage les plus efficaces.

Cette méthode a la particularité de prendre en compte les données biologiques disponibles, en particulier les résultats de mutagenèse dirigée. En RMN, les "contraintes ambigües" sont couramment utilisées au cours de la reconstruction des structures pour désigner des interactions dont les partenaires ne sont pas connus avec certitude. Les programmes utilisés en RMN permettent alors de chercher, parmi les réalisations possibles de ces contraintes, celles qui en satisfont le plus grand nombre.

Les auteurs de la méthode ont ainsi utilisé les contraintes ambigües pour "coder" le fait que certains acides aminés sur chacun des partenaires sont suceptibles de participer à l'interaction, puis ont utilisé les méthodes de traitement des données RMN pour rapprocher les deux partenaires de manière à satisfaire un maximum de ces contraintes.

Rosetta-Dock

La méthode Rosetta-Dock est également une méthode très performante [28, 53, 126]. De même que l'amarrage protéine-protéine, la modélisation *ab initio* de la structure d'une protéine requiert l'évaluation d'un grand nombre de conformations. En particulier, la méthode Rosetta, mise au point par Baker et collaborateurs [126] permet de construire, pour des protéines allant jusqu'à 150 acides aminés, des modèles ayant moins de 5 $\AA$ de RMSD[3] avec la structure native. Le principe de Rosetta-Dock est la modélisation des forces physiques sur des structures à haute résolution par des potentiels de champs moyens.

Un grand nombre de structures aléatoires sont générées en utilisant des potentiels à basse résolution. Les meilleures sont sélectionnées, puis affinées grâce aux potentiels à haute résolution.

Cette procédure a été adaptée avec succès à l'amarrage protéine-protéine.

[3]RMSD : **R**oot **M**ean **S**quare **D**eviation

1.3.3 CAPRI : une expérience à ne pas manquer

L'un des points critiques dans toute méthode de prédiction est la possibilité de la tester en aveugle. Ceci a été rendu possible, pour les algorithmes d'amarrage protéine-protéine, grâce à la mise en place de l'expérience CAPRI[4] [59, 60, 61, 62, 147].

Créée en 2001 par Joël Janin, Shoshanna Wodak, John Moult, Lynn Ten Eyck et Michael Sternberg, cette expérience a permis d'estimer la précision et la performance des différents algorithmes d'amarrage protéine-protéine existants par un test en aveugle [129]. Des cristallographes ayant déterminé la structure tridimensionnelle d'un complexe transmettent les coordonnées aux organisateurs avant la publication. Chaque participant est invité à prédire la structure de ce complexe à partir des structures tridimensionnelles des partenaires. Dix solutions peuvent être proposées par chaque participant et pour chaque complexe. Les prédictions sont alors comparées à la structure du complexe natif suivant un certain nombre de critères. L'ensemble de ces prédictions est réalisé sans connaissance de la réponse, et la structure du complexe obtenue expérimentalement n'est dévoilée aux participants qu'à l'issue des soumissions voire de la publication de celle-ci. Cette expérience, en presque 10 ans d'existence, a permis de réunir pas moins de 70 laboratoires participants pour la prédiction de 42 complexes.

Outre les méthodes précédemment citées, de nombreuses autres ont été utilisées : les algorithmes génétiques [39, 57, 110, 111, 112, 123, 125, 148], les harmoniques sphériques [120, 94], la dynamique moléculaire [17, 71, 133] ainsi que de nombreuses méthodes de minimisation d'énergie [53, 126] ou d'intégration de contraintes biologiques [29, 100, 140]. Si les conformations candidates, générées par les différents algorithmes de complémentarité de formes, sont relativement proches de la solution native [96], l'extraction de ces conformations, à savoir le tri de ces conformations putatives par la fonction de score, reste à améliorer.

Cette dernière constatation a amené les organisateurs de CAPRI à inclure, à partir de 2005 [79], une expérience supplémentaire d'évaluation des conformations candidates. Cette phase de "scoring" est réalisée à la suite de la phase de prédiction originale de CAPRI. Les "prédicteurs" sont invités à soumettre 100 conformations candidates pour chaque cible, les "scoreurs" cherchent à extraire, de cet ensemble, les 10 conformations les plus proches de la conformation native.

L'analyse globale des résultats [104] montre que les méthodes les plus performantes permettent d'obtenir une solution, au moins acceptable, parmi les 10 conformations soumises à l'évaluation dans 60% des cas. Ces performances, bien qu'elles représentent un progrès gigantesque par rapport à celles obtenues il y a seulement 10 ans, sont cependant encore très insuffisantes pour que l'amarrage devienne un outil dans les études biologiques.

De plus, les performances sont nettement inférieures lorsqu'on ne regarde que les cibles pour lesquelles les deux partenaires étaient donnés sous forme non-liée.

Enfin, les méthodes les plus performantes sont extrêmement consommatrices en temps de calcul, ce qui exclut toute utilisation à grande échelle.

[4] CAPRI : **C**ritical **A**ssessement of **PR**ediction of **I**nteractions

1.4 Le diagramme de Voronoï et les constructions dérivées

1.4.1 Un peu de géométrie...

Le diagramme de Voronoï est une décomposition particulière de l'espace en un sous-ensemble discret de points, qui doit son nom au mathématicien franco-russe Georgi Fedoseevich Voronoï qui généralisa la tessellation au cas à n dimensions en 1908 [142].

La tessellation de Voronoï d'un ensemble de points (ou nœuds) est un pavage de l'espace en régions qui peuvent être vues comme les zones d'influence de chacun des nœuds, régions appelées "cellules" ou polyèdres de Voronoï. Un point de l'espace appartient donc soit à une cellule, soit à une face commune entre des cellules. D'autre part, dans le cas d'un ensemble de points réguliers (et c'est le cas pour les protéines), ce diagramme est unique.

La triangulation de Delaunay est la construction la plus directement déduite de la tessellation de Voronoï. Elle est obtenue en reliant les nœuds dont les cellules partagent une face. Chaque tessellation de Voronoï correspond à une et une seule triangulation de Delaunay, et les deux constructions sont duales l'une de l'autre.

Les domaines dans lesquels le diagramme de Voronoï est utilisé sont nombreux, et particulièrement ceux visant à partitionner l'espace en zones d'influence. En météorologie, initialement par A.H. Thiessen, pour analyser les données de distributions spatiales, en planétologie pour le partitionnement des populations d'étoiles, mais aussi en physiologie pour l'analyse de la répartition des capillaires dans les muscles ou encore en cancérologie dans le diagnostique des cellules tumorales ainsi qu'en robotique dans le calcul de trajectoire mobile, et beaucoup d'autres encore.

Dans le cadre de l'analyse structurale des protéines, la tessellation de Voronoï a été utilisée, la première fois par Richards en 1974 pour évaluer dans les protéines globulaires les volumes des atomes, assimilés dans le cadre de cette étude au volume des polyèdres de Voronoï [77]. Dans cette première publication, Richards a identifié, et partiellement résolu, deux points particuliers. Le premier est que les atomes de la surface ont peu de voisins, leurs cellules ont des volumes qui sont très grands, voire infinis, et donc peu représentatifs de leurs propriétés. Par ailleurs tous les atomes, qui sont les nœuds des cellules de Voronoï, sont équivalents entre eux, là encore peu représentatifs de leur taille.

En ce qui concerne le premier point, Richards a proposé une première solution qui consistait à placer des molécules d'eau sur un réseau cubique autour de la protéine, puis à relaxer leurs positions. Cette méthode a ultérieurement été améliorée par Gerstein *et al.* [136, 137]. D'autres solutions ont été proposées, consistant à ne considérer que les atomes dont les cellules ont des volumes "raisonnables" [117], ou bien à placer les molécules d'eau en utilisant la dynamique moléculaire [19], ou enfin, plus récemment, en utilisant une représentation en union de sphères [95]. Certains auteurs utilisent un mélange de diagrammes de puissance et la représentation en union de sphères, la cellule d'un atome étant alors l'intersection de sa cellule et de sa sphère de Van der Walls.

Pour répondre au problème des différences de tailles entre les atomes, Richards a proposé d'introduire des poids dans la construction de Voronoï (voir Figure 1.11). Cette méthode, connue sous le nom de "méthode B de Richards" a été très largement utilisée. Cette méthode manque cependant de rigueur d'un point de vue mathématique, et la construction n'est plus un pavage de l'espace. Ainsi, des volumes sont perdus, parce que les plans ne se coupent plus en un point.

Richards a cependant démontré que ce volume perdu est très petit par rapport au volume des atomes. Cette approche a par la suite été affinée par Gellatly et Finney [47] qui ont défini de manière rigoureuse une construction de Voronoï pondérée dans laquelle des plans radicaux remplacent les plans médiateurs. Le diagramme qui en résulte est appelé diagramme de Laguerre ou diagramme de puissance (voir Figure 1.11). Dans cette construction il n'y a cependant pas de proportionnalité directe entre les poids attribués à chaque nœud et les distances aux plans radicaux. De ce fait, le choix des poids est assez délicat, et l'interprétation plus difficile. Goede *et al.* [50] ont proposé un autre type de diagramme de Voronoï pondéré, dans lequel les poids attribués aux atomes et les distances aux surfaces séparatrices sont reliés de manière linéaire. Cependant, dans ce cas les surfaces séparatrices ne sont plus des plans.

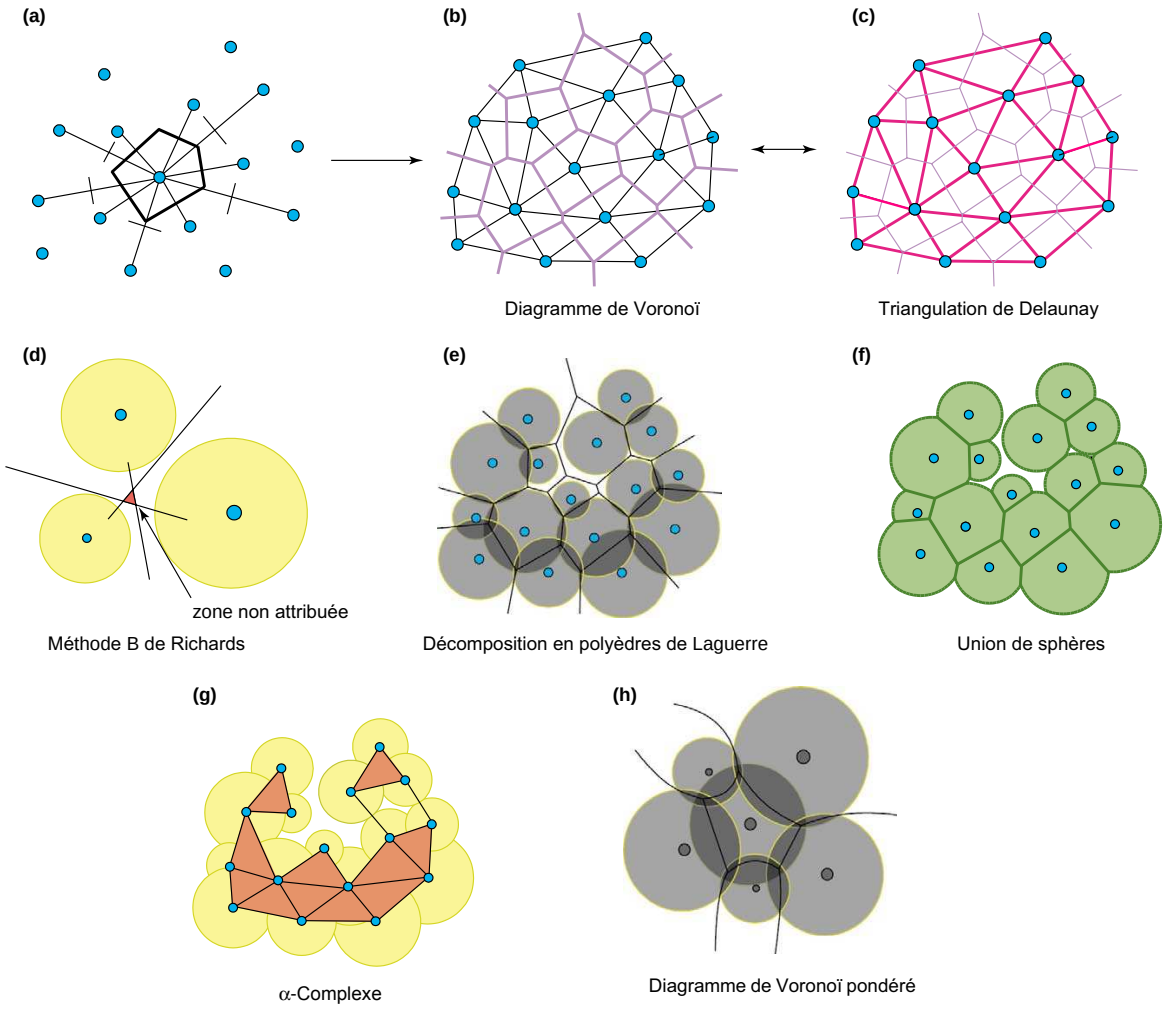

FIG. 1.11 – *Rappel des différentes tessellations.*

Une analyse formelle de ces applications a été réalisée par Edelsbrunner et ses collaborateurs [33, 34, 35, 84, 85]. Ces auteurs ont par la suite formalisé la notion d'α-shape (voir Figure 1.11). L'α-shape est constituée des arêtes de la triangulation de Delaunay qui sont contenues dans le volume défini par l'union de sphère. Cette méthode est particulièrement bien adaptée à l'étude des trous et cavités ainsi que la dynamique moléculaire [12, 15, 26, 107, 108, 86].

1.4.2 Analyse de la structure des protéines

Calcul de volumes

La première application, et probablement la plus utilisée, des constructions de Voronoï à l'étude de la structure des protéines a été le calcul des volumes atomiques. Les volumes atomiques jouent un rôle fondamental dans de nombreux phénomènes qui intéressent la structure des protéines : l'empilement, le repliement et la dynamique en particulier. Cependant, définir et calculer le volume d'un acide aminé est un problème complexe. Le volume de la cellule de Voronoï d'un acide aminé peut être considéré comme le volume disponible pour cet atome, et est généralement inférieur à son volume de Van der Waals.

Richards [77] a été le premier à calculer les volumes atomiques en utilisant un diagramme de Voronoï. Son estimation a été de nombreuses fois revisitée [136, 137, 19, 95, 38, 116, 135, 41], et des volumes de référence ont été proposés pour les groupements chimiques les plus souvent rencontrés dans les protéines [116, 135]. Deux groupes, Pontius *et al.* et Tsai *et al.* [116, 135, 138, 136], ont discuté les effets des différents paramètres sur les volumes calculés. Ils ont en particulier montré que :

- Le type de construction utilisé a une grande influence. En particulier, les constructions non-pondérées ont tendance à surestimer les volumes des petits atomes, et à sous-estimer ceux des gros atomes.
- Les méthodes pondérées sont très sensibles aux poids attribués aux atomes.
- Le jeu d'atomes utilisé dans le calcul est très important, en particulier les résultats sont nettement différents selon que des atomes partiellement exposés au solvant sont ou non utilisés.
- Les structures de faible résolution donnent des volumes plus grands, et la dispersion des volumes est une bonne mesure de la qualité des structures obtenues par cristallographie ou par RMN [116].

La première mesure des volumes des acides aminé a été réalisée en sommant les volumes des atomes qui composent l'acide aminé. Une solution alternative consiste à réduire l'acide aminé à un point, puis à construire une cellule de Voronoï pour chaque acide aminé. Soyer *et al.* [131] et notre groupe [6, 7] avons utilisé les centres géométriques des chaînes latérales comme nœuds pour la construction d'un diagramme de Voronoï classique. Comme attendu, les volumes calculés de cette manière sont surestimés pour les petits acides aminés et sous-estimés pour les gros acides aminés. Cependant, les volumes calculés par cette méthode présentent un cœfficient de corrélation de 0,93 avec les volumes calculés par Pontius *et al.* [116].

Densité et empilement

La détermination des volumes atomiques donne une estimation de la densité à l'intérieur de la protéine, et répond à la question de savoir si la densité à l'intérieur d'une protéine est homogène. De manière remarquable, toutes ces études concluent que la densité moyenne des atomes à l'intérieur d'une protéine est aussi élevée que dans les cristaux de petites molécules [48, 54, 49].

Soyer *et al.* [131] ont utilisé un diagramme construit à partir des centres géométriques des chaînes latérales, et analysé les résultats en termes de nombre de faces des polyèdres et de nombre de sommets par face. Le nombre de faces est également le nombre de voisins, le nombre de sommets donnant une estimation de la symétrie autour des arêtes. Les auteurs obtiennent des valeurs très proches de celles qui caractérisent les empilement aléatoires denses de sphères dures dans