Université Henri Poincaré - Nancy I **École doctorale SESAMES**
UFR Sciences et techniques de la matière et des procédés

Vers une nouvelle stratégie pour l'assemblage interactif de macromolécules

THÈSE

présentée et soutenue publiquement le 30 janvier 2009

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité Chimie Informatique et Théorique)

par

Matthieu CHAVENT

Composition du jury

Rapporteurs :	Gilbert DELÉAGE
	Joël JANIN
Examinateurs :	Jean-Paul BORG
	Daniel CANET
	Stéphane REDON
	Dave RITCHIE
Directeurs :	Bruno LÉVY
	Bernard MAIGRET

Équipe ORPAILLEUR
Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
UMR 7503 - Campus Scientifique - BP 239 - 54506 Vandœuvre-les-Nancy Cedex

Remerciements

Je voudrais tout d'abord remercier mes co-directeurs de thèse Bernard Maigret et Bruno Levy. Tous les deux m'ont permis de progresser considérablement durant ces trois ans.

Bernard m'a fait découvrir le monde de la recherche et m'a laissé libre de faire mes propres choix. Il m'a aussi permis de découvrir le monde "tout court" en m'envoyant présenter mes travaux dans diverses conférences autour du globe. Ceci m'a permis d'acquérir une certaine pratique de l'anglais mais, surtout, de vivre des moments inoubliables (je pense en particulier à Fortaleza...).

J'ai également beaucoup appris avec Bruno en informatique. Il m'a montré qu'il fallait toujours chercher à dépasser ses limites pour avancer... Ce ne fut pas toujours facile mais se fut une expérience très enrichissante. De plus, le programme MetaMol n'aurait jamais vu le jour sans son aide précieuse et ses encouragements. Enfin, je me rappelle avec plaisir des quelques mois passés dans son bureau où il m'a souvent fait profiter de son enthousiasme et de sa passion pour l'informatique et le graphisme.

Je souhaite également remercier Dave Ritchie pour tous les moments agréables passés à discuter (de "Docking" mais également de tout et de rien). J'aimerai aussi le remercier pour m'avoir permis de me lancer, avec lui, dans l'expérience CAPRI. J'espère continuer cette collaboration le plus longtemps possible.

Je remercie vivement les membres de mon jury d'avoir accepté de juger mon travail ; en particulier mes rapporteurs, Joël Janin et Gilbert Deléage, d'avoir pris le temps de relire attentivement mon manuscrit malgré des emplois du temps très chargés.

Je souhaite remercier Stephane Redon de m'avoir chaleureusement accueilli quelques jours dans son équipe à l'INRIA Grenoble - Rhone-Alpes et de m'avoir fait découvrir le programme SAMSON.

Je remercie Jean-Paul Borg et son équipe sans qui le projet "Erbin" n'aurait jamais abouti. Je souhaite tout particulièrement remercier Nadine Déliot pour tous les travaux qu'elle a dû mettre en oeuvre pour valider mon modèle ainsi que pour sa relecture attentive de la partie de ma thèse consacrée aux résultats biologiques.

Je voudrais particulièrement remercier Alex, avec qui j'ai passé des moments excellents dans notre bureau commun et que j'espère aller voir prochainement dans sa "cabane" au Canada. Je n'oublies pas les "anciens", JP et Jérôme, qui m'ont fait découvrir phi-science dans toute sa splendeur.

Un grand merci également à Laurent, Luc et Pilou pour leurs supports techniques qui m'ont permis de comprendre un peu mieux les joies de l'informatiques. Je les remercie également pour les pauses café (mention spéciale pour Laurent) et les soirées Chtimi (spéciale dédicace à Pilou et sa Quack).

Je remercie également toute l'équipe ORPAILLEUR de m'avoir accueilli chaleureusement et en particulier à Amedeo Napoli, responsable de celle-ci. Une petite dédicace pour le groupe bioinfo Lorraine : Marie-Dominique Devignes, Malika Smaïl-Tabbone et Michel Souchet pour leurs conseils (et relectures) avisés. Je n'oublie pas non plus les survivants du bureau B235 : un grand merci à Yesmine et Nazhia pour leur gentillesse, leur joie de vivre et surtout leur délicieux gâteaux, Léo pour les parties de tennis qui m'ont permis de me bouger un minimum et Vincent pour ses aides FORTRAN et ses contributions non négligeables au budget café.

Je tiens à remercier également les membres de l'équipe ALICE et en particulier Nicolas Ray, Bruno Vallet et Cécile Poisot, pour leur discussions toujours très instructives.

Je remercie également la région Lorraine et le CNRS qui ont financé mes trois ans de thèse via une Bourse de Docteur Ingénieur.

Enfin, je remercie ma famille et mes amis pour leur soutien au quotidien et plus particulièrement ma D&D qui a toujours été là pour moi...

Table des matières

Introduction

Bioinformatique génomique et structurale

Depuis la fin des années 60 de nombreuses recherches sont menées à l'aide des ordinateurs, que cela soit pour la visualisation de molécules (Levinthal, 1966), l'analyse de la structure des protéines (Levitt et Chothia, 1976) ou encore la recherche d'assemblages moléculaires (Wodak et Janin, 1978). Le terme le plus souvent employé à cette époque était biologie computationnelle (*computational biology*). Le terme bioinformatique semble, quant à lui, avoir été utilisé pour la première fois en 1978 par Paulien Hogeweg lors de l'étude de système biologique grâce à l'outil informatique.

La bioinformatique, telle que nous la connaissons actuellement, ne connut un réel développement qu'à la fin des années 90, époque où "l'ordinateur personnel" (*Personal Computer* ou PC) commença à se démocratiser. De plus, le besoin de traiter de grands volumes de données résultant de nouvelles méthodes expérimentales, comme le séquençage du génome (Lander et the International Human Genome Sequencing Consortium, 2001; Venter et Celera Genomics, 2001), les puces à ADN (Eisen *et al.*, 1998) ou le criblage double hybride (Uetz *et al.*, 2000), renforça ce développement.

Il y a encore quelques années, lorsque la génomique était en plein essor, la bioinformatique se bornait surtout à traiter les séquences d'ADN : les analyser, les classer et les stocker (Collins *et al.*, 1998). Mais, comme cela était prévisible, les chercheurs se rendirent rapidement compte qu'avoir la séquence complète d'un génome (humain ou d'une autre espèce) ne suffisait pas à élucider le fonctionnement biologique des cellules. Celles-ci sont, en effet, soumises à des phénomènes métaboliques et de régulation qui ne sont pas directement liés à l'expression d'un gène.

C'est dans cette ère de la post-génomique que la protéomique, l'étude des protéines au sens large, s'est développée afin de répondre aux questions que la génomique ne pouvait élucider (Pandey et Mann, 2000). Ces études biochimiques, souvent à haut débit, ont créé là encore une masse de données considérable. Ces données permettent de définir, par exemple, quelles protéines sont en interaction les unes avec les autres mais ne répondent que très superficiellement à la question : comment ? C'est pourquoi ces informations biochimiques peuvent être complétées par des études structurales des cibles biologiques (Russell *et al.*, 2004). Les techniques les plus souvent utilisées pour ces analyses sont la cristallographie aux rayons X et la spectroscopie par

Résonance Magnétique Nucléaire (RMN) qui permettent d'obtenir des structures très précises (résolution atomique souvent inférieure à 2 Å). A celles-ci s'ajoutent des techniques calculant des structures de plus basse résolution (de l'ordre de 5 Å). Il s'agit, par exemple, de la cryo-microscopie électronique (Grünewald *et al.*, 2003) ou de la diffraction aux rayons X à petits angles (*small angle X-ray scattering* ou SAXS) (Márquez *et al.*, 2003). Ces techniques ont, là aussi, engendré une quantité d'informations gigantesque qu'il a fallu analyser. La bioinformatique fut également utilisée pour sonder cette multitude de données.

Ainsi, de nombreuses équipes de biologistes ont besoin d'outils informatiques pour mener à bien leurs études. Toutes ces études peuvent se prévaloir du label "bioinformatique". Nous faisons, ici, la distinction entre la bioinformatique appliquée à l'analyse du génome et celle dédiée à l'analyse des données structurales. Nous nous intéressons plus particulièrement à cette dernière. La bioinformatique structurale reste elle aussi un terme générique regroupant de multiples voies de recherche. En effet, il est nécessaire de stocker toutes les structures dans diverses bases de données (Berman *et al.*, 1992, 2002; Natarajan *et al.*, 2005), de pouvoir visualiser celles-ci (Goddard et Ferrin, 2007) et surtout de les analyser (Levitt et Chothia, 1976; Lo Conte *et al.*, 1999).

Modéliser les systèmes macromoléculaires

A partir de toutes les données accumulées et des capacités grandissantes des ordinateurs pourra-t-on un jour modéliser la totalité du fonctionnement cellulaire ? Il n'y a pas de réponse définitive à cette question mais des équipes de recherche travaillent dans ce but (Loew et Schaff, 2001; Slepchenko *et al.*, 2003). Nous sommes encore loin de la cellule virtuelle mais, déjà, la combinaison des études structurales et biochimiques permet de recréer de nombreux systèmes (Aloy et Russell, 2002). Deux stratégies sont développées dans ce but (voir Aloy et Russell (2005) et figure 1) :

- Une approche descendante (*Zoom in*) où, à partir d'une visualisation abstraite d'un réseau d'interactions, on se focalise sur une voie de signalisation précise puis sur une interaction particulière. A cette étape, les informations structurales peuvent être ajoutées à l'information abstraite contenue dans les réseaux.
- Une approche ascendante (*Zoom out*) où, à l'inverse, à partir de structures de protéines isolées, il est possible de reconstituer un complexe binaire puis de replacer celui-ci dans une enveloppe basse résolution (SAXS ou microscopie électronique) (Gherardi *et al.*, 2006) voire même de le repositionner dans la cellule grâce à la tomographie cellulaire (Baumeister, 2005).

Quelle que soit l'approche choisie, l'élément essentiel en est le complexe binaire. En effet, celui-ci est, soit déduit à partir des informations du réseau par l'approche descendante, soit sert de base à l'approche ascendante. Ce complexe doit être construit dans la première approche et n'est pas toujours disponible dans la seconde, c'est pourquoi l'amarrage macromoléculaire *in silico* ou *docking* s'est développé (Aloy *et al.*, 2005). Cette méthode permet, à partir des structures séparées des deux macromolécules, de reconstituer un assemblage binaire. Cette thèse

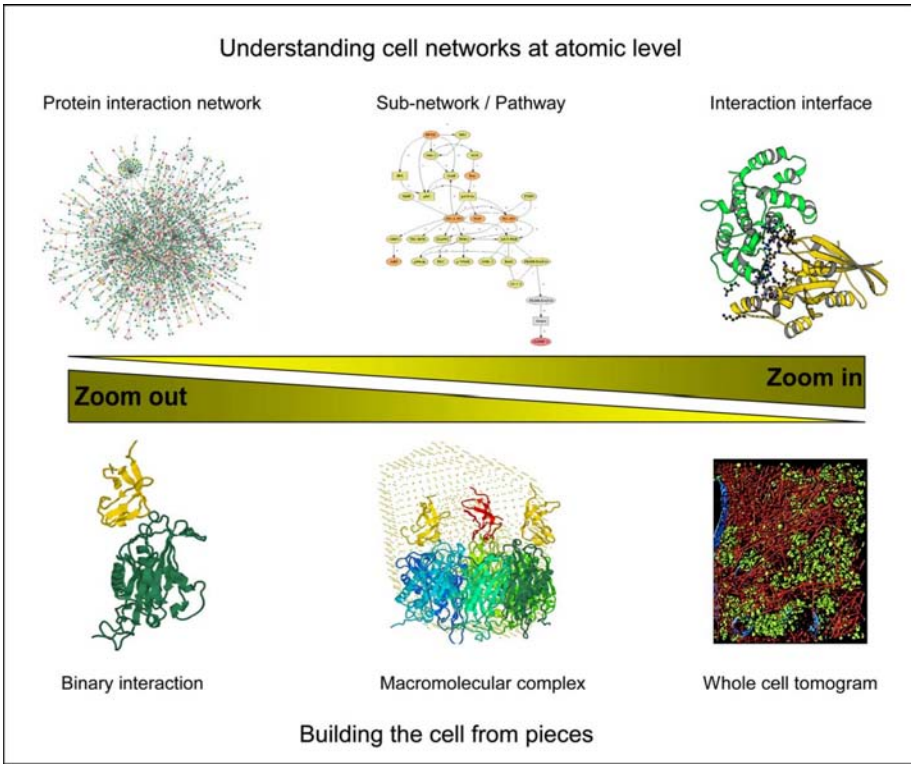


FIG. 1 – *Zoom in and out.* Il est possible de partir des informations contenues dans les réseaux de protéine pour ne se focaliser que sur une voie de signalisation et jusqu'à une interaction précise (*Zoom in*). A l'inverse, on peut commencer à assembler les structures de deux macromolécules en interaction puis replacer cet assemblage dans des enveloppes de plus basse résolution comme celles de cryo-microscopie électronique et enfin replacer l'ensemble dans des coupes tomographiques de la cellule (*Zoom out*). *Figure issue de (Aloy et Russell, 2005)*

décrit une stratégie originale qui pourrait être utilisée pour l'amarrage macromoléculaire *in silico*.

Présentation du mémoire

Ce manuscrit présente les recherches réalisées durant cette thèse. Celles-ci ont été dédiées à la prise en compte de la flexibilité des protéines au cours de l'association. Cette flexibilité a été abordée selon deux voies :

1. L'utilisation de la dynamique moléculaire pour analyser les résultats d'amarrage *in silico* de macromolécules.
2. Le développement d'un nouveau programme, nommé MetaMol, permettant la visualisation de la surface moléculaire.

En effet, la dynamique moléculaire est une méthode couramment employée pour analyser les changements conformationnels des macromolécules tandis que le programme MetaMol visualise les déformations de la surface moléculaire associées à ceux-ci.

La première partie de ce mémoire vise à définir notre domaine d'étude, à savoir les interactions macromoléculaires entre protéines mais aussi entre protéines et acides nucléiques et les programmes permettant de modéliser ces interactions.

Cette partie débutera par une analyse des diverses interactions macromoléculaires existantes entre protéines mais aussi entre protéines et acides nucléiques. Cette analyse se fera sur des complexes obtenus par cristallographie aux rayons X. Nous essaierons ensuite, à partir des données collectées dans les sections précédentes, de déduire les caractéristiques propres à chaque type d'assemblage. Nous comparerons également ces assemblages dits biologiques à des artefacts se formant lors de la cristallisation. Nous ferons ensuite un point sur les programmes permettant de modéliser les complexes macromoléculaires et nous verrons comment un challenge international (CAPRI : *Critical Assessment of PRedicted Interactions*) permet d'évaluer les avancées de chacun. Nous conclurons cette partie sur la nécessité de modéliser la flexibilité des partenaires lors de l'amarrage *in silico*.

Dans une deuxième partie, nous rapporterons les expériences menées au cours de cette thèse démontrant l'utilité de la dynamique moléculaire en solvant explicite pour modéliser la flexibilité des protéines et identifier les résidus clés à l'interface des complexes.

Après un bref rappel sur la cinétique mise en jeu lors de l'association macromoléculaire, nous énoncerons quelques principes de la dynamique moléculaire. Puis nous montrerons comment la dynamique en solvant explicite aide à l'identification de résidus clés pour le complexe PDZ d'Erbin/MH2 de Smad3. Nous verrons ensuite comment des simulations plus courtes permettent de faire converger des résultats de docking au départ différents. Enfin, nous montrerons comment des dynamiques de durée très courtes peuvent raffiner des résultats de docking rigide. Nous prendrons alors comme exemple un complexe protéine/ARN : la cible n° 34 du challenge CAPRI. Nous terminerons cette partie en évoquant les avancées récentes dans le domaine de la dynamique moléculaire.

La troisième partie de ce mémoire sera consacrée aux travaux réalisés lors du développement du programme MetaMol dédié à la visualisation moléculaire haute définition.

Nous ferons d'abord un court rappel sur la visualisation macromoléculaire et l'interactivité dans le domaine de la bioinformatique structurale. Nous décrirons ensuite le programme MetaMol dédié à la visualisation de la Skin Surface Moléculaire. Nous verrons comment la répartition des données sur les processeurs de l'ordinateur et de la carte graphique permet d'accélérer les calculs de la Skin Surface Moléculaire. Enfin, nous terminerons sur les futurs développements du programme.

En conclusion, nous présenterons le bilan des méthodes présentées et montrerons que celles-ci peuvent être combinées pour former une nouvelle stratégie d'amarrage *in silico*.

Chapitre 1

Les assemblages macromoléculaires : de l'analyse à la prédiction

Sommaire

Contexte

Pour passer de la séquence d'un gène à la protéine qui en résulte, de multiples mécanismes sont mis en jeu comme l'ouverture de la double hélice et la lecture d'un brin d'ADN, la création d'un ARN messager, la lecture de cet ARN par le ribosome, etc... Tous ces mécanismes font intervenir des associations macromoléculaires telles que les interactions protéine(s)-ADN, protéine(s)-ARN, protéine(s)-protéine(s). Ces interactions se déroulent en suivant une séquence d'événements particuliers, peuvent être de durée variable et permettent d'assurer le bon fonctionnement cellulaire. Il est intéressant de comprendre les mécanismes moléculaires de cette machinerie complexe afin d'en limiter les dérèglements. Ces derniers peuvent entraîner des effets délétères sur l'organisme tels que des cancers, des maladies auto-immunes,etc... Avec l'avènement de la cristallographie, de nombreuses structures de complexes impliqués dans ces phénomènes

ont été déterminées. Ces structures ont ensuite fait l'objet d'analyses statistiques afin d'en déduire des caractères communs à tous ces assemblages. Le but est d'établir des règles pouvant être utilisées par les programmes pour mimer les mécanismes du vivant et d'essayer de comprendre ceux-ci *in silico*.

Plusieurs études structurales sur divers complexes macromoléculaires (impliquant des protéines et des acides nucléiques ou exclusivement des protéines) ont montré qu'il y avait de nombreuses associations spécifiques à chaque type d'assemblage (Draper, 1999; Luscombe *et al.*, 2000). Ces études ne permettent malheureusement pas de comparer ces derniers. C'est pourquoi, au risque de rester généraliste, nous traiterons dans une première partie les complexes avec les mêmes outils d'analyse pour pouvoir comparer les diverses associations formées. Nous analyserons à travers divers travaux de référence les points suivants :
– la taille et la forme de l'interface formée par les partenaires macromoléculaires,
– les interactions ayant lieu à cette interface,
– le rôle des molécules d'eau dans la cohésion/dissociation du complexe.
Cette étude nous conduira ensuite à mettre en évidence les différences significatives entre des complexes formés exclusivement par des protéines et ceux impliquant des protéines et des acides nucléiques. Puis nous conclurons cette première partie en essayant de répondre à la question : comment caractériser un assemblage ayant une fonction biologique ? Pour cela nous comparerons des assemblages spécifiques à des artefacts de complexes formés pendant la cristallisation.

Dans une deuxième partie, nous verrons brièvement comment ces analyses sur ces différents complexes ont permis d'aboutir à la création de fonctions de score empiriques pour prédire les interfaces des complexes. Puis nous analyserons, à travers l'expérience CAPRI, les outils à notre disposition pour modéliser de tels assemblages biologiques.

1.1 Les interactions protéine-acide nucléique

Dans cette partie nous traiterons les interactions protéines-acides nucléiques dans leur ensemble en mettant en évidence les points communs et les différences entre les complexes impliquant des molécules d'ADN et des molécules d'ARN.

Cette étude est basée sur 15 publications de synthèse sur le sujet (voir tableau 1.1). La première, de Gutfreund *et al.* , est parue en 1998 ; la plus récente, de Bahadur *et al.* , a été publiée en 2008. Les jeux de données utilisés dans ces travaux vont de 26 complexes pour Jones *et al.* jusqu'à 240 pour Luscombe *et al.* . Ces données sont toutes issues de la *Protein Data Bank* (PDB) (Berman *et al.*, 2000, 2002) ou de la *Nucleic Acid Database* (NDB) (Berman *et al.*, 1992). Il est intéressant de noter qu'entre les années 1998 et 2002 ces études se focalisaient surtout sur les interactions protéine-ADN tandis qu'à partir de l'année 2005 et jusqu'à maintenant, ce sont les complexes protéine-ARN qui sont privilégiés. Ceci pourrait s'expliquer par le manque de données disponibles pour les structures protéine-ARN à la fin des années 90. En effet, si l'on regarde les statistiques données par la Protein Data bank [1], nous constatons qu'en 1998 il n'y avait que très peu de structures cristallisées de molécules d'ARN (≈100 contre 500 déjà pour les

[1] http ://www.rcsb.org/pdb/static.do ?p=general_information/pdb_statistics/index.html

molécules d'ADN). De plus, ces structures n'étant pas toutes sous la forme de complexes avec une protéine, il était donc difficile de faire des analyses poussées sur si peu de données. Nous pouvons néanmoins noter que des études sur les interactions protéine-ARN ont été réalisées avant 2005 mais sur des jeux de données assez faibles (Draper, 1999; Nadassy *et al.*, 1999; Jones *et al.*, 2001). Il faudra donc attendre les années 2004-2005 pour avoir un jeu de données plus conséquent (500 structures cristallisées impliquant des molécules d'ARN) afin de faire de études plus approfondies.

Ainsi, du fait de l'étalement dans le temps des études et de l'hétérogénéité du nombre et des types de complexes analysés, ces travaux peuvent donner des résultats très différents. Nous nous efforcerons donc de mettre en évidence les consensus tirés de ces études tout en évoquant les divergences entre les travaux. Pour cela, nous nous focaliserons, comme indiqué précédemment, sur l'analyse de caractères communs, à savoir :
– la taille et la géométrie des interfaces,
– les modes d'interaction des résidus (interactions chargées, hydrophobes, liaisons hydrogène),
– le rôle des molécules d'eau dans ces interfaces.

Tab. 1.1 – Synthèse des travaux sur les interactions protéines-acides nucléiques

Publication	Nb complexes prot-ADN	Nb complexes prot-ARN
Mandel-Gutfreund *et al.* (1998)	43	
Jones *et al.* (1999)	26	
Nadassy *et al.* (1999)	69	6
Luscombe *et al.* (2000)	240	
Pabo et Nekludova (2000)	100	
Jones *et al.* (2001)		32
Treger et Westhof (2001)		45
Selvaraj *et al.* (2002)	62	
Tolstorukov *et al.* (2004)	156	
Lejeune *et al.* (2005)	139	49
Kim *et al.* (2006)		86
Morozova *et al.* (2006)		41
Prabakaran *et al.* (2006)	62	
Ellis *et al.* (2007)		89
Bahadur *et al.* (2008)		81
Moyenne	100	60*

* : La valeur du jeu de données de Nadassy *et al.* étant très différente de celles des autres études, celle-ci n'a pas été prise en compte pour calculer la moyenne

1.1.1 Interface et géométrie des interactions

Taille de l'interface : Elle est très variable d'un auteur à l'autre : Jones *et al.* trouvent une interface entre protéine et ARN allant de 700 à 4900 Å^2. Celle mesurée par Bahadur *et al.* a une taille de 800 à 5200 Å^2. Enfin, chez Kim *et al.* elle varie de 100 à 7000 Å^2 de même que

pour Ellis *et al.* où elle varie de 240 à 14000 Å^2. On peut expliquer cette variation importante entre ces différents travaux par le type de données analysées : Ellis *et al.* et Kim *et al.* ont un jeu de données avec un grand nombre de complexes protéine-ARN impliqués dans le ribosome (respectivement : la moitié et le quart de leurs ensembles). Or ces complexes ont tendance à avoir des interfaces plus grandes que les autres types de complexes protéine-ARN. De plus, ceux-ci constituent une classe à part puisqu'ils peuvent former des interactions de manière permanente à l'inverse de la majorité des complexes protéine-ARN qui s'associent de façon transitoire. La taille moyenne de l'interface protéine-ARN pour Bahadur *et al.* est de 2530 Å^2, ce qui n'est pas trop éloigné de la moyenne trouvée par Jones *et al.* qui est de 2260 Å^2 mais reste bien en deçà de ce que trouve Ellis *et al.* : 3220 Å^2 (et ceci pour la même raison qu'expliqué ci-dessus). Cette interface implique, en moyenne, 43 acides aminés et 17,5 nucléotides (Bahadur *et al.*, 2008).
Dans le cas des interactions protéines-ADN, l'interface varie de 1120 Å^2 à 5800 Å^2 chez Nadassy *et al.* pour une interface moyenne de 3100 Å^2. Cette interface peut être divisée en modules de reconnaissance d'interface moyenne de 1600 Å^2 $\pm$ 400 Å^2 impliquant 24 $\pm$ 6 résidus protéiques et 12 $\pm$ 3 nucléotides. Ces valeurs correspondent à celles obtenues durant la même période par Jones *et al.* avec une interface variant de 1240 Å^2 à 5700 Å^2 pour une interface moyenne d'environ 3200 Å^2.

Ainsi, si l'on exclut les complexes impliqués dans le ribosome, en moyenne, l'interface des complexes protéine-ARN est plus petite que celle des complexes protéine-ADN double brin. Elle reste néanmoins plus grande que celles formées entre des protéines et des molécules d'ADN simple brin (Jones *et al.*, 2001).

Changements au niveau des protéines associées : Il existe 3 grands types de changements conformationnels lorsque la protéine se lie à une molécule d'acide nucléique (illustrés par la figure 1.1) :

1. Des changements dûs à des rotations de chaînes latérales ou à des mouvements limités des atomes du squelette protéique : ces changements sont nombreux (Pabo et Nekludova, 2000).
2. Des passages de structures de formes désordonnées à ordonnées : ceci est visible lorsque les structures cristallographiques de la protéine libre en solution et liée à la molécule d'acide nucléique sont disponibles. Ainsi, des morceaux non visibles dans la forme libre (car trop désordonnés et trop flexibles donc impossibles à cristalliser) sont visibles dans la forme liée et sont souvent ordonnés en structures secondaires comme des hélices α ou des feuillets β (Dyson et Wright, 2002).
3. Les protéines peuvent former des interfaces enserrant complètement l'acide nucléique (Jones *et al.*, 1999) (voir figure 1.8). Ceci peut conduire à de larges changements conformationnels au sein des protéines, que ce soit dans le cas d'interactions protéine-ADN (Nadassy *et al.*, 1999) ou protéine-ARN (Ellis et Jones, 2008). De plus, il est connu qu'un certain nombre de protéines s'associent à une molécule d'acide nucléique sous forme d'homodimères. Lorsqu'il y a fixation d'un homodimère, il arrive qu'il y ait des asymétries au niveau des zones de fixation entre les deux sous-unités et des molécules d'ADN ou d'ARN (Selvaraj *et al.*,

2002; Das *et al.*, 2004). Ceci peut conduire à des changements conformationnels au niveau de la structure quaternaire[2] de la protéine et amener à des spécificités de liaisons entre chaque sous-unité et la molécule d'ADN ; c'est le cas pour les sous-unités du répresseur γ, *Bam*H1 ou *Eco*RV (Selvaraj *et al.*, 2002).

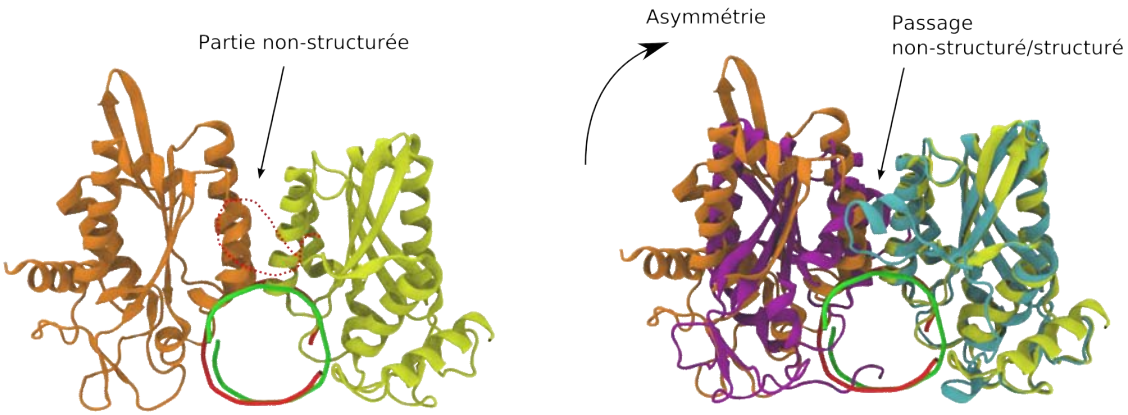


FIG. 1.1 – Changements conformationnels pour l'endonucléase *Bam*H1 : à gauche, dimère de la forme libre de *Bam*H1 (1bam). Ce dimère est obtenu en superposant une des deux sous-unités (la jaune) sur la sous-unité du dimère cristallisé, la seconde sous-unité (en orange) étant placée par symétrie par rapport à la molécule d'ADN (hélice en vert et rouge). A droite, superposition du complexe cristallisé (sous-unités en bleu clair et violet) avec le complexe symétrique. On peut voir les changements conformationnels impliqués dans la fixation de *Bam*H1 (1bhm) avec la molécule d'ADN : 1) déplacements limités d'atomes : ici, visibles car les sous-unités jaunes et cyan ne se superposent pas exactement, 2) passage de structures désordonnées et non visibles dans la structure libre à une forme ordonnée visible dans la forme liée, 3) rupture de la symétrie (ici, rotation vers le fond de l'image de la sous-unité violette par rapport à la sous-unité orange) et donc fixation spécifique de chaque sous-unité sur la molécule d'ADN. *Figure inspirée de (Nadassy* et al.*, 1999).*

Changements au niveau des acides nucléiques : L'association entre la protéine et la molécule d'acide nucléique engendre de fortes tensions au sein de cette dernière (Jones *et al.*, 1999; Dickerson, 1998). Dans le cas de l'ADN, ce changement conformationnel peut être mesuré en calculant le RMSD[3] entre la molécule d'ADN liée et la forme ADN-B[4] considérée comme la forme libre.

Entre 1,5 et 3 Å de RMSD, ces changements sont considérés comme limités. Ils peuvent être dûs à des phénomènes de courbure de la molécule d'ADN et/ou des phénomènes de déroulement

[2]La définition de la structure quaternaire est disponible en annexe A

[3]*Root mean square deviation* : de manière simplifiée, il s'agit de la distance moyenne quadratique entre atomes équivalents après superposition de deux structures macromoléculaires

[4]voir annexe A

de la double hélice permettant l'ouverture du grand sillon où la majorité des interactions ont lieu (Nadassy *et al.*, 1999). Il peut rester également des interactions dans le petit sillon, ces dernières sont d'ailleurs favorisées par des changements de position des sucres (Tolstorukov *et al.*, 2004). On se trouve alors dans une conformation A-ADN[5] qui semble favorisée lors des interactions avec les protéines (Lejeune *et al.*, 2005; Tolstorukov *et al.*, 2004). Elle permettrait à la protéine de se lier plus fortement et ainsi augmenterait la sélectivité de la liaison protéine-ADN (Tolstorukov *et al.*, 2004).

Lorsque le RMSD est supérieur à 3 Å, on peut aussi trouver des changements conformationnels importants. Dans ce cas de figure, la structure de la double hélice est très perturbée. Par exemple, des courbures de la molécule peuvent aller jusqu'à 160° pour le complexe avec le facteur d'intégration IHF, *integration host factor* -Rice *et al.* (1996). Dans le cas des molécules d'ARN, les modifications ne sont pas aussi bien définies. Du fait de sa structure tertiaire complexe, les mauvais appariements et les protubérances au sein de la molécule d'ARN augmentent l'ouverture des grands sillons. Ce phénomène engendre un meilleur accès pour la création de liaisons avec les protéines (Draper, 1999). De plus, la molécule d'ARN s'adapte à son partenaire protéique. Un exemple intéressant est le complexe ARN-U1A, où lorsqu'on isole la structure de l'ARN ; celle-ci décrit une conformation qui ne peut être stable sans l'interaction de la protéine (Leulliot et Varani, 2001).

Ces modifications de la structure jouent un rôle très important dans la reconnaissance des molécules d'acides nucléiques par les protéines qui leur sont associées (Shajani *et al.*, 2006; Spolar et Record, 1994).

Compacité des atomes à l'interface : La compacité des atomes à l'interface est un index de complémentarité géométrique entre les deux surfaces en contact. Nous avons répertorié 3 types de mesures de compacité :
1) l'indice de volume interfacial,
2) le pourcentage d'atomes enfouis,
3) la densité locale (*Local Density* ou L_D).
Ce dernier paramètre se retrouve dans diverses publications. Il représente le nombre moyen d'atomes à l'interface se trouvant jusqu'à 12 Å de la seconde interface. On peut aussi le relier au volume moyen d'atomes enfouis.

Les résultats mesurant cette compacité sont assez divergents. Pour Jones *et al.* , les complexes protéine-ARN ont une interface moins compacte que ceux formés de protéines et d'ADN (Jones *et al.*, 2001), tandis que Bahadur *et al.* obtiennent le résultat inverse. Néanmoins, si l'on regarde les valeurs de la densité locale pour les complexes ADN et ARN chez Bahadur *et al.* , celles-ci sont très proches : en moyenne, $L\_D_{prot} = 37$ et $L\_D_{nucl} = 42$ pour les complexes impliquant des molécules d'ARN et $L\_D_{prot} = 39$ et $L\_D_{nucl} = 46$ pour ceux impliquant des molécules d'ADN. Au vu des résultats, nous pouvons opter pour une vision plus nuancée en admettant que la compacité moyenne pour l'ensemble des assemblages protéine-acides nucléiques est à peu près la même. Cette bonne complémentarité géométrique serait due aux déformations engendrées lors de l'association avec les partenaires protéiques (Doudna, 2000; Westhof et Fritsch, 2000;

[5] voir annexe A

Bon *et al.*, 2008). On peut d'ailleurs noter une meilleure complémentarité des molécules d'ADN simple brin ayant peu de contraintes structurales (Jones *et al.*, 2001; Bahadur *et al.*, 2008). Cette complémentarité géométrique entre protéine et acide nucléique joue un rôle très important dans les processus de reconnaissance (Jones *et al.*, 1999, 2001).

Ainsi, à travers l'étude de l'interface entre protéines et acides nucléiques, un phénomène d'*induced fit*[6] a été mis en évidence (voir figure 1.1). Celui-ci semble d'autant plus important que la surface d'interaction est grande (Nadassy *et al.*, 1999). Ce type de changements conformationnels est, par exemple, un facteur important dans les phénomènes de reconnaissance impliquant des protéines et des molécules d'ARN (Williamson, 2000). Nous pouvons citer comme exemple l'interaction entre l'ARN de transfert et la protéine methionyl-tRNA formyltransferase (MTF) (Ramesh *et al.*, 1999). Des phénomènes équivalents ont aussi été démontrés quelques années plus tôt pour les complexes protéines-ADN (Spolar et Record, 1994).

La complémentarité d'assemblage due à l'organisation en structures secondaires (ou tertiaires) de l'ARN ou de l'ADN provoque de fortes tensions au sein de chaque structure impliquée dans l'interaction (Jones *et al.*, 2001). Ceci pourrait être un mécanisme permettant un mode d'association transitoire (Draper, 1999), les interactions protéines-acides nucléiques n'étant que très rarement permanentes (sauf peut-être au sein du ribosome).

1.1.2 Types de résidus présents à l'interface et chimie de l'interaction

A la fin des années 1970, Seeman *et al.* émettaient l'hypothèse que, comme pour le code génétique où à un triplet de bases correspond un type précis de résidu, un acide aminé spécifique devrait interagir avec une base particulière (Seeman *et al.*, 1976). Malheureusement, cette vision un peu simpliste (et pourtant si séduisante) n'existe pas : il n'y a pas de code simple pour la reconnaissance entre protéines et acides nucléiques (Pabo et Nekludova, 2000). Ceci peut s'expliquer par la flexibilité structurale des partenaires au niveau de l'ensemble de la structure ou au niveau des chaînes latérales.

Cependant, même si l'on ne peut établir de correspondance avec certitude entre un acide aminé et une base, on peut néanmoins dégager des tendances. Nous nous sommes basés sur la synthèse de 7 travaux de références (voir figure 1.2) pour mettre en évidence les acides aminés les plus représentés à l'interface des complexes protéines-acides nucléiques. Nous avons aussi synthétisé dans le tableau 1.2 les caractéristiques les plus remarquables de ces interactions.

Importance respective du squelette et des chaînes latérales dans les interactions : La majorité des études sur le sujet s'accordent à dire qu'au niveau des protéines, les chaînes latérales interagissent plus avec les molécules d'acide nucléique que le squelette. Il est possible d'expliquer ceci par le fait que les interactions à la surface des acides nucléiques se font principalement au niveau des sillons. Ceux-ci étant peu accessibles (du fait de la structure en double hélice pour l'ADN ou de la structure tertiaire complexe pour l'ARN), les seuls atomes pouvant réellement constituer des interactions sont ceux se trouvant au bout des chaînes latérales des

[6] il s'agit de l'adaptation réciproque de la structures des partenaires (Koshland, 1958)

acides aminés. Cette observation est confortée par la prévalence de résidus avec des chaînes latérales longues et flexibles (voir figure 1.2), exception faite du résidu glycine. Il est intéressant de noter que les résidus interagissant le moins fréquemment avec l'ARN le font principalement grâce à leur chaîne principale et ceci par l'intermédiaire de molécules d'eau tandis que les acides aminés interagissant le plus fréquemment utilisent leurs chaînes latérales (Treger et Westhof, 2001).

En ce qui concerne les acides nucléiques, cette fois, c'est le squelette constitué des groupements de sucre et de phosphate qui réalise le plus d'interactions. Nous pouvons ici faire une différence entre les molécules d'ADN pour lesquelles le phosphate est le groupe le plus impliqué dans les interactions et celles d'ARN où le sucre occupe le premier rang (Lejeune *et al.*, 2005). Ceci est directement lié à la structure même de ces molécules d'ADN et d'ARN, le désoxyribose de l'ADN étant remplacé par un ribose au niveau de l'ARN, ce qui lui permet de former plus de liaisons au niveau du groupement hydroxyle 2'OH (Jones *et al.*, 2001; Treger et Westhof, 2001; Lejeune *et al.*, 2005; Bahadur *et al.*, 2008). Cependant, même si le nombre d'interactions au niveau des bases nucléiques est moins élevé, elles n'en restent pas moins essentielles pour les phénomènes de reconnaissance.

Lysines, arginines et résidus chargés : Les résidus arginine, lysine et histidine sont surreprésentés à la surface des protéines interagissant avec les acides nucléiques (voir figure 1.2). Ces résidus vont pouvoir interagir avec les phosphates en formant des ponts salins mais aussi des

Tab. 1.2 – Comparaison des résultats principaux sur les interactions protéines-acides nucléiques issus de Nadassy *et al.* (1999); Jones *et al.* (2001); Treger et Westhof (2001); Lejeune *et al.* (2005); Ellis *et al.* (2007); Bahadur *et al.* (2008).

Paramètres	Nadassy *et al.*	Lejeune *et al.*	Jones *et al.*	Treger *et al.*	Ellis *et al.*	Bahadur *et al.*
Interactions	Prot. - ADN	Prot.-ADN/ARN	Prot. - ARN	Prot. - ARN	Prot. - ARN	Prot. - ARN
Contacts	VdW > l. H	VdW > l. H (ADN) l. H = VdW (ARN)	VdW > l. H	VdW > l. H	VdW > l. H	VdW > l. H (prot.) l. H > VdW (ARN)
Sql. / C. lat.	C. lat. > Sql.	--	--	C. lat. > Sql.	C. lat. > Sql.	C. lat. > Sql.
Sql. / Bases	Sql. > Bases	Bases > Sql.	Sql. > Bases	Sql. > Bases	Sql. > Bases	Sql. > Bases
Rés. Fav.	Lys, Arg, Thr, Ser, Asn	Arg, Lys, His, Asn, Thr	Lys, Tyr, Phe, Ile, Arg	Arg, Asn, Ser, Lys	Trp, Arg, His, Ser, Gly	Arg, Lys, His, Asn, Tyr
Rés. Défav.	Asp, Glu, Leu, Cys	Asp, Glu, Leu, Ile	Asp, Glu, His, Cys	Ala, Leu, Ile, Val	Asp, Glu, Leu, Cys	Asp, Glu, Leu, Val
Bases Préf.	T, G	--	G, U	Pas de Préf	G, A	U, C
Rés. - Bases	Arg-G, Lys-G, Asn-A, Gln-A	Arg-G, Arg-C (pour ADN/ARN)	Arg-U, Asn-G, Asn-U, Ile-G	Ser-A, Asp-G, Asn-U, Glu-G	Tyr-U, Phe-A, Trp-G	--

Ce tableau ne regroupe que les travaux pour lesquels des données suffisantes ont pu être extraites. En rouge, les résultats en contradiction avec le consensus. Abréviations : VdW : contacts de Van der Waals, l. H : liaisons hydrogène, Sql : squelette, C. lat. : chaîne latérales, Rés. Fav./Défav. : résidus favorisés/défavorisés, Bases Préf. : bases préférées, Rés-Bases : interactions résidus protéiques-bases acides nucléiques.

liaisons hydrogène. Cette mise en lumière est valable dans le cas des interactions avec l'ADN et l'ARN. De plus, ces résidus pourront aussi interagir avec les bases de façon préférentielle. Citons, pour exemple, les interactions arginine-guanine trouvées par Nadassy *et al.* et Lejeune *et al.* pour les complexes impliquant des molécules d'ADN et d'ARN. A l'inverse, les résidus acide aspartique et glutamique n'interagissent que très peu avec les acides nucléiques car ceux-ci possèdent des groupements phosphates chargés négativement. Seuls Lejeune *et al.* ont trouvé une forte propension (définition en annexe C) pour l'acide aspartique. Dans cette étude, ce résidu apparaît comme l'un des partenaires les plus fréquents pour la base guanine dans le cas des interactions protéine-ARN. Ce résultat se retrouve, à une moindre mesure, dans les travaux d'Ellis *et al.* et Treger*et al.* .

Résidus aromatiques et interactions hydrophobes : Les interactions non polaires sont les plus représentées au sein des complexes protéines-acides nucléiques (voir tableau 1.2). Ces interactions permettraient de stabiliser les complexes. Celles-ci n'impliquent pas tous les acides aminés non polaires de la même façon : les résidus aliphatiques sont sous représentés tandis que les acides aminés aromatiques semblent plutôt préférés (voir figure 1.2). Ces résidus aromatiques peuvent interagir avec le sucre du squelette, formant ainsi des liaisons C$\cdots$H (Lejeune *et al.*, 2005). Les interactions les plus observées sont les interactions tyrosine-guanine, tyrosine-thymine pour l'ADN et tyrosine-uracile (Lejeune *et al.*, 2005; Ellis *et al.*, 2007) ou triptophane-guanine (Ellis *et al.*, 2007; Baker et Grant, 2007) pour l'ARN. Les cycles aromatiques des bases et des résidus peuvent former des arrangements parallèles ou perpendiculaires dits en *stacking*[7] (Baker et Grant, 2007) et peuvent aussi se regrouper par paire pour faciliter le *stacking* (Kim *et al.*, 2006). Nous pouvons également noter des différences quant à la propension des acides aminés aromatiques pour les complexes impliquant des molécules d'ADN et ceux impliquant des molécules d'ARN. Il est possible de voir une baisse au niveau des résidus aromatiques dans le cas d'interaction avec les molécules d'ADN. En effet, ces interactions ne sont pas (ou très peu) possibles, dans le cas d'une molécule double brin de type B, car les bases sont enfouies. D'où la diminution de la propension des résidus aromatiques (en particulier pour la tyrosine et le tryptophane) à se trouver à l'interface des complexes protéine-ADN. A l'inverse, on peut voir une forte propension de résidus tryptophane, dans l'étude de Lejeune *et al.* , pour les complexes protéine-ARN. Bien que fortement accentuée dans cette étude, cette préférence pour le tryptophane pourrait, dans certains cas, jouer un rôle dans la différenciation entre ADN et ARN (Baker et Grant, 2007).

Liaisons hydrogène, résidus polaires et molécules d'eau : Le nombre moyen de liaisons hydrogène est de 1,4/100 Å^2 pour les complexes protéines-ADN (Jones *et al.*, 1999) et de 1/125 Å^2 pour les complexes protéines-ARN (Bahadur *et al.*, 2008). Il existe 3 types de liaisons hydrogène :

– La liaison hydrogène "classique" entre les groupements X-H$\cdots$Y (où X et Y peuvent être un atome d'oxygène ou d'azote). Ces liaisons se font principalement au niveau du phosphate pour les complexes protéine-ADN (pour 60% de ces interactions) (Nadassy *et al.*,

[7] terme général pour l'empilement des cycles aromatiques

1999) tandis que, dans le cas des interactions impliquant une molécule d'ARN, celles-ci sont mieux réparties entre le sucre et le phosphate (respectivement 36% et 33% des liaisons hydrogène) (Bahadur *et al.*, 2008). Ceci s'explique par la présence d'un groupement hydroxyle 2'OH au niveau du ribose de l'ARN permettant ainsi de créer plus de liaisons hydrogène. Nous pouvons aussi noter que, outre ces interactions chargées, l'arginine peut aussi créer des liaisons hydrogène avec, préférentiellement, la base uracile dans les interac-

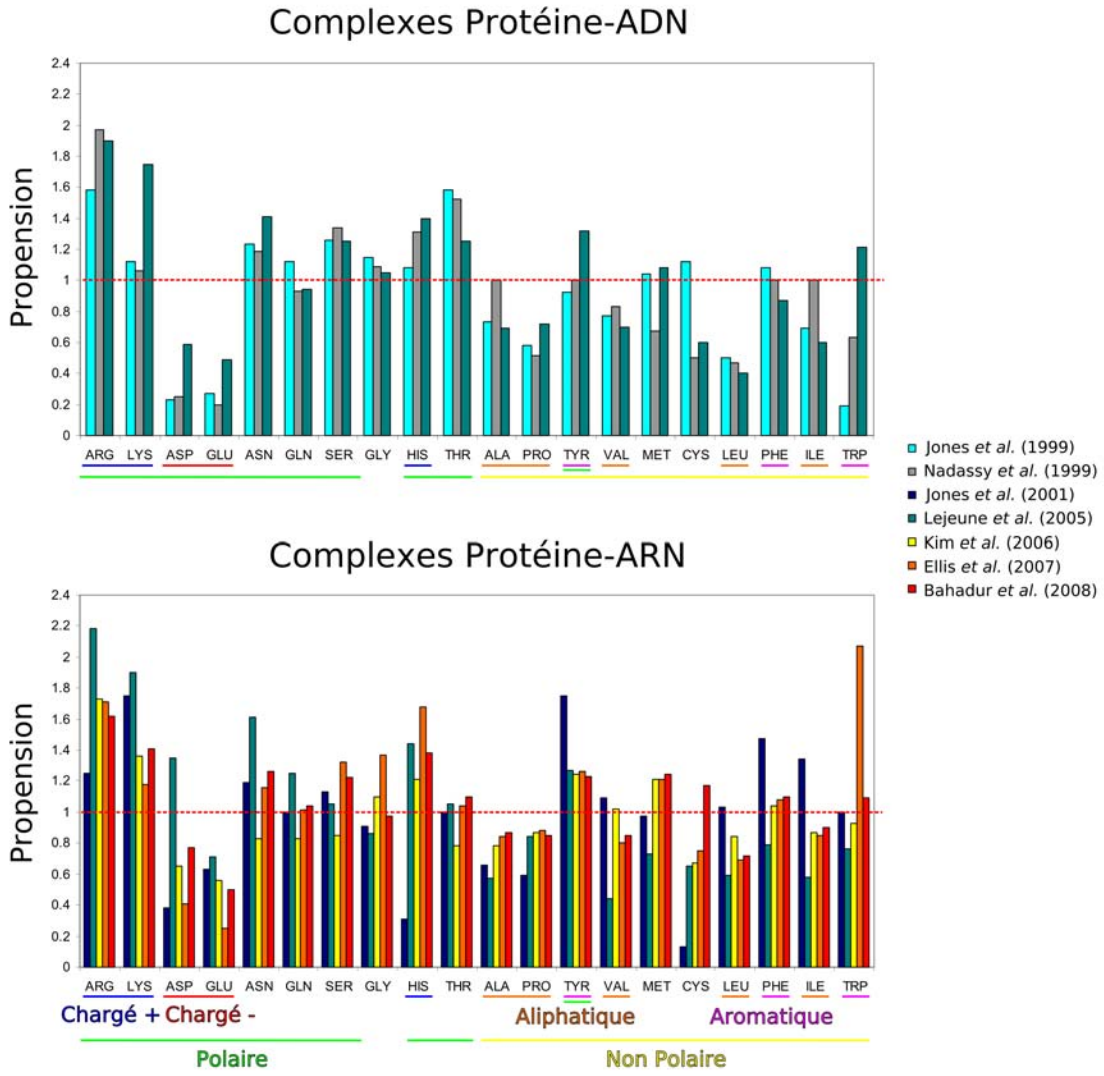


FIG. 1.2 – Synthèse des résultats de Jones *et al.* (1999); Nadassy *et al.* (1999); Jones *et al.* (2001); Lejeune *et al.* (2005); Kim *et al.* (2006); Ellis *et al.* (2007); Bahadur *et al.* (2008). Histogrammes montrant la propension de chaque résidu (voir définition exacte annexe C) à l'interface des complexes protéine-ADN et protéine-ARN. Une propension supérieure à 1 indique qu'un acide aminé se trouve plus fréquemment au niveau de la surface impliquée dans l'interface d'un complexe protéine-acide nucléique que sur le reste de la surface protéique.

tions protéine-ARN. L'asparagine, l'acide glutamique, la glycine, la thréonine ou la tyrosine créent également des liaisons hydrogène qui se font essentiellement avec les bases pour les 4 premiers acides aminés cités et plutôt avec le sucre dans le cas de la tyrosine (Jones *et al.*, 2001).

– La liaison hydrogène CH$\cdots$O impliquant le plus souvent l'atome C5 de la cytosine ou l'atome C5-met[8] de la thymine dans le cas interactions protéine-ADN (Mandel-Gutfreund *et al.*, 1998). Ces interactions CH$\cdots$O représentent plus de 33% des liaisons hydrogène formées entre la protéine et la molécule d'ARN (Treger et Westhof, 2001).

– Enfin, les liaisons hydrogène impliquant des molécules d'eau. Du fait de la nature polaire des acides nucléiques les molécules d'eau sont présentes à l'interface des complexes protéine-acides nucléiques et y jouent un rôle important (Nadassy *et al.*, 1999; Jayaram et Jain, 2004). Ces interactions permettent parfois de ponter des atomes de part et d'autre de l'interface et minimisent les interactions entre acides aminés de même charge.

En conclusion, il n'existe pas de code simple pour les reconnaissances protéine-acides nucléiques (Pabo et Nekludova, 2000). En effet, chaque type de complexe a des spécificités propres. Les interactions protéines-ADN impliquent majoritairement des atomes provenant du phosphate alors que les interactions protéine-ARN impliquent plus fréquemment les bases ou le sucre. Les liaisons hydrogène sont possibles avec le groupe 2'OH pour les complexes impliquant des molécules d'ARN et absentes dans les complexes protéine-ADN.

De plus, pour les interactions protéine-ARN ou protéine-ADN on trouve au sein de chaque famille des modes d'association avec des structures spécifiques comme hélice-boucle-hélice, en doigts zinc... De même, dans le cas de sous-unités identiques se liant sous forme de dimères, les modes d'association peuvent différer (Selvaraj *et al.*, 2002). Le but de cette étude n'étant pas de rentrer plus dans les détails, je laisse le soin au lecteur voulant aller plus en avant dans cette recherche de consulter les revues Luscombe *et al.* (2000) pour approfondir le sujet des interactions protéine-ADN et Draper (1999) comme base de travail pour les interactions protéine-ARN.

Il reste des caractéristiques communes à l'ensemble des complexes protéines-acides nucléiques. D'un point de vue géométrique, on remarque l'importance des interactions au niveau du grand sillon (indispensable au phénomène de reconnaissance et de spécificité) ainsi que la mise en place d'*induced fit* entraînant une meilleure complémentarité des partenaires. D'un point de vue chimique, nous pouvons noter l'importance des résidus chargés positivement à la surface des protéines allant de pair avec la surface chargée négativement au niveau des acides nucléiques. Nous avons aussi remarqué une fréquence élevée des acides aminés aromatiques. Ceux-ci pourraient jouer un rôle dans la sélectivité entre les molécules d'ADN ou d'ARN. Enfin, au niveau des acides nucléiques, nous avons constaté que certaines bases étaient préférées comme la guanine, l'adénine ou l'uracile.

[8] positions C5 présentées en annexe A

1.2 Les associations protéine-protéine

De nombreuses protéines s'associent en complexes multimériques. Nous n'évoquerons pas cet aspect dans cette partie mais nous nous concentrerons sur les assemblages protéiques binaires : les dimères (pour plus de détails sur les assemblages oligomériques voir Ponstingl *et al.* (2005)). Nous verrons, à travers cette partie, qu'il existe plusieurs types d'associations tels les complexes homodimériques et hétérodimériques. Les homodimères sont formés de deux sous-unités identiques. Dans la plupart des cas, ces sous-unités ne se trouvent pas dans un état stable lorsqu'elles sont isolées et leur structuration se fait en même temps que la formation du complexe. Il arrive cependant que ces homodimères puissent se dissocier en monomères ou former des assemblages d'oligomères plus grands suivant la concentration du pH, la fixation d'un ligand ou d'autres paramètres pouvant moduler les interactions. Les hétérodimères sont formés de deux sous-unités différentes. Ces sous-unités sont, la plupart du temps, stables de manière indépendante et peuvent chacune interagir avec d'autres partenaires pour remplir une fonction particulière (voir figure 1.3). De nombreuses fonctions cellulaires se font à travers de tels complexes ayant des temps de vie très courts. Dans cette partie, nous essaierons de différencier les homodimères des hétérodimères en nous basant sur les paramètres dont nous nous sommes déjà servi :
– la taille et la géométrie de l'interface,
– les composantes électrostatiques, hydrophobiques et polaires de ces interfaces,
– les molécules d'eau présentes à l'interface.

Les travaux sur les assemblages protéiques sont bien plus avancés que ceux sur les complexes protéine-acide nucléique. De ce fait, nous trouvons des milliers de références à notre disposition dans les bases de données bibliographiques. Au 15 août 2008, il existait ≈15000 références dont ≈1100 publications de synthèse trouvées dans PubMed pour les mots clés : "protein complex". Nous pouvons aussi noter que, contrairement aux assemblages protéine-acide nucléique pour lesquels les résultats peuvent être divergents, les résultats traités dans cette partie (travaux publiés à partir de 1996) se recoupent dans la majorité des cas.

1.2.1 Interface et géométrie des interactions

Géométrie et taille de l'interface : La forme et la courbure de l'interface sont intéressantes à étudier. Le caractère planaire d'une interface est analysé en calculant le RSMD de tous les atomes à l'interface par rapport au plan des moindres carrés passant par tous les atomes. Si tous les atomes se trouvent sur le plan, l'index de planarité devrait être égal à 0. L'index de planarité est de 3,5 Å pour les homodimères et 2,8 Å pour les hétérodimères (Jones et Thornton, 1996). Les interfaces d'homodimères sont donc un peu moins planes que celles des hétérodimères, ce qui peut être corrélé à une meilleure complémentarité de forme (voir figure 1.3). Ces valeurs indiquent surtout que les interfaces protéine-protéine sont planes quel que soit le type d'association, à l'exception de complexes enzyme-inhibiteur. Dans ce cas particulier, les sites de fixation côté inhibiteur peuvent prendre une forme convexe tandis que les sites de fixation côté enzyme seront de forme concave.

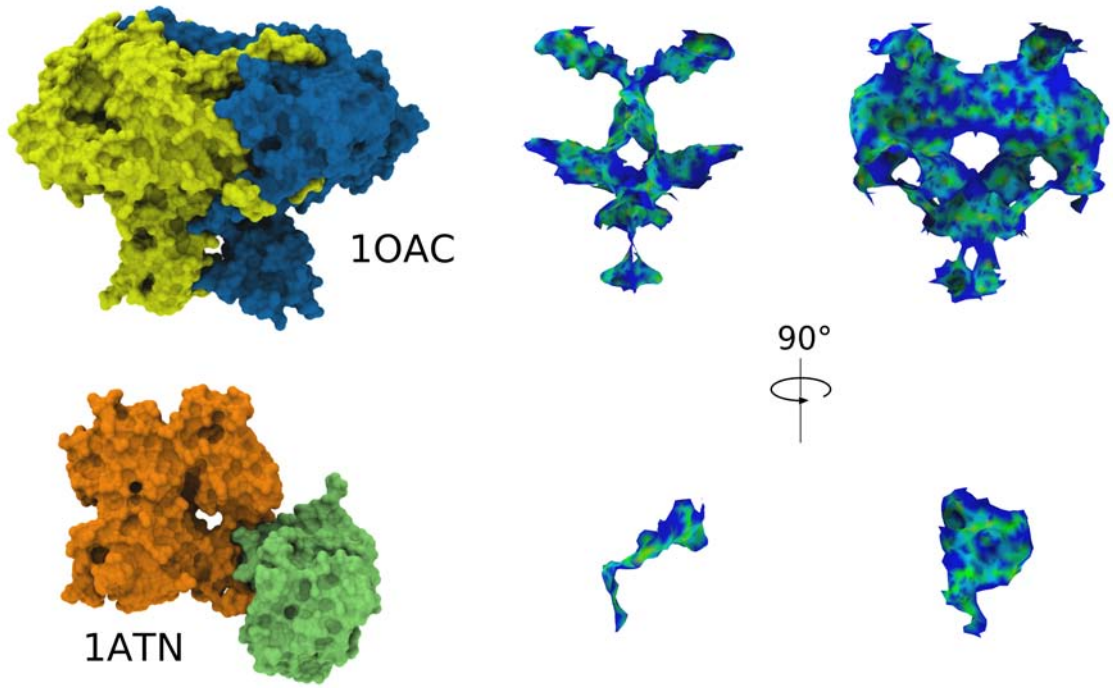


FIG. 1.3 – Exemple d'assemblages homodimériques et hétérodimériques. En haut à gauche, la cuivre amine oxydase (1AOC), homodimère ayant une surface enfouie de 14300 Å^2. En bas, le complexe hétérodimèrique, Actine-DNase (1ATN) ayant une surface enfouie de taille standard : 1770 Å^2. A côté de chaque complexe, visualisation de son interface colorée en fonction de la distance entre chaque partenaire, jaune entre 2 et 3 Å, vert entre 3 et 4 Å et bleu supérieur à 4 Å. *Valeurs des surfaces issues de (Chakrabarti et Janin, 2002; Bahadur* et al.*, 2003). Interface réalisée avec le programme Intersurf (voir Annexe B).*

Un autre facteur, donné par Jones et Thornton, est la circularité de l'interface. Ce paramètre est le ratio entre les longueurs des axes principaux de l'interface (un ratio de 1 définissant une interface pratiquement circulaire). Pour l'ensemble des complexes protéiques, cette interface n'est pas parfaitement circulaire. Elle est égale à 0,71 pour les homodimères et 0,73 pour hétérotérodimères.

En moyenne, les interfaces d'homodimères sont 2 fois plus grandes que celles des hétérodimères (Bahadur *et al.*, 2003) et, au delà de 4000 Å^2, Nooren et Thornton ne trouvent plus que des homodimères. La taille de l'interface peut donc permettre de différencier ces deux types d'assemblages pour les cas extrêmes (interfaces de petite ou très grande taille). Par contre, il existe une plage de valeurs d'interface qui peut correspondre à des assemblages homodimériques ou hétérodimériques.

Changements au niveau des protéines : Comme dans le cas des interactions protéines-acides nucléiques, on trouve 3 types de changements conformationnels. Ces changements peuvent être mesurés en calculant la valeur de RMSD entre la structure liée de la protéine et sa structure

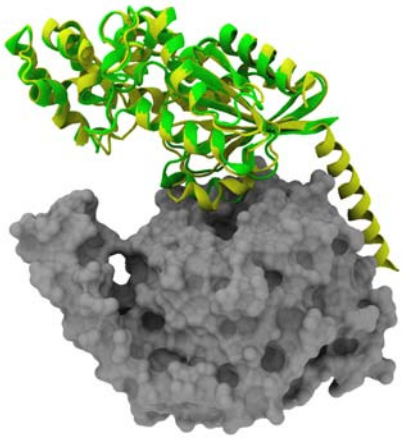


FIG. 1.4 – Exemple de changements conformationnels pour le complexe Gα-G$\beta\gamma$. Superposition de la forme libre de la sous-unité α (en vert) sur la forme liée de celle-ci (en jaune). On peut voir les changements de faible amplitude, les mouvements plus importants de l'hélice α se trouvant à l'interface. On remarque aussi le passage de structure ordonnée à désordonnée : l'hélice alpha la plus à droite est présente dans la forme liée (en jaune) et absente (car trop désordonnée pour être structurée) dans la forme libre de la protéine (en vert).

libre si celle-ci a été cristallisée (voir figure 1.4) :

1. Si le RMSD < 2 Å, les mouvements conformationnels sont limités. Ceux-ci impliquent des réorientations de chaînes latérales ou de légers mouvements de la chaîne principale. Ce type de mouvements est impliqué dans la plupart des associations protéase-inhibiteur ou antigènes-anticorps.
2. Si le RMSD > 2 Å, les mouvements sont de plus grande ampleur, transformant la forme et la composition chimique de l'interface : c'est le phénomène d'*induced fit*. Il apparaît pour des interfaces > 2000 Å^2, entraînant d'importants changements conformationnels. Les déformations peuvent se localiser au niveau des régions de boucles et affecter la structure des protéines entraînant, par exemple, le passage d'un état désordonné à ordonné. Ceci est valable pour de nombreux complexes et permet aux protéines de réaliser diverses fonctions biologiques (Dunker *et al.*, 2001) ou de se lier de différentes manières (Dyson et Wright, 2002). Il est intéressant de noter que ce type de passage de structure désordonnée-libre à ordonnée-liée est à la base des interactions homodimériques.
3. Ces mouvements peuvent aussi engendrer des changements dans la structure quaternaire, c'est à dire de larges mouvements de sous-domaines permettant à l'assemblage protéique d'adopter une nouvelle conformation (Lo Conte *et al.*, 1999).

Compacité des atomes à l'interface : Lo Conte *et al.* ne trouvent pas de différence entre la compacité des atomes au sein des protéines et celle au niveau de l'interface (Lo Conte *et al.*,

1999). Ceci a été nuancé par Ponstingl *et al.* pour qui les atomes semblent moins densément compactés à l'interface qu'à l'intérieur de la protéine. Par contre, si l'on se limite à l'étude des atomes enfouis, ceux-ci sont plus densément compactés (Ponstingl *et al.*, 2005). De plus, Halperin *et al.* ont montré qu'au niveau des interfaces, les résidus ayant un rôle important sont entourés de régions plus denses (Halperin *et al.*, 2004). De même, on constate un fort compactage des atomes hydrophobes à l'interface des homo-oligomères (Jones et Thornton, 1997).

Néanmoins, faire la différence entre les homodimères et les hétérodimères par rapport au seul critère de compacité à l'interface reste délicat. C'est le constat que font Bahadur *et al.* dans une étude se basant sur divers paramètres mesurant la densité atomique au niveau de l'interface des homodimères et hétérodimères (Bahadur *et al.*, 2004).

1.2.2 Types de résidus présents à l'interface et chimie de l'interaction

De nombreux travaux ont été réalisés sur l'étude des résidus à l'interface, que ce soit du point de vue de la structure ou du point de vue de la séquence (Ofran et Rost, 2007), dans le but de caractériser les interactions. Nous reportons dans cette partie les résultats d'études sur la fréquence des résidus à l'interface (voir figure 1.5) et complétons ceux-ci par des travaux plus spécifiques sur l'étude des caractères électrostatiques ou hydrophobiques des résidus.

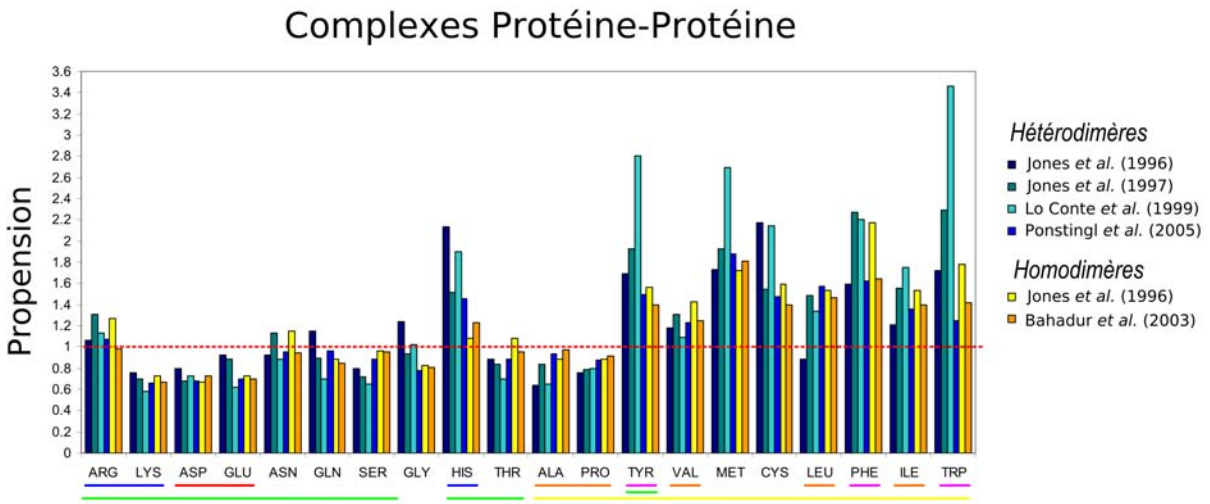


FIG. 1.5 – Synthèse des résultats de Jones et Thornton (1996, 1997); Lo Conte *et al.* (1999); Bahadur *et al.* (2003); Ponstingl *et al.* (2005). Histogrammes montrant la propension de chaque résidu (voir définitions annexe C) à l'interface des complexes protéine-protéine. Une propension supérieure à 1 indique qu'un acide aminé se trouve plus fréquemment au niveau de la surface impliquée dans l'interface d'un complexe protéine-protéine que sur le reste de la surface protéique. Code couleur équivalant à la figure 1.2

Résidus aliphatiques et interactions hydrophobes : A travers les publications de synthèse sur le sujet, il est clair que les résidus non polaires ont un rôle central dans les interactions protéine-protéine. La figure 1.5 montre une fréquence plus élevée des résidus hydrophobes à l'interface. Ce sont plus exactement les résidus aliphatiques (à l'exception de l'alanine et de la proline) et aromatiques qui sont surreprésentés.

On peut également trouver à l'interface des paires préférées entre résidus aromatiques : par exemple, la paire tryptophane-tyrosine. De même, on trouve des paires préférées entre résidus aromatiques et résidus aliphatiques comme la paire tyrosine-leucine (Glaser *et al.*, 2001). Il y a aussi une forte propension en résidus cystéine : en effet, ce résidu peut créer des ponts disulfides ainsi que des liaisons chimiques fortes permettant la stabilisation de l'interface. Ce résidu est d'ailleurs très conservé à l'interface des protéines (Hu *et al.*, 2000). Nous pouvons noter ici que la plupart des études montrent que le caractère hydrophobe de l'interface est plus prononcé pour les homodimères que pour les hétérodimères (Bahadur *et al.*, 2003; Jones et Thornton, 1996). Dans notre cas, la compilation des résultats figure 1.5 ne montre pas de différences marquées.

Ainsi, la composition en acides aminés des interfaces protéiques est enrichie en résidus hydrophobes. De plus, la densité atomique au niveau des interfaces est proche de celle trouvée au sein des protéines. La question qui se pose est donc de déterminer s'il existe le même effet hydrophobe au sein de l'interface qu'à l'intérieur des protéines. Ceci est d'autant plus intéressant dans le cas des complexes homodimériques formant des assemblages stables puisque l'association entre domaines ressemble à l'association entre chaînes lors de la structuration de la protéine. Or il n'en est rien : l'effet hydrophobe au niveau des interfaces protéiques n'est pas le même qu'au sein des protéines. Cet effet est plus fort au sein des protéines qu'à leur interface (Tsai *et al.*, 1997). L'interface est donc une structure intermédiaire entre la surface de la protéine et l'intérieur de celle-ci (Tsai *et al.*, 1997; Lo Conte *et al.*, 1999; Ma *et al.*, 2003). Cette différence entre l'intérieur et l'interface des protéines est liée à la composition en acides aminés polaires chargés et neutres.

Arginine, histidine et résidus chargés : La figure 1.5 montre que les résidus chargés ont une propension bien inférieure à celle des résidus non polaires. Ainsi, on pourrait penser que les résidus chargés jouent un rôle moins important pour les interactions protéiques (si l'on compare par exemple aux interactions protéine-acides nucléiques). Il faut en fait nuancer ceci car les résidus chargés peuvent créer des ponts salins : 2 par interface (Sheinerman *et al.*, 2000), ainsi que des liaisons hydrogène. De plus, ceux-ci permettraient une reconnaissance à longue distance des partenaires (Sheinerman *et al.*, 2000). Nous verrons d'ailleurs dans la suite de cette thèse un exemple concret de ce phénomène pour l'interaction entre le domaine PDZ d'Erbin et le domaine MH2 de Smad3.

Si l'on détaille la propension pour chaque résidu chargé, on remarque que les résidus chargés négativement ainsi que la lysine sont peu représentés à l'interface. Par contre, les résidus arginine et histidine sont préférés. En effet, l'arginine peut servir de "point d'attache" (*anchor residue*) pour les interactions (Rajamani *et al.*, 2004). De plus, la différence avec la lysine est due à une meilleure capacité du groupe guanidinium de l'arginine à établir des liaisons hydrogène comparé au groupe amine de la lysine (Bahadur *et al.*, 2008). Le résidu arginine est d'ailleurs

l'un des résidus les plus conservés (Hu *et al.*, 2000). Nous notons également des interactions moins courantes comme les interactions impliquant une arginine et un résidu aromatique formant des interactions cation-π (Crowley et Golovin, 2005).

En ce qui concerne le résidu histidine, une différence entre homodimères et hétérodimères se fait quant à la propension de celle-ci, ce résidu étant plus abondant à l'interface des hétéro-complexes. L'histidine, comme l'arginine, peut se lier avec d'autres résidus par des ponts salins ou des liaisons hydrogène dans les complexes impliquant deux protéines différentes (Sheinerman et Honig, 2002). Elle joue néanmoins un rôle dans les interfaces d'homodimères comme pour l'enzyme cuivre amine oxydase où l'on trouve 4 histidines conservées impliquées dans l'interface (Parsons *et al.*, 1995).

Les résidus chargés sont donc, à plus d'un sens, importants pour l'association protéique. Une compilation de résultats de mutagénèse par alanine (*alanine-scanning mutagenesis*) a d'ailleurs démontré que, dans la majorité des cas, les mutations d'acides aminés chargés entraînent une déstabilisation du complexe (Bogan et Thorn, 1998).

Résidus polaires, liaisons hydrogène et molécules d'eau : La figure 1.5 ne montre pas de préférence particulière pour les acides aminés polaires, les valeurs de propension étant pratiquement celles de la surface protéique, c'est à dire proches de 1. Bien que moins présents à l'interface des protéines, les résidus polaires sont importants pour l'interaction. Ils sont d'ailleurs (si l'on inclut aussi les résidus chargés arginine et acide aspartique) les plus conservés à l'interface (Hu *et al.*, 2000). Ces résidus peuvent former des liaisons hydrogène à l'interface. Néanmoins, on trouve aussi des liaisons hydrogène impliquant les autres résidus. D'ailleurs, 61% des interactions polaires se font avec la chaîne principale (Bahadur *et al.*, 2003). On trouve en moyenne 10 liaisons hydrogène pour 1000 $Å^2$ d'interface (Ponstingl *et al.*, 2005) ou 1 liaison hydrogène pour 75 $Å^2$ d'interface polaire (Bahadur *et al.*, 2003). Comme pour les complexes protéine-acide nucléique, nous retrouvons trois types de liaisons hydrogène :

1. La liaison hydrogène classique du type : X-H$\cdots$Y (où X et Y peuvent être un atome d'oxygène ou d'azote).
2. La liaison hydrogène CH$\cdots$O.
   Elle est plus faible que la liaison hydrogène conventionnelle. On peut voir apparaître ce type d'interaction entre 2 protéines, au niveau des brins β parallèles et antiparallèles, créant ainsi un feuillet β intermoléculaire (Jiang et Lai, 2002).
3. Les liaisons hydrogène formées grâce à des molécules d'eau.
   En moyenne on trouve 10 molécules d'eau pour 1000 $Å^2$ dans le cas des homodimères et des hétérodimères (Rodier *et al.*, 2005). Par contre, la répartition des molécules d'eau ne se fait pas de la même manière. En général, les interfaces des homodimères sont constituées d'un coeur hydrophobe entouré par une périphérie hydrophile où peuvent venir s'intercaler des molécules d'eau (Bahadur *et al.*, 2003). Dans le cas des complexes hétérodimériques, il arrive que l'on retrouve des molécules d'eau réparties à travers toute l'interface, on parle alors d'interfaces *humides* (Janin, 1999). Ces molécules d'eau peuvent former des liaisons hydrogène avec la chaîne principale de la protéine (45% des interactions) ou avec les chaînes latérales chargées ($\approx$29%) ou neutres ($\approx$27%) (Rodier *et al.*, 2005). Pour une re-

vue plus détaillée sur le rôles des molécules d'eau dans la reconnaissance moléculaire, nous vous conseillons la publication de synthèse suivante : *Water Mediation in Protein Folding and Molecular Recognition* (Levy et Onuchic, 2006).

Ainsi, même s'il est possible d'ébaucher quelques caractères distinctifs entre les homodimères et hétérodimères, il reste difficile de faire une réelle distinction entre ces assemblages. Et ce, pour une raison évidente, il s'agit d'interactions soumises aux mêmes règles physiques et biologiques de la reconnaissance. De plus, on ne peut pas non plus classer précisément les homodimères comme des molécules liées et les hétérodimères commme des molécules non liées. Il existe plutôt un continuum et la stabilité du complexe va dépendre des conditions physiologiques de l'environnement (Nooren et Thornton, 2003a). On peut identifier 3 types de contrôle :

1. La rencontre.
   L'interaction est due à la rencontre entre les surfaces d'interaction des deux partenaires, ce qui demande une co-localisation dans le temps et l'espace. Il faut donc que les partenaires se situent dans un même lieu. Si ce n'est pas le cas, des phénomènes de diffusion ou de transport (par exemple vasculaire) sont nécessaires.
2. La concentration locale.
   Des mécanismes de contrôle comme l'expression des gènes, les niveaux de sécrétion, la dégradation des protéines, le stockage etc.. altèrent la concentration des partenaires.
3. L'environnement physiologique local.
   L'affinité des partenaires peut être altérée par la présence d'une autre molécule (ATP) ou par un changement des conditions physiologiques (concentration en ions).

Enfin, l'interface entre les protéines ne représente pas un ensemble indivisible mais devrait plutôt être vue comme une réunion de parcelles unitaires (voir figure 1.6). Ces parcelles peuvent être définies comme un noyau hydrophobe entouré par une périphérie hydrophile (Lo Conte *et al.*, 1999; Chakrabarti et Janin, 2002; Bahadur *et al.*, 2003). Cette périphérie peut être accentuée par des molécules d'eau à l'interface venant ponter les partenaires (Rodier *et al.*, 2005).

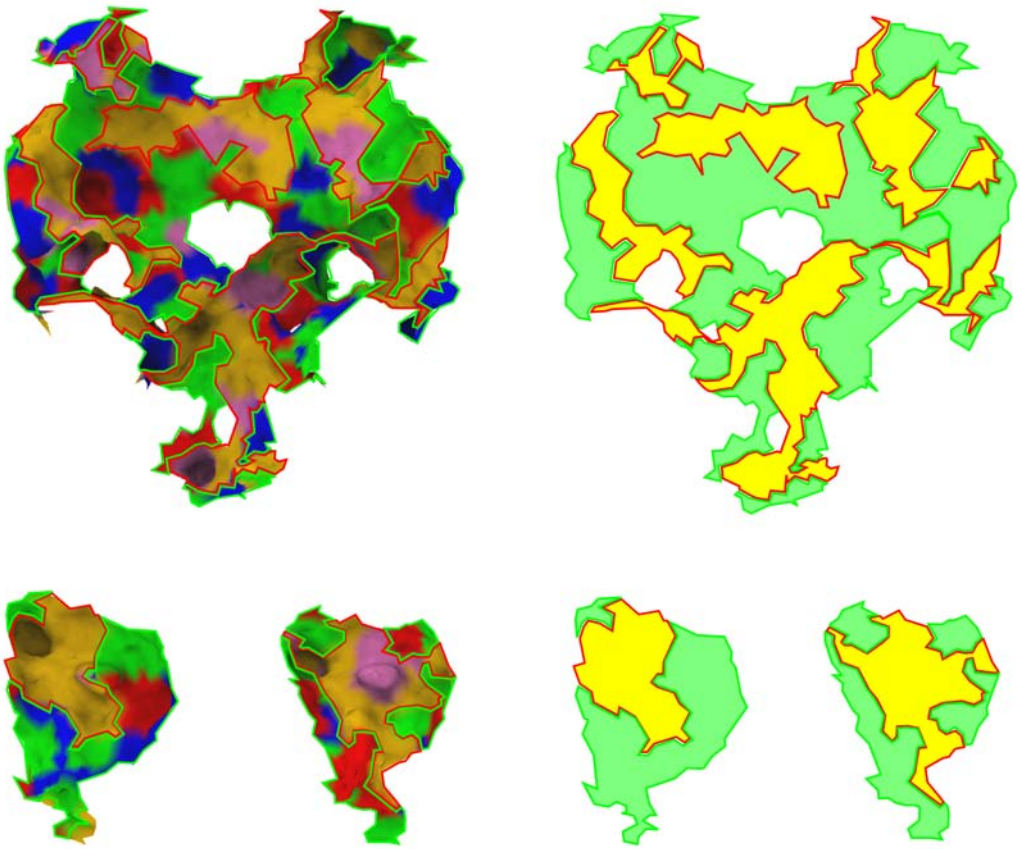


FIG. 1.6 – Interfaces des complexes 1ATN et 1OAC présentés figure 1.3 colorées en fonction des types de résidus. Pour la parcelle du complexe hétérodimérique (1ATN), l'interface est colorée en fonction des résidus de chaque partenaire. Si l'on regarde l'interface schématique, à droite, il apparaît des parcelles non polaires (en jaune) entourées par des parcelles polaires (en vert). On peut voir que les modules hydrophobes et hydrophiles ne sont pas identiques de part et d'autre de l'interface du complexe hétérodimérique (1OAC). Les résidus non polaires (Ala, Cys, Ile, Leu, Met, Pro et Val) sont représentés en orange, les résidus aromatiques (Phe, Trp, Tyr) en violet, les résidus chargés négativement (Asp, Glu) en rouge, les résidus chargés positivement (Arg, His, Lys) en bleu et les résidus polaires neutres (Asn, Gln, Ser, Thr) et Glycine sont représentés en vert.

1.3 Comment caractériser un assemblage macromoléculaire ?

Après avoir analysé séparément les interactions protéines-protéines et protéines-acides nucléiques, nous allons, dans cette partie, confronter les divers résultats trouvés afin d'essayer de caractériser chaque type d'interaction. De plus, nous tenterons de différencier les interactions que nous qualifierons de *biologiques* des interactions dites *physiques*, correspondant à des contacts cristallins.

1.3.1 Assemblages protéiques *vs* assemblages protéine-acide nucléique

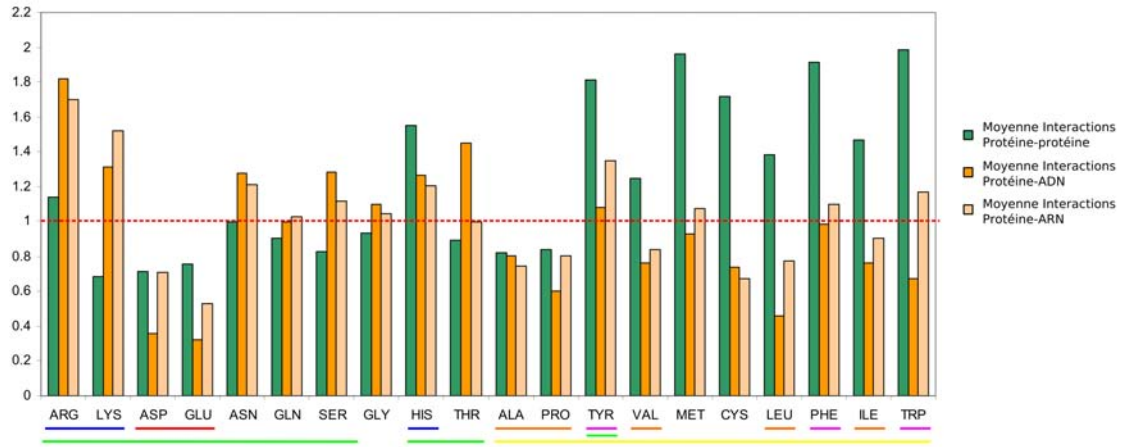


FIG. 1.7 – Synthèse des parties précédentes. Les valeurs de chaque histogramme correspondent aux moyennes des valeurs présentées en figures 1.2 et 1.5.

Nous pouvons donc maintenant faire un bilan sur les interactions impliquant des protéines avec des acides nucléiques ou d'autres protéines. Ce bilan, comme les analyses précédentes, se fera autour de deux thèmes principaux : les caractères géométriques de l'interface ainsi que sa composition en acides aminés.

Nous pouvons voir sur la figure 1.7 des tendances communes aux divers types de complexes tandis que d'autres sont spécifiques d'un type donné. Ainsi, il apparaît que le résidu arginine est surreprésenté à l'interface de tous les types de complexes, avec cependant une préférence pour ceux impliquant des acides nucléiques. De même, la lysine est préférée pour ces mêmes complexes tandis qu'elle est défavorisée à l'interface des complexes protéiques. Pour expliquer cela au niveau des interactions protéiques, une hypothèse pourrait être une compétition entre les résidus arginine et lysine. En effet, l'arginine peut être favorisée par rapport à la lysine, du fait de sa plus grande flexibilité et de son groupement guanidinium plus apte à créer des liaisons hydrogène et des ponts salins que le groupement amine de la lysine. Dans le cas des complexes protéine-acides nucléiques, la surface chargée négativement de ces derniers n'aurait pas entraîné de sélection vis-à-vis d'un des deux acides aminés. Nous pouvons aussi noter une diminution des acides aminés chargés négativement à l'interface de tous les complexes comparé à la surface des protéines. Ce résultat est compréhensible au niveau des complexes protéine-acide nucléique

où la surface chargée négativement de ces derniers défavorise de tels acides aminés. Ce résultat est un peu plus surprenant au niveau des interfaces protéiques. Pourtant, ces acides aminés ont un rôle notable au sein des interfaces. Ils peuvent créer des ponts salins mais aussi des liaisons hydrogène. De mêmes, ils interviennent dans la reconnaissance des complexes à longue distance. Enfin, on trouve souvent ce type de résidu proche d'un résidu lysine ou arginine. Ils forment alors un dipôle (+-) se liant avec d'autres dipôles de charges opposées (-+), ce qui renforce les liaisons à l'interface. Ce phénomène est visible pour tous les types de complexes (voir figure 1.6 ou Kim *et al.* (2006)). Enfin, le résidu histidine est lui aussi favorisé ; il cumule deux caractères importants pour se trouver à l'interface : sa capacité à être chargé positivement et sa chaîne latérale aromatique.

Il apparaît aussi un consensus quant à la distribution des résidus aromatiques (à l'exception du tryptophane) : ceux-ci sont favorisés dans tous les types de complexes (avec une préférence pour le résidu tyrosine). Ces résidus peuvent interagir entre eux en formant des structures en *stacking* mais aussi avec des résidus chargés formant ainsi des interactions cation-π (Gallivan et Dougherty, 1999; Crowley et Golovin, 2005). De plus, en interagissant avec les bases, ces résidus jouent un rôle important dans les phénomènes de reconnaissance des acides nucléiques (Draper, 1999; Baker et Grant, 2007).

La composition en acides aminés polaires neutres et non polaires varie très fortement d'un type de complexe à l'autre. Dans le cas des interactions protéiques, on trouve une forte propension en acides aminés non polaires (à l'exception de l'alanine et de la proline) tandis que les acides aminés neutres ne sont pas favorisés. Si l'on reprend la définition de parcelle unitaire évoquée précédemment, avec un noyau fortement hydrophobe et une périphérie hydrophile, des contacts entre zones hydrophobes vont se former pour réduire leurs accessibilités respectives au solvant ; ceci est énergétiquement favorable. Cet assemblage de *patchs* hydrophobes peut d'ailleurs entraîner de forts changements conformationnels lors de la formation du complexe dû à des phénomènes de désolvatation. A l'inverse, les résidus polaires neutres sont favorisés au sein des interfaces protéine-acide nucléique en raison de la surface polaire de ceux-ci. Ces résidus participent à la formation de liaisons hydrogène importantes dans les phénomènes de reconnaissance et de spécificité des acides nucléiques. Nous pouvons noter ici que cette vision est, encore une fois, globale et qu'il existe toujours des exceptions à la règle. Ainsi, on trouve aussi des interfaces constituées par un noyau hydrophobe entouré d'une périphérie hydrophile au niveau des complexes protéines-acides nucléiques ; c'est le cas, par exemple, de la protéine se liant à la boîte TATA pour les interactions protéine-ADN.

Enfin les molécules d'eau jouent aussi un rôle important dans les interactions (Janin, 1999). Au niveau des interfaces protéine-protéine, ces molécules sont aussi nombreuses que les liaisons hydrogène (10 liaisons pour 1000 $Å^2$). Elles stabilisent l'interface en créant des liaisons hydrogène supplémentaires mais également les interactions entre résidus de même charge en se plaçant à l'interface de ceux-ci. Les molécules d'eau participent aussi à renforcer la complémentarité géométrique au niveau de l'interface ; ceci est valable autant pour les interactions protéiques au niveau des interfaces *humides* que pour des interactions protéine-acide nucléique où l'interface, plus polaire, favorise les interactions avec les molécules d'eau.

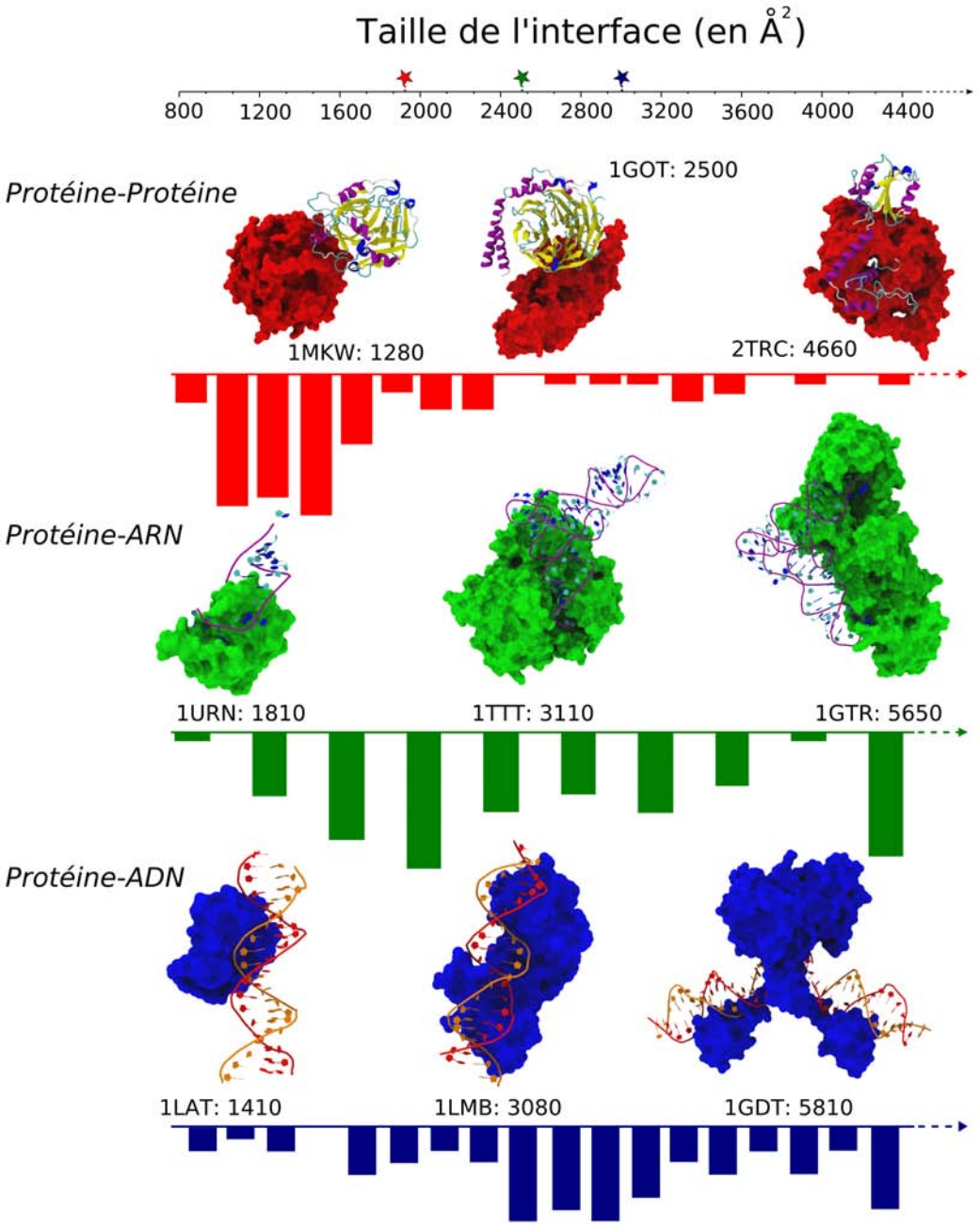


FIG. 1.8 – Taille de l'interface pour les complexes protéiques et protéines-acide nucléiques illustrée par quelques exemples. Pour chaque exemple, il est indiqué son identifiant PDB et la taille de son interface (en $Å^2$). Sous les exemples, on peut voir la distribution des valeurs des interfaces issues de Bahadur *et al.* (2008); Chakrabarti et Janin (2002); Lo Conte *et al.* (1999); Nadassy *et al.* (1999). Pour chaque distribution de valeur, le dernier élément correspond au nombre de complexes dont l'interface est supérieure à 4400 $Å^2$. Les étoiles, donnant la taille de l'interface moyenne et la surface des partenaires protéiques, sont colorées en fonction du type de complexe : rouge, pour les complexes protéiques ; vert, pour les complexes protéine-ARN et bleu, pour les complexes protéine-ADN.

La figure 1.8 fait une synthèse de la taille de l'interface pour les complexes protéine-protéine, protéine-ADN et protéine-ARN. En moyenne, les interfaces des dimères de protéines sont plus petites que les interfaces des assemblages protéine-ARN qui sont elles-mêmes plus petites que les interfaces des complexes protéine-ADN. Néanmoins, chaque type d'interface peut couvrir un large spectre (comme le montre la distribution de valeurs des interfaces) avec des tailles d'interfaces communes au trois types d'interactions. Nous pouvons découper ces interfaces en parcelles unitaires. Chaque parcelle possède également une taille variable mais il est admis de considérer qu'une parcelle standard a une aire d'interface de 1600 ± 400 Å^2. Ceci est valable pour les complexes protéiques et protéines-ADN (Janin et Chothia, 1990; Nadassy *et al.*, 1999). Aucune étude n'a pour l'instant été réalisée sur le découpage des interfaces protéines-ARN en parcelles unitaires. Il est aussi intéressant de comparer la stabilité des complexes en fonction de la taille de l'interface. Dans le cas des protéines, au delà de 4400 Å^2, les hétérodimères se font plus rares ; il ne reste pratiquement que des assemblages homodimèriques qui sont généralement stables. Or, pour la même taille d'interface, nous trouvons un certain nombre de complexes protéine-acide nucléique. Ce type de complexe est le plus souvent transitoire (*i.e.* les deux partenaires peuvent être stable de façon indépendante) et implique donc des phénomènes de dissociation. Comment expliquer qu'une même taille définisse un assemblage stable pour un type de complexe et non pour d'autres ? Ceci s'expliquerait, dans le cas des assemblages protéines-acides nucléiques, par des interactions chargées ou polaires qui faciliteraient la dissociation des partenaires en comparaison des interactions hydrophobes au niveau des complexes protéiques. De plus, même si la compacité au niveau de toutes les interfaces est proche de celle que l'on trouve à l'intérieur des protéines, les partenaires ne subissent pas les mêmes contraintes suivant le type de complexe. Ainsi, pour les assemblages protéine-acide nucléique, la structure secondaire ou tertiaire des acides nucléiques entraîne de fortes tensions au sein de ces molécules lors de l'association avec des protéines. Les liaisons polaires, un peu plus faibles, et les fortes tensions au sein des molécules d'ADN ou d'ARN favoriseraient donc la dissociation des complexes protéine-acide nucléique. Enfin, s'il est assez difficile de définir une taille maximum pour qu'une association soit transitoire, toutes les études s'accordent sur une taille minimale de 800 Å^2 en dessous de laquelle il ne peut y avoir d'interaction spécifique (Nooren et Thornton, 2003b; Bahadur *et al.*, 2008).

1.3.2 Assemblages physiques *vs* assemblages biologiques

Nous allons maintenant comparer les résultats obtenus précédemment avec ceux obtenus pour des assemblages protéiques résultant de contacts cristallins. A l'inverse des contacts biologiques, qui reflètent une interaction spécifique entre deux molécules, les contacts cristallins sont considérés comme des artefacts liés à la cristallisation. Or, ces artefacts peuvent ressembler à des complexes ayant un sens biologique ; il faut donc faire la différence entre ces deux types de complexes. Pour cela, nous avons réutilisé les outils qui nous ont servi à différencier les complexes protéiques et protéines-acides nucléiques.

On trouve des contacts cristallins avec une interface très petite (*i.e.* inférieure à 800 Å^2), ce qui permet de discriminer facilement ceux-ci. Néanmoins, il existe aussi des contacts cristallins avec des tailles d'interfaces équivalentes à celles que l'on peut trouver au sein des assemblages

biologiques (Bahadur *et al.*, 2004; Janin *et al.*, 2007). Dans ce cas, d'autres caractères distinctifs doivent être trouvés. Une possibilité est d'étudier la compacité des atomes à l'interface. Les études réalisées sur ce sujet montrent que les assemblages résultant de contacts cristallins sont moins compacts que ceux impliqués dans des processus biologiques (Nooren et Thornton, 2003b; Bahadur *et al.*, 2004).

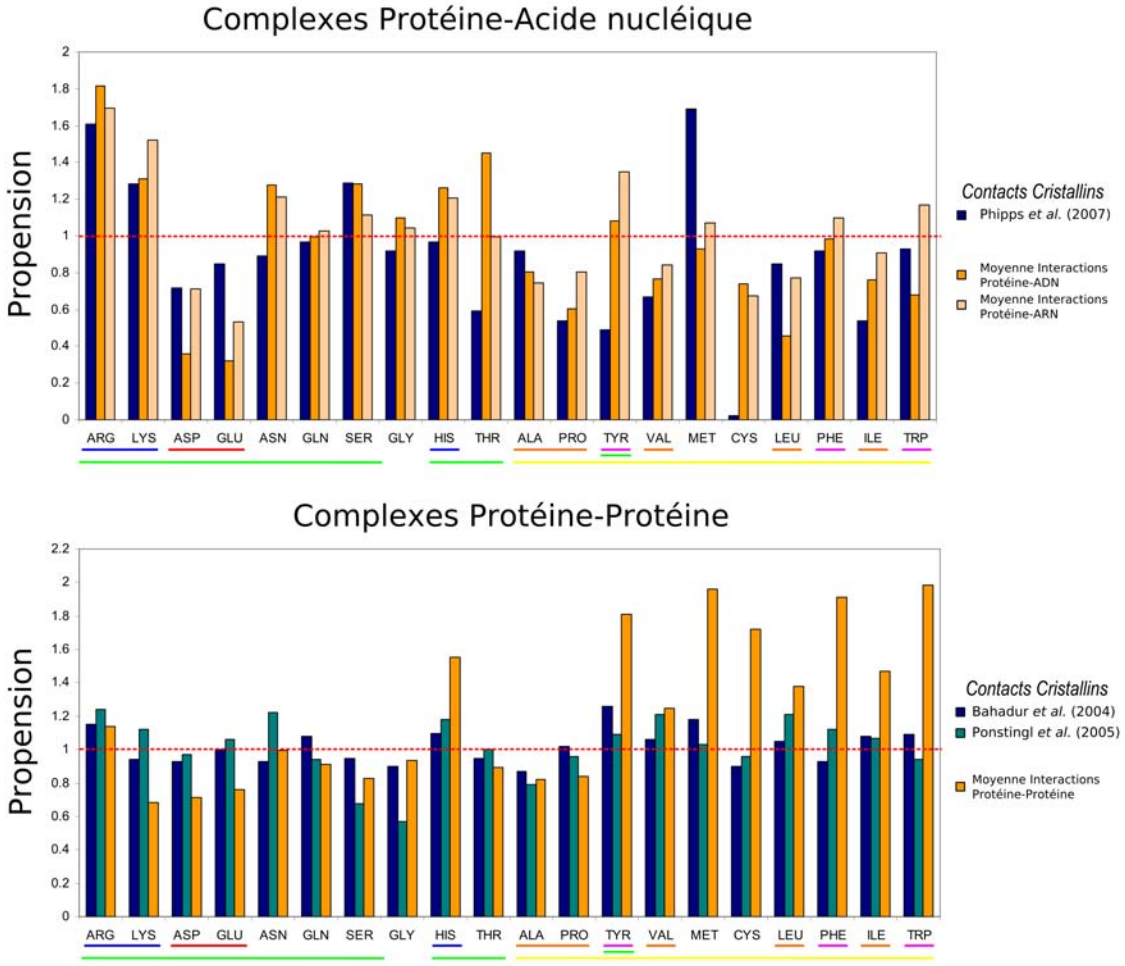


FIG. 1.9 – Comparaisons des moyennes de propension pour les résidus impliqués dans des interactions protéine-protéine ou protéine-acides nucléiques avec celles trouvées pour des contacts cristallins.

Dans le cas des contacts cristallins, nous pouvons aussi voir une différence quant à la propension des atomes à l'interface (voir figure 1.9). On remarque que la propension des résidus à l'interface des contacts cristallins, n'impliquant que des protéines, est proche de 1 (à l'exception de la sérine, de la glycine et de l'alanine). Pour ces contacts, la composition en résidus de l'interface est donc pratiquement équivalente à la composition en résidus de la surface accessible au solvant. Il n'y a donc aucun résidu favorisé contrairement aux complexes biologiques ; c'est pourquoi ces assemblages sont aussi appelés des complexes non spécifiques. Le nombre de liaisons hydrogène est aussi plus faible que pour les complexes protéiques biologiques : il est d'une

liaison hydrogène pour 120 $Å^2$ dans le cas des contacts cristallins contre une liaison hydrogène pour 75 $Å^2$ dans le cas des complexes protéiques.

Cette vision n'est malheureusement pas confirmée pour les contacts cristallins impliquant des molécules d'ARN (à ce jour, aucune étude de ce type sur les complexes protéine-ADN n'a été publiée). Loin d'avoir une propension égale pour chaque résidu, on trouve, dans l'étude de Phipps et Li, des résidus favorisés et d'autres non. Cette préférence ne suit d'ailleurs pas obligatoirement celle que l'on peut trouver à l'interface des complexes protéine-ARN spécifiques. Ainsi, les résidus tyrosine, thréonine ou cystéine sont peu représentés à l'interface de ces assemblages cristallins alors qu'ils sont préférés dans le cas des assemblages biologiques. Dans les cas de la tyrosine et de la thréonine, cette différence mettrait en évidence le rôle considérable de ces types de résidus dans la spécificité des interactions protéine-ARN (Phipps et Li, 2007). A l'inverse, les résidus chargés négativement ainsi que la méthionine sont surreprésentés. Les résidus chargés créeraient plus de liaisons hydrogène avec les molécules d'ARN (en particulier au niveau du sucre et des bases) que dans le cas d'assemblages spécifiques. Enfin, des résidus préférés à l'interface des assemblages spécifiques, peuvent l'être aussi dans le cas des complexes non spécifiques comme l'arginine, la lysine ou la sérine.

Comment expliquer cette différence de propension, au niveau des contacts cristallins, entre les complexes protéiques et les complexes protéine-ARN ? L'explication la plus simple est de considérer le jeu de données de Phipps et Li comme n'étant pas suffisamment représentatif (50 structures) pour pouvoir réellement conclure sur le sujet. Il est vrai que ce jeu reste assez faible face à celui de ≈200 structures analysées pour Bahadur *et al.*. Néanmoins, le jeu de données de Ponstingl *et al.* ne contient pas beaucoup plus de conformations (56 structures) et leurs résultats sont proches de ceux de Bahadur *et al.*. Il faut donc chercher une autre explication. Celle-ci pourrait venir de la nature même de chaque type d'interaction : dans le cas des complexes protéiques, une grande partie des interactions implique des résidus non polaires. Or, ces interactions apparaissent en nombre plus faible dans les contacts cristallins. Il ne reste alors plus que des interactions polaires se faisant au niveau de surfaces non spécifiques. Dans le cas des interactions protéine-ARN, bien que la majorité des interactions soit non polaire, les interactions polaires sont beaucoup plus nombreuses que dans les complexes protéiques. Les interactions au sein du cristal sont assez semblables aux interactions spécifiques. Ainsi, il apparaît que les interactions spécifiques entre protéines peuvent être plus facilement discernables de contacts non spécifiques. Ceci semble beaucoup plus délicat pour les complexes protéine-ARN et sûrement aussi pour les complexes protéine-ADN.

Il existe aussi d'autres critères que nous n'avons pas utilisés dans cette étude, comme l'analyse de la conservation de séquence, qui peuvent aider à discriminer les complexes fonctionnels des artefacts cristallins (Elcock et McCammon, 2001; Valdar et Thornton, 2001; Ofran et Rost, 2007).

1.3.3 Conclusion sur les associations macromoléculaires

Ainsi, à travers des caractères facilement identifiables, nous avons pu comparer pratiquement tous les types de structures connues : les interactions spécifiques protéine-protéine et protéine-acides nucléiques mais aussi les interactions non spécifiques. Ces caractères nous ont permis

d'identifier certaines spécificités propres à chaque type d'assemblage. Bien sûr, cette analyse reste assez succincte et des études plus approfondies sur chaque type de complexes, citées au cours de cette partie, peuvent apporter des compléments d'analyse. De plus, cette compilation de résultats a pu lisser certaines différences et n'a pas mis en valeur la très grande diversités des complexes biologiques. Néanmoins, cette méthodologie nous a aussi évité de tomber dans les pièges d'une étude restreinte à un jeu de données, comme cela a été le cas, par exemple, pour Phipps et Li. En effet, l'analyse du seul jeu de données de Jones *et al.* a amené Phipps et Li à des conclusions un peu trop hâtives sur la propension de l'arginine dans les complexes biologiques et les contacts cristallins. Notre étude sur différents résultats montre que les valeurs de Jones *et al.* pour l'arginine sont plus basses que la moyenne générale (voir figure 1.5). De la même manière, la phénylalanine semble vraiment sous-représentée dans l'étude de Phipps et Li alors que la moyenne des études sur le sujet montre un résultat plus nuancé (voir figure 1.9). Enfin, l'analyse de ces paramètres peut ensuite être utilisée pour prédire les zones potentielles d'amarrage comme nous allons le voir dans la section suivante.

1.4 Méthodes *in silico* pour l'assemblage macromoléculaire

Nous avons vu, dans la partie précédente, comment l'analyse d'un grand nombre de complexes a permis de mettre en évidence certaines caractéristiques structurales ou chimiques des assemblages biologiques. Nous allons voir maintenant comment ces analyses et les données qui en résultent ont pu être utilisées pour prédire les sites potentiels d'interactions entre les macromolécules et raffiner les résultats de docking. Pour commencer, nous effectuerons une description des programmes utilisés pour l'assemblage macromoléculaire *in silico*.

1.4.1 Les programmes d'assemblage moléculaire

Principes du docking : Avant de détailler les divers programmes de docking existants, nous allons présenter les étapes de base de ces programmes. Il existe d'illeurs plusieurs publications de synthèse très intéressantes sur le sujet (Halperin *et al.*, 2002; Smith et Sternberg, 2002; Vajda et Camacho, 2004; Gray, 2006; Ritchie, 2008).

Le principe des programmes de docking est d'essayer d'assembler deux structures (nous évoquerons brièvement le cas des prédictions d'oligomères en fin de partie) macromoléculaires, protéines et/ou acides nucléiques, afin d'obtenir un complexe. Ce processus se déroule en plusieurs étapes comme présenté figure 1.10.

Recherche globale des assemblages potentiels : Cette première étape est basée sur la complémentarité de forme entre les partenaires. En règle générale, la plus grosse molécule (le récepteur) est fixée dans l'espace tandis que la plus petite molécule (le ligand) explore l'espace des conformations autour du récepteur. Nous pouvons noter 3 grandes techniques pour effectuer cette recherche.

La plus largement utilisée est la transformée de Fourier rapide ou *Fast Fourier Transform* - FFT (Harrison *et al.*, 1994). Avec cette méthode, la molécule est discrétisée en voxel (petit cube unitaire d'une grille). Cette technique a l'avantage de fournir des solutions très rapidement. De nombreux programmes de docking utilisent cette approche : comme ZDock, SmoothDock, MolFit, DOT, FTDock, 3D_Dock, GRAMM ou encore BiGGER (voir tableau 1.3). Une autre approche possible est d'utiliser les harmoniques sphériques pour le calcul de la surface moléculaire et la recherche rapide de complexes potentiels : il s'agit de la transformée de Fourier polaire ou *Polar FT* (Ritchie et Kemp, 2000) qui est utilisée dans le programme HEX.

L'utilisation des techniques de Monte-Carlo ou de Monte-Carlo Pseudo Brownien sont utilisées par les programmes RosettaDock, Haddock et ICM-Disco.

Il est aussi possible d'utiliser un hachage géométrique (Fischer *et al.*, 1995). Cette technique répertorie les points critiques (creux, bosses et points de selle) des deux partenaires pour ne définir que quelques zones spécifiques à comparer. Cette comparaison se fait à l'aide d'une table de hachage.

Reclassement des complexes trouvés : Toutes ces techniques permettent d'obtenir de quelques centaines à quelques milliers d'assemblages potentiels. Parmi ces assemblages, un grand

nombre constitue des "faux-positifs", c'est à dire des assemblages qui semblent correspondre à un résultat correct en fonction, par exemple, de la complémentarité de formes mais qui ne représentent, en réalité, aucunement le complexe natif. Il faut alors pouvoir discriminer ce type de complexe des solutions potentielles. Il est important de rajouter des informations comme le potentiel électrostatique à la surface des partenaires, la prise en compte de la formation de liaisons hydrogène ou des termes de désolvatation dans le but de reclasser les assemblages potentiels. Tous ces paramètres sont le plus souvent inclus dans des fonctions de score.

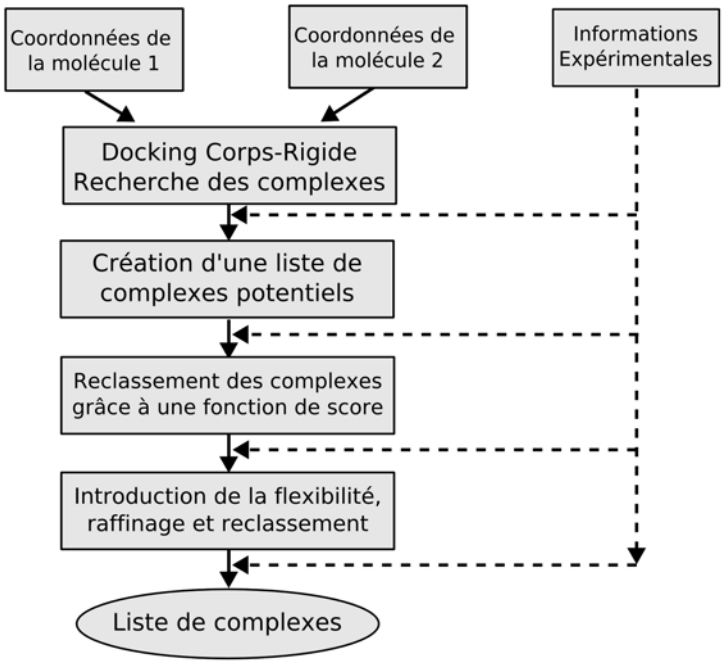


FIG. 1.10 – Les étapes du Docking. *Modifié de Smith et Sternberg (2002)*

Deux autres phases cruciales du docking sont : l'incorporation d'informations (expérimentales ou issues de la littérature) et le traitement de la flexibilité. Une description complète des techniques employées pour remplir ces étapes sera faite par la suite (respectivement dans la partie consacrée aux fonctions de score et en introduction du chapitre 2). Ainsi, nous avons présenté le schéma global des programmes de docking. Chaque programme peut néanmoins utiliser ces étapes de différentes manières, par exemple, en combinant la recherche de candidats potentiels avec l'utilisation de fonctions de score pour guider les recherches, ou en modélisant la flexibilité avant, après ou pendant cette même phase de recherche.

Les divers programmes de docking : A travers cette partie, nous allons développer plus en détails les programmes et les techniques que nous avons présentées dans le paragraphe précédent. Nous nous focaliserons sur les programmes les plus remarquables. Le tableau 1.3 résume les différentes spécificités de chacun.

Les programmes utilisant la technique de FFT ou assimilée : Comme expliqué dans la partie précédente, un grand nombre de programmes utilisent la FFT pour la première étape de la recherche de complexes. Katchalski-Katzir *et al.* furent les premiers à employer cette technique pour l'amarrage moléculaire au travers d'un algorithme qui devint par la suite le programme MolFit (Katchalski-Katzir *et al.*, 1992; Berchanski *et al.*, 2004). Le récepteur et le ligand sont d'abord représentés sur une grille (voir figure 1.11). On attribue ensuite un score à chaque voxel de la grille. Le récepteur (R) et le ligand (L) sont alors définis par les équations :

$$R_{l,m,n} = \begin{cases} 1 & \text{à la surface du récepteur} \\ \phi & \text{à l'intérieur du récepteur} \\ 0 & \text{à l'extérieur du récepteur} \end{cases} \tag{1.1}$$

$$L_{l,m,n} = \begin{cases} 1 & \text{à la surface du ligand} \\ \delta & \text{à l'intérieur du ligand} \\ 0 & \text{à l'extérieur du ligand} \end{cases} \tag{1.2}$$

où l, m et n sont les indices de la grille de dimensions $N \times N \times N$; $l, m, n = (1, \ldots, N)$.

La complémentarité de surface se fait en calculant la fonction de corrélation à partir des équations (1.1) et (1.2) :

$$C_{\alpha,\beta,\gamma} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} R_{l,m,n} \cdot L_{l+\alpha,m+\beta,n+\gamma}$$

où α, β et γ sont les pas de la grille permettant le déplacement du ligand par rapport au récepteur.

Suivant le positionnement du ligand, les valeurs de $C_{\alpha,\beta,\gamma}$ sont nulles lorsqu'il n'y a pas de contact entre les deux molécules ; elles sont positives si celles-ci sont en contact ; voire négatives quand il y a trop d'interpénétrations. En effet, pour éviter des interpénétrations trop importantes, qui n'ont pas de sens physique, il est possible d'assigner aux coefficients ϕ et δ, respectivement une grande valeur négative et une petite valeur positive. Lors du calcul de la fonction de corrélation ces valeurs sont multipliées, entraînant une contribution négative de l'ensemble. Les modèles sont ensuite classés en fonction des résultats obtenus par la fonction de corrélation : plus la complémentarité de surface est bonne, plus les valeurs sont élevées.

Cette recherche est utilisée par de nombreux programmes comme ZDock, SmoothDock, DOT, FTDock, 3D_Dock ou GRAMM. La différence entre ces programmes se fait alors sur d'autres étapes comme le reclassement des résultats grâce à la fonction de score ou l'incorporation de la flexibilité au niveau des chaînes latérales ou du squelette protéique.

Une approche un peu différente a été développée par Ritchie et Kemp avec le programme HEX (Ritchie et Kemp, 2000). La surface des molécules est représentée par des harmoniques

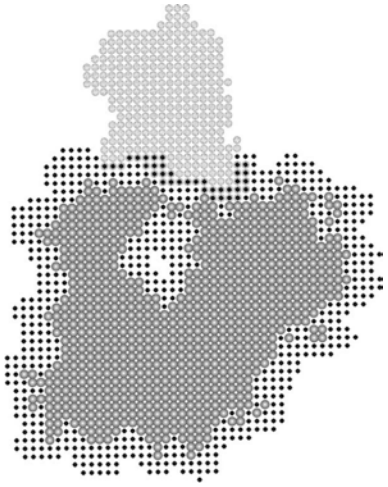


FIG. 1.11 – Vue en deux dimensions de la discrétisation du récepteur et du ligand sur une grille. L'intérieur du récepteur, molécule la plus grosse , est représenté par des sphères grisées tandis que sa surface est représentée par des cercles noirs. Le ligand, molécule la plus petite, est visualisé par des cercles gris. Sur ce dessin il n'est pas fait de distinction entre l'intérieur et la surface du ligand. *Figure issue de Eisenstein et Katchalski-Katzir (2004)*

sphériques. Alors que les méthodes précédentes accélèrent les recherches par translation, l'approche développée par D. Ritchie favorise les recherches par rotation en utilisant une transformée de Fourier à une dimension (les translations étant aussi prises en compte). Cette implémentation, ajoutée à partir de la version 3.1 de HEX, permet un gain de temps de 30 à 50% par rapport aux versions précédentes (Ritchie, 2003).

Enfin, il existe aussi une méthode où la protéine est représentée implicitement sur deux grilles par les chiffres 0 ou 1 représentant la surface ou l'intérieur de celle-ci. Cette méthode, implémentée dans le programme BiGGER, utilise des techniques de manipulations de bits permettant une comparaison très rapide des grilles (Palma *et al.*, 2000).

PatchDock et ATTRACT, représentations simplifiées de la protéine : A la différence des programmes précédents, qui représentent les partenaires sur une grille tridimensionnelle et font une recherche exhaustive des assemblages potentiels, de nouvelles méthodes simplifient la recherche d'assemblage. Pour cela, l'espace de recherche est limité à certains points critiques ou la structure de la protéine est approximée (voir figure 1.12).

Le programme ATTRACT, développé par Martin Zacharias, représente la structure de la protéine par un modèle réduit (Zacharias, 2003). Celui-ci simplifie la représentation des résidus par deux ou trois pseudo-atomes suivant la taille de ceux-ci (voir figure 1.12A). Ce modèle réduit permet un calcul plus rapide de l'assemblage des interactions protéiques et diminue le nombre de minima énergétiques par rapport à une représentation tout atome.

Le programme PatchDock, développé par l'équipe de Haim Wolfson et Ruth Nussinov, limite la recherche à certaines zones de la surface moléculaire. Pour cela, dans une première étape, la

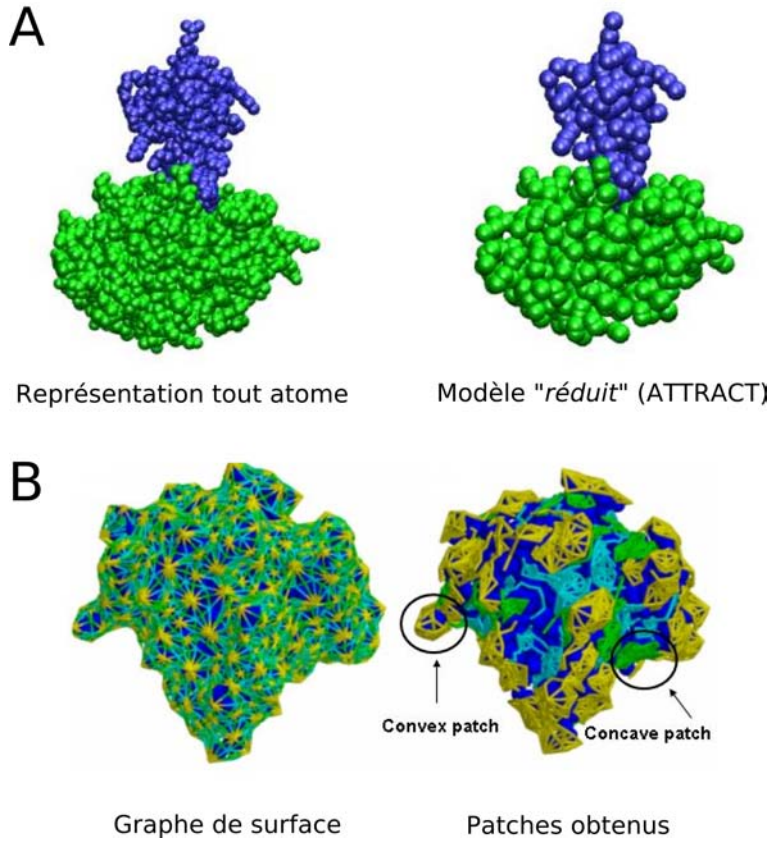


FIG. 1.12 – A- Représentation du complexe trypsine-BPTI avec tous les atomes (à gauche) et avec seulement les pseudos atomes (à droite). B- Graphe de points critiques de la molécule de l'inhibiteur de trypsine (à gauche) et visualisation des patchs associés (à droite) : en jaune, les patchs convexes ; en vert, les patchs concaves et en bleu clair les zones planes. La molécule est colorée en bleu foncé. *Figures issues de Zacharias (2003) et de Duhovny* et al. *(2002).*

surface moléculaire (Connolly, 1983) est discrétisée et les points de celle-ci sont filtrés pour ne garder que les points au niveau des creux, des bosses et de certaines zones planes. Ces points sont ensuite reliés pour former un graphe en trois dimensions puis sont regroupés en patchs convexes, concaves et plats (voir figure 1.12B). Les patchs de surface de chaque partenaire sont ensuite comparés les uns aux autres pour trouver des complémentarités de forme entre les deux protéines. Pour cela, une paire de patchs au niveau d'un partenaire est comparée à une paire de patchs de l'autre partenaires suivant la règle : les patchs concaves ne sont comparés qu'aux patchs convexes tandis que les patchs plats peuvent être comparés à tous les types de patchs. Les patchs sont comparés à l'aide d'une fonction de hachage permettant d'accélérer les calculs. Les assemblages obtenus avec cette méthode sont ensuite reclassés en fonction de leur complémentarité de forme et les complexes présentant trop d'interpénétrations sont éliminés.

Les méthodes de Docking haute résolution : Contrairement aux programmes précédents, les programmes décrits dans cette section utilisent une représentation tout atome des

molécules à assembler (Gray, 2006). Nous avons recensé trois programmes utilisant ce type de représentation : les programmes RosettaDock, ICM-DISCO et HADDOCK.

Le programme RosettaDock est développé par deux équipes menées par Jeffrey Gray et David Baker. Ce programme s'inscrit dans une structure plus globale nommé *Rosetta Commons*[9] regroupant de nombreuses équipes de recherche à travers le monde autour du programme Rosetta. Cette *organisation* de recherche travaille sur de nombreux aspect de la biologie structurale comme le repliement des protéines (avec le programme RosettaAbinitio), les bibliothèques de fragments (avec RosettaFragment ou RosettaRNA) mais aussi le docking protéine-petites molécules (avec le programme RosettaLigand) et protéine-protéine (RosettaDock).

La méthode employée pour le programme RosettaDock est divisée en deux grandes étapes (Gray *et al.*, 2003) :

- D'abord, une recherche basse résolution par Monte Carlo permet de placer les deux partenaires de façon approximative. Les structures de basse résolution ne sont constituées que par les atomes du squelette peptidique auxquels est ajouté un pseudo-atome pour chaque chaîne latérale.
- Des chaînes latérales récupérées dans une bibliothèque de rotamères sont ensuite ajoutées au squelette peptidique. Cette étape est suivie d'un raffinage des chaînes latérales et du squelette peptidique grâce à une minimisation par Monte-Carlo. Il s'agit d'un raffinage itératif consistant en de légers mouvements de rotation-translation suivis d'une minimisation.

Cette étape de raffinage a permis d'obtenir de très bons résultats sur de nombreux exemples (Wang *et al.*, 2005). C'est pourquoi elle sert aussi à raffiner des complexes obtenus à partir d'autres programmes de docking rigide (Kozakov *et al.*, 2008; Pierce et Weng, 2008).

Le programme ICM-DISCO, développé par l'équipe de Ruben Abagyan, est aussi constitué de deux étapes (Fernández-Recio *et al.*, 2002). La première étape consiste en un docking rigide qui échantillonne les différentes positions du ligand autour du récepteur. Cet échantillonnage est guidé par une procédure de Monte-Carlo pseudo-Brownienne (Abagyan *et al.*, 2004). Cette étape est suivie par un amarrage fin des chaînes latérales à l'interface par une minimisation de Monte-Carlo biaisée (Abagyan *et al.*, 2004).

Enfin, le programme HADDOCK, développé par Alexander M. J. J. Bonvin, est, quant à lui, composé de trois phases (Dominguez *et al.*, 2003). Ce programme, comme les deux précédents, essaie de mimer la dynamique d'assemblage des protéines (présentée en partie suivante). Dans une première étape de docking rigide les deux protéines, placées au départ à 150 Å l'une de l'autre, sont rapprochées en minimisant l'énergie du complexe. Ensuite, les 200 solutions de plus basse énergie sont raffinées par recuit simulé ce qui permet de réorienter le squelette et les chaînes latérales tout en passant des barrières énergétiques. Enfin, une dynamique en solvant explicite permet de relaxer les complexes. Notons qu'une extension de ce programme permet de modéliser les molécules d'eau à l'interface du complexe (van Dijk et Bonvin, 2006).

[9] `http://www.rosettacommons.org/main.html`

Les programmes présentés dans cette section différent donc par leurs méthodologies mais aussi par le temps de calcul associé à celles-ci. En effet, suivant le degré de précision demandé, les temps de calcul varient de quelques minutes à plusieurs jours (voir tableau 1.3). Néanmoins, les moyens de calcul se développent et l'utilisation de clusters ou de grilles d'ordinateurs permettent de réduire considérablement ces temps. De plus, de nouvelles stratégies utilisant le potentiel des cartes graphiques commencent à se développer (Ritchie, 2008).

Ces programmes de docking sont donc des outils très puissants pour modéliser les complexes protéiques. Cependant, leur prise en main est parfois difficile et une connaissance minimale en programmation est recommandée. C'est pourquoi une alternative est l'utilisation des versions simplifiées de ces programmes, disponibles sur Internet : les serveurs de docking.

TAB. 1.3 – Présentation des programmes de docking les plus utilisés. *Inspiré de Méndez* et al. *(2005)*

Groupes	Programme[a]	Recherche[b]	Fonction de score[c]	Flexibilité[d]	Oligomère[e]	Temps	Références[f]
Baker	RosettaDock(mod.) http://graylab.jhu.edu/docking/rosetta	MC	VdW + désolvatation + électro.	C. lat.	C_n et D_n	semaine	Wang *et al.* (2005)
Camacho	SmoothDock http://structure.pitt.edu/servers/smoothdock/	FFT	forme + désolvatation + électro.	C. lat. (DM)	C_n et D_n	heures	Camacho et Gatchell (2003)
Eisenstein	MolFit http://www.weizmann.ac.ic.il/Chemical_Research_Support/molfit/	FFT	forme + non polaire + électro.	C. lat.	C_n ou D_n	heures	Berchanski et Eisenstein (2003)
Weng*	ZDock http://zdock.bu.edu/software.php	FFT	forme + électro. + désolvatation + champ de force (Zrank)	C. lat. (RDOCK)	C_n (M-ZDock)	heures	Chen *et al.* (2003a)
Abagyan	ICM-DISCO http://www.molsoft.com	PSB-MC	potentiel sur grille	C. lat.	–	heures	Fernández-Recio *et al.* (2003)
Gray	RosettaDock http://graylab.jhu.edu/docking/rosetta	MC	VdW + désolvatation + électro.	C. lat.	–	semaine	Gray *et al.* (2003)
Bonvin	HADDOCK http://www.nmr.chem.uu.nl/haddock	MC-DM	champ de force	Sqlt, C. lat. (DM,RMN)	–	jours	Dominguez *et al.* (2003)
Ten Eyck	DOT http://www.sdsc.edu/CCMS/DOT/	FFT	forme + electro.	–	–	heures	Mandell *et al.* (2001)
Ritchie	HEX http://www.csd.abdn.ac.uk/hex/	PFT	forme + electro.	C. lat.	–	minutes	Ritchie et Kemp (2000)
Wolfson*	PatchDock http://bioinfo3d.cs.tau.ac.il/PatchDock/patchdock.html	géométrie	forme + électro. (FireDock)	Sqlt, C. lat. (FlexDock)	C_n (Symm-Dock)	minutes	Schneidman-Duhovny *et al.* (2005b)

Suite du tableau page suivante

Groupes	Programme[a]	Recherche	Fonction de score	Flexibilité[b]	Oligomère[c]	Temps	Références[d]
Bates	FTDOCK(mod.) + DM http://www.sbg.bio.ic.ac.uk/docking/ftdock.html	FFT + DM	forme + électro. + champ de force	Sqlt, C. lat. (MD)	–	jours	Smith *et al.* (2005b)
Sternberg	3D_Dock http://www.sbg.bio.ic.ac.uk/docking/	FFT	forme + électro.	C. lat.	–	heures	Aloy *et al.* (2001)
Zacharias	ATTRACT (non disponible)	Vecteurs propres	potentiel réduit	Sqlt, C. lat.	–	heures	Zacharias (2003)
Vakser	GRAMM http://www.chem.ac.ru/Chemistry/Soft/GRAMM.en.html	FFT	forme + électro.	–	–	heures	Tovchigrechko *et al.* (2002)
Palma	BiGGER (plus disponible)	grille de bits	forme + électro.	C. lat.	–	heures	Palma *et al.* (2000)
Valencia	–	homologie	séquences - mutations	–	–	–	Valencia et Pazos (2002)

[a] Lien internet si programme disponible pour la communauté.

[b] MC : Monte-Carlo ; FFT : Fast Fourier Transform ; PSB : Pseudo Brownien ; DM : Dynamique Moléculaire ; PFT : Polar Fourier Transform.

[c] électro. : prise en charge des interactions électrostatiques.

[d] Traitement de la flexibilité au niveau du squelette protéique (Sqlt.) ou des chaînes latérales (C. lat.) ou des deux. – : aucun traitement disponibles.

[e] C_n : symétrie cyclique et D_n : symétrie dièdre.

[f] Références du programe de docking.

[*] Les noms entre parenthèse font référence à des programmes spécifiques indépendants du programme principal.

Les groupes ont été classés en fonction des résultats des équipes de recherche à CAPRI dans le but de pouvoir plus facilement identifier les méthodes donnant de bons résultats (voir tableau 1.8).

Les serveurs de Docking : Il existe divers serveurs de docking accessibles sur internet (voir tableau 1.4). Ceux-ci ont une interface utilisateur simplifiée permettant aux personnes non expertes de construire des assemblages macromoléculaires.

Le premier serveur, CLUSPRO, a été développé en 2004 par l'équipe de Sandor Vajda et Carlos J. Camacho (Comeau *et al.*, 2004b,a). Ce serveur utilise les programmes d'amarrage DOT, GRAMM ou ZDOCK pour l'amarrage corps-rigide. Les résultats sont ensuite reclassés par une fonction de score prenant en compte l'énergie libre de désolvatation et les interactions électrostatiques. L'originalité de CLUSPRO repose sur sa dernière étape de reclassement qui consiste à grouper les solutions pour former des amas (*cluster*) et à compter le nombre de solutions par groupe. Chaque groupe est ensuite reclassé par ordre décroissant du nombre de solutions. Les résultats donnés par CLUSPRO correspondent à la structure moyenne des solutions de chaque groupe. Nous pouvons noter que CLUSPRO dérive de l'algorithme SmoothDock (Camacho et Gatchell, 2003) disponible également sous forme de serveur.

Tab. 1.4 – Liste non exhaustive des serveurs de docking et leurs sites web associés

Nom du serveur	Adresse Internet	Recherche	Fonction de score	Référence
CLUSPRO	`http://nrc.bu.edu/cluster/`	FFT	forme + électro	Comeau *et al.* (2004b)
Gramm-X	`http://vakser.bioinformatics.ku.edu/resources/gramm/grammx`	FFT	forme adoucie + electro	Tovchigrechko et Vakser (2006)
HEX	`http://www.csd.abdn.ac.uk/hex_server/`	voir tableau 1.3		
HADDOCK	`http://haddock.chem.uu.nl/Haddock/haddock.php`	voir tableau 1.3		
PatchDock	`http://bioinfo3d.cs.tau.ac.il/PatchDock/`	voir tableau 1.3		Schneidman-Duhovny *et al.* (2005)
SKE-DOCK	`ske-dock@pharm.kitasato-u.ac.jp`	géométrie	CIRCLE + intervention humaine	Terashi *et al.* (2007)
SmoothDock	`http://structure.pitt.edu/servers/smoothdock/`	voir tableau 1.3		
RosettaDock	`http://rosettadock.graylab.jhu.edu/`	voir tableau 1.3		Lyskov et Gray (2008)
ZDock	`http://zdock.bu.edu/`	voir tableau 1.3		

SKE-DOCK est un serveur particulier car il ne possède pas d'interface graphique sur internet où l'utilisateur peut donner les structures à assembler. Il est nécessaire d'envoyer un e-mail aux chercheurs responsables de ce "serveur" contenant les fichiers des protéines à assembler (voir tableau 1.4). L'assemblage corps rigide des protéines est basé sur programme un prenant en compte la complémentarité de surface des partenaires. Pour cela, chaque surface est représentée par un ensemble de creux, de bosses et de points de selle (équivalent à PatchDock). Les complexes obtenus sont ensuite reclassés par une fonction de score interne au laboratoire nommée CIRCLE et par inspection visuelle Terashi *et al.* (2007).

La majorité des autres serveurs dérivent des programmes de docking cités précédemment comme HADDOCK, HEX, Gramm-X, PatchDock, RosettaDock ou ZDOCK. Certaines équipes ont amélioré leurs programmes lors du passage de ceux-ci en serveurs. C'est le cas du serveur Gramm-X dont le fonction approximant le potentiel de Lennard-Jones a été adoucie (Tovchigre-

chko et Vakser, 2006). A l'inverse, le serveur RosettaDock a subi quelques restrictions par rapport à la version exécutable. En effet, le temps de calcul très long empêche une recherche globale des assemblages, celle-ci est donc limitée à une recherche locale (Lyskov et Gray, 2008). C'est pourquoi ce serveur est surtout destiné à raffiner les résultats obtenus avec d'autres programmes de docking.

L'oligomérisation, nouvelle voie de recherche : Les progrès dans le domaine de la biologie structurale ont mis en évidence des assemblages constitués d'un grand nombre de sous-unités identiques. Ces sous-unités s'organisent souvent de manière symétrique (Lawson *et al.*, 2008). Le nouveau défi posé aux programmes de docking est de construire de tels assemblages à partir d'une seule sous-unité.

Des méthodes ont été spécialement conçues pour répondre à cette problématique. En général, elles imposent des contraintes de symétrie aux programmes de docking (Berchanski et Eisenstein, 2003). Les symétries les plus communes sont les symétries cycliques (C_n) et les symétries dièdres (D_n) (Voet *et al.*, 1998). Certaines approches ne permettent de modéliser que les symétries C_n (Pierce *et al.*, 2005; Schneidman-Duhovny *et al.*, 2005a) tandis que d'autres modélisent les symétries C_n ou D_n (Berchanski *et al.*, 2005). Enfin, il existe quelques méthodes prenant en compte les deux types de symétries et leurs combinaisons (Comeau et Camacho, 2005; André *et al.*, 2007).

1.4.2 Incorporation d'informations pour guider l'amarrage

Les données expérimentales ou issues de la littérature sont très importantes pour cibler les zones d'interactions afin d'obtenir des assemblages plus précis (Méndez *et al.*, 2005; Lensink *et al.*, 2007). Nous allons voir, maintenant, comment les informations collectées grâce aux analyses statistiques des interfaces (présentées dans le chapitre précédent) peuvent être réemployées pour limiter la zone de recherche lors du docking. Nous verrons aussi quelles sont les méthodes utilisées pour guider les programmes de docking durant la phase d'amarrage ainsi que les fonctions de score utilisées pour classer les complexes obtenus.

Avant la phase de docking, pour limiter la zone de recherche : Les paramètres, pris en compte dans le premier chapitre, comme la propension d'un acide aminé à l'interface ou la géométrie de celle-ci, peuvent être exploités pour créer des règles de prédiction des zones d'associations entre macromolécules (Bordner et Abagyan, 2005). Outre ces informations statistiques, des règles peuvent aussi être déduites de l'analyse exhaustive de mutations réalisées à l'interface de complexes (Bogan et Thorn, 1998; Moreira *et al.*, 2007) ou de l'étude de la conservation des acides aminés au cours de l'évolution (Marcotte *et al.*, 1999; Hu *et al.*, 2000).

Ces dernières années, de nombreuses méthodes se sont développées pour essayer de prédire l'interface de complexes (Zhou et Qin, 2007). Certaines méthodes utilisent des techniques d'apprentissage comme les machines à vecteurs de support (*Support Vector Machine* ou SVM) ou les réseaux neuronaux (Bradford et Westhead, 2005; Porollo et Meller, 2007) ; d'autres utilisent diverses combinaisons de paramètres (de Vries *et al.*, 2006; Kufareva *et al.*, 2007). L'approche

paramétrique a l'avantage d'être plus transparente et donne de bons résultats dans le cas d'une fonction bien réglée tandis que les techniques d'apprentissage ne requierent pas d'intervention humaine et peuvent mettre en évidence des règles inattendues (de Vries et Bonvin, 2008).

Ces différentes méthodes ont été comparées dans deux publications récentes (Zhou et Qin, 2007; de Vries et Bonvin, 2008).

Pendant la phase de docking, pour guider le programme : A notre connaissance, parmi les programme de docking cités précédemment, le programme HADDOCK est celui qui prend le mieux en compte les informations biologiques données par l'utilisateur pendant la phase de docking. Le programme HADDOCK (*High Ambiguity driven docking*) a été créé pour exploiter de nombreux types d'informations expérimentales comme les effets NOE (*Nuclear Overhauser Effect*) obtenus par spectroscopie RMN ou les données obtenues par mutagénèse (van Dijk *et al.*, 2005a). Ces informations sont traduites en termes de contraintes de distance notées AIRs, pour *Ambiguous Interaction Restraints*, incorporées dans le champ de forces pour augmenter le score des assemblages pour lesquels des atomes importants pour l'interaction se trouvent à l'interface (Dominguez *et al.*, 2003). L'inconvénient de ce programme est la nécessité d'introduire des informations sur le site actif pour que celui-ci puisse prédire un assemblage. Néanmoins, la nouvelle version de ce programme semble s'être affranchie de cette limitation (de Vries *et al.*, 2007).

D'autres équipes de recherche utilisent des informations biochimiques pour guider l'assemblage comme celle de Rebecca Wade (Motiejunas *et al.*, 2008) ou de Jeffrey Gray (Chaudhury *et al.*, 2007). Enfin, plusieurs serveurs de docking permettent à l'utilisateur de définir des résidus en interaction (Comeau *et al.*, 2004b; Schneidman-Duhovny *et al.*, 2005).

Nous pouvons enfin noter que des données expérimentales comme les enveloppes basse résolution (SAXS ou Cryo-EM) ou les mesures de fluorescence, jusque-là peu exploitées, commencent à être employées comme contraintes pour le docking de macromolécules (Wriggers et Chacón, 2001; Knight *et al.*, 2005; Frankenstein *et al.*, 2008).

Après la phase de docking, pour classer les solutions potentielles : Ce classement se fait à l'aide de fonctions de score. Le tableau 1.3 récapitule les paramètres des fonctions de score utilisées par les divers programmes de docking.

La complémentarité de forme : La majorité des programmes de docking utilise en premier lieu une fonction calculant la complémentarité de forme entre les deux partenaires assemblés. Nous avons présenté cette fonction pour les programmes de docking utilisant la FFT (voir paragraphe "Les programmes utilisant la technique de FFT ou assimilée"). L'équipe de Zhipping Weng a amélioré cette fonction en rajoutant une partie imaginaire aux équations (1.1) et (1.2). Ces paramètres supplémentaires permettent de prendre en compte des paires d'atomes à l'interface et non plus de simples points sur la grille (Chen et Weng, 2003). Cette fonction de score ne suffit cependant pas à classer finement les complexes ; il est donc nécessaire d'ajouter d'autres grandeurs (Gabb *et al.*, 1997).

Le potentiel électrostatique : Le deuxième élément le plus souvent codé dans les fonctions de score est le potentiel électrostatique. Pour les programmes utilisant la FFT, la complémentarité électrostatique peut être codée dans la partie imaginaire des équations (1.1) et (1.2) (Eisenstein et Katchalski-Katzir, 2004). Pour les programmes n'utilisant pas la FFT, comme PatchDock, les contributions électrostatiques peuvent être approximées par un modèle de Coulomb (Andrusier *et al.*, 2007). De nombreux cas montrent que l'ajout de la contribution électrostatique améliore le classement des résultats de docking (Gabb *et al.*, 1997; Mandell *et al.*, 2001). Néanmoins, certaines études ont nuancé ce résultat en constatant que, même si dans la majorité des cas l'ajout de la composante électrostatique améliore les résultats, il arrive que celle-ci détériore le classement d'assemblages proches de la structure native du complexe (Heifetz *et al.*, 2002; Ritchie et Kemp, 2000). Sheinerman *et al.* explique que les groupements chargés peuvent déstabiliser l'assemblage final à cause des phénomènes de désolvatation ayant lieu à l'interface (Sheinerman *et al.*, 2000). Il est donc nécessaire de prendre en compte ce phénomène de désolvatation.

L'hydrophobicité ou l'énergie de désolvatation : Depuis que Zhang *et al.* estimèrent les énergies de contact atomique (*atomic contact energies* ou ACE) à partir de structures cristallisées de protéines (Zhang *et al.*, 1997), de nombreuses autres études démontrèrent l'intérêt de calculer l'énergie de désolvatation nécessaire à l'assemblage de complexes (Camacho *et al.*, 1999, 2000; Camacho et Vajda, 2001). Ce paramètre est donc maintenant inclus dans plusieurs fonctions de score, comme celle du programme SmoothDock (Camacho *et al.*, 2006), ICM-Disco (Fernández-Recio *et al.*, 2005) ou encore MolFit (Berchanski *et al.*, 2004). Dans ce dernier cas, il ne s'agit pas explicitement de l'énergie de désolvatation du système mais plus d'un terme de complémentarité hydrophobe.

La combinaison de ces trois contributions (complémentarité de forme, potentiel électrostatique et désolvatation) permet d'approximer des fonctions d'énergie libre (Fernández-Recio *et al.*, 2002; Camacho *et al.*, 2006).

Les champs de forces : Enfin, en complément ou en remplacement de la fonction d'énergie libre certains programmes utilisent des champs de forces (une définition sera donnée dans le deuxième chapitre). Par exemple, pour le programme HADDOCK, l'énergie d'interaction entre les protéines est calculée par le champ de forces OPLS (Dominguez *et al.*, 2003) tandis que la fonction de score ZRANK utilise le champ de forces CHARMM19 pour les interactions électrostatiques à une distance de moins de 5 Å (Pierce et Weng, 2007). Nous pouvons aussi noter que le programme ATTRACT utilise un potentiel simplifié adapté à la représentation des protéines en modèle réduit (Zacharias, 2003).

1.4.3 Évaluation des méthodes : *Benchmarks* et challenge CAPRI

Le développement d'un programme de docking est un processus long et complexe. Tous les programmes de docking utilisés actuellement sont développés depuis de nombreuses années. Chaque programme est constitué de différentes étapes qu'il faut valider. Il existe pour cela plusieurs méthodes :

– L'utilisation des jeux de données tests pré-établis ou *benchmarks*. Ceux-ci sont très utiles pour tester les méthodes d'amarrage sur des exemples précis et classés en fonction de degrés de difficulté ou de types de complexes.
– La participation au challenge CAPRI. Dans ce challenge, les équipes de recherche et leurs programmes de docking sont confrontés à des cas concrets de complexes dont ils ne connaissent pas la solution. Ce système permet de tester de façon impartiale et sans *a priori* les méthodes de docking.

Utilisation de *Benchmarks* : L'utilisation de jeux de données permet d'entraîner et de tester les programmes de docking sur des cas bien définis.

L'un des premiers *benchmarks* disponibles fut celui développé par l'équipe de J. Janin et Z. Weng (Chen *et al.*, 2003b). Celui-ci est constitué d'un ensemble non-redondant de structures de complexes pour lesquelles les formes libres des partenaires sont disponibles. Ce *benchmark* est constitué de 59 complexes classés en quatre catégories : 29 complexes enzyme-inhibiteur, 19 complexes antigènes-anticorps, 11 complexes divers et 7 complexes considérés comme "difficiles" car impliquant d'importants changements structuraux entre les structures des protéines libres et celles des protéines complexées.

Ce jeu de données a fait l'objet de 2 mises à jour successives[10] (Mintseris *et al.*, 2005; Hwang *et al.*, 2008). Elles prennent en compte un nombre plus important de cas et offrent une grande variété de complexes. Dans la dernière version, le nombre de complexes a plus que doublé par rapport à la version 1.0, passant à 124 assemblages. De plus, l'ensemble des partenaires se trouvent à la fois sous la forme liée (au sein du complexe) et sous la forme libre (issue de la *Protein Data Bank* mais superposée à la forme liée pour faciliter l'évaluation des résultats). Ces complexes furent classés par degrés de difficulté allant d'un niveau "corps-rigide" pour les complexes où les structures des protéines à l'interface ne subissent que peu de déformations à un niveau difficile pour lequel on trouve un RMSD à l'interface supérieur à 3 Å (entre les formes libres et liées des protéines partenaires). Comme dans la version 1.0, les dernières versions proposent aussi un classement des complexes par rapport à leur fonction biologique. On trouve ainsi trois catégories : enzyme-inhibiteur, anticorps-antigène et "autres" rassemblant des complexes variés.

Ce *benchmark* a été très largement exploité puisqu'il a fait l'objet de plus de 150 citations (si l'on regroupe les citations des *benchmarks* 1.0 et 2.0). On l'utilise pour entraîner divers programmes de docking (Tovchigrechko et Vakser, 2006; Andrusier *et al.*, 2007), pour développer de nouvelles fonctions de score (Huang et Zou, 2008; Martin et Schomburg, 2008) ou valider des recherches de minima globaux (Kozakov *et al.*, 2008).

Cette démarche a été reprise par d'autres équipes. La base de données DOCKGROUND[11] (Douguet *et al.*, 2006; Gao *et al.*, 2007) met à la disposition de la communauté un nombre conséquent d'assemblages protéiques mais, à la différence des *benchmarks* 1.0 à 3.0, celle-ci possède un grand nombre de complexes dont les partenaires ne se trouvent que sous la forme liée. De ce fait, le nombre de complexes disponibles est beaucoup plus important que pour le

[10] Ces 3 versions sont disponibles à l'adresse : `http://zlab.bu.edu/zdock/benchmark.shtml`
[11] accessible via `http://dockground.bioinformatics.ku.edu/`

benchmark 3.0 (plusieurs milliers de complexes pour DOCKGROUND contre un peu plus d'une centaine pour le *benchmark* 3.0). En contrepartie, cette base de données contient un grand nombre de structures redondantes ainsi que de nombreuses structures modélisées.

Outre les ensembles généralistes comme le *benchmark* 3.0 ou la base de données DOCKGROUND, il existe des jeux de données plus spécifiques dédiés aux interactions antigène-anticorps (Ponomarenko et Bourne, 2007) ou protéine-ADN (van Dijk et Bonvin, 2008).

Le challenge CAPRI :

Historique : CAPRI (*Critical Assessment of PRedicted Interactions*) est une expérience internationale permettant d'évaluer, de façon impartiale, les algorithmes et méthodes dédiés à l'assemblage *in silico* de complexes protéiques. Le challenge CAPRI est basé sur le modèle de CASP (*Critical Assessment of methods of protien Structure Prediction*). CASP a débuté en 1992 à l'initiative de John Moult. Cette expérience a pour but d'encourager le développement des méthodes de prédiction du repliement des protéines (prédiction de la structure 3D des protéines uniquement à partir de leurs séquences) et de tester celles-ci par des prédictions en aveugle (sans connaître la structure finale). CASP en est actuellement à sa $8^{\text{ème}}$ édition. Durant la deuxième édition de CASP, en 1996, une tentative de prédiction du complexe *Hemagglutinin*/Anticorps à partir des structures non liées de chaque partenaire a été mise en place (Dunbrack *et al.*, 1997). Cet essai, ainsi qu'un précédent quelques années plus tôt (Strynadka *et al.*, 1996), a fait prendre conscience aux chercheurs travaillant dans le domaine de l'assemblage protéiques *in silico* de l'utilité d'une expérience équivalente à CASP mais dédiée au *docking* de protéines. Quelques années plus tard, durant l'été 2001, ce projet fut entériné lors de la conférence intitulée « *Conference on Modeling Protein Interactions in Genomes* »à Charleston en Caroline du Sud (Vajda *et al.*, 2002). Ainsi naquit l'expérience CAPRI.

L'expérience CAPRI, contrairement à CASP où les dates de chaque édition sont fixées à l'avance, débute quand une cible (ou un groupe de cibles) est offerte par un (ou des) expérimentateur(s), car il est difficile de trouver des assemblages cibles adéquats. En effet, l'utilisation des structures des deux partenaires cristallisés (sous forme liée) pour prédire l'interaction biaiserait les résultats obtenus car il s'agirait de trouver une simple complémentarité de forme. Or nous avons vu, dans la partie consacrée aux assemblages macromoléculaires, que le phénomène d'association entre macromolécules est souvent couplé à un réarrangement structural plus ou moins important des partenaires. C'est pourquoi, pour que les algorithmes prennent en compte ce facteur, il est préférable de transmettre la structure d'au moins un partenaire dans sa forme libre (non liée). Les complexes retenus comme cible pour l'expérience CAPRI doivent donc avoir au moins un des partenaires sous forme libre dans la *Protein Data Bank*, ce qui est relativement rare (Janin *et al.*, 2003). Une autre solution est de modéliser un des partenaires par homologie à partir de la structure d'une protéine de la même famille issue de la *Protein Data Bank*. Ceci permet de prendre plus de complexes en compte tout en augmentant la difficulté des cibles. De plus, au cours de ces derniers cycles, le caractère nouveau de l'interaction fut aussi un facteur important (Janin, 2007).

TAB. 1.5 – Calendrier du challenge CAPRI : rounds 1 à 12 (2001-2006). *Inspiré de Janin (2005a)*

Round		Durée	Cibles	Predicteurs	*scoreurs*	Soumissions
1	Juil.-Sept. 2001	≈3 mois	3	19	–	276
2	Jan.-Mars 2002	≈3 mois	4	16	–	281
3	6 Jan.-2 Fév. 2003	≈1 mois	2	22	–	344
4	1 Sept.-12 Oct. 2003	6 semaines	4	25	–	820
5	19 Jan.-21 Mars 2004	≈2mois	4	29	–	850
6	17 Jan.-6 Fév. 2005	3 semaines	1	35	–	266
7	9 Mai-22 Mai 2005	2 semaines	1	37	–	337
8*	5 Sept.-10 Sept. 2005	5 jours	2	40	12	138
9	20 Mars-2 Avril 2006	2 semaines	2	36	3	677
10	15 Mai-4 Juin 2006	3 semaines	1	40	8	366
11	27 Nov.-10 Déc. 2006	2 semaines	1[a]	40	16	681[a]
12	12 Fév.-11 Mars 2006	1 mois	1	41	10	372

* Annulé car résultats donnés sur Internet avant la fin du délai imparti.
[a] Une seule cible mais deux associations possibles

CAPRI est actuellement dans sa 4ème édition. Les 3 premières éditions se sont déroulées de Juillet 2001 à Mars 2006. Chaque édition comprend plusieurs cycles (appelé *rounds*). La première édition comprenait 7 cibles réparties sur deux cycles ; celle-ci s'est déroulée sur 6 mois de Juillet 2001 à Mars 2002. La seconde édition comprenait 10 cibles réparties sur 3 cycles s'étendant de Janvier à fin Mars 2003 soit 4 mois et demi. Enfin, la troisième édition regroupant les cycles de 6 à 12 et comprenant 9 cibles a duré 4 mois et 3 semaines (voir tableau 1.5). La durée d'un cycle varie entre 2 et 6 semaines suivant le nombre d'assemblages à modéliser. Les délais de soumission de prédictions sont donc relativement courts, ceci pour deux raisons :

1. Les complexes cibles sont, la plupart du temps, impliqués dans des interactions essentielles pour la machinerie cellulaire et font donc l'objet de recherches intensives. Les expérimentateurs ne peuvent donc perdre la primeur de leurs travaux en laissant trop de temps entre le moment de leur découverte et le moment où les résultats sont publiés.
2. Le but est de tester la qualité des méthodes de docking en temps limité afin que celles-ci puissent prédire convenablement un grand nombre d'assemblages de manière routinière.

Chaque édition se clôture par un meeting présentant un bilan des résultats obtenus et analysant les échecs et les réussites de l'édition (Janin et Wodak, 2007).

Au fil des éditions, l'expérience CAPRI a évolué. Ceci est visible dans le nombre de participants (présenté tableau 1.5 et figure 1.13) : celui-ci a augmenté au fil des premiers rounds, montrant l'engouement des équipes de recherche pour cette nouvelle "expérience". À partir du round 8, ce nombre s'est stabilisé à une quarantaine de participants. Comme pour son aînée CASP, de nouvelles catégories ont été mises en place au fil des rounds, reflétant l'évolution des recherches et des moyens de prédiction dans le domaine. Ainsi, la catégorie *scoreurs* permet aux équipes de recherche de focaliser leur attention sur le reclassement des modèles déjà soumis tandis que la catégorie serveurs ne prend en compte que les programmes donnant des prédictions sans intervention humaine. Ces deux catégories ont été mises en place à partir du round 8. C'est

d'ailleurs à partir de ce round que le nombre de participants des catégories prédicteurs et serveurs s'est stabilisé (à respectivement ≈40 participants et ≈6 serveurs). Le nombre de *scoreurs* a fluctué pendant quelques rounds mais tend désormais, lui aussi, à se stabiliser. Cette tendance à la stabilité se confirme dans les cycles actuels (round 13 à 15) avec une quarantaine de groupes prédicteurs, une moyenne de 20 groupes de *scoreurs* et entre 6 et 7 serveurs par round[12].

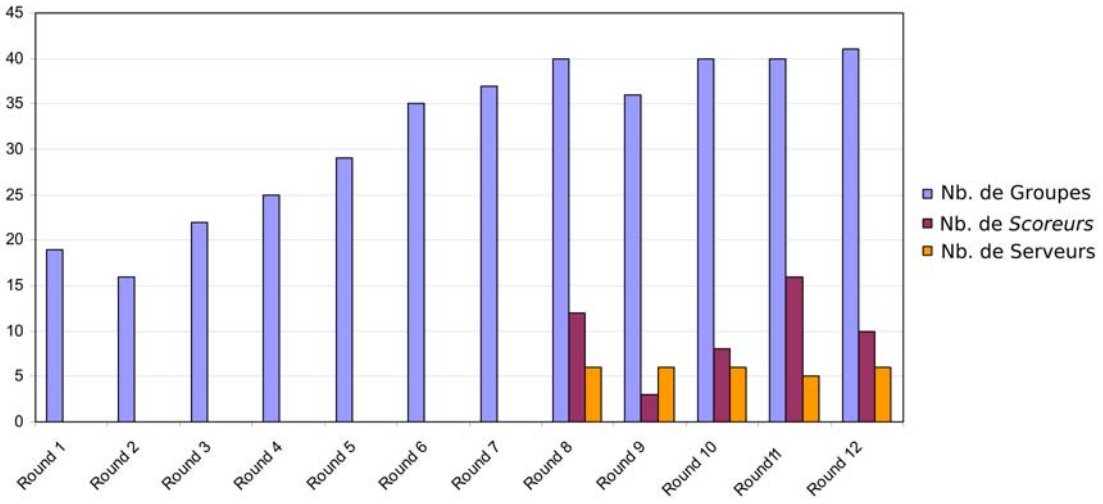


FIG. 1.13 – Nombre de groupes prédicteurs, *scoreurs* et nombre de serveurs automatiques de *Docking* pour les 12 derniers rounds de CAPRI.

Les cibles de CAPRI : Le tableau 1.6 présente les différents complexes cibles proposés depuis le commencement de l'expérience CAPRI. Durant les premiers cycles (1 et 2 regroupant les cibles 1 à 7 - voir tableau 1.5), la majorité des assemblages étaient des complexes anticorps-antigènes. Les partenaires présentés aux prédicteurs se trouvaient sous la forme lié-non lié. Les partenaires des cibles 1 et 7 étaient sous la forme non lié-non lié.

[12]voir le site de CAPRI : http ://www.ebi.ac.uk/msd-srv/capri/

TAB. 1.6 – Les divers complexes présentés à CAPRI durant les rounds 1 à 12. *inspiré de Janin et Séraphin (2003), Janin (2005a) et Janin et Wodak (2007)*

Cible	PDB	Complexe	Référence	Fonction[a]	Type[b]
T01	1kkl	HPr kinase/HPr	Fieulaine *et al.* (2002)	phosphorylation	Nl-Nl
T02	—	Rotavirus VP6/Fab	Thouvenin *et al.* (2001)	sys. immunitaire	Nl-L
T03	1ken	Flu hemagglutinin/Fab	Barbey-Martin *et al.* (2002)	sys. immunitaire	Nl-L
T04	1kxv	Amylase/Camel V_{HH}	Desmyter *et al.* (2002)	sys. immunitaire	Nl-L
T05	1kxt	Amylase/Camel V_{HH}	Desmyter *et al.* (2002)	sys. immunitaire	Nl-L
T06	1kxq	Amylase/Camel V_{HH}	Desmyter *et al.* (2002)	sys. immunitaire	Nl-L
T07	1l0x	Superantigen/TCRβ	Sundberg *et al.* (2002)	sys. immunitaire	Nl-Nl
T08	1npe	Nidogen/laminin	Takagi *et al.* (2003)	prot. mbn.	L-Nl
T09	1tlv	LicT dimer	Graille *et al.* (2005c)	rég. transc.	Oligo.
T10	1urz	trimère TBE virus E	Bressanelli *et al.* (2004)	prot. virale	Oligo.
T11	1ohz	Cohesin/dockerin (*non lié*)	Carvalho *et al.* (2003)	cellulosome	Nl-H
T12	1ohz	Cohesin/dockerin (*lié*)	Carvalho *et al.* (2003)	cellulosome	Nl-L
T13	1ynt	SAG1/Fab	Graille *et al.* (2005b)	sys. immunitaire	Nl-L
T14	1s70	Phosphatase 1δ/MYPT1	Terrak *et al.* (2004)	phosphatases	H-L
T15	1v74	Colicin D/Imm D	Graille *et al.* (2004)	inhibition prot.	L-L
T18	1t6g	Xylanase/TAXI	Sansen *et al.* (2004)	inhibition prot.	Nl-L
T19	1tpx	Ovine prion-FAB	Eghiaian *et al.* (2004)	sys. immunitaire	H-L
T20	2b3t	HemK/RF1	Graille *et al.* (2005a)	fin traduction	Nl-H
T21	1zhi	Orc1/Sir1	Hou *et al.* (2005)	rég. transc.	Nl-Nl
T22	1syx	U5-15k/U5-52k	Nielsen *et al.* (2007)	splicing de l'ARN	Nl-Nl
T23	2b8w	dimère hGBP1	Ghosh *et al.* (2006)	transd. du signal	Oligo.
T24	2j59	Arf1/ARHGAP21	Ménétrey *et al.* (2007)	transd. du signal	Nl-H
T25	2j59	Arf1/ARHGAP21	Ménétrey *et al.* (2007)	transd. du signal	Nl-L
T26	2hqs	TolB/Pal	Bonsor *et al.* (2007)	prot. mbn.	Nl-Nl
T27.1	2o25	HIP2/Ubc9	Walker[c]	traitement prot.	Nl-Nl
T27.2	2o25	HIP2/Ubc9	Walker[c]	traitement prot.	Nl-Nl
T28	2oni	dimère NEDD4L	Walker[d]	traitement prot.	Oligo.

[a] sys. immunitaire : anticorps-antigènes, prot. mbn. : protéines membranaires, rég. transc. : complexe intervenant dans la régulation de la transcription, prot. virale : protéine virale, cellulosome : protéines constitutive du cellulosome, inhibition prot. : complexe enzyme-inhibiteur, transd. du signal : complexe intervenant dans la transduction du signal, tranport mb. : complexe , traitement prot. : complexes intervenant dans le traitement des protéines (*protein processing*).

[b] Nl : non lié, L : lié, H : homologie, Oligo. : prédiction d'un oligomère.

[c] Communication personnelle : J.R. Walker, G.V. Avvakumov, S. Xue, E.M. Newman, F. Mackenzie, J. Weigelt, M. Sundstrom, C.H. Arrowsmith, A.M. Edwards, A. Bochkarev, S. Dhe-Paganon.

[d] Communication personnelle : J.R. Walker, G.V. Avvakumov, S. Xue, C. Butler-Cole, J. Weigelt, M. Sundstrom, C.H. Arrowsmith, A.M. Edwards, A. Bochkarev, S. Dhe-Paganon.

Durant les cycles 3 à 5 (cibles 8 à 19), le nombre de complexes anticorps-antigène a drastiquement diminué laissant la part belle aux complexes protéase-inhibiteur (cibles 15 à 18). Malheureusement, les cibles 16 et 17 furent annulées car les structures cristallines des complexes

furent rapidement disponibles sur Internet. De même, la structure de la cible 15 fut accessible quelques jours avant la date finale de soumission, ne permettant pas à tous les prédicteurs de déposer leurs modèles. En plus des cibles dites classiques : antigène-anticorps et protéase-inhibiteur où l'interface d'interaction peut être assez facilement identifiée, de nouvelles cibles sont apparues durant les cycles 3 à 5. Il s'agit des cibles 8 et 14, respectivement impliquées dans la structuration de la membrane cellulaire et la transduction du signal. Enfin, de nouveaux protocoles de docking ont vu le jour durant cette $2^{\text{ème}}$ édition. Pour certaines cibles (T11, T14 et T19), il fut nécessaire de modéliser un des partenaires par homologie, ajoutant encore un degré de difficulté supplémentaire. Ce protocole fut d'ailleurs analysé pour le complexe cohesine-dockerine qui fut proposé, pour la cible 11, avec la molécule de dockerine à modéliser tandis que, pour la cible 12, cette même molécule était issue du complexe cristallisé. Les cibles 9 et 10 proposèrent un nouveau défi aux prédicteurs puisqu'il ne s'agissait plus de prédire l'assemblage d'un dimère mais la création d'un oligomère à partir de sa sous-unité.

Les cibles des cycles 6 à 12 ont reflété l'intérêt grandissant des biologistes structuraux pour les complexes impliqués dans des mécanismes cellulaires jusque-là peu étudiés au niveau structural comme : la transcription des gènes, la transduction du signal, le traitement des protéines ou de l'ARN (Janin, 2007). Ces nouvelles cibles ont constitué un nouveau défi pour les prédicteurs qui avaient entraîné leur programmes surtout sur des complexes anticorps-antigènes et protease-inhibiteur. Comme dans les cycles précédents, au moins l'un des partenaires des cibles 20 à 28 était sous la forme non-lié et certains devaient être modélisés par homologie.

Une description plus détaillée de l'ensemble des complexes cibles présentés à CAPRI est disponible dans les publications : Janin et Séraphin (2003); Janin (2005b, 2007).

Évaluation des résultats : Chaque groupe participant (prédicteurs, serveurs ou *scoreurs*) est autorisé à déposer 10 modèles par cible depuis le cycle 3 (5 cibles pour les précédents cycles). Chaque modèle est ensuite comparé à la structure cristallisée du complexe et évalué suivant plusieurs critères (voir figure 1.14 et Méndez *et al.* (2003)).

Trois paramètres permettent de classer les modèles prédits : le RMSD de l'interface, le RMSD du ligand et le pourcentage de contacts natifs (voir figure 1.14).

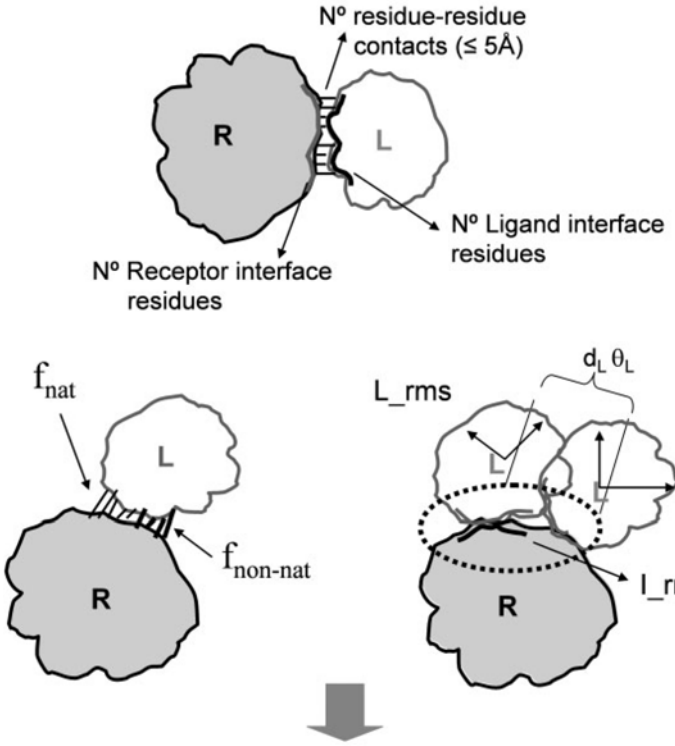
1. Pour calculer les fractions de contacts natifs et non natifs, il faut d'abord définir le terme de contact. Pour les évaluateurs de CAPRI, deux résidus sont considérés en contact à l'interface si l'on ne trouve pas d'atomes d'autres résidus à une distance de 5 Å autour des résidus en contact. La fraction de contacts natifs (f_{nat}) est définie comme le rapport entre le nombre de contacts corrects prédits par le modèle et le nombre de contacts au sein du complexe cristallisé. La fraction de contacts non-natifs ($f_{non-nat}$) est le nombre de contacts incorrects, prédits par le modèle, divisé par le nombre total de contacts prédits dans le complexe. Même si cette dernière mesure n'est pas prise en compte pour le classement, elle est intéressante pour étudier la qualité du modèle. En effet, si le pourcentage de contacts natifs est faible par rapport à celui de contacts non-natifs, cela démontre que l'interface peut être trop grande ou mal positionnée et que le modèle perd de son intérêt. Un but subsidiaire des prédicteurs est donc de maximiser la fraction de contacts natifs tout en minimisant la fraction de contacts non-natifs.

2. Pour évaluer la concordance géométrique entre le modèle prédit et le complexe cristallisé, il est nécessaire de calculer le RMSD du ligand. Celui-ci est calculé en superposant la structure de la molécule dite "récepteur" du modèle (la plus grande des deux protéines formant le complexe) sur celle de la cible. On calcule ensuite le RMSD du ligand (la plus petite des deux molécules) en se limitant aux atomes formant le squelette protéique (N,Cα,C,O).
3. La taille du ligand peut fortement varier d'un complexe cible à un autre. Or, parfois, le ligand peut être constitué de plusieurs parties dont une interagit avec le récepteur tandis que l'autre peut subir de forts changements conformationnels durant l'association (voir cible 13 Méndez *et al.* (2005)). Dans ce genre de cas, le modèle peut décrire une interaction correcte mais avoir un RMSD pour le ligand très élevé si, par exemple, la position du domaine n'interagissant pas n'a pas a été correctement modélisée. Pour pallier ce problème, il est possible de calculer une autre valeur de RMSD : le RMSD de l'interface (ou I_rmsd). Il s'agit de limiter le calcul du RMSD aux résidus constituant l'interface. Dans ce cas, l'interface est définie comme tout résidu possédant un atome à une distance de moins de 10 Å de la protéine partenaire.

D'autres paramètres sont pris en compte, non pas pour classer les résultats, mais pour en éliminer certains. Il s'agit du pourcentage d'identité entre le complexe prédit et cristallisé et le nombre d'interpénétrations (*clashes*) présentes au niveau de l'interface prédite. Dans le premier cas, un taux d'identité trop bas entre la cible et la prédiction (construite par homologie) entraîne une élimination du complexe prédit. Le second cas s'explique par le fait que trop d'interpénétrations peuvent augmenter artificiellement la fraction de contacts natifs. Il est donc nécessaire de fixer une limite qui corresponde à la moyenne des interpénétrations de tous les complexes prédits à laquelle on ajoute deux fois la déviation standard. Les *clashes* sont définis comme des contacts entre atomes (autres que les atomes d'hydrogène) de moins de 3 Å.

Tous ces paramètres permettent ensuite de classer les modèles en fonction de la fraction de contacts natifs puis des divers RMSDs en quatre catégories : high, medium, acceptable et incorrect (voir figure 1.14).

Les enseignements tirés du challenge CAPRI : Il existe plusieurs manières de tirer des informations des tableaux 1.7 et 1.8. D'abord, nous pouvons faire un bilan, édition par édition, sur les progrès réalisés par l'ensemble des prédicteurs. Ainsi, si l'on regarde la première édition de CAPRI (regroupant les cibles 1 à 7), les résultats obtenus n'étaient pas homogènes bien que la majorité des cibles soient du type anticorps-antigène (voir tableau 1.6). En effet, les deux dernières cibles de cette édition ont été particulièrement bien prédites tandis que les cibles 4 et 5 n'ont fait l'objet d'aucun résultat correct. En effet, le mode d'association de ces cibles ne suit pas le schéma classique d'association antigène-anticorps impliquant, dans ce cas, le domaine VHH (*VH Heavy chain*) alors que l'on s'attendait à une interaction avec la région CDR (*Complementary Determining Region*)(Méndez *et al.*, 2003). Pour le seul complexe n'impliquant pas un système anticorps-antigène (TO1) les résultats furent très moyens (8 modèles acceptables). Le bilan de cette première édition fut que les programmes de docking ne pouvaient être utilisés de façon routinière pour prédire précisément les assemblages protéiques (Méndez *et al.*, 2003). Néanmoins, ils pouvaient être utilisés afin de délimiter des zones d'interactions pour guider les



Ranking	Conditions based on Capri computed parameters
High	$f_{nat} \geq 0.5$ AND (L_rms ≤ 1.0 OR I_rms ≤ 1.0)
Medium	($f_{nat} \geq 0.3$ AND $f_{nat} < 0.5$) AND (L_rms ≤ 5.0 OR I_rms ≤ 2.0) OR $f_{nat} \geq 0.5$ AND L_rms > 1.0 AND I_rms > 1.0
Acceptable	($f_{nat} \geq 0.1$ AND $f_{nat} < 0.3$) AND (L_rms ≤ 10.0 OR I_rms ≤ 4.0) OR $f_{nat} \geq 0.3$ AND L_rms > 5.0 AND I_rms > 2.0
Incorrect	$f_{nat} < 0.1$ OR (L_rms > 10.0 AND I_rms > 4.0)

FIG. 1.14 – Illustrations du calcul des RMSD de l'interface et du ligand. L_rmsd : RMSD de l'interface ; L_rmsd : RMSD du ligand ; f_{nat} : fraction de contacts natifs ; $f_{non-nat}$: fraction de contacts non natifs. Tableau présentant le classement des modèles en fonction de I_rmsd, L_rmsd et f_{nat}. *Reproduit de Méndez* et al. *(2005)*

.

recherches dans le cas, par exemple, de mutagénèse dirigée. Enfin, il apparaissait que l'utilisation d'informations issues de la bibliographie avait permis de cibler les bonnes zones d'interactions et avait ainsi facilité la recherche de complexes.

Les résultats de la deuxième édition (cibles 8 à 19) furent réellement impressionnants : l'ensemble des cibles proposées fut prédit correctement (*i.e.* il a été prédit au moins un modèle acceptable pour chaque cible) contrairement à l'édition précédente. Ces résultats sont d'autant plus remarquables que la majorité des cibles étaient proposées sous la forme lié-non lié et qu'il fallait, pour certains complexes, créer un partenaire par homologie (voir tableau 1.6). Au cours

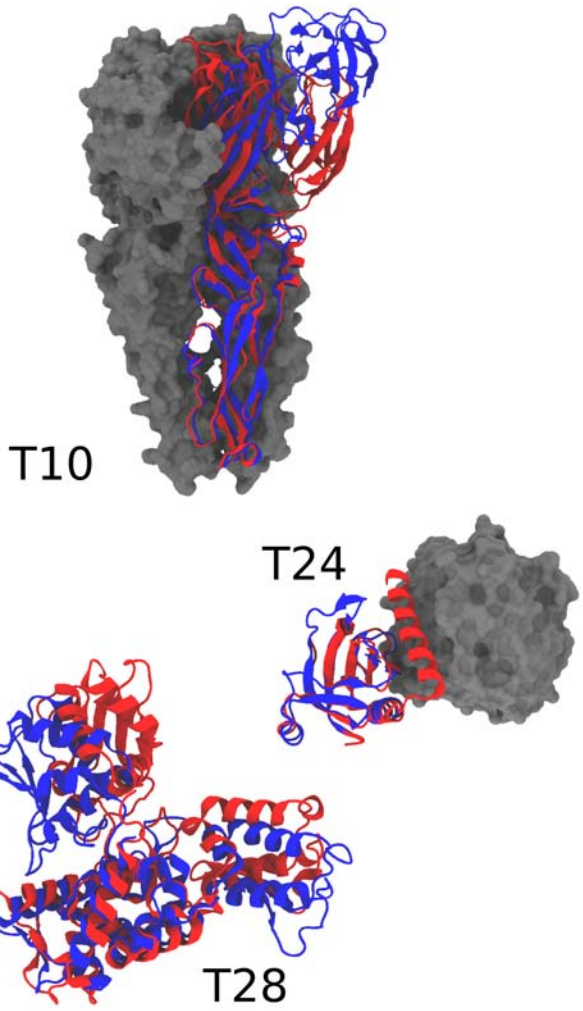
de cette édition, de nouvelles méthodes ont été mises en place, en particulier pour prendre en compte la flexibilité du squelette protéique et des chaînes latérales (Méndez *et al.*, 2005). Il restait néanmoins certaines difficultés pour les programmes de docking à prédire des modèles d'une grande précision lorsque les changements conformationnels étaient importants, comme ce fut le cas pour les cibles 9 et 10 (voir tableau 1.7). Enfin, lors de cette édition, est apparu un nouveau type de programme de docking : le serveur automatique nommé CLUSPRO (Comeau *et al.*, 2005). Il fut le précurseur de nombreux serveurs utilisés lors de l'édition suivante.

Au cours de la troisième édition de CAPRI, deux nouvelles catégories ont été créées : la catégorie serveur et la catégorie *scoreurs*. Cette dernière est très intéressante pour analyser la réussite des fonctions de score. Les résultats obtenus lors de cette édition furent un peu moins bons que pour l'édition précédente : aucune équipe n'a proposé de modèle convenable pour la cible 28. De même, les modèles de qualité high furent peu nombreux. Ceci est dû à l'originalité des complexes proposés, plus difficiles à prédire que les complexes classiques antigène-anticorps et protéase-inhibiteur de l'édition précédente. Les serveurs de docking automatiques ont aussi donné des résultats très corrects pour certaines cibles montrant ainsi qu'il est possible de prédire la structure d'assemblages protéiques sans intervention humaine. Enfin, les résultats des *scoreurs* lors de cette édition ont montré l'importance des fonctions de score pour reclasser les complexes. En effet, en utilisant les modèles proposés par les groupes de prédicteurs, les *scoreurs* obtinrent un taux plus élevé de résultats de meilleure qualité (Lensink *et al.*, 2007). Notons que ces *scoreurs* ont correctement reclassé les résultats mis à leur disposition mais n'ont pas amélioré la qualité de ceux-ci (Lensink *et al.*, 2007).

Nous pouvons aussi faire un bilan par rapport aux cibles proposées : les tableaux 1.7 et 1.8 montrent certaines cibles particulièrement difficiles à prédire : il s'agit des cibles 4, 5, 9, 10, 20, 24 et 28. Les résultats sur les cibles 4 et 5 s'expliquent par une zone d'interaction différente de ce qu'attendaient les équipes de prédicteurs. En ce qui concerne les cibles 9 et 10, la difficulté était double : prédire un dimère ou un trimère à partir d'une sous-unité qui subissait d'importantes variations structurales (voir l'exemple de la cible 10 tableau 1.7). Cette problématique s'est aussi retrouvée pour la cible 28 où, là encore, aucune prédiction correcte n'a été soumise. Les partenaires des cibles 20 et 24 subissaient eux aussi de forts changements conformationnels, en particulier pour la cible 24 où il manquait une hélice C-terminale dans la forme libre du domaine PH de la protéine ARHgap21 (voir tableau 1.7). Ce complexe Arf1-ARHGap21 a été soumis à nouveau mais, cette fois, avec la forme liée du domaine PH (cible 25) : les résultats ont été, dans ce cas là, bien meilleurs. Il est possible d'arriver à la même conclusion pour les cibles 11 et 12 où la structure de la dockerine a été soumise sous sa forme non-liée (T11) puis sous sa forme liée (T12). Il apparaît donc, à travers ces exemples, qu'il reste encore très difficile de prédire la structure de complexes impliquant une forte variation structurale des partenaires.

TAB. 1.7 – Résultats obtenus pour chaque cible durant les rounds 1 à 12. *inspiré de Janin (2005a) et Janin et Wodak (2007)*

		Qualités des modèles [a]		
Cible	Type [b]	*High*	*Medium*	*Acceptable*
T01	NI-NI	0	0	8
T02	NI-L	0	1	6
T03	NI-L	0	2	0
T04	NI-L	0	0	0
T05	NI-L	0	0	0
T06	NI-L	4	4	0
T07	NI-NI	5	7	8
T08	L-NI	2	9	16
T09	Oligo.	0	0	1
T10	Oligo.	0	1	3
T11	NI-H	0	11	31
T12	NI-L	21	0	14
T13	NI-L	6	6	7
T14	H-L	16	20	32
T15 [c]	L-L	4	5	7
T18	NI-L	0	6	4
T19	H-L	1	10	9
T20	NI-H	0	0	3
T21 [d]	NI-NI	0	4	7
T24	NI-H	0	0	4
T25	NI-L	1	13	20
T26	NI-NI	0	22	20
T27.1	NI-NI	0	0	0
T27.2	NI-NI	0	2	55
T28	Oligo.	0	0	0



[a] Résultats issus de Méndez *et al.* (2003, 2005); Lensink *et al.* (2007)
[b] Nl : non lié, L : lié, H : homologie, Oligo. : prédiction d'un oligomère.
[c] Seules prédictions soumises avant publication de la structure cristallographique du complexe sur Internet.
[d] Pour les cibles T22 et T23 les prédictions ont été annulées car une image de l'assemblage était visible sur Internet.
A droite du tableau, quelques exemples de cibles avec un fort changement structural entre la forme libre et complexée d'un des partenaires. En rouge la forme liée et en bleu, superposée, la forme libre.

TAB. 1.8 – Résultats des prédicteurs et des serveurs de docking pour les rounds 1 à 12. *inspiré de Méndez* et al. *(2003), Méndez* et al. *(2005) et Lensink* et al. *(2007)*

Groupes	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13	T14	T18	T19	T20	T21	T24	T25	T26	T27.2	T28	Récapitulatif
Prédicteurs																								
Baker[a]	0	0	0	0	0	**	***	–	0	0	**	***	**	***	0	***	*	0	0	0	**	*	0	4***/4**/2 (10/22)
Camacho	*	0	0	0	0	***	***	**	0	0	0	***	***	**	**	*	0	0	*	*	0	*	0	4***/3**/5 (12/23)
Eisenstein	*	*	0	0	0	0	***	***	0	0	*	***	0	**	0	0	0	0	0	***	*	*	0	4***/1**/5 (10/23)
Weng	0	**	0	0	0	0	**	**	0	0	*	***	***	***	**	**	0	*	*	**	*	**	0	3***/7**/4 (14/23)
Abagyan	0	0	**	0	0	***	**	**	0	*	**	***	*	***	**	**	–	–	–	–	–	–	–	3***/6**/2 (11/16)
Gray[a]	0	0	0	0	0	**	***	***	–	–	**	***	0	0	0	**	0	**	0	0	0	**	0	3***/5**/0 (8/21)
Bonvin	–	–	–	–	–	–	–	–	–	**	**	0	***	***	0	0	0	**	0	*	**	*	0	2***/4**/2 (8/14)
Ten Eyck	*	*	0	0	0	**	0	0	0	0	0	***	***	**	0	0	–	**	0	0	0	*	0	2***/3**/3 (8/22)
Ritchie	0	0	**	0	0	***	0	0	0	0	**	***	*	*	0	0	0	0	0	0	*	0	0	2***/2**/3 (7/23)
Wolfson	*	0	0	0	0	0	***	**	*	*	*	*	0	**	**	*	0	0	0	0	0	*	0	1***/3**/7 (11/23)
Bates	–	–	–	0	0	0	***	*	0	*	**	*	0	**	**	*	0	0	0	*	0	0	0	1***/3**/5 (9/20)
Sternberg	0	*	0	0	0	***	*	**	0	0	*	*	0	**	0	*	–	–	–	–	–	–	–	1***/2**/5 (8/16)
Zacharias	–	–	–	–	–	–	–	**	0	–	–	–	–	***	0	**	0	0	0	0	*	*	0	1***/2**/2 (5/12)
Vakser	0	*	0	–	–	–	–	–	–	0	–	–	–	**	**	–	0	0	–	–	**	*	0	0***/3**/2 (5/11)
Palma	–	0	–	0	0	**	*	0	0	0	–	0	0	0	0	0	–	–	–	–	–	–	–	0***/1**/1 (2/13)
Valencia	*	–	–	–	–	–	–	*	0	0	*	*	–	0	0	–	–	–	–	–	–	–	–	0***/0**/4 (4/7)
Serveurs																								
Cluspro	–	–	–	–	–	–	–	**	0	0	0	***	*	0	0	*	0	0	0	*	0	*	0	1***/1**/4 (6/16)
PatchDock	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	**	*	–	–	0***/1**/1 (2/3)
Gramm X	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0	**	–	0	0	0***/1**/0 (1/5)
SKE Dock	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0	**	0	0	0	0***/1**/0 (1/6)
SmoothD.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0	0	0	**	0	0	0	0***/1**/0 (1/7)
Res.	5	5	2	0	0	7	10	11	1	4	11	13	7	13	6	9	1	4	2	10	8	11	0	
cible		1**	2**			4***	6***	2***		1**	6**	9***	4***	5***	6**	1***		3**		1***	3**	2**		
						3**	2**	7**					1**	7**		4**				5**				

[a] Gray et Baker formaient une même équipe pour les cibles 1 à 7 puis ils ont ensuite donné chacun leurs propres résultats : les résultats de Baker et Gray sont donc dupliqués dans ce tableau pour les cibles 1 à 7. La légende du tableau est présentée page suivante.

Le tableau de la page précédente présente les résultats obtenus par la majorité des participants à l'expérience CAPRI. Pour une raison de place, nous n'avons gardé que les groupes ayant participé à au moins deux éditions (à l'exception des serveurs de docking). Les groupes cités dans ce tableau ont donné au moins une prédiction de qualité acceptable ou supérieure. La première colonne indique le nom du responsable de chaque groupe ou le nom du serveur de docking. Les 23 colonnes suivantes indiquent les résultats obtenus par chaque participant : 0 indique qu'aucun des modèles soumis n'a été acceptable. – indique que le groupe participant n'a pas donné de modèle pour la cible. *, ** ou *** indiquent la qualité du résultat du meilleur modèle proposé (respectivement acceptable, medium et haut - voir partie "évaluation des résultats" pour plus d'information). La dernière colonne récapitule les résultats de chaque groupe participant en donnant le nombre de résultats hauts suivi du nombre de résultats mediums et acceptables suivis, entre parenthèses, du nombre de cibles prédites correctement (d'une précision acceptable ou supérieure) et le nombre de cibles tentées. La dernière ligne indique, pour chaque cible, le nombre de modèles corrects (d'une précision acceptable ou supérieure) suivi du nombre de modèles hauts ou mediums.

Nous pouvons enfin faire un bilan quant à la réussite de chaque équipe tout au long de l'expérience CAPRI . A travers le tableau 1.8, nous voyons que six équipes se distinguent de l'ensemble du reste des participants : celles de David Baker, de Carlos J. Camacho, de Miriam Eisentein, de Zhiping Weng, de Ruben Abagyan et celle de Jeffrey J Gray. Toutes ces équipes (à l'exception de celle de Ruben Abagyan) ont participé à la majorité des cycles de CAPRI. Elles ont donc pu, à travers les différents complexes prédits, améliorer leurs programmes de docking. Ces équipes, outre le fait de présenter certains modèles d'une très grande qualité, ont su adapter leurs programmes pour obtenir des résultats corrects lors des différentes éditions. Nous trouvons ensuite un autre groupe constitué par des équipes ayant obtenu entre 1 et 2 modèles d'une précision haute. Ce groupe est assez hétérogène puisqu'il est constitué à la fois d'équipes "établies" ayant participé à l'ensemble de l'expérience CAPRI (comme les équipes de Lynn Ten Eyck, de David Ritchie, de Paul Bates, de Haim Wolfson ou de Michael Sternberg) mais aussi d'équipes plus récentes arrivées au cours de la deuxième édition (comme l'équipe de Alexandre Bonvin ou de Martin Zacharias). Certaines équipes ont eu de très bons résultats lors de la deuxième édition mais des résultats plus mitigés lors des cycles 6 à 12 : c'est le cas, par exemple, des équipes de Paul Bates ou de Haim Wolfson. A l'inverse, les équipes de Martin Zacharias ou d'Alexandre Bonvin ont obtenu de bons (voire très bons) résultats durant ces mêmes cycles, compte tenu de leur arrivée plus tardive au sein du challenge CAPRI. Enfin, il reste un groupe de prédicteurs n'ayant pas encore obtenu de modèles de haute qualité : il est constitué par les équipes d'Alfonso Valencia, de P. Nuno Palma et de Ilya Vakser. Si les deux premières équipes ne semblent plus vraiment soumettre de prédictions (à la vue de la dernière édition de CAPRI), l'équipe de Ilya Vakser a obtenu des résultats tout à fait honorables lors des derniers rounds.

Il reste un paramètre très difficile à évaluer : le degré connaissance de chaque équipe des cibles présentées. Il est certain que les équipes présentes depuis la première édition ont acquis une expérience notable quant à la manière dont se déroule chaque round et, de par leur expérience dans le domaine de l'assemblage macromoléculaire, peuvent analyser plus rapidement les zones d'interactions potentielles pour une cible donnée. Il faut aussi noter que l'expérience personnelle de chaque participant, comme son cursus professionnel, peut apporter une plus va-

lue non négligeable dans l'analyse des données issues de publications ou dans la construction d'un complexe. Ainsi, nous pouvons noter que des équipes ayant de bons résultats ont à leur tête des personnalités ayant des compétences poussées dans des domaines touchant à l'expérimental : comme Zhipping Weng qui a un diplôme en ingénierie biomédicale, Alexander Bonvin qui maîtrise parfaitement le domaine de la Résonance Magnétique Nucléaire (RMN) ou Martin Zaccharias ayant des diplômes en biochimie et biophysique. De même, les expériences de David Baker ou Carlos J. Camacho dans le challenge CASP ont aidé ceux-ci à modéliser les modèles par homologie (Méndez *et al.*, 2005). Ce facteur "humain" est très important et nous verrons, dans le chapitre suivant, qu'il l'a été aussi dans notre cas pour construire le modèle de complexe pour la cible 34. Un moyen de mettre en exergue ce facteur et de développer des programmes automatiques de docking. Nous pouvons faire un premier bilan des résultats de ces serveurs de docking : la tendance amorcée par CLUSPRO lors de la deuxième édition de CAPRI a été poursuivie par celui-ci ainsi que par des nouveaux serveurs lors de la 3$^{\text{ème}}$ édition, à savoir obtenir de bons résultats pour la prédiction de complexes protéiques sans intervention humaine.

Du fait du caractère multi-factoriel de l'analyse des résultats (expertise, stratégie employée, connaissance du complexe à modéliser...), il est très difficile de classer les équipes de recherche dans le domaine du docking ; d'ailleurs ceci n'a qu'un intérêt limité. Ce qu'il faut surtout retenir est le nombre important d'équipes engagées dans ce challenge et que celles-ci, à travers les différentes approches testées et les divers résultats obtenus, permettent de mieux cerner les difficultés intrinsèques à l'assemblage de complexes protéiques *in silico*.

Ainsi, CAPRI, à l'instar de CASP pour le *folding* de protéines, est devenu une expérience nécessaire et stimulante pour le développement de nouvelles méthodologies de docking macromoléculaire. Tous les deux ans, elle permet de faire le point sur les avancées et les échecs dans le domaine de l'assemblage *in silico*. Les conclusions établies au cours des différentes éditions de CAPRI restent toujours d'actualité :

1. Le reclassement des résultats reste un point sensible : il est encore difficile de discriminer entre un résultat de bonne qualité et un moins bon (Lensink *et al.*, 2007). C'est pourquoi il est très important d'intensifier les efforts dans ce domaine et la création d'une catégorie spéciale *scoreurs* durant les rounds 6-12 de CAPRI permet de structurer cela.
2. Le traitement de la flexibilité est encore une étape limitante. Même si les programmes de docking actuels commencent à incorporer un certain degré de flexibilité, comme les mouvements de chaînes latérales, il reste encore très difficile de prédire de grands changements structuraux se déroulant lors de l'association macromoléculaire. Les échecs ou les résultats peu convaincants des diverses équipes de recherche sur certaines cibles sont à imputer à ces changements conformationnels difficiles à prévoir (comme pour les cibles 10, 24 ou 28 : voir tableau 1.7).
3. Enfin, l'incorporation de données issues, par exemple, de la littérature reste un critère important pour l'obtention d'un modèle correct mais peu de programmes ont développé

d'outils spécifiques pour exploiter au mieux les informations fournies par l'utilisateur.

1.5 Conclusion

Ce premier chapitre nous a donc permis de faire un point sur les assemblages macromoléculaires. Nous avons pu voir comment les outils de docking pouvaient exploiter les analyses faites sur des ensembles de complexes. L'utilisation de *benchmarks* ou la participation à l'expérience CAPRI ont permis d'améliorer considérablement les programmes de docking. Il reste cependant encore des limitations comme la prise en compte de données biologiques non structurales ou la modélisation de la flexibilité des protéines. Dans la partie suivante, nous allons voir, au travers d'exemples concrets, comment l'utilisation de la Dynamique Moléculaire en Solvant Explicite et la combinaison de divers programmes de docking peuvent aider à dépasser ces problèmes.

Chapitre 2

La dynamique moléculaire pour modéliser la flexibilité des assemblages

Sommaire

Contexte

Comme nous avons pu le voir à travers la partie précédente, les programmes de docking peuvent obtenir de très bon résultats ou à l'inverse n'apporter aucune solution. Une des raisons de ces échecs est la modélisation approximative de la flexibilité par les programmes de docking. En effet, les éléments d'entrée de ces programmes sont des structures rigides obtenues le plus souvent par cristallographie. Il faut alors mimer cette flexibilité de diverses manières.

Dans ce chapitre, nous ferons un point sur la cinétique mise en place lors de l'amarrage de macromolécules. Puis, nous présenterons les techniques qui permettent de modéliser les changements structuraux ayant lieu lors de cet amarrage. Enfin, nous nous intéresserons à une méthode en particulier : la dynamique moléculaire (plus particulièrement en solvant explicite). Nous montrerons comment celle-ci peut aider à mimer la flexibilité des protéines. Nous prendrons pour cela trois cas concrets traités durant cette thèse :
– La création d'un modèle de complexe entre les protéines Erbin et Smad3 : nous verrons d'abord comment une dynamique assez longue (16 ns) permet la mise en évidence de résidus importants pour l'interaction.
– Puis nous montrerons comment des simulations de dynamique moléculaire de durée un peu plus courte (10 ns) peuvent aider au raffinage d'assemblages obtenus grâce à divers serveurs de docking.
– Enfin, nous étendrons ce principe à des dynamiques très courtes (0,5 ns) afin de vérifier si celles-ci améliorent les résultats de docking rigide. Nous étayerons ceci à travers l'exemple de la cible 34 du challenge CAPRI.

2.1 Dynamique de l'association : cinétique et flexibilité

La majorité des études structurales sur les complexes protéiques obtenus, par exemple, par cristallographie ne dépeignent qu'un état figé de l'assemblage. Elles ne prennent pas en compte la cinétique de l'association : c'est à dire les mécanismes de formation ou de dissociation du complexe. Or, ces mécanismes sont aussi d'un intérêt capital pour la compréhension des systèmes macromoléculaires. Nous traiterons brièvement, dans cette partie, de l'aspect cinétique de l'assemblage. Nous nous focaliserons ensuite sur les réarrangements structuraux des protéines

partenaires lors de la phase finale de l'association. Enfin, nous présenterons les méthodes permettant de mimer cette flexibilité.

2.1.1 Cinétique de l'association

L'association entre deux protéines est le plus souvent décrite comme une réaction en deux étapes :

$$P_1 + P_2 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} P_1 : P_2 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} P_1P_2$$

où P_1 et P_2 sont des protéines à l'état libre, $P_1 : P_2$ le complexe de rencontre et P_1P_2 le complexe final.

Suivant ce schéma, deux protéines diffusent de manière aléatoire en solution jusqu'à atteindre une zone, appelée la *steering region*, où les partenaires vont s'orienter en fonction des interactions électrostatiques jusqu'à former un complexe dit de rencontre (voir figure 2.1). Cette étape de diffusion a été largement étudiée d'un point de vue expérimental et d'un point de vue théorique (Schreiber, 2002; Tang *et al.*, 2006). Du point de vue théorique, une méthode utilisée est la dynamique Bronwnienne (Gabdoulline et Wade, 1999; Elcock *et al.*, 2001; Gabdoulline et Wade, 2001). Ces études permettent de calculer la constante d'association qui est généralement comprise entre 10^5 et 10^6 $M^{-1}s^{-1}$ (Schreiber, 2002). Cette valeur peut être considérablement augmentée si des interactions électrostatiques fortes sont mises en jeu et ce jusqu'à $\approx 7.10^9$ $M^{-1}s^{-1}$ (Janin, 1997). Ce phénomène de diffusion guidée par les forces électrostatiques a d'ailleurs été modélisé en utilisant le champ de forces CHARMM22 (Fitzjohn et Bates, 2003).

Une fois le complexe de rencontre créé, des interactions plus spécifiques peuvent se former. L'association aboutissant au complexe final requiert parfois des réarrangements structuraux des chaînes latérales et du squelette protéique (voir figure 2.1) afin de compenser les pertes entropiques des partenaires dues aux phénomènes de désolvatation (Camacho *et al.*, 1999; Grünberg *et al.*, 2004).

Nous pouvons noter que des travaux récents ont été entrepris pour modéliser l'ensemble du phénomène "diffusion-association" en utilisant, par exemple, la dynamique brownienne guidée par des informations biochimiques suivie d'étapes de classement-raffinage (Motiejunas *et al.*, 2008), des calculs d'énergie libre (Camacho et Vajda, 2001) ou la dynamique moléculaire avec modélisation du solvant (Grünberg *et al.*, 2004).

2.1.2 Mise en évidence de la flexibilité des protéines

Durant la phase d'association, des réarrangements structuraux au niveau des parties flexibles des protéines se produisent. C'est pourquoi, il est nécessaire de repérer ces zones flexibles pour

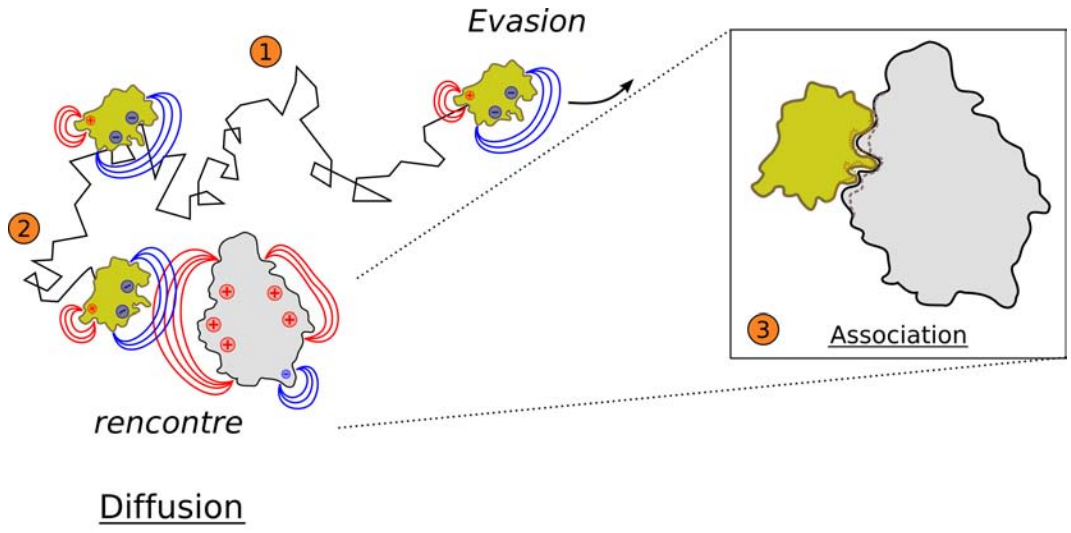


FIG. 2.1 – Deux protéines partenaires diffusent aléatoirement. Si elles sont mal orientées, elles ne se rencontreront pas (1). Si elles sont bien orientées, les interactions électrostatiques permettront le positionnement de ces protéines l'une par rapport à l'autre pour former un complexe de rencontre (2). Ce complexe peut ensuite évoluer en un assemblage final grâce à des changements conformationnels au niveau de l'interface (3). *Figure inspirée de (Elcock* et al.*, 2001)*

pouvoir ensuite les modéliser.

Analyse de la flexibilité des protéines grâce aux données expérimentales : Il est possible d'utiliser les structures des protéines déterminées par cristallographie aux rayons X ou par Résonance Magnétique Nucélaire (RMN) pour identifier les zones flexibles (Gerstein et Echols, 2004).

Si la protéine a été cristallisée sous sa forme libre et en complexe (sous sa forme liée), il est possible de comparer ces deux formes pour identifier les régions ayant subi des changements structuraux (Betts et Sternberg, 1999). Le facteur thermique B donné dans un fichier PDB (Berman *et al.*, 2000) peut être également analysé. Ce facteur est un indicateur de l'oscillation de l'atome autour de la position du modèle. La spectroscopie RMN permet d'obtenir les structures des protéines en solution et de visualiser les mouvements internes de celles-ci (Wand, 2001; Mittermaier et Kay, 2006).

Ces données seront ensuite être utilisées pour générer des trajectoires en utilisant, par exemple, des techniques de *morphing* (Zeev-Ben-Mordehai *et al.*, 2003). Ces trajectoires sont ensuite stockées dans des bases de données de mouvements (Echols *et al.*, 2003).

Analyse de la flexibilité des protéines grâce aux modes normaux : Les modes normaux de vibration sont considérés comme des oscillations harmoniques autour d'un minimum d'énergie. Chaque mode décrit un état du système où toutes les particules oscillent avec les

mêmes caractéristiques de fréquence. En théorie, les modes normaux ne devraient pas pouvoir s'appliquer aux protéines qui possèdent plusieurs états conformationnels (les minimums locaux). En pratique, ces modes normaux modélisent très correctement les changements de structures entre la forme libre et liée des protéines (Petrone et Pande, 2006). Les modes normaux les plus utilisés sont ceux de basses fréquences. Tama et Sanejouand ont montré, sur un exemple de 10 protéines, que les changements conformationnels observés pouvaient être représentés par un seul mode de vibration, le plus souvent l'un des trois plus bas (Tama et Sanejouand, 2001).

Les modes normaux sont utilisés depuis une trentaine d'années pour étudier la flexibilité des protéines : les premiers travaux se limitaient à de petits polypeptides (Moore et Krimm, 1976) pour ensuite être appliqués sur des protéines de plus grande taille (Tasumi *et al.*, 1982; Brooks et Karplus, 1983). Récemment, une analyse de modes normaux appliquée sur un grand jeu de données (134 complexes protéiques) a montré son efficacité pour mettre en évidence les changements structuraux (Dobbins *et al.*, 2008). Les données recueillies peuvent ensuite être directement exploitées par les programmes de docking pour orienter l'amarrage (May et Zacharias, 2008).

Pour plus d'informations sur les modes normaux, un livre a récemment été publié dans la série *Mathematical and Computational Biology* intitulé : *Normal Mode Analysis : theory and applications to biological and chemical systems* (Chapman et Hall-CRC, 2006).

Analyse de la flexibilité des protéines grâce la dynamique moléculaire : La dynamique moléculaire a été largement utilisée pour modéliser les mouvements des macromolécules (Karplus et McCammon, 2002). Elle est d'autant plus employée que l'utilisateur peut contrôler un grand nombre de paramètres comme l'environnement autour du système étudié, la température ou le champ de forces utilisé (tous ces points seront abordés dans la section suivante).

Les mouvements de nombreux systèmes ont été étudiés par dynamique moléculaire comme les mouvements de la protéine chaperone GroEL (Ma *et al.*, 2000) ou les mouvements concertés des sous-unités de protéines kinase (Young *et al.*, 2001). De plus, les mouvements de systèmes de plus en plus massifs peuvent maintenant être calculés. Les travaux les plus marquants sont, pour l'instant, la modélisation de la totalité du virus satellite de la mosaïque du tabac, système comprenant 1 million d'atomes (voir figure 2.2), pour une durée de 50 ns (Freddolino *et al.*, 2006) ou la modélisation du ribosome contenant 2,64 millions d'atomes, pour un temps de 20 ns (Sanbonmatsu *et al.*, 2005).

Les simulations de dynamique moléculaire demandent des temps de calcul importants. C'est pourquoi elles sont uniquement utilisées pour modéliser des mouvements relativement restreints ayant lieu dans une fenêtre de temps de quelques picosecondes à quelques centaines de nanosecondes.

2.1.3 Comment modéliser cette flexibilité ?

La modélisation de la flexibilité peut être introduite avant, après ou pendant la phase de docking.

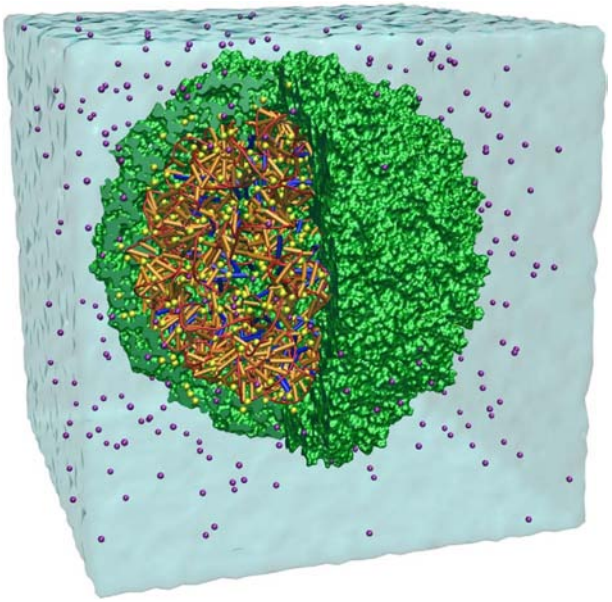


FIG. 2.2 – Représentation du virus satellite de la mosaïque du tabac (Capside + ARN) simulé en solvant explicite. *Figure issue de (Freddolino* et al.*, 2006)*

Avant la phase de docking, pour générer un ensemble de structures : Il s'agit le plus souvent de générer un ensemble de conformères afin de simuler la flexibilité de la protéine. Cet ensemble est ensuite utilisé pour réaliser un amarrage croisé. Ceci consiste, pour deux protéines partenaires P_1 et P_2, à amarrer chaque structure de l'ensemble des conformères de P_1 avec chaque structure de l'ensemble des conformères de P_2.

L'ensemble de conformères est en général obtenu grâce aux techniques utilisées pour l'analyse de la flexibilité. Ainsi, Mustard et Ritchie ont utilisé les vecteurs propres (*eigen vectors*) issus de simulations de dynamiques essentielles[13] pour une étude *a posteriori* des cibles 11 à 14 du challenge CAPRI. Ils ont montré que l'utilisation pour le docking rigide des structures générées à partir des premiers vecteurs propres permettait d'obtenir de meilleures prédictions (Mustard et Ritchie, 2005). Il est aussi possible d'utiliser la dynamique moléculaire pour générer l'ensemble des conformères (Grünberg *et al.*, 2004; Król *et al.*, 2007b). Cette méthode ne semble néanmoins réellement intéressante que dans le cas où les protéines partenaires subissent d'importants changements conformationnels durant l'association (Król *et al.*, 2007a). De plus, cette technique peut engendrer des solutions éloignées de la structure native mais pourtant bien classées par les fonctions de score (Smith *et al.*, 2005a).

Comme elle oblige à amarrer de façon exhaustive la totalité des conformères de chaque ensemble, cette technique reste néanmoins coûteuse en temps de calculs.

Pendant la phase de docking, pour guider l'assemblage : Durant cette étape, la surface protéique peut être approximée pour ne pas trop contraindre les programmes de docking. La

[13] dynamique contrainte à certains mouvements harmoniques

recherche peut être également guidée par des données recueillies lors de la phase d'analyse de la flexibilité.

Traitement implicite de la flexibilité : Comme nous l'avons vu précédemment, de nombreux programmes de docking représentent la structure des partenaires sur une grille. Dans ce cas, il est possible de modéliser la flexibilité des protéines en autorisant un certain degré d'interpénétration entre le récepteur et le ligand (Vakser, 1995; Vakser *et al.*, 1999). On parle alors de docking adouci ou *Soft docking* (Jiang et Kim, 1991). Cette technique est l'un des fondements du programme BiGGER (Palma *et al.*, 2000).

Il est aussi possible de représenter la protéine de façon simplifiée par une surface lissée. Ceci a été implémenté dans le programme HEX (Ritchie et Kemp, 1999) et a été testé pour des résolutions très basses (Sumikoshi *et al.*, 2005).

Une autre méthode consiste à simplifier la représentation des résidus de surface. Par exemple, Heifetz et Eisenstein ont "taillé" (*trimming*) les chaînes latérales de résidus très flexibles comme l'arginine ou la lysine (Heifetz et Eisenstein, 2003). Il existe aussi des représentations "gros grains" (*coarse grain*) des chaînes latérales (Levitt et Warshel, 1975; Wodak et Janin, 1978). Dans ce cas, elles ne sont plus représentées que par un (Gray *et al.*, 2003; Li *et al.*, 2003b) ou deux pseudo-atomes (Li *et al.*, 2003a; Zacharias, 2003).

Traitement explicite de la flexibilité : Pendant la phase d'assemblage, il est possible de modéliser les régions charnières (*hinge*). Ces régions constituent des zones peu structurées entre deux parties plus organisées (comme les feuillets-β ou les hélices-α). De plus, ces régions subissent en général de forts changements conformationnels et entraînent des mouvements de grande amplitude au niveau des parties plus structurées (en général des domaines). Le programme FlexDock modélise la flexibilité de ces régions à condition que l'utilisateur définisse lui-même les positions charnières (Schneidman-Duhovny *et al.*, 2005a). Emekli *et al.* ont amélioré cette méthode avec l'algorithme HingeProt. Celui-ci identifie, à partir d'une structure, les zones charnières ainsi que leurs mouvements modélisés grâce aux modes normaux (Emekli *et al.*, 2008).

Il est aussi possible de modéliser des boucles. Il s'agit là encore de régions désorganisées mais, à l'inverse des régions charnières, elles n'entraînent pas de mouvements de grande amplitude. Pour modéliser la flexibilité de ces boucles, il est possible de les représenter par un ensemble de conformations (Zacharias, 2003). Contrairement aux techniques d'amarrage de l'ensemble des structures des partenaires, cette méthode se limite à refaire l'amarrage de la boucle flexible et non de la totalité de la protéine. Les ensembles de conformations des boucles peuvent être obtenus par multi-copie (Bastard *et al.*, 2006).

Après la phase de docking, pour raffiner les résultats : Au niveau de la dernière étape du docking, correspondant à la phase de raffinage, la majorité des programmes introduisent un certain degré de flexibilité. Cette dernière peut être appliquées au niveau des chaînes latérales et du squelette protéique.

Flexibilité des chaînes latérales : Introduire de la flexibilité au niveau des chaînes latérales permet, en général, d'optimiser les conformations de celles-ci à l'interface. Pour cela, il est possible d'utiliser des bibliothèques de rotamères. Celles-ci sont obtenues par des analyses statistiques des structures des protéines. Ces bibliothèques permettent d'obtenir des informations sur les angles de rotation des chaînes latérales en fonction de la conformation du squelette protéique (Dunbrack et Karplus, 1993). Wang *et al.* ont utilisé cette méthode couplée à une minimisation des rotamères. Celle-ci, appliquée dans le cadre du challenge CAPRI, a prouvé son efficacité (Wang *et al.*, 2005).

Une autre solution est d'utiliser des simulations courtes de Monte-Carlo. Cette méthode, appliquée à la chaîne latérale du ligand et guidée par un paramètre d'énergie (Abagyan et Totrov, 1994), a été implémentée dans le programme ICM-Disco (Fernández-Recio *et al.*, 2003).

Enfin, il est aussi possible de conduire des simulations courtes de dynamique moléculaire afin d'optimiser les positions des chaînes latérales (Camacho, 2005).

Flexibilité du squelette protéique et des chaînes latérales : Durant cette dernière phase, modéliser la flexibilité du squelette protéique seul n'aurait pas un grand intérêt puisque ce changement conformationnel entraînerait en même temps le déplacement des chaînes latérales. Il faut donc relaxer l'ensemble du système squelette + chaînes latérales.

La méthode la plus utilisée dans ce cas est la dynamique moléculaire. Cette technique de raffinage a été utilisée par le groupe de Paul Bates (Król *et al.*, 2007b). Elle fait aussi partie du protocole du programme HADDOCK combinée à la méthode de recuit simulé qui consiste en une augmentation de la température pour passer les barrières énergétiques (Dominguez *et al.*, 2003).

En conclusion, les programmes de docking comme ICM-Disco, HADDOCK ou RosettaDock ont obtenu de très bons résultats au challenge CAPRI (voir tableau 1.8). La modélisation de la flexibilité des chaînes latérales et du squelette protéique par ces programmes peut expliquer, en partie, de tels résultats. Pour plus de détails sur l'analyse et la modélisation de la flexibilité, nous avons recensé 4 publications de synthèse sur le sujet : Segal et Eisenstein (2005); Bonvin (2006); Andrusier *et al.* (2008); Ritchie (2008).

Nous avons vu que la dynamique moléculaire était un outil particulièrement utilisé pour modéliser la flexibilité des protéines. Nous allons maintenant expliquer cette technique plus en détails.

2.2 La dynamique moléculaire

2.2.1 Principe de la dynamique

La dynamique moléculaire, DM, (van Gunsteren et Berendsen, 1990; Leach, 2001) est une technique couramment utilisée pour la simulation de biomolécules (Karplus et Kuriyan, 2005).

Son but est d'étudier l'évolution d'un système moléculaire au cours du temps en intégrant les équations de Newton relatives au système : $m_i\vec{a}_i = \vec{F}_i$ où m_i est la masse d'un atome i, $\vec{a}_i$ son accélération et $\vec{F}_i$ la somme des forces qui lui sont appliquées du fait de son interaction avec les autres atomes de l'environnement.

Lors de la simulation de DM, le système subit des changements conformationnels et cinétiques qui permettent d'explorer l'espace des phases espace-temps accessibles par le système. A chaque particule, à tout temps t, on associe un couple (position $\vec{r}_i(t)$,vitesse $\vec{v}_i(t)$). L'ensemble des coordonnées sur la totalité de l'espace temporel exploré constitue la *trajectoire*. Suivant l'hypothèse ergodique, l'étude d'une trajectoire infiniment longue d'un système par DM revient à échantillonner tout l'espace des phases de ce système. Il est alors possible d'accéder à des grandeurs thermodynamiques (coefficients de diffusion, fonctions de distributions radiales, énergie libre...) afin de relier la simulation à l'échelle microscopique aux expérimentations à l'échelle macroscopique.

Les équations du mouvement de Newton s'écrivent de la manière suivante :

$$m_i \frac{d\vec{v}_i(t)}{dt} = -\frac{d\vec{V}(\vec{r}_1, \vec{r}_2, ..., \vec{r}_n)}{d\vec{r}_i} = \vec{F}_i(t)$$

où m_i est la masse d'un atome i, $\vec{v}_i(t)$ sa vitesse à l'instant t et $\vec{r}_i$ sa position dans l'espace. $\vec{V}(\vec{r}_1, \vec{r}_2, ..., \vec{r}_n)$ est le potentiel d'interaction (champ de force) entre les atomes du système.
A partir d'un système dont les conditions initiales ont été fixées, la DM consiste en la répétition de deux opérations. Elle évalue d'abord la force agissant sur chaque atome au temps t, puis détermine les coordonnées et les vitesses des atomes au temps $t+\Delta t$ (avec Δt, le pas d'intégration) en fonction des forces subies par chacun d'entre eux.

2.2.2 Intégration des trajectoires

Il n'existe pas de solution analytique exacte aux équations du mouvement. Leur résolution analytique peut être fastidieuse, voire impossible, en particulier si les mouvements des particules sont couplés (problème à N-corps). Différents algorithmes ont été développés afin de les résoudre numériquement. Pour cela, une approximation usuelle consiste à diviser l'évolution du système en intervalles de temps, appelée discrétisation temporelle. À l'issue de chacun de ces pas de temps, le potentiel de chaque particule est recalculé. L'erreur induite par cette approximation est négligeable quand les intervalles de temps utilisés sont suffisamment petits. Dans la pratique, le pas d'intégration Δt, adapté pour la simulation de systèmes biologiques, est de l'ordre de la femtoseconde (10^{-15} s).

Différents intégrateurs sont disponibles, chacun se caractérisant par un rapport spécifique entre précision et efficacité. L'algorithme de Verlet (Verlet, 1967) est parmi les plus utilisés. Dans cet algorithme, les coordonnées sont développées en séries de Taylor au troisième ordre, au temps $t + \Delta t$ et $t - \Delta t$:

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \vec{v}_i(t)\Delta t + \frac{\vec{F}_i(t)}{m_i}\frac{\Delta t^2}{2!} + \frac{d\vec{F}_i(t)}{m_i dt}\frac{\Delta t^3}{3!}$$

$$\vec{r}_i(t-\Delta t) = \vec{r}_i(t) - \vec{v}_i(t)\Delta t - \frac{\vec{F}_i(t)}{m_i}\frac{\Delta t^2}{2!} - \frac{d\vec{F}_i(t)}{m_i dt}\frac{\Delta t^3}{3!}$$

La différence de ces deux séries permet d'obtenir les vitesses des atomes :

$$v_i(t) = \frac{\vec{r}_i(t+\Delta t) - \vec{r}_i(t-\Delta t)}{2\Delta t}$$

Des variantes optimisées de l'algorithme de Verlet existent. On peut citer, par exemple, les algorithmes de *leap frog* (Hockney, 1970), de Beeman (Beeman, 1976) et le plus utilisé : *velocity Verlet* (Swope *et al.*, 1981).

La modélisation d'un système macromoléculaire dans des conditions de solvatation explicite requiert la gestion d'un nombre de degrés de liberté tel qu'il faut souvent réduire artificiellement celui-ci. Pour accélérer le calcul, on peut également avoir recours à l'algorithme *SHAKE* (van Gunsteren et Berendsen, 1977; Ryckaert *et al.*, 1977; Kräutler *et al.*, 2001) qui permet d'augmenter modérément le pas d'intégration (par exemple $\Delta t = 2$ fs au lieu de 1 fs). Ceci en corrigeant les oscillations trop importantes qui peuvent alors apparaître sur les degrés de liberté les plus "rapides" du système.

Pour améliorer l'efficacité de l'intégration, des intégrateurs à pas multiples, tels que l'algorithme r-RESPA (*reversible Reference System Propagator Algorithm*) (Grubmüller *et al.*, 1991; Tuckerman *et al.*, 1992; Humphreys *et al.*, 1994) peuvent être employés afin de réduire substantiellement le temps de la simulation (d'environ un ordre de grandeur), sans perte de précision notable. Une telle optimisation diffère ici de la contrainte de type *SHAKE*, indiquée précédemment, car ce sont les degrés de liberté les plus lents qui sont périodiquement figés[14], tandis que le pas de calcul reste invariant.

2.2.3 Description de l'environnement

Représentation du solvant : Les premières applications de simulations de protéines ont été réalisées, il y a 30 ans, dans le vide (McCammon *et al.*, 1977). Bien que les résultats obtenus donnaient un aperçu de la flexibilité de la macromolécule étudiée, ils ne pouvaient pas rendre compte précisément des propriétés dynamiques de celle-ci en milieu solvaté (Levitt et Sharon, 1988). Il est pourtant primordial de tenir compte des effets du solvant lors de l'étude des biomolécules car ils jouent un rôle essentiel dans la stabilisation de ces dernières. Pour considérer les effets de solvant lors d'une simulation, deux approches sont possibles, la solvatation implicite ou la solvatation explicite.

Le traitement implicite du solvant repose sur une forme de potentiel qui traite les molécules du solvant comme un continuum diélectrique (par exemple via le modèle de Born généralisé) (Mackerell *et al.*, 2004). Les divers modèles de solvatation implicite se révèlent utiles pour réduire de façon significative le temps de calcul des simulations. Toutefois, bien qu'ils reproduisent fidèlement l'effet global du solvant, ils ne peuvent pas, par définition, reproduire les interactions

[14] Les forces qui varient le plus lentement, par exemple l'électrostatique à longue distance, sont recalculées moins fréquemment que les plus rapides, comme par exemple les interactions entre atomes liés.

des molécules du solvant au niveau local (par exemple à la surface des molécules) (Mackerell, 2004).

Le traitement explicite correspond, lui, à la modélisation d'un nombre suffisant de molécules d'eau autour des systèmes biomoléculaires d'étude au sein d'une cellule qui constitue une boîte d'eau placée dans les conditions périodiques (voir paragraphe suivant). Bien que ce traitement s'avère très coûteux en temps de calcul (les molécules d'eau ajoutées augmentant considérablement le nombre total d'atomes à considérer), il permet de modéliser de manière plus réaliste le milieu physiologique dans lequel les biomolécules évoluent.

Conditions périodiques aux limites : Une méthode particulièrement adaptée pour conduire une simulation dans des conditions de solvatation explicite, tout en réduisant les effets de bord, est celle des conditions de limites périodiques (Metropolis *et al.*, 1953). Ceci consiste à répliquer implicitement l'ensemble fini de particules du système d'étude, réparties dans une boîte centrale (en général cubique ou parallélépipédique) selon les trois directions de l'espace.

Les atomes dans les cellules images reproduisent les mouvements des atomes correspondants dans la cellule centrale (voir figure 2.3). La simulation par DM s'effectue pour les atomes de la cellule centrale en tenant compte de la présence des cellules images uniquement lors du calcul du potentiel. Le caractère pseudo-infini du système ainsi simulé contraint à effectuer certaines approximations concernant le traitement des interactions entre molécules, en particulier celle dite de "l'image minimale" qui suppose que chaque particule i de la cellule centrale n'interagit qu'avec l'image la plus proche de toutes les autres particules j.

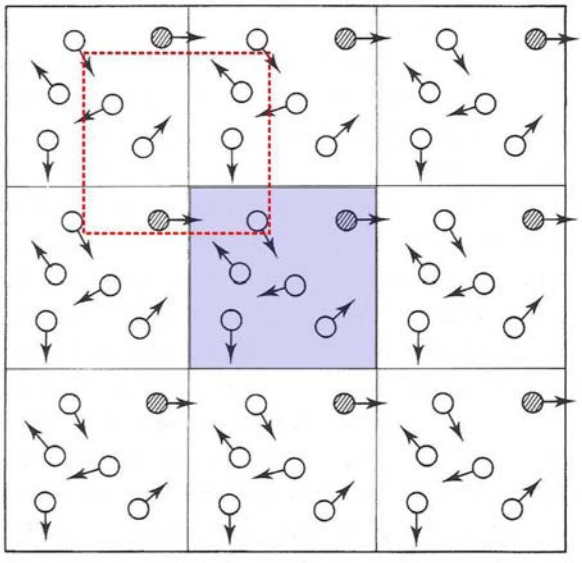


FIG. 2.3 – Conditions périodiques illustrées en 2 dimensions avec une boîte cubique. La boîte en bleu représente la cellule centrale répliquée. La boîte en pointillés rouges symbolise l'image minimale.

Prise en compte des interactions entre atomes non-liés : La complexité des algorithmes de DM, en l'absence d'optimisation, est en $O(n^2)$[15] ; ceci limite rapidement l'intervalle d'espace et de temps accessible des simulations. En effet, alors qu'à un ensemble de particules liées ne correspond qu'un nombre limité d'interactions à prendre en compte dans le calcul, le nombre de paires de particules non-liées croît, lui, en $O(n^2)$. L'optimisation du calcul des interactions entre atomes non-liés est, par conséquent, centrale afin d'améliorer l'efficacité des algorithmes de DM.

Afin de limiter le temps nécessaire au calcul des potentiels entre paires d'atomes non-liés s'exerçant à longue distance, il est courant de limiter la prise en compte de ces interactions, pour un atome donné, à ses voisins les plus proches, c'est à dire à ceux inclus dans une sphère dite de troncature (ou *Cut-off*) centrée sur l'atome en question.

Dans le cas de l'utilisation des conditions périodiques aux limites, le rayon de troncature doit être inférieur ou égal à la moitié du plus petit côté de la cellule centrale, afin qu'il n'y ait pas plus d'une image de chaque atome prise en compte. Ainsi, la complexité du calcul n'est plus que de $O(n)$ car le nombre total d'atomes au sein de la cellule centrale est borné. Mais, si l'emploi d'un *cut-off* de l'ordre de 8 à 10 Å est acceptable pour les interactions de Van der Waals, ce n'est pas le cas de l'interaction Coulombienne en 1/r, même dans le cas de faibles charges atomiques partielles. Ainsi, le problème des interactions entre atomes non-liés se réduit principalement à celui du traitement des interactions électrostatiques (Koehl, 2006).

Calcul des interactions électrostatiques : L'emploi d'un *cut-off* abrupt sur les interactions non-liées provoque des artefacts importants dans les calculs de DM (Saito, 1994). Ceci peut être atténué par l'emploi d'un *cut-off* progressif, défini en réalité par deux seuils de *cut-off*, délimitant ainsi l'intervalle de prise en charge des interactions électrostatiques de façon décroissante de 100% à 0%. Dans les conditions périodiques, la technique de sommation d'Ewald (Ewald, 1921) permet un calcul exact, mais au prix d'une complexité en $O(n^{3/2})$[16]. Pour pallier cela, une des solutions les plus utilisées de nos jours est la variante SPME (*Smooth Particle Mesh Evald*) (Essmann *et al.*, 1995), ayant une complexité en $O(n \log(n))$, qui est d'ailleurs implémentée dans le programme NAMD (Phillips *et al.*, 2005).

2.2.4 Paramétrisation du champ de forces

Enfin, avant toute simulation de DM, il est nécessaire de procéder à l'application d'un champ de force. Ce processus relie la structure du système (liste des atomes, de leur nature, de leur coordonnées, des liaisons qu'ils forment et de la nature de celles-ci) à sa topologie (liste des groupes d'atomes que chaque terme du potentiel doit considérer). Celle-ci est ensuite associée aux paramètres prédéfinis dans le champ de forces. Les paramètres sont, au cours d'une simulation de mécanique moléculaire, des constantes qui doivent être définies rigoureusement avant toute simulation. La détermination des paramètres du champ de forces s'effectue sur la base des résultats expérimentaux (données spectroscopiques, calorimétriques, structurales...), ou lorsque

[15] algorithme ayant une complexité en n^2

[16] Cette technique introduit comme contrainte supplémentaire la neutralité électrique de l'ensemble du système. Si nécessaire, cette neutralité est obtenue artificiellement par l'emploi de contre-ions répartis au sein de la cellule centrale.

ceux-ci ne sont pas disponibles (ou trop imprécis), à partir de résultats de mécanique quantique *ab initio* (Neumaier, 1997).

La transférabilité des paramètres d'une molécule vers un groupement structuralement similaire d'une autre molécule est l'une des hypothèses fondamentales de la mécanique moléculaire. Bien que cette approche puisse paraître assez rudimentaire, les multiples applications effectuées depuis son introduction ont permis de démontrer son efficacité. La validation d'un champ de forces correspond à la reproduction *in silico* de résultats expérimentaux. Notons que chaque champ de forces constitue un compromis entre simplicité conceptuelle et précision. L'amélioration d'un champ de forces est délicate du fait, par exemple, que différents paramètres peuvent être couplés (van Gunsteren et Berendsen, 1990). Toutefois, l'augmentation des performances informatiques (loi de Moore) permet d'y incorporer des termes de plus en plus sophistiqués sans que cela ne rallonge trop les temps de calcul.

Pour conduire une simulation, le choix du champ de forces est à faire sur la base de résultats déjà obtenus et disponibles dans la littérature, en utilisant les champs de forces déjà éprouvés sur des molécules analogues à celle que l'on souhaite étudier. En effet, l'emploi d'un champ de forces paramétrisé pour décrire une certaine catégorie de molécules (comme les protéines) se limite à une classe de molécules ayant des ressemblances structurales et fonctionnelles. Les champs de forces couramment utilisés pour la modélisation de biomolécules (Ponder et Case, 2003), tels que CHARMM (MacKerell *et al.*, 1998; MacKerell et Banavali, 2000), AMBER (Weiner *et al.*, 1984; Cornell *et al.*, 1995; Wang *et al.*, 2004), GROMOS (Daura *et al.*, 1998; Oostenbrink *et al.*, 2004) et OPLS (Jorgensen et Tirado-Rives, 1988; Jorgensen *et al.*, 1996), fruits de nombreuses années de recherches spécifiques, peuvent être considérés comme matures. Ils permettent donc, couplés à des programmes appropriés tels NAMD (Kale et Skeel, 1999; Phillips *et al.*, 2005), GROMOS (Christen *et al.*, 2005), CHARMM (Brooks *et al.*, 1983) ou AMBER (Weiner et Kollman, 1981), de conduire des simulations de dynamique moléculaire considérées comme fidèles, a priori, pour des systèmes classiques comme les protéines ou les acides nucléiques ainsi que les assemblages impliquant ces macromolécules.

2.2.5 Paramètres utilisés pour les simulations de dynamique moléculaire

Nous avons donc décrit les éléments clés de la dynamique moléculaire. Nous allons maintenant décrire brièvement les paramètres utilisés pour conduire les simulations de dynamique moléculaire réalisées durant cette thèse.

- Nous avons utilisé le programme NAMD avec le champ de forces CHARMM27 (MacKerell *et al.*, 2000).
- Les équations de mouvement furent intégrées avec un pas de 1 fs en utilisant l'algorithme r-RESPA dans l'ensemble de nos simulations excepté pour les dynamiques conduites sur les résultats les plus intéressants des serveurs de docking (voir section 2.4). Dans ce cas particulier, pour accélérer les calculs, nous avons augmenté le pas d'intégration à 2 fs. Ce

pas de temps fut couplé à l'algorithme *SHAKE*.

- Nous avons aussi utilisé une représentation explicite du solvant avec des conditions périodiques aux limites. Pour cela nous avons créé une boîte d'eau avec des molécules d'eau du type TIP3P. Nous avons ajouté des ions Na^+ et Cl^- pour atteindre, soit une concentration physiologique de 0,1M dans le cas des simulations menées sur le complexe Erbin/Smad3, soit la neutralité du système dans les autres cas.
- La température et la pression ont été maintenues respectivement à 300 Kelvins et 1 atmosphère (équations de Langevin) (Tuckerman *et al.*, 1992).
- Les interactions entre atomes non liés ont été prises en compte jusqu'à une distance limite de 11 Å. Les interactions électrostatiques à longue distance ont été traitées en utilisant l'algorithme *Smooth Particle Mesh Ewald*.

Nous allons maintenant montrer l'utilisation des simulations de dynamique moléculaire pour relaxer les complexes moléculaires que nous avions créés, soit par homologie, soit en utilisant divers programmes de docking.

2.3 La dynamique moléculaire pour mettre en évidence les résidus en interaction : exemple du complexe Erbin PDZ/Smad3 MH2

Nous présentons, ici, les travaux réalisés en collaboration avec des chercheurs biologistes dirigés par le Professeur Jean-Paul Borg du centre Paoli-Calmettes à Marseille. Nous avons montré comment l'utilisation de la modélisation par homologie et de la dynamique moléculaire a pu aider à l'identification d'un nouveau site de fixation entre le domaine PDZ d'Erbin et le domaine MH2 de Smad3. Les expérimentations biochimiques présentées dans cette partie ont été réalisées par Nadine Déliot post-doctorante au sein de l'équipe du Professeur Borg. Ce travail a fait l'objet d'une publication dans le journal : "*Biochemical and Biophysical Research Communications*" (voir annexe E). Du fait des contraintes du journal, nous n'avons pu détailler tous les aspects de notre méthode. C'est pourquoi, nous présentons ici une version étendue de cet article.

2.3.1 Erbin et la voie du TGF-β

Erbin protéine de régulation et de la polarité : Erbin, pour *ErbB2-interacting protein*, a été identifiée par le laboratoire du professeur Borg en 2000 lors de recherches d'interacteurs du récepteur ErbB2/HER2 par un crible double hybride chez la levure. Le récepteur ErbB2/HER2 est surexprimé dans plus de 30% des cancers du sein (Borg *et al.*, 2000). ErbB2 fait partie de la famille des récepteurs ErbB à activité tyrosine kinase. Cette famille comprend 4 membres (ErbB1/EGFR, ErbB2/HER2, ErbB3 et ErbB4) mais seul le récepteur ErbB2 interagit avec Erbin. Erbin fait partie de la famille des protéines LAP (*LRR and PDZ*). Cette protéine comporte trois régions caractéristiques (voir figure 2.4) : une région en amino-terminale, formée par 16 domaines LRR pour *Leucine Rich Repeat*, suivie d'une partie *LAP specific domain* (ou LAPSD) puis d'un domaine PDZ (*PSD-95/Discs-large/ZO-1*) en carboxy-terminal.

Le domaine PDZ est un module structural très commun. En effet, on trouve dans le génome humain 440 domaines PDZ présents dans, environ, 260 protéines ou enzymes (Schultz *et al.*, 2000). Ces domaines sont constitués de ≈90 résidus et adoptent une structure commune organisée en 6 brins β (βA à βF) encadrés par deux hélices α (αA et αB) (Nourry *et al.*, 2003; Kim et Sheng, 2004; Harris et Lim, 2001). Nous pouvons noter que le domaine PDZ d'Erbin se différencie des autres domaines PDZ puisqu'il ne possède qu'une seule hélice α (αB) (Birrane *et al.*, 2003). Le plus souvent, les domaines PDZ reconnaissent les derniers résidus de la partie C-terminale des protéines (Kornau *et al.*, 1995; Skelton *et al.*, 2003; Laura *et al.*, 2002; Wiedemann *et al.*, 2004).

Les domaines PDZ sont classés en fonction de leurs spécificités d'interaction avec les derniers résidus de la partie carboxy-terminale de leurs ligands : les PDZ de classe I se lient avec le motif S/TxV, les PDZ de classe II avec le motif ΦxΦ et les PDZ de classe III avec le motif D/ExV où S=sérine, V=valine, Φ est un résidu hydrophobe, D=acide aspartique, E=acide glutamique

et x est un résidu quelconque (Bezprozvanny et Maximov, 2001). La fixation de cette partie C-terminale se fait principalement au niveau d'une poche hydrophobe formée entre le brin βB et l'hélice αB. L'interaction Erbin/ErbB2 fait intervenir le domaine PDZ de Erbin et les sept derniers résidus carboxy-terminaux de ErbB2. La fixation principale se fait au niveau du motif VPV de ErbB2. Ainsi, le domaine PDZ d'Erbin peut se lier au motif de la classe II. La phosphorylation de ErbB2 de la tyrosine en position -7 provoque la dissociation du complexe (Borg *et al.*, 2000). De plus, ce dernier peut aussi interagir avec des motifs de classe I (Laura *et al.*, 2002; Appleton *et al.*, 2006; Jaulin-Bastard *et al.*, 2002). Cette faculté d'adaptation s'explique en partie par sa structure à laquelle il manque une hélice α (Skelton *et al.*, 2003; Birrane *et al.*, 2003).

Erbin et ErbB2 colocalisent à la membrane basolatérale dans les cellules épithéliales. L'équipe du Professeur Borg a montré qu'une délétion ou une mutation du site de liaison au domaine PDZ d'Erbin provoque une mauvaise localisation du récepteur (Borg *et al.*, 2000). De plus, les 15 derniers résidus de ErbB2 sont suffisants pour induire une localisation basolatérale de la protéine (Jaulin-Bastard *et al.*, 2001). Cependant, le rôle de Erbin dans la localisation basolatérale de ErbB2 reste discuté : en effet, un autre groupe de recherche a montré qu'Erbin n'était pas nécessaire à cette localisation (Dillon *et al.*, 2002; Kolch, 2003).

En présence de son ligand, le récepteur EGF (EGFR) se dimérise avec ErbB2 qui va à son tour s'autophosphoryler et activer la voie Ras/MAPK. La découverte de l'interaction entre Erbin et ErbB2 a posé la question du rôle d'Erbin dans la voie RAS/MAPK. Huang *et al.* ont montré que Erbin avait un rôle de régulateur négatif de la voie Ras/MAPK en inhibant la phosphorylation de Erk (*Extracellular signal-regualted kinase*), une MAP kinase, par Ras (Huang *et al.*, 2003). Erbin, en interagissant avec la protéine Sur8 via sa région LRR, inhibe la formation du complexe raf/Ras entraînant l'inhibition de la voie de signalisation Ras/Mapk (Dai *et al.*, 2006).

Erbin n'est pas seulement impliquée dans des réseaux moléculaires de signalisation, mais est également au coeur d'intéressants complexes protéiques participant à l'adhérence cellulaire (voir figure 2.4). Ainsi, la protéine des hémidesmosomes BPAG1 (*Bullous pemphigoid antigen-1*) et l'Intégrine β4 interagissent toutes les deux avec Erbin (Favre *et al.*, 2001). Ces deux protéines interviennent dans la formation des interactions cellule-matrice extracellulaire. Erbin pourrait donc faire le lien entre la formation des hémidesmosomes et la signalisation par la voie du récepteur ErbB2, bien que sa présence au niveau des hémidesmosomes n'ait pas été prouvée. Une autre interaction impliquant p0071/Plakaphiline-4 a été observée dans les cellules épithéliales (Jaulin-Bastard *et al.*, 2002). p0071 appartient à la famille des p120-Caténines qui permettent la connexion des E-Cadhérines au niveau des jonctions adhérentes avec le cytosquelette. Erbin est donc, par l'intermédiaire de p0071, connectée au cytosquelette (voir figure 2.4). L'affinité d'Erbin pour p0071 est plus forte que pour ErbB2 et la dissociation de ce complexe entraîne une perte de l'intégrité épithéliale. D'autres protéines de la famille des p120-Caténines sont impliquées dans le recrutement d'Erbin aux jonctions adhérentes. Il s'agit de δ-Caténine et ARVCF (*armadillo repeat gene deletes in velocardiofacial syndrome*) (Laura *et al.*, 2002). Toutes ces interactions avec la famille des p120-Caténines nécessitent la présence du domaine PDZ d'Erbin.

Enfin, Erbin interagit également avec les protéines Smad, maillons incontournables de la signalisation du TGF-β impliqué, entre autre, dans la transition épithéliale-mésenchyme (Warner *et al.*, 2003).

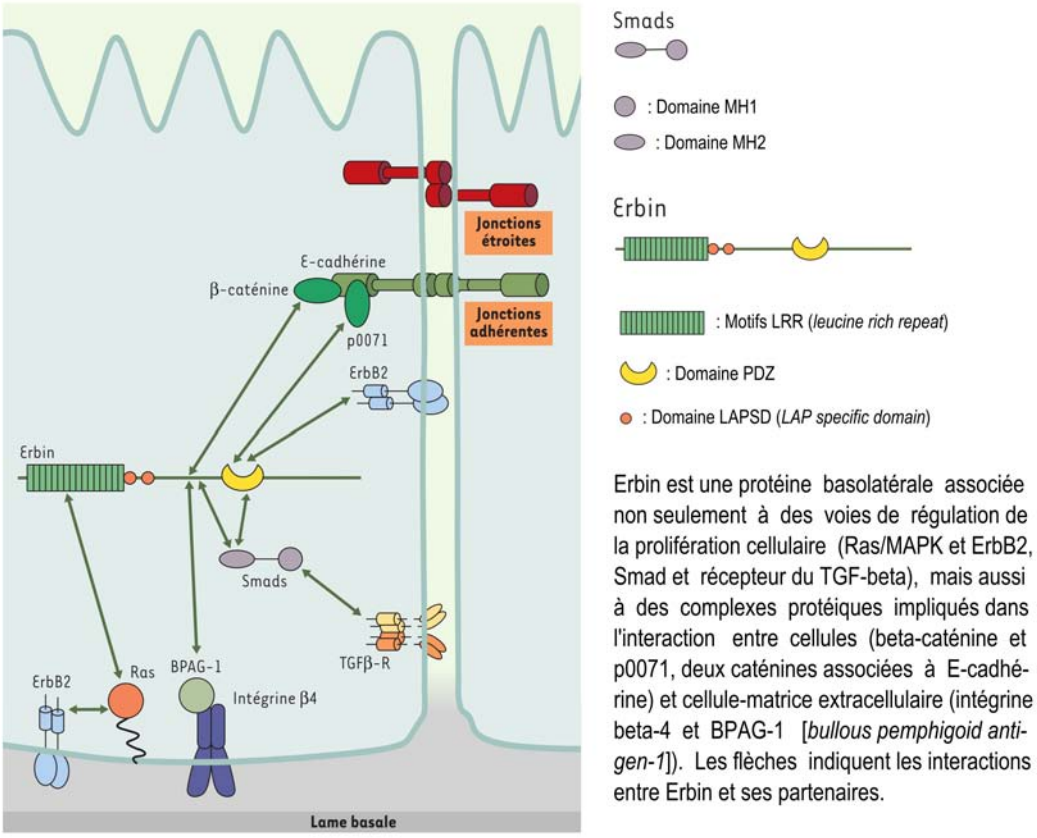


FIG. 2.4 – Réseaux moléculaires associés à Erbin. *Figure issue de (Jaulin-Bastard* et al.*, 2005).*

La famille Smad et la voie du TGF-β : Les protéines Smad sont des transducteurs des récepteurs du TGF-β et sont classées en trois groupes distincts. Ces groupes sont basés sur les fonctions et les séquences de ces protéines (ten Dijke et Hill, 2004) : les R-Smad, les Co-Smad et les I-Smad.

- Les *Receptor regulated* Smad, ou R-Smad, sont phosphorylées par les récepteurs du TGF-β de type I. Ces protéines sont subdivisées en 2 groupes : une partie est régulée par le TGF-β ou les activines (c'est le cas pour les Smad 2 et 3) tandis que l'autre partie est régulée par les BMP, *Bone Morphogenic Proteins* (Smad 1, 5 et 8).
- Les *Common* Smads, il s'agit de Smad4 mais aussi de Smad4β/Smad10 découverte récemment chez le Xénope (Masuyama *et al.*, 1999). Smad4 peut former des hétérodimères ou des hétérotrimères avec d'autres R-smad pour permettre la translocation du complexe dans le noyau et ainsi agir au niveau des gènes cibles du TGF-β.
- Enfin, les *Inhibitory* Smad, ou I-Smad, sont constituées de Smad 6 et 7. Comme leur nom

l'indique, ces protéines ont un rôle d'inhibiteur dans la voie de signalisation du TGF-β : Smad6 inhibe l'interaction Smad1/Smad4 dans la voie BMP tandis que Smad7 intervient à la fois dans la signalisation des BMP et celle du TGF-β en formant un complexe avec les protéines Smurf (Shi et Massagué, 2003).

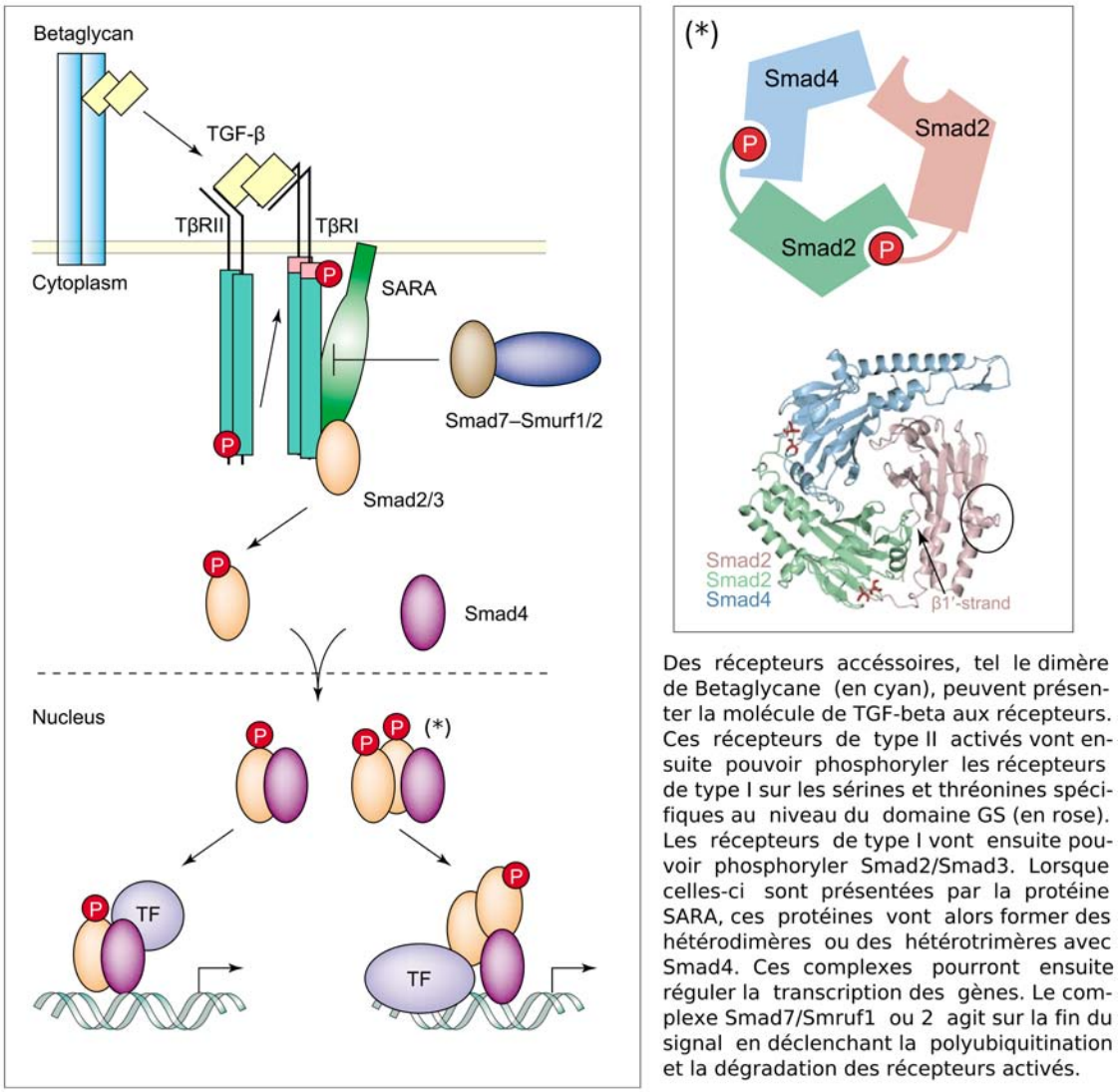


FIG. 2.5 – La voie classique d'activation des Smads par le TGF-β. NB : sur ce schéma, l'hétérotrimérisation (à droite) n'implique que les domaines MH2 de Smad2 et Smad4 mais elle peut tout aussi bien impliquer de la même manière la protéine Smad3. TF : *Transforming Factor*. *Figure modifiée de ten Dijke et Hill (2004).*

La figure 2.5 présente la voie d'activation des Smads par le TGF-β. Les protéines Smad2 et Smad3 peuvent interagir avec les récepteurs de type I et de type II. Pour cela, elles peuvent être recrutées à la membrane par la protéine SARA (*Smad Anchor for Receptor Activation*) via un domaine d'interaction nommé SBD (*Smad Binding Domain*) qui reconnaît le domaine MH2 des

protéines Smad. Cette interaction est nécessaire à la phosphorylation de Smad2 - et donc ensuite à son activité transcriptionnelle (Tsukazaki *et al.*, 1998). Celle-ci est néanmoins facultative dans le cas de Smad3 ; en effet, des mutations dans le domaine MH2 de Smad3 abolissent l'interaction avec le SBD de SARA mais n'empêchent pas Smad3 d'avoir une activité transcriptionnelle (Goto *et al.*, 2001). L'interaction avec les récepteurs de type I permettra ensuite la phosphorylation de Smad2 ou 3. Cette phosphorylation va rompre la liaison entre SARA et les Smads libérant celles-ci dans le cytoplasme. Ceci autorisera la formation de complexes homodimèriques de Smad2 ou 3 ou hétérodimèriques avec Smad4 (Wu *et al.*, 2000; Shi et Massagué, 2003). Nous pouvons noter que des interactions électrostatiques entre Smad2 ou 3 et Smad4 jouent un rôle significatif dans la formation de cet hétérotrimère (Chacko *et al.*, 2004).

Les complexes formés par Smad2, 3 et/ou 4 vont ensuite s'accumuler dans le noyau où ils vont interagir avec des facteurs de transcription pour cibler des gènes spécifiques (Moustakas *et al.*, 2001). Nous pouvons citer, pour exemple, le facteur de transcription FoxH1 (chez le Xenope) qui interagit avec le complexe Smad2-Smad4 activé via une région SID (*Smad Interacting Domain*) dans sa partie C-terminale (Chen *et al.*, 1996, 1997). Un autre site de fixation commun à FoxH1 et à la famille Mix (autres facteurs de transcription) a été découvert et nommé SIM (*Smad Interacting Motif*). Il s'agit d'un motif riche en prolines qui est nécessaire et suffisant pour interagir avec le domaine MH2 de Smad2 (et de Smad3) et permettre ainsi le recrutement des complexes de Smads actifs vers l'ADN. Ce motif est intéressant puisqu'il partage des homologies de séquences avec la région SBD de SARA (Wu *et al.*, 2000).

Ces complexes vont ensuite pouvoir réguler la transcription, soit négativement, soit positivement (Moustakas *et al.*, 2001; Shi et Massagué, 2003). Pour y parvenir, ils doivent tout d'abord se lier à la molécule d'ADN. Les domaines MH1 de Smad3 et Smad4 possèdent la capacité de se lier directement à l'ADN au niveau de séquences spécifiques appelées SBE (*Smad Binding Element*) : 5'-AGAC-3' (Shi *et al.*, 1998). En revanche, Smad2 ne peut pas se lier directement à l'ADN en raison de la présence d'un exon supplémentaire dans son domaine MH1 (Dennler *et al.*, 1999). Notons que les Smads n'ont qu'une affinité assez faible pour l'ADN c'est pourquoi il leur est nécessaire de se lier avec des facteurs de transcription pour assurer la spécificité de leur liaison avec l'ADN (Zhang et Derynck, 1999; ten Dijke *et al.*, 2000). L'activation ou la répression de la transcription se fera ensuite à l'aide de co-activateurs, comme p300 ou CBP (*CREB-Binding Portein*), ou de co-répresseurs, comme c-Ski, SnoN ou TGIF. Ces complexes de Smad s'accumulent dans le noyau où ils semblent rester plusieurs heures. Néanmoins, des études plus poussées ont fait état d'une déphosphorylation des Smads R de façon très lente. Celle-ci permet ensuite leur dissociation de smad4. Les protéines R-smad seront ensuite réexportées dans le cytoplasme. Si les récepteurs sont toujours actifs, elles pourront être rephosphorylées pour former à nouveau les complexes avec Smad4, sinon elles s'accumuleront dans le cytoplasme (Inman *et al.*, 2002).

Ainsi, l'interaction entre Erbin et Smad3 se trouve à l'intersection de deux voies de signalisations importantes pour le développement de l'organisme et la régulation cellulaire.

2.3.2 Mise en évidence de l'interaction entre le domaine PDZ d'Erbin et le domaine MH2 de Smad3

Smad3 interagit avec la région Carboxy-terminale d'Erbin : Afin d'identifier de nouveaux partenaires de Smad3, plusieurs groupes de recherche, dont l'équipe du Professeur Borg, ont réalisé des criblages double hybride chez la levure (*yeast two hybrid screening*) (Warner *et al.*, 2003). Cette technique a permis de mettre en évidence une interaction directe entre Smad3 et la protéine Erbin. Nous avons voulu, dans un premier temps, confirmer ces résultats à l'aide de tests biochimiques. Pour cela, des protéines extraites de lignées cellulaires épithéliales de mammifères (MFC10.2A) furent sujettes à immunoprécipitation en utilisant soit un anticorps anti-Smad, capable de reconnaître Smad1, Smad2 et Smad3 (Smad1/2/3) soit un anticorps anti-Myc. Après un *Western Blot* avec les anticorps appropriés, Erbin coimmunoprécipite avec les protéines Smad. De plus, ces résultats ont montré que Lano et Scrib, deux membres de la famille LAP, n'interagissaient pas avec les protéines Smad (voir figure 2.6A). Des résultats identiques ont été obtenus en utilisant des anticorps spécifiques anti-Smad3.

Nous avons ensuite essayé d'identifier les régions impliquées dans cette interaction. Pour cela, nous avons d'abord cotransfecté les HA-Smad3 avec les constructions Myc-Erbin entière (*Full Length* -FL) ou avec les partie contenant : Myc-Erbin$^{1\text{-}853}$ (résidus 1-853) ou Myc-Erbin$^{853\text{-}1371}$ (résidus 853-1371) dans des cellules COS (voir figure 2.6B). L'immunoprécipitation avec des anticorps anti-Myc montre que Smad3 interagit avec la partie Carboxy-terminale d'Erbin (853-1371) qui contient le domaine PDZ (voir figure 2.6C). Dans le but d'évaluer le rôle du domaine PDZ dans cette interaction, nous avons exprimé un mutant d'Erbin incapable de se lier avec les ligands du domaine PDZ (Myc-Erbin mutPDZ). Pour arriver à ce résultat, deux mutations (les résidus $H_{1347}G_{1348}$ changés en $Y_{1347}D_{1348}$) ont été introduites dans l'hélice αB du domaine PDZ. Ces mutations provoquent un fort affaiblissement de la liaison entre le domaine PDZ et les ligands de classe I et II, notamment les protéines ErB2 ou β-caténine (Borg *et al.*, 2000). Cette mutation n'abolit pas l'interaction avec Smad3, suggérant que, soit le domaine PDZ n'est pas impliqué dans cette interaction, soit la poche de fixation habituelle des ligands du domaine PDZ n'est pas requise pour cette interaction (voir figure 2.6C).

Pour identifier la région d'Erbin impliquée dans l'interaction avec Smad3, nous avons aussi utilisé différents fragments d'Erbin fusionnés avec la protéine de fusion GST, *Glutation S-Transferase* (voir figure 2.6D). La figure 2.6D montre que Erbin$^{853\text{-}1257}$ (853-1257) et Erbin$^{1257\text{-}1372}$ (1257-1371) interagissent avec Smad3. Erbin$^{853\text{-}1257}$ contient la région SID (*Smad Interacting Domain*) récemment décrite comme interagissant avec Smad3 (Dai *et al.*, 2007). Il n'existe malheureusement aucune information structurale sur cette région qui pourrait nous aider à modéliser cette partie. D'un autre côté, la région (1257-1371) contenant le PDZ interagit aussi avec Smad3 et ceci même si cette protéine ne contient pas de motif consensus de fixation au domaine PDZ dans sa partie Carboxy-terminale. De plus, comme il sera démontré plus tard, le domaine PDZ seul, fusionné avec la protéine GST (GST-Erbin$^{1280\text{-}1371}$), est capable de se lier avec la protéine Smad3 (voir figure 2.15). Ainsi, ces résultats suggèrent des interactions au niveau de zones multiples entre la protéine Smad3 et la région Carboxy-terminale d'Erbin.

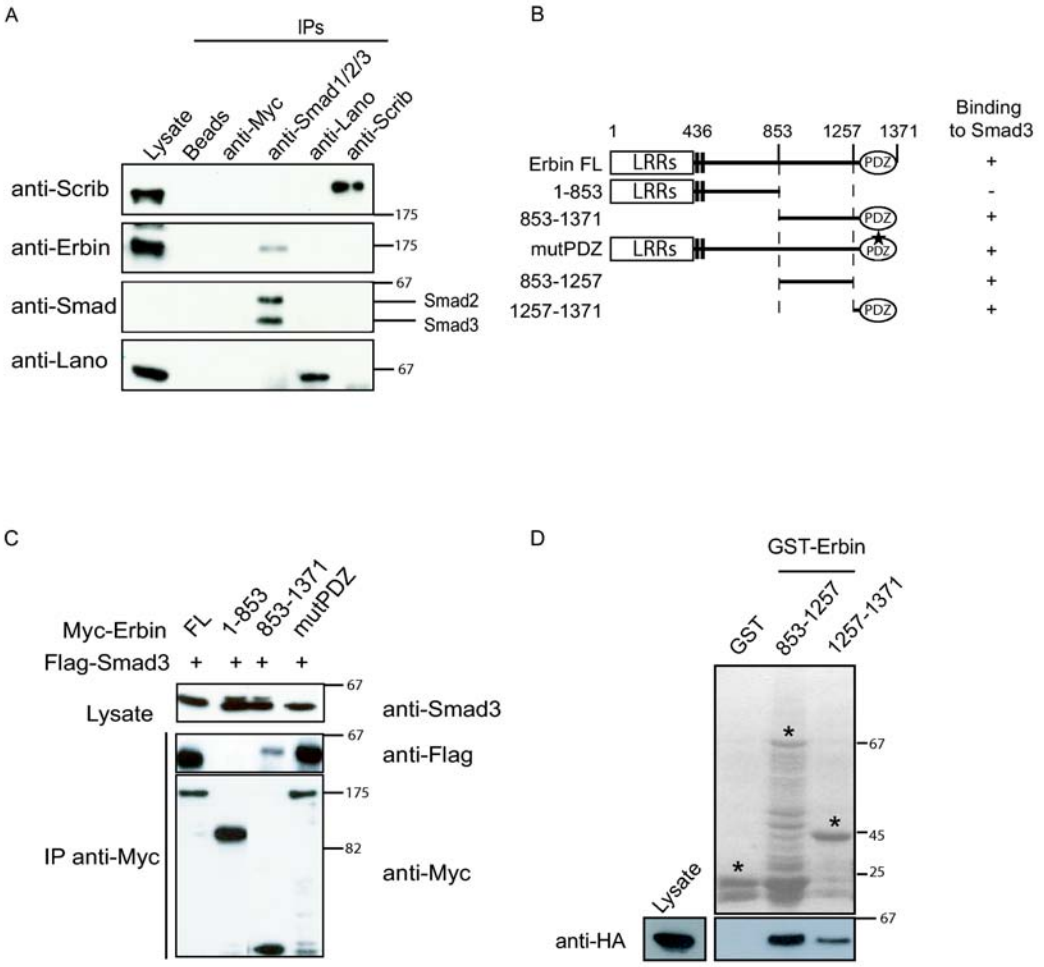


FIG. 2.6 – A- Les extraits cellulaires de MFC10.2A ont été sujets à immunoprécipitation avec des anticorps anti-Myc et anti-Smad1/2/3. Les anticorps anti-Smad reconnaissent Smad2 et Smad3. Le niveau d'expression de Smad2 et Smad3 endogènes est trop faible pour détecter celles-ci dans le lysat cellulaire. B- Schéma des diverses constructions d'Erbin. C- Des cellules COS furent transfectées avec des constructions Flag-Smad3 et Myc-Erbin. Les lysats cellulaires furent immunoprécipités avec des anticorps anti-Myc puis une analyse *Western Blot* a été réalisée. D- *Pull-down assay* sur des cellules COS exprimant HA-Smad3. Les lysats furent incubés avec les constructions GST-Erbin indiquées. Les astérisques indiquent la position des protéines GST révélée par ponceau rouge. Les poids moléculaires (en kD) sont indiqués à côté des panneaux.

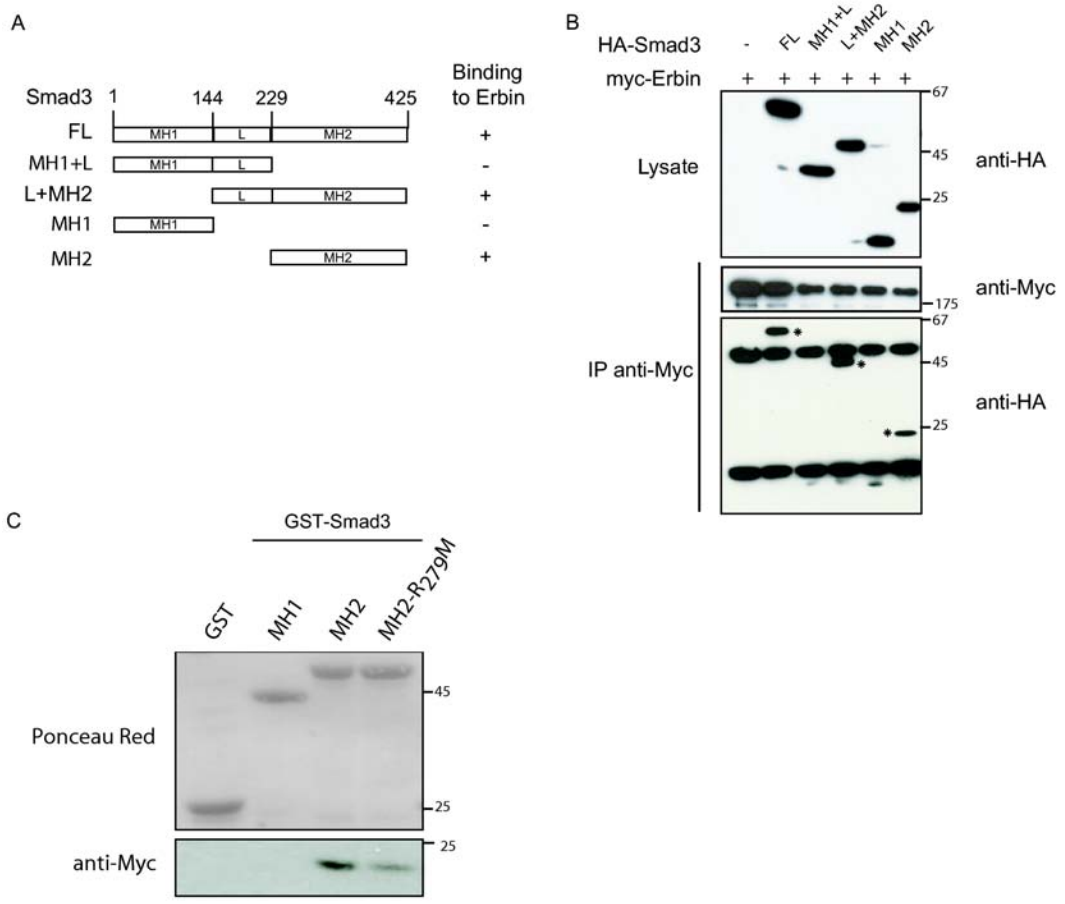


FIG. 2.7 – A- Schéma des constructions de Smad3. B- Des cellules COS furent transfectées avec les différentes constructions HA-Smad3 et Myc-Erbin. Le lysat cellulaire fut immunoprécipité avec des anticorps anti-Myc, action suivie d'un *Western blot* avec les anticorps indiqués. Les astérisques indiquent les constructions HA-Smad3 co-immunoprécipités avec Erbin (FL, L+MH2 et MH2). C- Les protéines GST mentionnées furent utilisées pour les *pulldown assays* avec Myc-Erbin. La mutation $R_{279}M$ diminue l'affinité entre le domaine MH2 de Smad3 et Erbin. Le poids moléculaire (kD) est indiqué à côté des panneaux.

Le domaine MH2 de Smad3 est nécessaire pour l'interaction Erbin/Smad3 : Nous avons ensuite voulu caractériser la région de Smad3 impliquée dans l'interaction avec Erbin. Pour cela, des constructions HA-Smad3 correspondant à différentes régions de la protéine (voir figure 2.7A) ont été transfectées dans des cellules COS avec des assemblages Myc-Erbin. Les lysats cellulaires ont été immunoprécipités en utilisant des anticorps anti-Myc. Nous avons confirmé que le domaine MH2 de Smad3 pouvait interagir avec Erbin (figure 2.7B), ce résultat est en accord avec les travaux réalisés précédemment (Warner *et al.*, 2003; Dai *et al.*, 2007). Nous avons aussi produit des constructions GST-MH1 (1-144) et GST-MH2 (229-425) dans des bactéries.

Seule la construction GST-MH2 est capable de se lier à Myc-Erbin (figure 2.7C). Pour décrire plus en détails l'interaction, nous avons tronqué la partie C-terminale du domaine MH2 de Smad3. La suppression des 58 derniers résidus du domaine MH2 ne rompt pas l'association entre Erbin et Smad3 (données non visibles). De plus, les 4 résidus terminaux SSVS ne sont pas connus comme étant un site de fixation aux domaines PDZ. Ainsi, nous avons montré que Erbin et Smad3 s'associaient via le domaine MH2 de Smad3.

2.3.3 Modélisation du complexe PDZ d'Erbin et MH2 de Smad3

Ces premiers résultats biochimiques suggèrent que le domaine PDZ d'Erbin est impliqué dans une interaction inhabituelle pour les domaines PDZ (voir figure 2.6D). D'abord, des mutations dans la poche de fixation du domaine PDZ n'altèrent pas l'interaction avec Smad3 (figure 2.6C). Ensuite, la délétion de l'extrémité Carboxy-terminale de Smad3 n'a aucun effet sur l'interaction Smad3-Erbin. Enfin, le domaine PDZ seul est toujours capable de s'associer à Smad3 (figure 2.15). Pour mieux comprendre cette interaction inhabituelle impliquant un domaine PDZ, nous avons modélisé un assemblage permettant de mettre en évidence l'interaction entre la partie 1257-1371 d'Erbin et le domaine MH2 de Smad3.

Modélisation des deux partenaires : Les résultats expérimentaux révèlent que la région d'Erbin située entre les résidus V_{1257} et S_{1371} interagit avec le domaine MH2 de Smad3. Pour modéliser cette partie, nous avons utilisé comme point de départ la structure du domaine PDZ d'Erbin issu de la *Protein Data Bank* : 1MFG (Birrane *et al.*, 2003). Cette structure comprend les résidus A_{1277} jusqu'au résidu S_{1371}. Il a donc fallu modéliser la partie 1257-1276 pour avoir la région complète d'Erbin qui interagit avec le domaine MH2 (figure 2.8A). Pour trouver des modèles à cette région manquante, nous avons d'abord comparé la structure du domaine PDZ d'Erbin avec les autres structures de domaine PDZ présentes dans la PDB en utilisant le serveur web : *Secondary Structure Matching* (Krissinel et Henrick, 2004). Nous avons trouvé que le domaine PDZ de PAR6 (Garrard *et al.*, 2003), identifiant PDB : 1NF3, pouvait servir de patron pour modéliser la partie N-terminale du domaine PDZ d'Erbin. En effet, les taux d'identité et de similarité entre le domaine PDZ d'Erbin et de PAR6 sont, respectivement, de 28% et 44% (en utilisant la matrice d'alignement BLOSUM62 - Henikoff et Henikoff (1992); Eddy (2004)). Les structures secondaires sont bien conservées entre les deux domaines, à l'exception d'une hélice α absente au niveau du domaine PDZ d'Erbin (figure 2.8A). Nous avons trouvé suffisamment de résidus identiques dans la partie N-terminale de chaque domaine pour modéliser la partie manquante du domaine PDZ d'Erbin à partir de celle du domaine PDZ de PAR6.

Le domaine MH2 de Smad3 a été lui aussi cristallisé : identifiant PDB 1MJS (Qin *et al.*, 2002). Malheureusement, cette structure n'était pas complète : les boucles (324-328) et (381-387) n'étaient pas résolues car elles devaient être trop flexibles. Nous avons considéré que ces 2 régions n'étaient pas importantes pour la reconnaissance entre Erbin et Smad3, c'est pourquoi nous avons décidé de les omettre et de construire un modèle de domaine MH2 où les résidus 323 et 329 sont liés ainsi que les résidus 380 et 388. Néanmoins, nous avons ensuite réalisé une minimisation moléculaire autour de ces régions pour stabiliser l'ensemble.

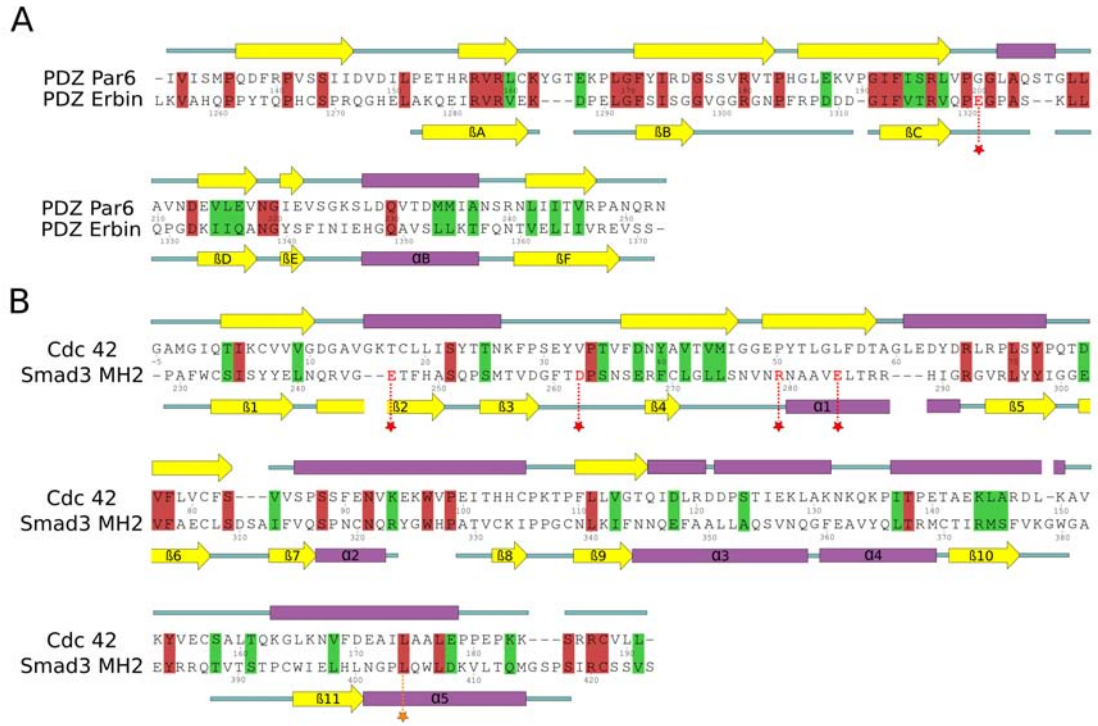


FIG. 2.8 – A- Alignement des domaines PDZ de PAR6 et d'Erbin. B- Alignement de la protéine Cdc42 avec le domaine MH2 de Smad3. Les résidus identiques sont colorés en rouge, ceux similaires en vert (résultats donnés par ClustalW). Les résidus mutés dans cette étude (E_{246}, D_{262}, R_{279}, E_{284}) sont indiqués par une étoile rouge. La leucine 404 utilisée comme point d'ancrage pour la superposition du domaine MH2 sur la protéine Cdc42 est indiquée par une étoile orange. Les hélices α sont représentées en violet et les feuillets β en jaune.

***Docking* des domaines PDZ et MH2 :** Nous avons utilisé le complexe PAR6/Cdc42 (Garrard *et al.*, 2003) comme modèle pour assembler les domaines PDZ d'Erbin et MH2 de Smad3. Garrard *et al.* décrivent un assemblage inhabituel entre le domaine PDZ de PAR6 et la protéine Cdc42 où la partie C-terminale de Cdc42 n'est pas impliquée. Nous avons donc voulu savoir comment ces 2 protéines interagissaient et quelles étaient les régions de chacune impliquées dans l'interaction. Le domaine PDZ de PAR6 interagit au niveau d'une région appelée semi-CRIB (CRIB pour *Cdc42/Rac interactive Binding*).

L'alignement de la partie N-terminale avec les régions CRIB ou semi-CRIB de PAR6, ACK et WASP (tous partenaires de Cdc42) montre plusieurs zones identiques et conservées (figure 2.11A). Pour être plus exact, les résidus prolines 136 et 141 au niveau de PAR6 ou l'histidine 83 au niveau de PAK (conservée aussi chez ACK et WASP), qui jouent un rôle important dans l'interaction avec Cdc42, sont conservés au niveau d'Erbin (*i.e.* P_{1261}, P_{1266} et H_{1267}).

L'alignement global de Cdc42 avec le domaine MH2 de Smad3 montre un faible taux d'iden-

tité : 11,3% (24,1% de similarité) et les structures secondaires ne sont pas toutes similaires (figure 2.8B). Néanmoins, il convient de se focaliser sur la région de Cdc42 impliquée dans l'interaction avec PAR6. Pour identifier les résidus impliqués dans cette interaction, nous avons utilisé le serveur SCOPPI (Winter *et al.*, 2006) (voir figure 2.11A). En nous limitant à la zone indiquée par SCOPPI, nous avons trouvé une région plus conservée que le reste de la protéine. Cette zone se situe entre les résidus 34 et 48 pour Cdc42, ce qui correspond à la séquence (263-277) au niveau du domaine MH2 de Smad3. Nous pouvons aussi remarquer que c'est une zone où une partie de la structure secondaire est conservée (voir figures 2.8B et A). Ces acides aminés conservés se trouvent à l'interface entre PAR6 et Cdc42. Par ailleurs, les acides aminés conservés au niveau de Smad3 sont exposés à la surface de la protéine (voir figure 2.9). De plus, dans cette région, nous trouvons des résidus conservés entre Cdc42 et Smad3 qui jouent un rôle important dans l'association de Cdc42 avec ses partenaires. Il s'agit des résidus D_{38} et Y_{40} (respectivement E_{267} et F_{269} chez Smad3) qui, lorsqu'ils sont mutés, affaiblissent voire abolissent l'association de Cdc42 avec ses partenaires (Owen *et al.*, 2000; Garrard *et al.*, 2003).

Réunies, ces informations nous permettraient de penser que la région ainsi mise en évidence est une région "chaude" (*hot region*) (Keskin *et al.*, 2005; Keskin et Nussinov, 2007). Nous faisons donc l'hypothèse qu'il existe une possible analogie quant aux modes d'association entre les complexes PDZ-PAR6/Cdc42 et PDZ-Erbin/MH2-Smad3. Par conséquent, le complexe PAR6/Cdcd42 a été utilisé comme patron pour modéliser l'interaction entre le domaine PDZ d'Erbin et le domaine MH2 de Smad3. Pour cela, la structure du domaine PDZ d'Erbin a été superposée au domaine PDZ de PAR6 en utilisant programme STAMP, *Structural Alignment of Multiple Proteins* (Russell et Barton, 1992). De même, le domaine MH2 de Smad3 a été superposé sur la protéine Cdc42 en utilisant les résidus conservés dans la région chaude (*i.e.* dans la région 263-277 pour Smad3 et 34-48 pour Cdc42) et la leucine 404 au niveau du domaine MH2 de Smad3 (équivalent à la leucine 174 au niveau de Cdc42). Enfin, les structures de PAR6 et de Cdc42 ont été supprimées pour ne laisser que le complexe entre les domaines PDZ d'Erbin et MH2 de Smad3 (voir figure 2.9).

Dynamique moléculaire du complexe PDZ-MH2 : La figure 2.10A montre les positions relatives entre le domaine MH2 de Smad3 et le domaine PDZ d'Erbin après minimisation. La taille de l'interface est équivalente à celle d'une parcelle unitaire standard : 793 $Å^2$, calculée avec le programme Intersurf (voir Annexe B), ce qui est équivalent à une valeur d'interface enfouie (B) de 1600 $Å^2$. Cette zone de contact est divisée en deux parties (voir figure 2.10A) : une partie entre le peptide (1257-1280) d'Erbin et le domaine MH2 de Smad3 d'aire d'interface $\approx$ 600 $Å^2$ (équivalent à B $\approx$ 1230 $Å^2$) et une zone reliant directement les domaines PDZ d'Erbin et MH2 de Smad3 d'aire d'interface $\approx$ 200 $Å^2$ (équivalent à B $\approx$ 420 $Å^2$). Pour tester la stabilité de l'assemblage et mettre en évidence les résidus critiques pour l'interaction, nous avons ensuite réalisé une dynamique moléculaire d'une durée de 16 ns.

Cette dynamique a été réalisée sur le complexe Erbin PDZ/Smad3 MH2 minimisé (6400 pas de gradient conjugué). Nous avons ensuite créé une boîte d'eau autour de l'assemblage d'une taille 76,5 Å$\times$ 67,4 Å$\times$ 107,6 Å. Les molécules d'eau utilisées sont du type TIP3P. Nous avons

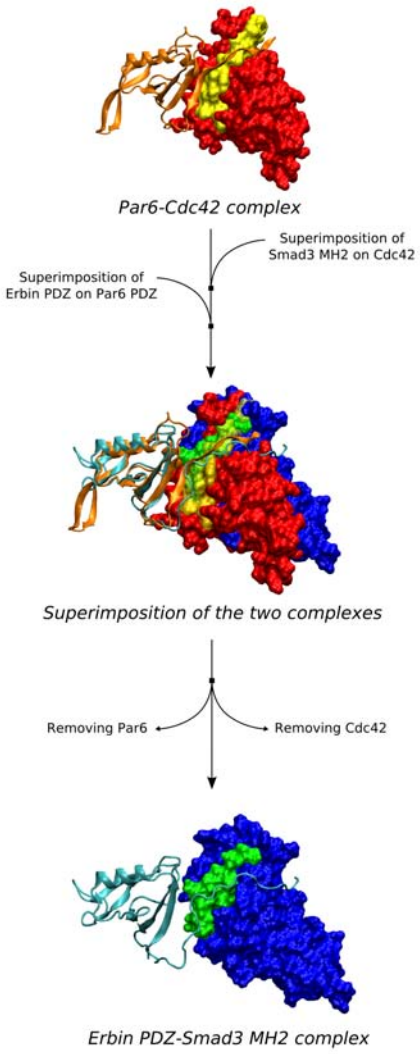


FIG. 2.9 – Vue schématique de la stratégie de *docking*. Le domaine PDZ de PAR6 est de couleur orange ; le domaine PDZ d'Erbin est en cyan ; Cdc42 est représentée en rouge. La zone d'interaction entre PAR6 et Cdc42 est en jaune. Le domaine MH2 de Smad3 est coloré en bleu. La zone conservée entre Cdc42 et Smad3 est représentée en vert.

ajouté des ions Na^+ et Cl^- pour atteindre une concentration physiologique de 0,1M. Le programme de dynamique moléculaire utilisé est NAMD (Phillips *et al.*, 2005) avec le champ de forces CHARMM27 (MacKerell *et al.*, 2000). La température et la pression ont été maintenues respectivement à 300 Kelvins et 1 atmosphère (équations de Langevin). Les équations de mouvement furent intégrées avec un pas de 1 fs en utilisant r-RESPA (Tuckerman *et al.*, 1992). Les interactions électrostatiques à longue distance ont été traitées en utilisant *Particle Mesh Ewald* avec une distance limite de 11 Å.

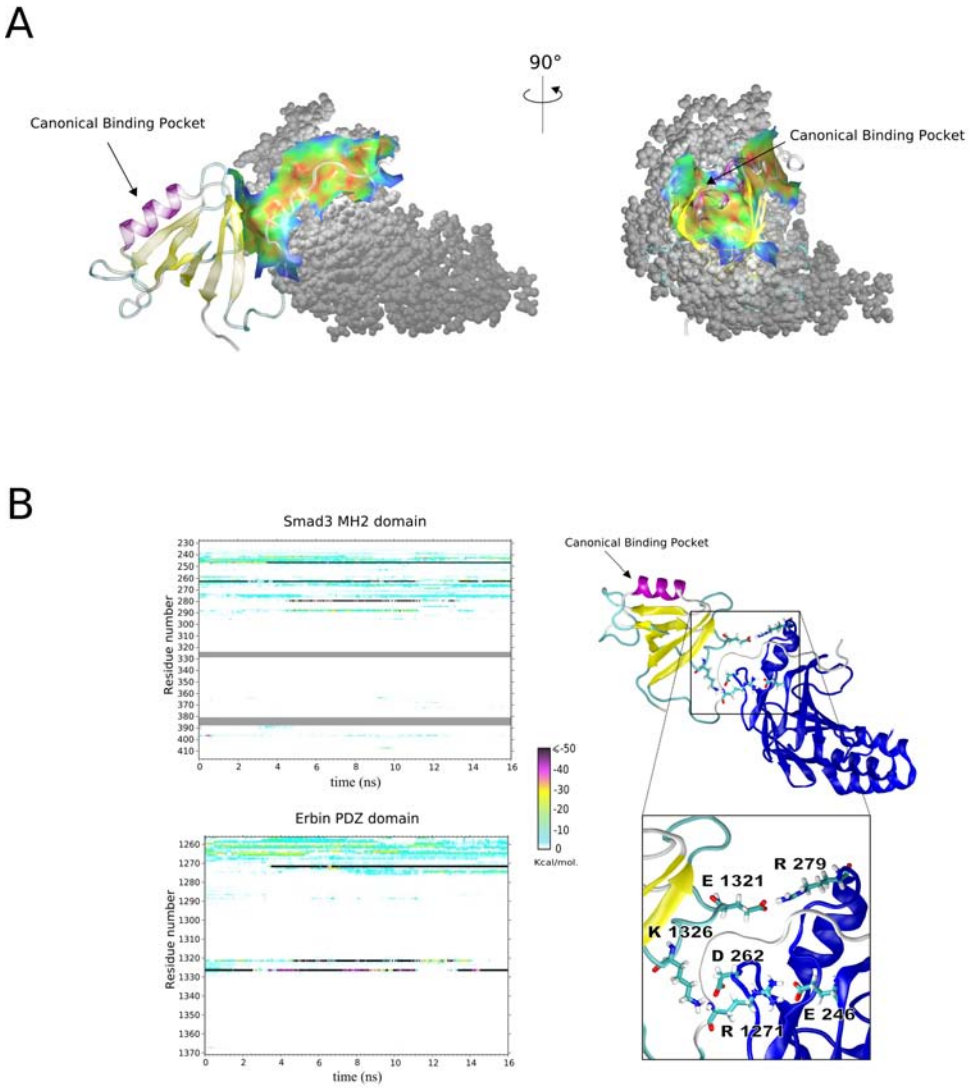


FIG. 2.10 – A- Surface de contact entre les domaines PDZ et MH2. Le domaine MH2 est représenté en gris. Le domaine PDZ est représenté en structures secondaires. Ce modèle d'interaction présente 2 zones d'interaction : une petite (≈200 Å^2) impliquant des résidus chargés et une région plus large de ≈600 Å^2 au sein de laquelle on trouve plutôt des résidus hydrophobes. L'interface est colorée en fonction de la distance : de rouge (pour une distance de 2 Å) jusqu'à bleu (pour une distance supérieure à 6 Å). B- Profils d'énergie d'interaction entre les domaines PDZ et MH2 et représentation du complexe avec les interactions principales. Les lignes colorées représentent les profils d'énergie allant du bleu clair (pour de faibles interactions) au noir (pour des interactions fortes). Les zones grisées au niveau du diagramme pour le domaine MH2 représentent les parties non cristallisées. Il faut préciser ici que les valeurs obtenues ne sont pas des mesures d'énergie libre et ne peuvent donc pas être utilisées pour déduire les caractéristiques thermodynamiques de l'interaction. La représentation du complexe a été réalisée à 11 ns lorsque toutes les interactions chargées étaient présentes. Le domaine MH2 de Smad3 est représenté en bleu. Le domaine PDZ est coloré en fonction de sa structure : violet pour les hélices α et jaune pour les feuillets β. Les flèches signalent la zone habituelle d'interaction des domaines PDZ.

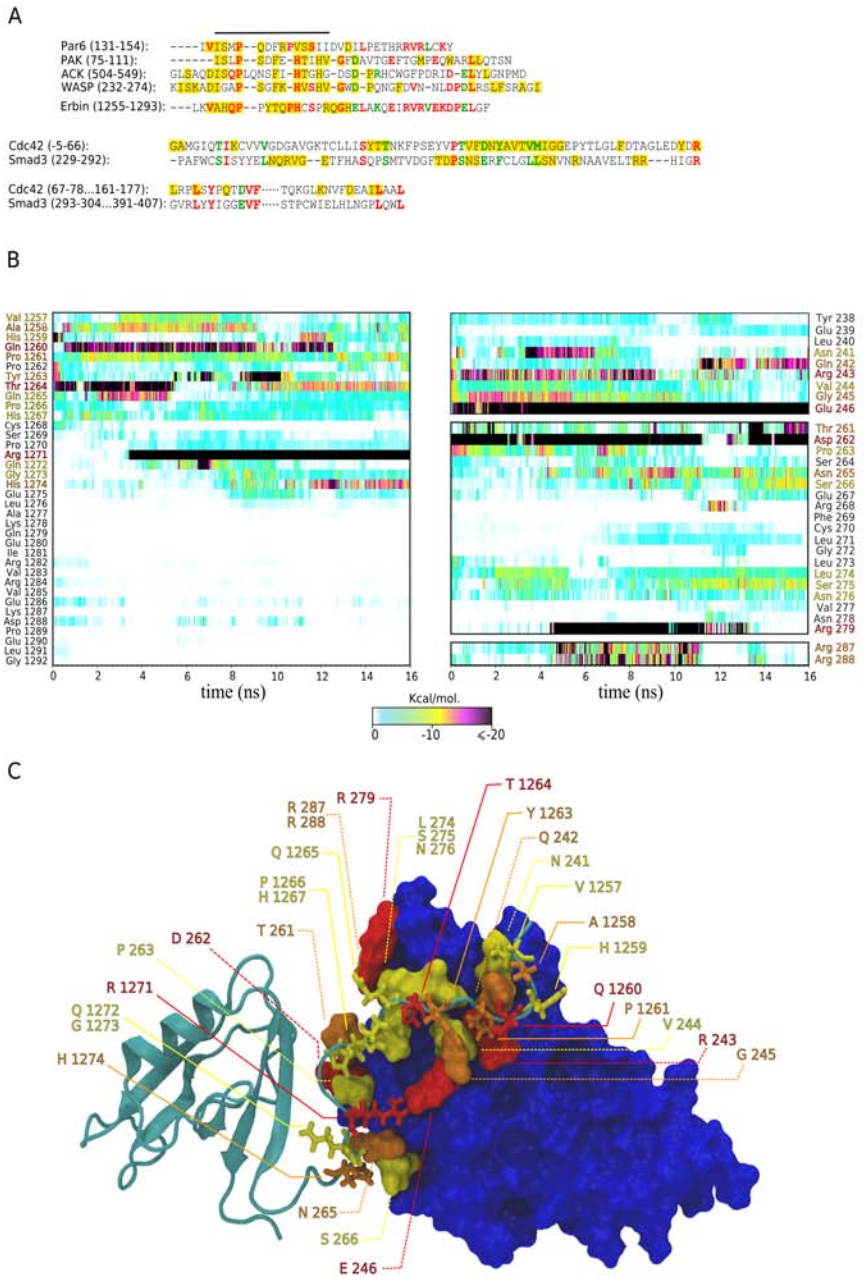


FIG. 2.11 – A- Alignements de la séquence (1257-1280) d'Erbin avec les régions CRIB et semi-CRIB de PAR6, PAK, ACK et WASP. Les résidus impliqués dans ces interactions sont encadrés en jaune : informations extraites de (Garrard *et al.*, 2003) pour les partenaires de Cdc42 et issues de la dynamique moléculaire (voir B) pour Erbin. La ligne noire au dessus de l'alignement indique la région CRIB. Pour l'alignement entre Cdc42 et Smad3, les résidus impliqués dans l'interaction sont encadrés en jaune ; ces résidus ont été mis en évidence par SCOPPI dans le cas de Cdc42 et en fonction des interactions mises en évidence durant la dynamique. Les résidus identiques sont colorés en rouge et les résidus similaires en vert. B- Profils d'énergie d'interaction pour les domaines PDZ et MH2. Les noms des résidus sont colorés en fonction de la moyenne de l'énergie d'interaction. Noir : énergie inférieure à 2Kcal/mol ; jaune : énergie entre 2 et 5 Kcal/mol ; orange : énergie entre 5 et 8 Kcal/mol et rouge : énergie supérieure à 8 Kcal/mol. C- Zoom sur l'interaction entre le peptide (1257-1280) d'Erbin et le domaine MH2. Le domaine PDZ est coloré en cyan et le domaine MH2 en bleu foncé. Les résidus importants pour l'interaction sont présentés en *Licorice* pour Erbin et en utilisant la surface moléculaire pour Smad3. Les résidus sont colorés en fonction de la moyenne de l'énergie d'interaction.

L'analyse des résultats de dynamique moléculaire au niveau de l'interface entre les partenaires révèle l'évolution des deux zones d'interaction. Au niveau de la région de 600 $Å^2$ décrite précédemment, nous avons constaté une majorité d'interactions hydrophobes. Ces interactions se trouvent entre la partie (1257-1280) d'Erbin et le domaine MH2 de Smad3 (voir figure 2.11). Nous avons identifié 3 zones, au niveau du peptide, qui interagissent avec le domaine MH2 : les résidus de 1257 à 1261, de 1263 à 1267 et de 1271 à 1274. Ces résidus sont en interaction avec 3 régions au niveau de MH2 : les régions 241-246, 261-266 (à l'exception de S_{264}) et 274-276 (figure 2.11A et B). On peut noter que les résidus prolines 1261 et 1266 ainsi que l'histidine 1267 interagissent avec le domaine MH2 : P_{1261} interagit avec Q_{242} et V_{244}, P_{1266} et H_{1267} interagissent avec T_{261} et P_{263} et de manière plus faible avec L_{274} et S_{275}.

Nous avons aussi mis en évidence des interactions polaires entre la glutamine 1260 et la glutamine 242 ou l'arginine 243. Ensuite, nous avons identifié des interactions plus fortes entre les résidus chargés au niveau de la petite région de $\approx 200Å^2$ (figures 2.10B et 2.11B). Pour identifier les résidus interagissant, il faut regarder les résidus avec les profils d'énergie similaires : ainsi, nous observons 3 couples de résidus en interaction forte : E_{246} au niveau de Smad3 avec l'arginine R_{1271} au niveau d'Erbin, D_{262} avec K_{1326} et R_{279} avec E_{1321} (figure 2.10B). De plus, la figure 2.11B montre, qu'au niveau de Smad3, les arginines R_{287} et R_{288} ont le même profil d'énergie que l'arginine R_{279} : il y aurait donc un "patch" d'arginines interagissant de concert avec le résidu E_{1321} (voir figure 2.12). Nous pouvons néanmoins noter que les arginines R_{287} et R_{288} interagissent de façon moins prononcée avec E_{1321} que R_{279}. Nous pouvons aussi noter que l'interaction entre le patch d'arginines et l'acide glutamique 1321 n'est pas complètement stable durant tout le temps de la simulation.

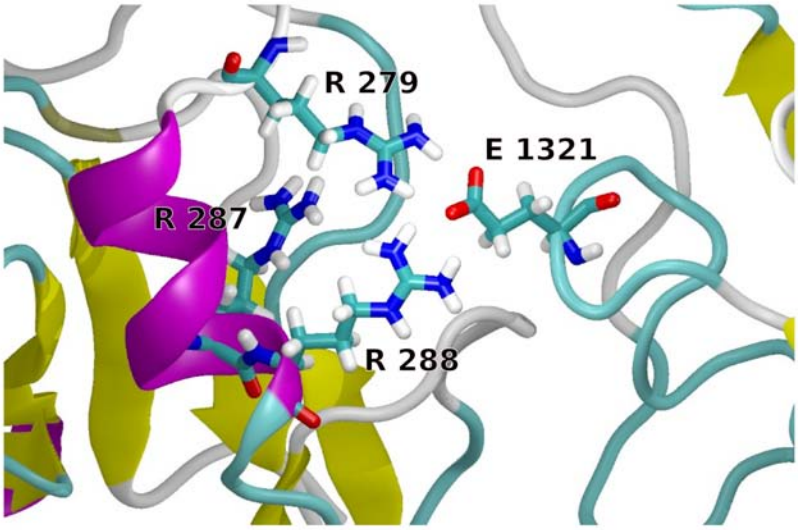


FIG. 2.12 – Zoom sur l'interaction entre E_{1321} et le patch d'arginines : R_{279}, R_{287} et R_{288} à t = 10,2 ns.

Ainsi, grâce à la simulation de dynamique moléculaire, nous avons mis en évidence 2 régions

d'interaction : une région plutôt hydrophobe présente du début à la fin de la dynamique. Cette région recoupe en partie la région d'interaction entre Cdc42 et ses partenaires arborant un motif CRIB, ainsi que la zone SID nouvellement identifiée. La dynamique moléculaire a aussi fait apparaître une nouvelle zone, peu visible dans le modèle avant simulation. Cette zone met en jeu des résidus chargés et elle s'est vraiment mise en place au cours de la dynamique (à t = 5 ns). La mise en place de cette zone d'interaction a impliqué une reconnaissance longue distance entre les résidus chargés ; ce qui a entraîné des changements conformationnels afin de rapprocher le patch d'arginines (et en particulier R_{279}) de l'acide glutamique 1321 (voir figure 2.13). Ces changements conformationnels ont surtout eu lieu au niveau du domaine PDZ. Il est aussi intéressant de remarquer que la zone la plus stable durant la dynamique (entre 5 et 13 ns) est présente quand les résidus R_{279} et E_{1321} sont en contact (distance entre les résidus $\approx$ 4 Å). Une perte minime de cette interaction (pour t $\approx$ 11,5 ns) visible par un léger éloignement de ces résidus semble d'ailleurs entraîner un nouvel état transitoire (pour t compris entre 12,5 et 13 ns). Ces zones d'interaction ont été identifiées à l'aide d'un modèle de complexe créé par homologie. Il faut maintenant valider ce modèle : la région hydrophobe a déjà été mise en évidence par des résultats biologiques antérieurs issus de la bibliographie mais la région chargée n'a jamais été identifiée. C'est pourquoi nous nous sommes focalisés sur cette zone et nous avons réalisé des mutations dirigées afin de valider la totalité du modèle.

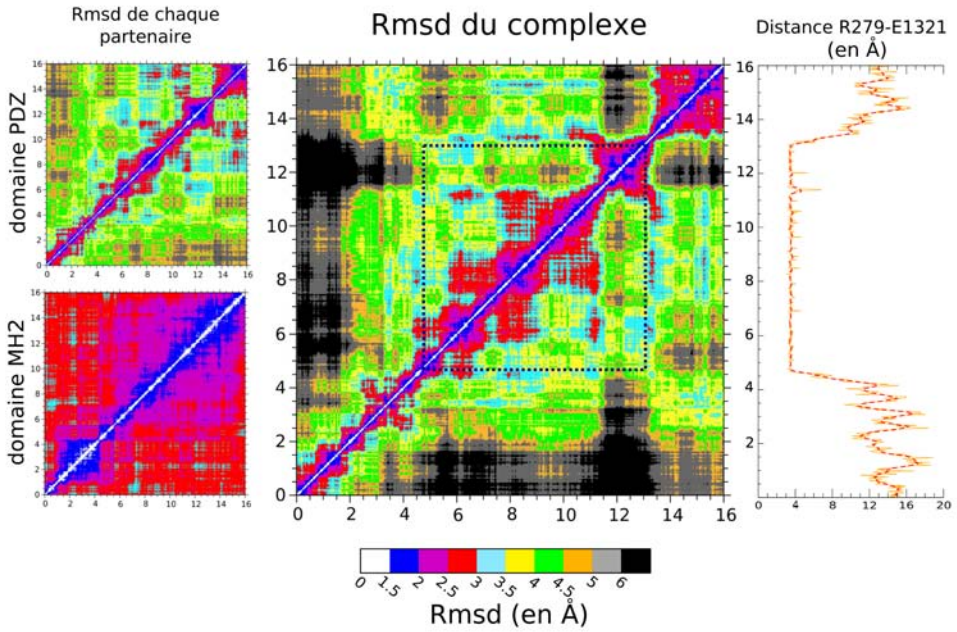


Fig. 2.13 – Évolution du RSMD au cours de la trajectoire de la dynamique pour chaque partenaire et pour le complexe. En abscisse et ordonnée des diagrammes, le temps en nanosecondes. À droite, distance entre les résidus R_{279} et E_{1321} au cours de la dynamique. En orange, les valeurs de distance brutes et en rouge pointillé la moyenne de ces valeurs sur 10 pas de temps.

2.3.4 Validation du modèle par mutations et *charge swap*

E_{246} et R_{279} sont importants pour l'interaction Smad3-Erbin : Pour confirmer notre modèle, nous avons introduit expérimentalement des mutations au niveau du domaine MH2 de Smad3 (mutagénèse dirigée). Ces mutations ciblent plus particulièrement les résidus chargés mis en évidence lors de la dynamique moléculaire. Ainsi, E_{246} a été muté en leucine, D_{262} en alanine et R_{279} en méthionine afin de supprimer la charge des résidus tout en respectant au maximum les contraintes stériques liées à la taille des chaînes latérales. Nous avons utilisé l'acide glutamique 284 comme témoin, nous l'avons muté en alanine et nous avons regardé si cet acide aminé chargé, proche des résidus prédits, avait une action sur l'interaction. Les constructions du domaine MH2 de Smad3 marquées HA furent transfectées dans des cellules COS. L'interaction entre ces différents fragments de Smad3 et Erbin a été étudiée par des expérience de *GST pulldown* en utilisant des constructions GST-Erbin$^{1257\text{-}1371}$. L'assemblage GST-Erbin$^{1257\text{-}1371}$ a précipité les protéines sauvages (sans mutation) et celles ayant la mutation $D_{262}A$ et $E_{284}A$. Ceci prouve que les mutations $D_{262}A$ et $E_{284}A$ n'ont pas eu d'effet sur l'interaction. Par contre, l'assemblage GST-Erbin$^{1257\text{-}1371}$ a moins d'affinité pour le mutant $R_{279}M$ (voir figure 2.14A). Nous avons utilisé l'interaction avec la β-caténine (Ress et Moelling, 2006) comme contrôle de la liaison avec Erbin.

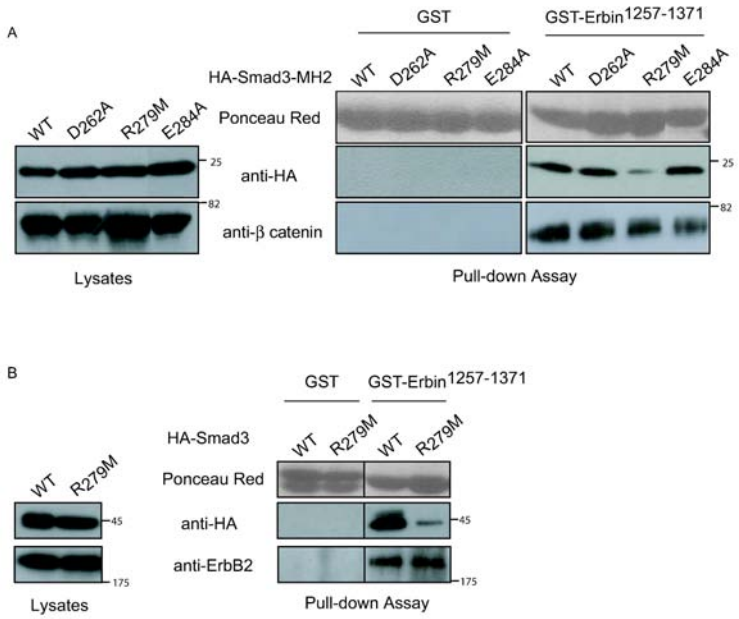


FIG. 2.14 – A- *Pull-down assays* en présence d'assemblages GST-Erbin$^{1257\text{-}1371}$. Les protéines sauvages (*Wild Type* : WT) ou les mutants ($D_{262}A$, $R_{279}M$ et $E_{284}A$) HA-Smad3 MH2 ont été exprimés dans des cellules COS. Les *pull-down* furent analysés en utilisant des anticorps anti-HA et des anticorps ciblés β-caténine comme contrôle. B- Même protocole que précédemment en utilisant, cette fois, des protéines Smad3 complètes. Les poids moléculaires (kD) sont indiqués à côté des panneaux.

De même, une construction GST-MH2 avec la mutation $R_{279}M$ n'a précipité que faiblement la protéine Myc-Erbin$^{1257\text{-}1371}$ en comparaison de la construction GST-MH2 sauvage (voir figure 2.7C). Nous avons aussi obtenu des résultats similaires lorsque nous avons exprimé une

version complète de la protéine Smad3 avec la mutation $R_{279}M$ (HA-Smad3 $R_{279}M$) dans des cellules COS en présence des protéines GST-Erbin$^{1257\text{-}1371}$ (figure 2.14B). Nous avons aussi remplacé l'acide glutamique E_{246} par une leucine au niveau du domaine MH2 de Smad3. Des *pulldown assays* ont montré que la construction GST-Erbin$^{1257\text{-}1371}$ avait une affinité moindre pour la construction HA-Smad3-MH2 avec la mutation $E_{246}L$ (voir figure 2.14C).

Ainsi, les tests *in vitro* confirment la majorité des interactions prévues par le modèle en mettant en évidence l'importance des résidus E_{246} et R_{279} pour l'association PDZ d'Erbin / MH2 de Smad3. De plus, R_{279} a été montré comme un résidu important pour l'hétérodimérisation et donc pour la régulation des protéines Smad (Chacko *et al.*, 2004). C'est pourquoi nous nous sommes ensuite focalisés sur l'interaction R_{279}-E_{1321} prédite par le modèle.

Le résidu E_{1321} du domaine PDZ est impliqué dans l'interaction avec Smad3 : Nous avons donc muté l'acide glutamique 1321 (E_{1321}) au niveau du domaine PDZ d'Erbin. Ce résidu a été changé en Leucine ($E_{1321}L$) au sein d'une construction GST-Erbin$^{1257\text{-}1371}$. Un *pull-down* a été fait en utilisant des lysats de cellules COS exprimant les protéines HA-Smad3 MH2. Nous avons utilisé l'interaction avec le récepteur ErbB2 (Borg *et al.*, 2000) comme contrôle pour vérifier l'intégrité structurale du domaine PDZ d'Erbin. La mutation $E_{1321}L$ a diminué l'interaction avec HA-Smad3-MHé mais pas avec ErbB2 (voir figure 2.15A).

De même, la mutation de la construction GST-Erbin$^{1257\text{-}1371}$ au sein de la région du domaine PDZ impliquée habituellement dans les liaisons entre les domaines PDZ et leurs ligands (notée mutPDZ) a rompu l'interaction avec ErbB2, comme attendu, mais pas avec Smad3 (figure 2.15A). Dans le même but d'évaluer le rôle de la poche classique d'interaction du domaine PDZ, nous avons aussi produit une construction fusionnant la protéine GST et le domaine PDZ seul (GST-Erbin$^{1280\text{-}1371}$).

Cette protéine recombinante a été difficile à exprimer au niveau des bactéries. Nous avons vu que cette construction ne pouvait plus se lier à ErbB2 (figure 2.15A). Nous pouvons expliquer ceci par l'absence du peptide N-terminal (séquence 1257-1279) situé avant le domaine PDZ. Cette absence a pu empêcher la bonne structuration et/ou diminuer la stabilité du domaine PDZ . Malgré cela, la construction GST-Erbin$^{1280\text{-}1371}$ pouvait encore se lier à Smad3 (figure 2.15A) suggérant que le domaine PDZ possédait deux interfaces bien distinctes pour lier Smad3 et ErbB2.

Ainsi, ces résultats renforcent notre modèle puisqu'il semble que l'interaction entre Erbin et Smad3 ne mette pas en jeu la poche classique d'interaction pour un domaine PDZ et son ligand. De plus, les mutations de E_{1321} d'Erbin et de R_{279} de Smad3 peuvent toutes les deux diminuer l'interaction entre le domaine PDZ d'Erbin et le domaine MH2 de Smad3.

Mise en évidence de l'interaction entre E_{1321} et R_{279} : Nous avons ensuite voulu savoir si les résidus E_{1321} et R_{279} impliqués dans l'interaction entre le domaine PDZ et le domaine MH2

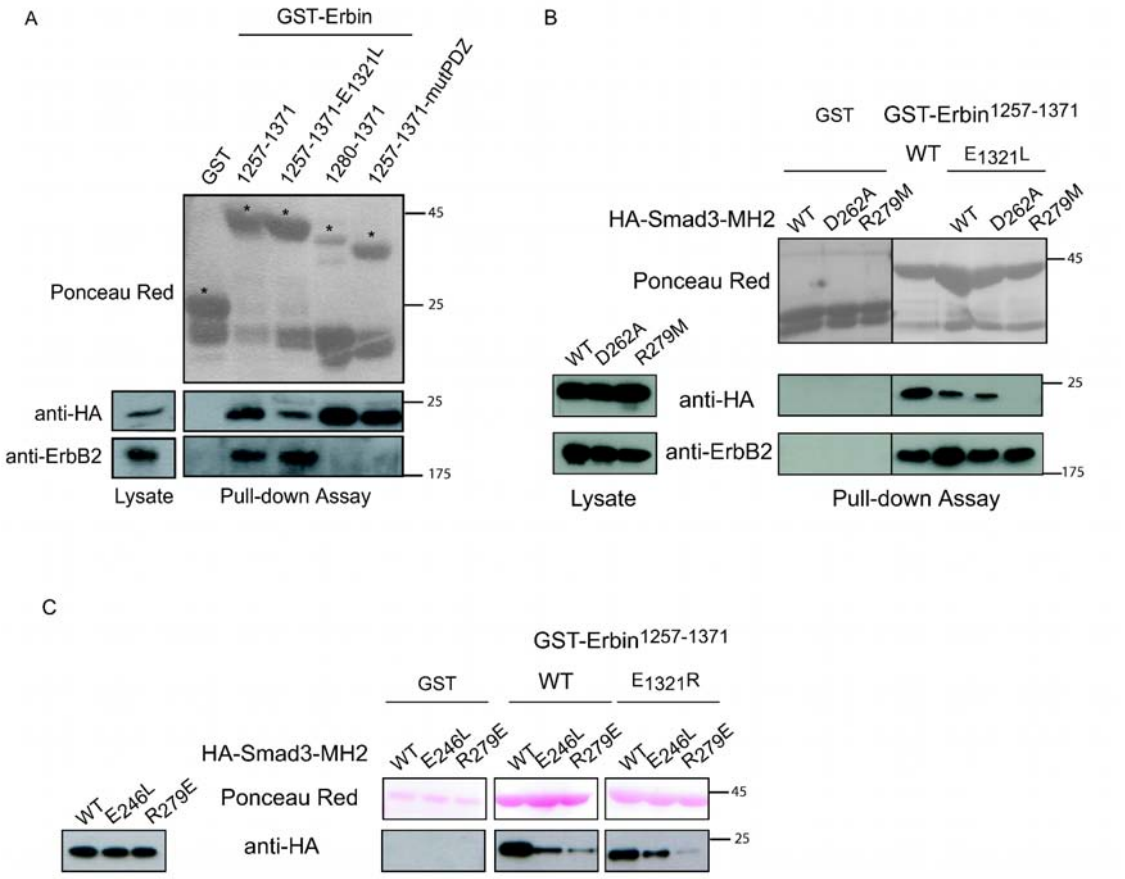


FIG. 2.15 – A- Les cellules COS ont été transfectées avec des constructions HA-Smad3-MH2 et le lysat a été incubé avec différentes version de GST-Erbin$^{1257\text{-}1371}$ (sauvage, muté au niveau de l'acide glutamique : $E_{1321}L$ ou au niveau de la zone d'interaction avec ErbB2 : mutPDZ) ou avec la construction GST-Erbin$^{1280\text{-}1371}$. Les constructions HA-Smad3-MH2 ou les protéines ErbB2 furent révélées grâce aux anticorps appropriés. Les astérisques indiquent la position des protéines GST révélée par Ponceau Rouge. B- Réalisation des mêmes tests en utilisant les mutants de HA-Smad3-MH2 décrits précédemment précipités avec le mutant $E_{1321}L$ de la construction GST-Erbin$^{1257\text{-}1371}$. C- Les constructions GST-Erbin$^{1257\text{-}1371}$ (sauvage - WT - et $E_{1321}R$) furent incubées dans des lysats de cellules COS transfectées avec HA-Smad3-MH2 WT, $E_{246}L$ et $R_{279}E$. Les poids moléculaires (kD) sont indiqués à côté des panneaux.

interagissaient ensemble comme le prévoit le modèle. Pour cela, nous avons isolé la construction Smad3-MH2 $R_{279}M$ avec GST-Erbin$^{1257\text{-}1371}$ $E_{1321}L$. Comme le montre la figure 2.15A, l'interaction est complètement détruite en présence des mutations combinées. Nous pouvons noter que le mutant $D_{262}A$ du domaine MH2 n'a eu aucun effet sur l'interaction comme nous l'avons démontré précédemment (voir figure 2.14A). La liaison entre ErbB2 et le mutant $E_{1321}L$ n'est pas affectée (figure 2.15B). Ces résultats suggèrent donc une interaction directe entre l'arginine 279 de Smad3 et l'acide glutamique 1321 d'Erbin comme l'a prédit le modèle.

Pour confirmer ce résultat nous avons ensuite réalisé un "échange de charges". Ce test s'est déroulé en 2 étapes :

1. Au niveau de Smad3, le résidu arginine chargé positivement a été remplacé par un acide glutamique chargé négativement : E_{279} (HA-Smad3-MH2 $R_{279}E$). Cette protéine a été exprimée avec un marqueur HA dans des cellules COS.
2. Au niveau d'Erbin, le résidu acide glutamique chargé négativement a été remplacé par une arginine chargée positivement : R_{1321}. Ce mutant fut fusionné à la protéine GST (GST-Erbin$^{1257\text{-}1371}$-$E_{1321}R$).

Le but de cette manipulation était de vérifier la spécificité de l'interaction R_{279}-E_{1321}. En effet, l'échange des deux résidus devrait rompre l'interaction entre les mutants $R_{279}E$ ou $E_{1321}R$ avec leurs partenaires sauvages : les résidus de charges identiques à l'interface (E_{279}-E_{1321} ou R_{279}-R_{1321}) devant déstabiliser l'interface. A l'inverse, l'interaction entre les deux mutants devrait être conservée puisque l'on recrée une interaction entre deux résidus de charges opposées (E_{279}-R_{1321}). Nous avons réalisé des *pulldown assays* avec les constructions : GST seules, GST-Erbin$^{1257\text{-}1371}$ et GST-Erbin$^{1257\text{-}1371}$-$E_{1321}R$ en utilisant des extraits cellulaires contenant des constructions HA-Smad3-MH2 et HA-Smad3-MH2 $R_{279}E$ (figure 2.15C). Les résultats obtenus sont les suivants :

– Nous constatons une diminution de l'interaction entre GST-Erbin$^{1257\text{-}1371}$ et HA-Smad3-MH2 $R_{279}E$, confirmant le rôle de ce résidu au niveau de l'interaction. Néanmoins, nous ne voyons pas une disparition totale de l'interaction. Notre modèle peut expliquer ce résultat : en effet, durant l'analyse de la dynamique moléculaire, nous avions remarqué qu'il y avait un patch d'arginine qui interagissait avec l'acide glutamique 1321. Il pourrait donc encore y avoir des interactions entre cet acide glutamique et les deux arginines restantes R_{287} et R_{288} (voir figure 2.12), ce qui ne romprait pas totalement l'interaction.

– GST-Erbin$^{1257\text{-}1371}$-$E_{1321}R$ descend HA-Smad3-MH2 de manière efficace : il n'y a donc pas de perte d'interaction comme on pouvait le supposer au départ. Néanmoins, le modèle peut encore expliquer ce phénomène : le remplacement de l'acide glutamique par une arginine aurait permis de créer un patch d'arginines renforcé à l'interface. En effet, il a été démontré que les résidus arginines pouvaient s'associer grâce à leurs groupes guanidinium de manière stable (Boudon *et al.*, 1990; Soetens *et al.*, 1997). Ce type d'interaction se trouve communément au sein des complexes protéiques (Magalhaes *et al.*, 1994).

– GST-Erbin$^{1257\text{-}1371}$-$E_{1321}R$ montre une forte diminution de son affinité pour HA-Smad3-MH2 $R_{279}E$. Ceci va à l'encontre de notre hypothèse de départ : une récupération de l'interaction entre les domaines MH3 de Smad3 et PDZ d'Erbin si l'on échange les deux charges. Néanmoins, si l'on regarde l'interaction du patch d'arginine avec l'acide glutamique en figure 2.12, nous pouvons penser que la mutation de l'arginine en acide glutamique peut créer une déstabilisation du patch d'arginines formé précédemment entre GST-Erbin$^{1257\text{-}1371}$-$E_{1321}R$ descend HA-Smad3-MH2. De plus, le résidu E_{279} pourrait créer des liaisons intra-moléculaires avec les arginines R_{287} et R_{288}. Ces interactions intra-moléculaires pourraient être favorisées par rapport à l'interaction inter-moléculaire attendue entre E_{279} et R_{1321}.

Ainsi, même si la spécificité de l'interaction entre R_{279} et E_{1321} ne peut être complètement prouvée, le modèle de complexe présenté peut expliquer tous les résultats biologiques obtenus. Il apparaît donc que le domaine PDZ d'Erbin interagit avec le domaine MH2 de Smad3 par

l'intermédiaire d'une interface indépendante de la poche d'interaction classique des domaines PDZ. Les résidus R_{279} et E_{1321} semblent impliqués dans cette interaction.

2.3.5 Discussions sur la validité du modèle et le rôle d'Erbin dans la voie du TGF-β

Nous avons ainsi identifié des segments capables de se fixer à Smad3 dans la partie C-terminale d'Erbin : 2 régions ont été identifiées au niveau d'Erbin, les régions 853-1257 et 1257-1371 (figure 2.6D). Ces données recoupent en partie les informations obtenues dans les deux derniers travaux publiés sur ce sujet. Ainsi, Warner *et al.* ont identifié la région 1004-1280 comme importante pour l'interaction tandis que Dai *et al.* ont mis en évidence la région 1172-1282 (Warner *et al.*, 2003; Dai *et al.*, 2007). En réunissant toutes ces données, y compris les nôtres, nous concluons que la région 1172-1257 est impliquée dans l'interaction avec Smad3.
Warner *et al.* ont démontré que le domaine PDZ seul n'était pas suffisant pour se lier à Smad3 grâce à des tests double hybride chez la levure. Cependant, la délétion de ce domaine diminue l'interaction, suggérant la contribution de celui-ci (Warner *et al.*, 2003), ce qui concorde avec notre étude : à savoir que le domaine PDZ seul (région 1280-1371) peut aussi interagir avec Smad3. Cette interaction directe n'a jamais été testée dans aucune autre étude (Dai *et al.*, 2007). En conclusion, nous définissons les régions 1172-1257 et 1257-1371 d'Erbin comme interagissant directement avec le domaine MH2 de Smad3.

Nous nous sommes ensuite intéressés plus précisément à l'interaction entre la région 1257-1371 d'Erbin contenant le domaine PDZ, et le domaine MH2 de Smad3. Pour cela, nous nous sommes servis des données cristallographiques des domaines MH2 et PDZ et de l'analogie d'interaction avec le complexe PAR6/Cdc42. Bien que le taux d'identité entre les partenaires de chaque complexe ne soit pas élevé (en particulier entre Cdc42 et le domaine MH2 de Smad3 qui est de 11%), des zones conservées ainsi que des similarités de forme ont pu être identifiées au niveau de l'interface (voir figure 2.11 et figure 2.9). Une approche similaire à la notre a d'ailleurs été publiée recemment (Günther *et al.*, 2007).

Notre approche bioinformatique a donc permis de mieux comprendre cette interaction. Nous avons prédit que la poche de fixation classique des domaines PDZ n'était pas impliquée dans cette interaction et qu'elle se situait à l'opposé de la zone de d'ancrage de ErbB2 et de la β-caténine. En accord avec cette prédiction, les mutations abrogeant l'interaction entre le domaine PDZ et la protéine ErbB2 n'ont eu aucun effet sur l'interaction entre les domaines PDZ et MH2 (voir figure 2.15A). Nous avons aussi prédit que les résidus 1257-1280 se trouvant en amont du domaine PDZ étaient importants pour l'interaction (voir 2.10). Des travaux menés parallèlement aux nôtres ont montré l'importance de tels résidus (Dai *et al.*, 2007).

De plus, la mutation de l'arginine 279 au niveau du domaine MH2 a entraîné une diminution de l'interaction avec Erbin (voir figure 2.14). Cette arginine est située au niveau de la boucle L2 de Smad3 qui est impliquée dans l'hétérodimérisation et la régulation de la famille des Smads (Chacko *et al.*, 2004; Prokova *et al.*, 2007). Dai *et al.* ont montré que, dans le groupes des R-

smad, seules Smad2 et Smad3 pouvaient interagir avec Erbin (Dai *et al.*, 2007). L'alignement du domaine MH2 des différentes protéines Smads a révélé que l'arginine 279 était conservée dans la famille Smad mais que des divergences apparaissaient dans la région entourant ce résidu. Ceci expliquerait en partie la spécificité d'interaction de Smad2 et Smad3 par rapport à Erbin. Des résidus de Smad3 ont été remplacés dans Smad4 par des acides aminés avec des propriétés différentes. C'est le cas de l'asparagine 278 au niveau de Smad3 remplacée par une histidine ou la leucine 273 par une glutamine (voir figure 2.16A). Notre modèle a enfin mis en évidence les résidus R_{297} chez Smad3 et E_{1321} chez Erbin comme étant importants pour l'interaction. Des mutations réalisées sur ces résidus ont provoqué une diminution de l'interaction entre ces protéines et les mutations combinées ont complètement rompu l'interaction (voir figure 2.15B). L'utilisation de tests biochimiques nous a permis de mettre en évidence une interaction directe entre différentes zones d'Erbin et le domaine MH2 de Smad3. Cette liaison est spécifique car ni Densin, ni Scrib ou Lano, 3 homologues d'Erbin, ne coimmunoprécipitent avec Smad3 (voir figure 2.6A). L'absence d'interaction avec Lano peut s'expliquer par l'homologie limitée au niveau de la partie C-terminale de cette protéine par rapport à Erbin : en effet, Lano ne possède pas de domaine PDZ. Pour le cas de Scrib et Densin, des différences au sein du peptide situé avant le domaine PDZ ainsi que dans la région entourant le résidu 1321 peuvent expliquer ce manque d'affinité (voir figure 2.16B).

Contrairement au site de fixation classique des domaines PDZ où les contributions majeures se font via des résidus hydrophobes (Basdevant *et al.*, 2006), notre modèle met en évidence des interactions chargées au niveau de l'interface des domaines MH2 de Smad3 et PDZ d'Erbin. C'est pourquoi les résidus R_{279} et E_{1321} mutés en des résidus hydrophobes (respectivement par une méthionine et une leucine) diminuent de façon drastique cette interaction. Les profils d'énergie créés à partir de la dynamique moléculaire (présentés figure 2.10 et 2.11B) montrent que E_{1321} interagit avec le patch d'arginines formé par les résidus R_{279}, R_{287} et R_{288} de manière très similaire au résidu D_{493} au niveau de Smad4 (Chacko *et al.*, 2004). Une étude récemment publiée a montré que la mutation de R_{287} au niveau de Smad3 inhibe la voie de signalisation du TGF-β (Prokova *et al.*, 2007).

Nous postulons que l'engagement d'Erbin avec Smad3 entrerait en compétition avec l'interaction Smad3-Smad4 et empêcherait l'hétéromérisation de Smad3 et Smad4, ce qui inhiberait l'activation de la voie du TGF-β. Cette hypothèse est en accord avec des résultats récents montrant que la surexpression d'Erbin inhibe la translocation de Smad3 au niveau du noyau cellulaire mais aussi son activité transcriptionnelle en diminuant la dimérisation de Smad3 et Smad4 (Dai *et al.*, 2007). L'interaction entre le domaine PDZ d'Erbin et le domaine MH2 de Smad3 laisse libre la région classique de fixation des domaines PDZ. Il n'existe que très peu d'exemples de telles associations, nous citons l'interaction PAR6/Cdc42 mais aussi le complexe nNOS-syntrophine (Hillier *et al.*, 1999).

Ainsi, la combinaison de résultats théoriques et expérimentaux ont fait ressortir un nouveau type d'interaction pour le domaine PDZ d'Erbin. Celle-ci se fait via des résidus chargés mais aussi hydrophobes. Elle est d'autant plus intéressante qu'elle met en évidence une capacité jus-

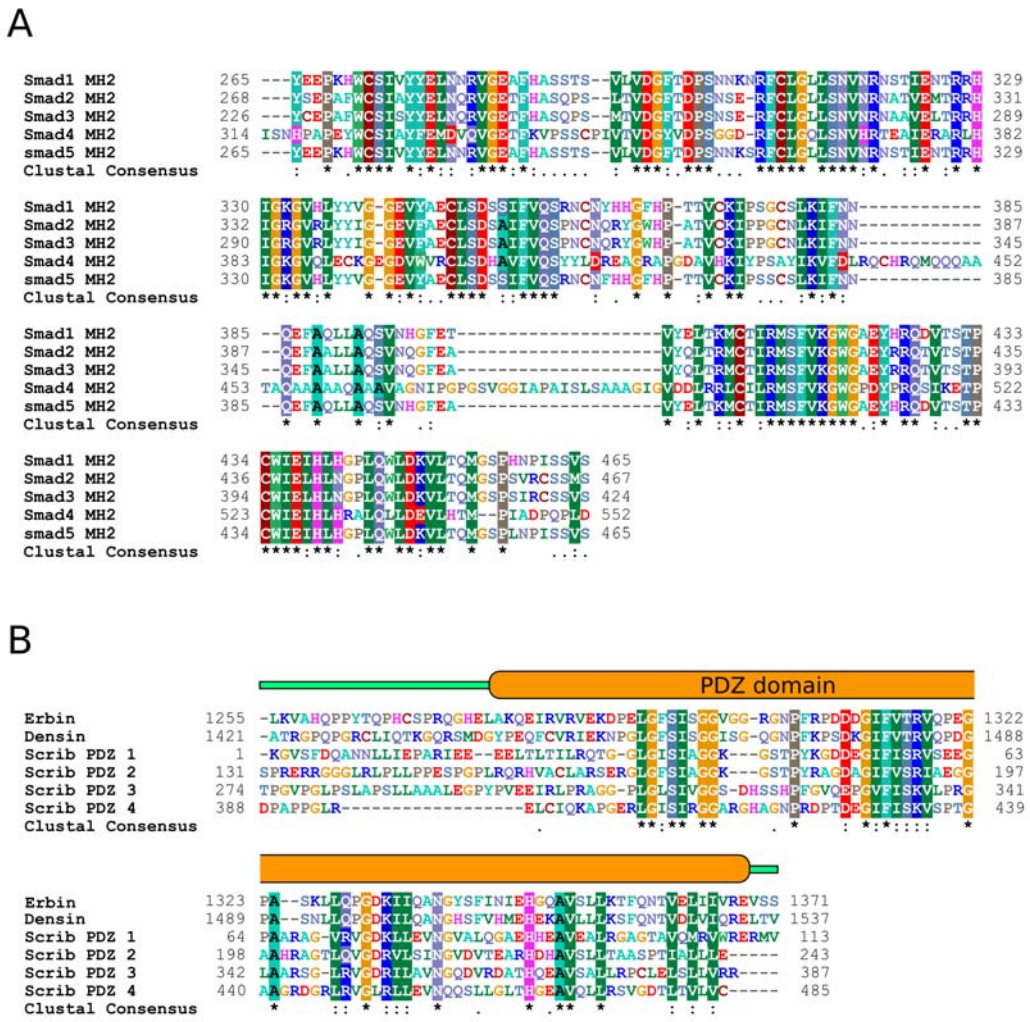


FIG. 2.16 – A- Alignements des séquences de Smad R et communes disponibles. B- alignements des séquences d'Erbin, Densin et Scrib autour du domaine PDZ. Les résidus identiques ou similaires sont encadrés en couleur. La dernière ligne représente l'analyse de ClustalW. [*] indique les résidus identiques, [:] les résidus très similaires et [.] les résidus similaires.

qu'alors insoupçonnée du domaine PDZ d'Erbin qui est de réguler la voie du TGF-β en limitant l'association Smad3/smad4. Erbin régule également la voie des MAPK en aval de nombreuses tyrosines kinases comme ErbB2. Erbin se trouve donc au carrefour de deux voies de signalisation cruciales pour le développement celulaire. La double capacité d'Erbin d'interagir physiquement et fonctionnellement avec les acteurs principaux de ces différentes voies de signalisation montre l'importance de cette protéine dans la transduction du signal.

2.4 La dynamique moléculaire pour affiner les résultats consensus de docking rigide

Nous avons montré dans la partie précédente comment la dynamique moléculaire en solvant explicite (DMSE) a permis de déterminer des résidus clés intervenant dans la formation du complexe Erbin PDZ/ Smad3 MH2. Ce modèle d'assemblage fut créé par analogie avec le complexe PAR6/cdc42. C'est pourquoi, dans la suite de ce chapitre, nous le nommons modèle par analogie. Dans l'étude précédente, il restait un problème majeur inhérent à ce modèle : le taux d'identité entre la protéine Cdc42 et le domaine MH2 de Smad3 était très bas (≈11%). De ce fait, bien que les résidus clés prédits par le modèle semblent jouer un rôle dans l'interaction, il était concevable de se poser la question : ce modèle par analogie est-il représentatif de l'assemblage réel ?

Pour répondre à cette problématique, nous avons décidé d'utiliser des programmes de docking afin de valider notre modèle. En effet, les résultats présentés lors du challenge CAPRI montrent la capacité de tels programmes à trouver, dans certains cas, une solution convenable pour l'assemblage protéique. Ainsi, si nous trouvons des solutions de docking suffisamment proches de notre résultat, nous admettrons que celui-ci constitue un modèle plausible d'assemblage pour le complexe PDZ Erbin/Smad3 MH2.

Nous avons décidé de développer un protocole que l'on peut qualifier de "consensus" : à la manière des serveurs de recherche de structures secondaires à partir de la séquence, nous avons comparé les résultats de divers serveurs de docking pour ne garder que les solutions communes. Des démarches similaires ont déjà été employées mais celles-ci se limitaient à chaque fois à un programme de docking spécifique (Fernández-Recio *et al.*, 2004; Sacquin-Mora *et al.*, 2008). Cette méthode a aussi un intérêt conceptuel puisqu'elle nous permettra de comparer les résultats des divers serveurs sur le complexe Erbin PDZ/ Smad3 MH2. Si nous nous basons sur les critères de Vajda et Camacho, ce complexe serait de type III. En effet, il possède une aire d'interface standard (entre 1400 $Å^2$ et 2000 $Å^2$) et on trouve à l'interface de ce complexe de fortes interactions électrostatiques. Ces critères font de ce système un assemblage difficile à prédire (Vajda et Camacho, 2004). Conscients de cette difficulté, nous avons ensuite réalisé des simulations de DMSE pour tester la stabilité des modèles. Nous avons ensuite comparé les trajectoires de dynamique de ces modèles avec celle du modèle par analogie afin de vérifier s'il existait une convergence de tous les modèles (voir figure 2.17 pour une vision globale de la stratégie employée). Enfin, les données de mutagénèse ont été utilisées pour guider le docking. Nous avons ainsi pu comparer les résultats des programmes de docking avec et sans ajout d'informations.

2.4.1 Choix des serveurs de docking

Comme nous avons pu le voir dans la première partie, les serveurs de docking permettent d'obtenir de bons, voire de très bons, résultats pour la prédiction de complexes protéiques. Pour notre expérience, il fallait que les programmes de docking puissent produire des résultats avec et sans ajout d'informations. De ce fait, le serveur de docking HADDOCK, bien que donnant de très bonnes prédictions à CAPRI, a été éliminé de nos choix puisque celui-ci nécessite obli-

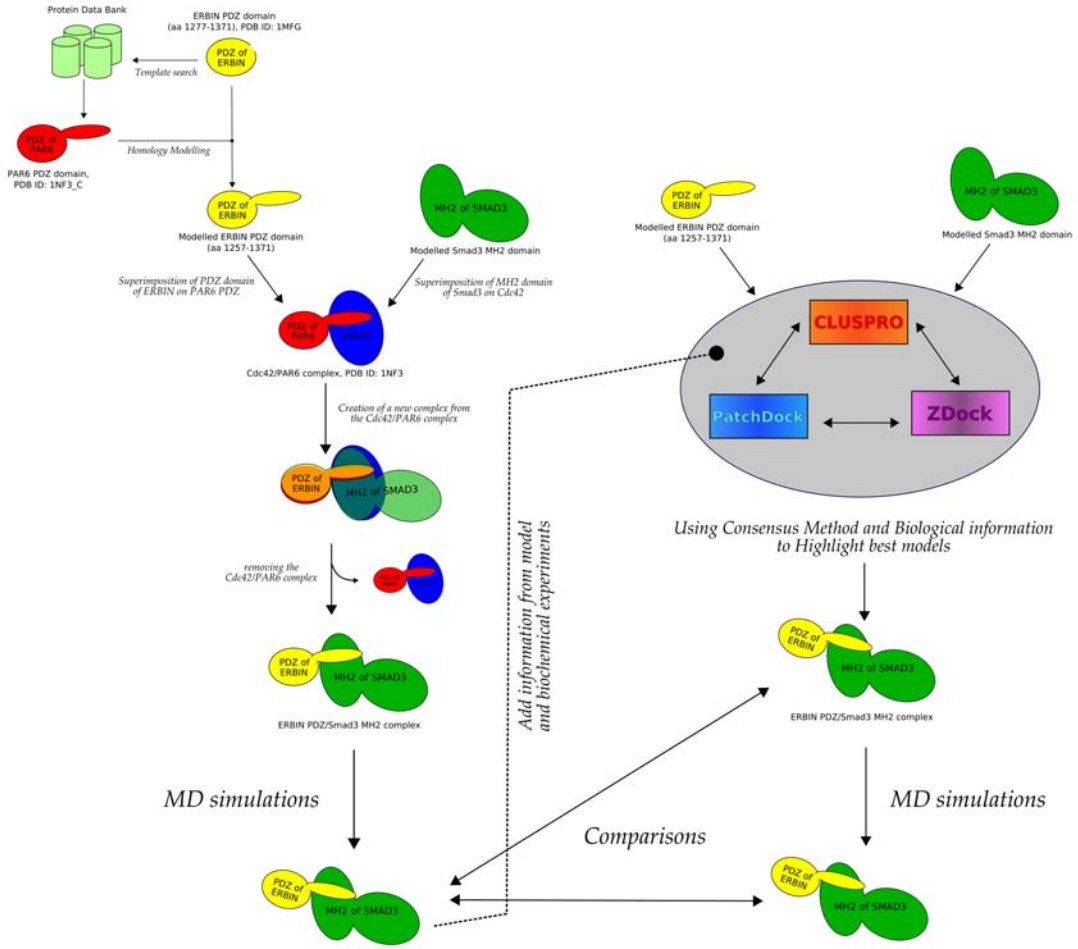


FIG. 2.17 – A gauche, stratégie adoptée pour la création du complexe Erbin PDZ / Smad3 MH2 à partir du complexe PAR6/Cdc42. A droite, présentation de la stratégie "consensus" : utilisation des structures modélisées du domaine MH2 et PDZ comme structures d'entrée pour les serveurs PatchDock, CLUSPRO et Zdock. Les résultats consensus furent ensuite comparés à la trajectoire de la dynamique du modèle obtenu à partir du complexe PAR6/Cdc42 puis ont été relaxés par dynamique moléculaire en solvant explicite. Les pas de chaque dynamique ont été ensuite comparés.

gatoirement des informations de départ pour modéliser un complexe (van Dijk *et al.*, 2005b). Une amélioration de ce programme existe : elle permet de modéliser un complexe sans ajout d'informations préalables (de Vries *et al.*, 2007). Malheureusement ce programme n'est pas encore disponible sur Internet. Nous avons également éliminé les serveur SKE-dock et SmoothDock pour les mêmes raisons. Le serveur GRAMM-X, sélectionné au départ car il possédait toutes les caractéristiques pour faire partie de cette expérience (ajout ou non d'informations), a été exclu après de multiples essais sans résultats (le programme ayant crashé plusieurs fois ou n'ayant jamais donné de réponse). Outre les serveurs de docking dédiés, il existe des serveurs de docking

dérivant de programmes dit *stand alone* (que l'on peut télécharger et utiliser directement sur un ordinateur personnel), Zdock *server* est l'un d'entre eux. Ce programme a donné de très bon résultats lors des derniers cycles de CAPRI (Wiehe *et al.*, 2005, 2007) ; nous avons donc testé le serveur automatique de prédictions dérivant de celui-ci.

Au final, nous avons sélectionné les trois serveurs de docking suivants : CLUSPRO (utilisé avec le programme DOT), PatchDock et Zdock-*server*. Nous avons aussi utilisé les fonctions de score de Patchdock (FireDock [Andrusier *et al.* (2007)]) et Zdock (Rdock-Zrank [Li *et al.* (2003b); Wiehe *et al.* (2007)]) pour reclasser les résultats de ceux-ci. Nous nous sommes servis des paramètres par défaut de chaque serveur afin de se placer dans les conditions les plus neutres possibles. Nous avons choisi le plus petit nombre maximal de solutions disponibles pour avoir un jeu de données homogène à comparer. Ce nombre a été défini comme le nombre maximal de solutions proposées par CLUSPRO qui est de 30. Nous avons, enfin, conduit deux cycles de prédictions avec et sans informations biologiques. Dans le cas d'ajout d'informations, nous avons précisé que E_{1321} au niveau d'Erbin et R_{279} au niveau de Smad3 étaient en interaction.

2.4.2 Comparaison des résultats des serveurs de docking

Évaluation des solutions : Les résultats de chaque serveur ont été comparés entre eux. Pour cela, nous avons choisi des critères équivalents à ceux utilisés pour classer les prédictions lors de l'expérience CAPRI, à savoir, le RSMD du ligand, L_{rmsd} (voir la partie *évaluation des résultats*, chapitre précédent, pour la définition du RMSD du ligand) et la fraction d'interfaces natives (f_{Inat}). Ce dernier paramètre diffère de la fraction de contacts natifs. La fraction d'interface se définit comme le nombre de résidus présents à la fois à l'interface du modèle par analogie et à l'interface d'un modèle prédit par un serveur, divisé par le nombre total de résidus présents à l'interface du modèle par analogie :

$$f_{Inat} = \frac{\text{Rés. analogie} \cap \text{Rés. prédit}}{\text{Rés. analogie total}}$$

Où Rés. analogie : résidus présents à l'interface du modèle par analogie et Rés. prédit : résidus présents à l'interface du modèle prédit par un serveur. Les résidus présents à l'interface des complexes ont été mis en évidence en utilisant le serveur PROTORP (Reynolds *et al.*, 2008)[17].

Nous avons, dans un premier temps, comparé le RMSD sur les carbones α des ligands des résultats de chaque serveur après avoir superposé les récepteurs. Ici, le ligand correspond au domaine PDZ d'Erbin et le récepteur au domaine MH2 de smad3.

Analyse des résultats des divers serveurs de docking L'analyse des résultats présentés figure 2.18, fait apparaître des solutions proches d'un programme de docking à un autre. Au niveau des résultats de docking sans ajout d'informations, il y a 25 solutions que nous nommerons "consensus". Une solution consensus est une solution d'un serveur de docking dont la valeur L_{rmsd}

[17] http ://www.bioinformatics.sussex.ac.uk/protorp/

est inférieure à 10 Å par rapport à une solution d'un autre serveur de docking. La valeur de 10 Å a été choisie car il s'agit de la limite de RMSD pour le challenge CAPRI (Méndez *et al.*, 2003).

Ces solutions consensus sont réparties aléatoirement à travers l'ensemble des résultats de chaque serveur (voir figure 2.18). Lorsque l'on ajoute l'information suivante, E_{1321} en contact avec R_{279}, on remarque que les solutions consensus se retrouvent exclusivement dans les cinq premiers résultats de chaque serveur de docking à l'exception d'une solution dans la comparaison des résultats entre Zdock et Patchdock. Par contre, le nombre de solutions consensus a légèrement diminué, passant à 22 (voir tableau 2.1).

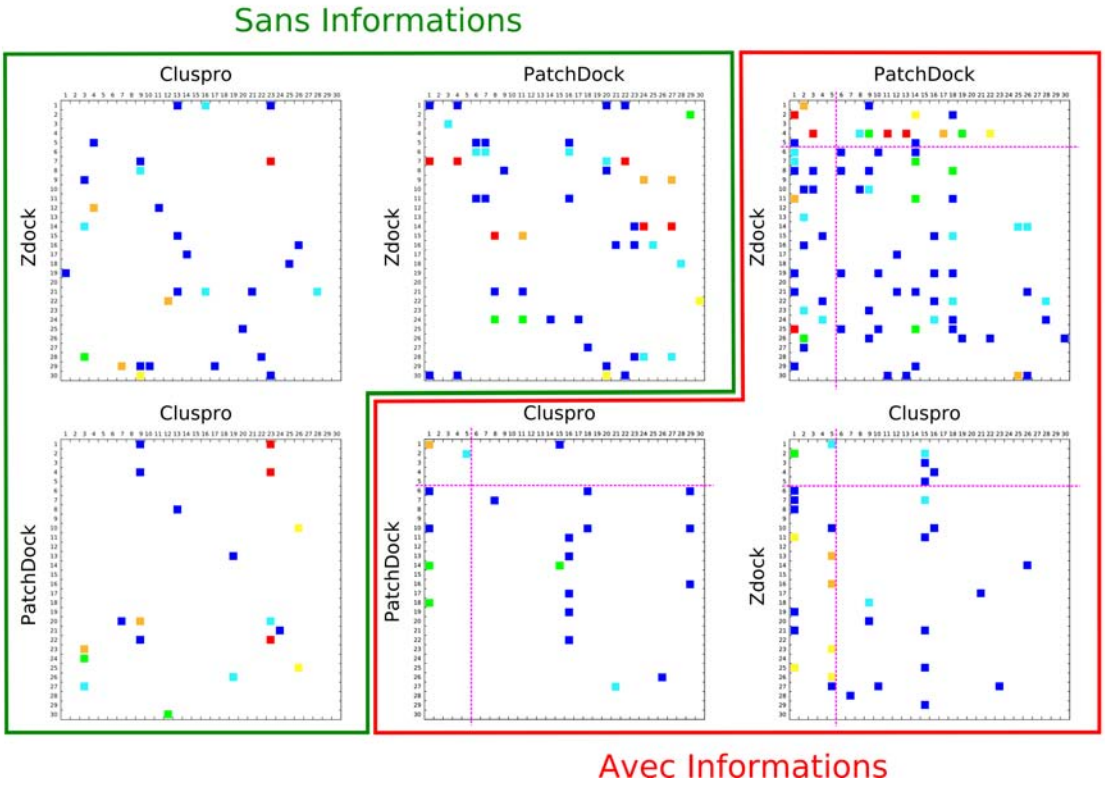


FIG. 2.18 – Graphiques montrant les résultats de chaque serveur de docking comparés à ceux des autres serveurs. Les couleurs correspondent aux RMSD des ligands : en rouge, solutions ayant un RMSD inférieur à 5 Å ; en orange, RMSD compris entre 5 et 7,5 Å ; en jaune, RMSD compris entre 7,5 et 10 Å ; en vert, RMSD compris entre 10 et 12,5 Å ; en cyan, RMSD compris entre 12,5 et 15 Å ; en bleu, RSMD compris entre 15 et 20 Ået en blanc, RMSD supérieur à 20 Å.

La répartition des résultats autour du récepteur et du ligand évoluent différemment en fonction de l'ajout ou non d'information (voir figures 2.19 et 2.20). Dans le cas d'un docking sans information, les solutions sont réparties tout autour du récepteur ou du ligand. Seul Patchdock donne un résultat assez ciblé pour le ligand, résultat d'autant plus intéressant qu'il met en évidence une région, au niveau du domaine PDZ, qui recoupe la zone identifiée comme importante par le modèle par analogie (zone face au résidu E_{1321}, comparer les figures 2.19 et 2.10B).

Nous avons ensuite regardé la répartition des solutions de docking avec ajout d'information (voir figure 2.20). On peut noter, en comparant les figures 2.19 et 2.20 que, pour ces solutions, la distance entre les résidus E_{1321} et R_{279} a diminué, preuve que l'ajout de contrainte, pour les serveurs Zdock et PatchDock a bien été prise en compte. En effet, pour ces serveurs, on trouve des solutions dont la distance entre les résidus E_{1321} et R_{279} est inférieure à 7,5 voir 5 Å, plaçant ceux-ci dans des positions où des interactions électrostatiques sont possibles. Le serveur CLUSPRO donne des résultats un peu moins bons que les deux autres serveurs lorsque l'on ajoute des contraintes. De plus, les solutions de ce serveur sont très éloignées les unes des autres : ceci s'explique par le protocole utilisé qui fixe des groupements de structure (aussi appelé *clusters*) éloignés de 9 Å de RMSD (Comeau *et al.*, 2004b). Cette distance, fixée à 9 Å par défaut, peut être néanmoins modifiée sur la page d'accueil du serveur.

En conclusion, des solutions consensus ont été trouvées entre les différents serveurs. Ces solutions ont été affinées en utilisant des contraintes comme des résidus potentiellement à l'interface. Ces dernières sont très importantes puisqu'elles permettent de limiter les recherches, dans notre cas, aux 5 meilleures solutions de chaque serveur.

2.4.3 Comparaison des résultats des serveurs à la dynamique moléculaire du modèle par analogie

Nous avons ensuite voulu savoir si, parmi les solutions données par les serveurs, certaines se rapprochaient de notre modèle par analogie. Celui-ci ayant subi d'importants changements conformationnels durant la dynamique de 16 ns (voir figure 2.13) nous avons donc décidé de comparer les solutions à l'ensemble de la trajectoire de la dynamique moléculaire et non pas seulement au modèle de départ créé à partir du complexe Cdc42/PAR6. Pour cette étude *a posteriori*, nous avons calculé le RMSD du ligand de chaque pas de la trajectoire de la dynamique de 16 ns avec l'ensemble des solutions des serveurs de docking.

Nous définissons, ici, le terme de solution isolée comme une solution d'un serveur de docking qui a une structure proche d'un pas de la dynamique moléculaire mais qui n'est pas considérée comme une solution consensus : c'est à dire dont la structure à un $L_{rmsd} > 10$ Å par rapport à toutes les solutions des autres serveurs de docking. Il s'agit donc d'une solution spécifique à un serveur de docking donné.

Les résultats présentés dans le tableau 2.1 montrent que, sans ajout d'information, l'ensemble des solutions ayant une structure proche d'un pas de la dynamique du modèle par analogie (c'est à dire ayant un $L_{rmsd} < 15$ Å) se retrouvent dans les solutions consensus. Lorsque l'on ajoute des contraintes, on voit apparaître des solutions isolées, ces solutions sont peut-être dues à la manière d'ajouter des contraintes spécifiques à chaque serveur. Néanmoins, nous retrouvons encore une fois des solutions proches d'un pas de la dynamique moléculaire, preuve qu'en partant d'un modèle de docking fait par analogie et qu'en relaxant celui-ci par une dynamique moléculaire,

on peut retrouver des structures proches de celles obtenues avec des programmes de docking.

TAB. 2.1 – Solutions consensus et isolées ayant une structure proche d'un pas de la dynamique moléculaire.

Serveurs	Nb. consensus	Consensus $15 \geq L_{rmsd} > 10$	Consensus $L_{rmsd} \leq 10$	Isolée $15 \geq L_{rmsd} > 10$	Isolée $L_{rmsd} \leq 10$
Sans Information					
Zdock	8	2	0	0	0
PatchDock	10	0	3	0	0
CLUSPRO	7	0	1	0	0
Total	25	2	4	0	0
Avec Information					
Zdock	10	4	1	1	1
PatchDock	10	0	1	0	0
CLUSPRO	2	1	0	1	0
Total	22	5	2	2	1

Nous avons ensuite été plus loin en analysant les solutions de docking ayant un $L_{rmsd} < 10$ Å d'un pas de la dynamique moléculaire. Ces résultats sont présentés tableau 2.2.

TAB. 2.2 – Solutions consensus et isolées ayant une structure proche d'un pas de la dynamique moléculaire (pour un L_{rmsd} inférieur à 10 Å).

Serveurs	n° sol.	Consensus	Isolée	L_{rmsd}	Temps (ns)	f_{Inat}
Sans Information						
PatchDock	23	X	-	9.3	13.1	85.3
	24	X	-	9.4	11.8	80
	27	X	-	10	11.8	74.2
CLUSPRO	3	X	-	8.8	13.1	82.3
Avec Information						
Zdock	1	X	-	9.1	11.8	72.7
	27	-	X	7.5	2.4	75
PatchDock	2	X	-	9.6	11.8	71.4

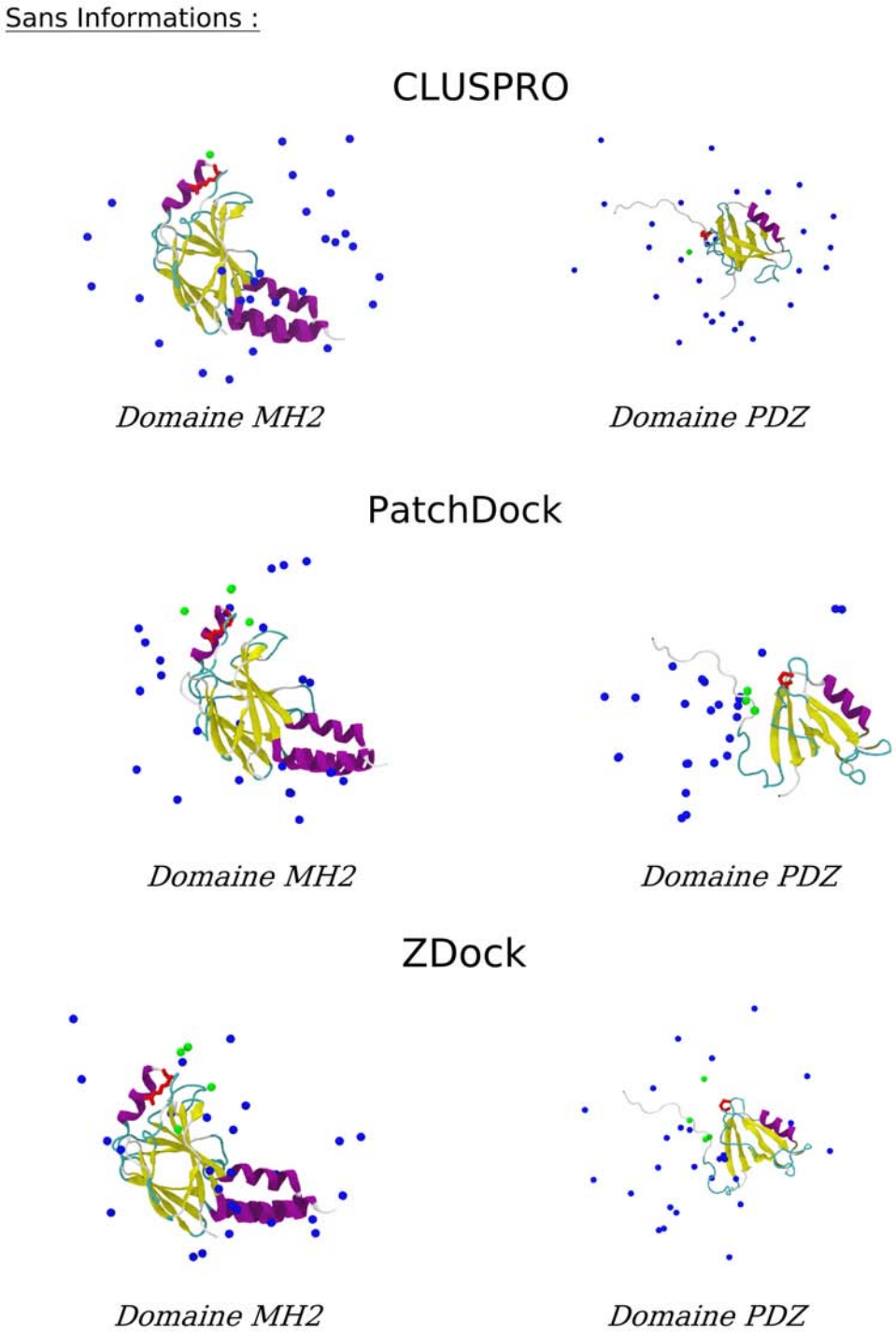


FIG. 2.19 – Graphique montrant la répartition du ligand par rapport au récepteur et inversement dans le cas où aucune information n'est donné aux serveurs. Les sphères représentent les positions du résidu R_{279} ou E_{1321} pour chaque solution. Les sphères sont colorées en fonction de la distance R_{279}-E_{1321} : bleu pour une distance supérieure à 15 Å ; vert : pour une distance comprise entre 10 et 15 Å ; jaune : pour une distance comprise entre 7,5 et 10 Å ; orange : pour une distance comprise entre 5 et 7,5 Å et rouge : pour une distance inférieure à 5 Å. En rouge, le résidu R_{279} au niveau de Smad3 et E_{1321} au niveau d'Erbin.

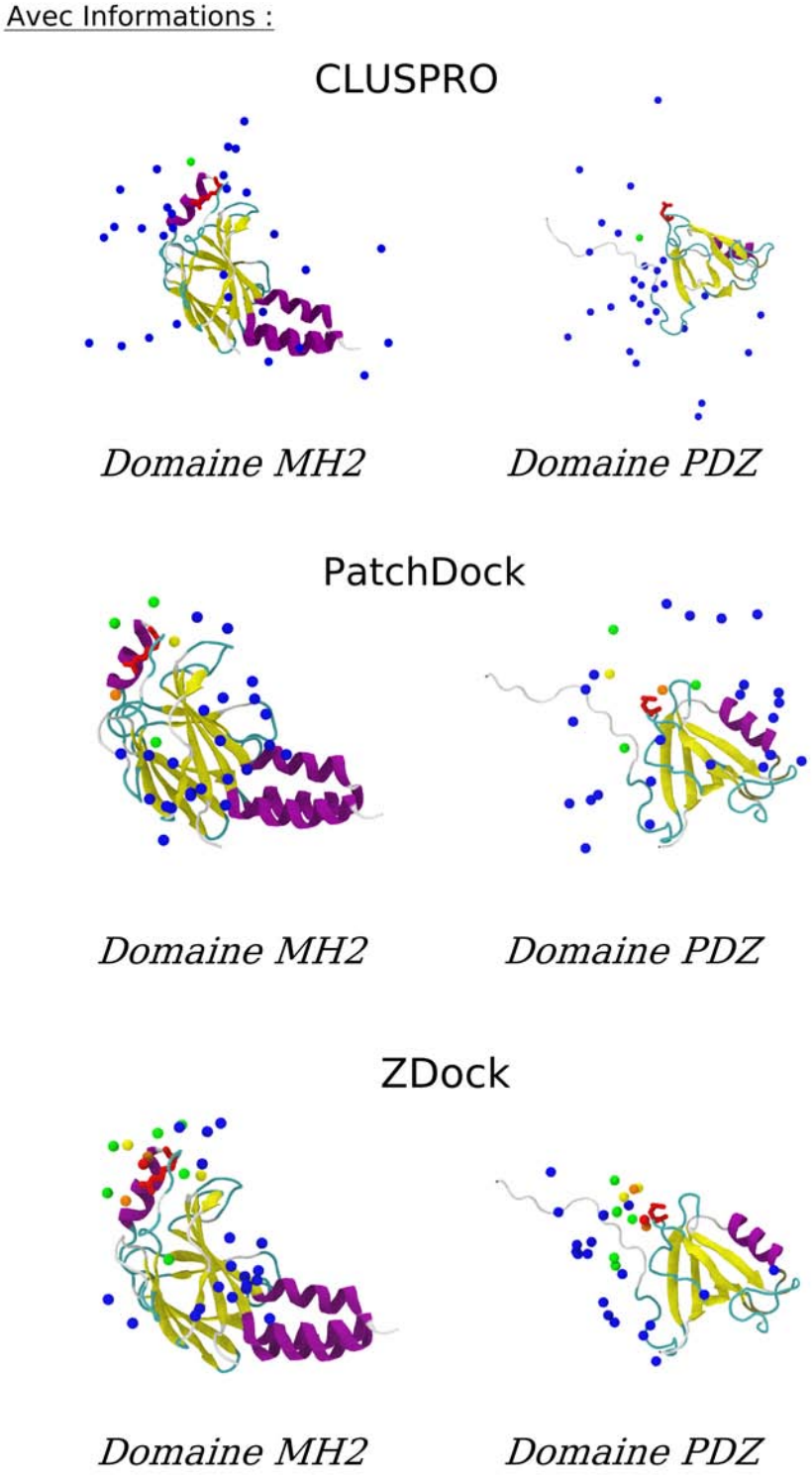


FIG. 2.20 – Graphique montrant la répartition du ligand par rapport au récepteur et inversement dans le cas où l'on donne R_{279} en contact avec E_{1321} comme information aux serveurs. Les sphères représentent les positions du résidu R_{279} ou E_{1321} pour chaque solution. Les sphères sont colorées en fonction de la distance R_{279}-E_{1321} : bleu pour une distance supérieure à 15 Å ; vert : pour une distance comprise entre 10 et 15 Å ; jaune : pour une distance comprise entre 7,5 et 10 Å ; orange : pour une distance comprise entre 5 et 7,5 Å et rouge : pour une distance inférieure à 5 Å. En rouge, le résidu R_{279} au niveau de Smad3 et E_{1321} au niveau d'Erbin.

Il apparaît très clairement une focalisation de toutes les solutions consensus sur deux pas de temps bien spécifiques de la dynamique qui sont $t = 11,8$ ns et $t = 13,1$ ns. Or, si l'on reprend les profils d'énergie figures 2.10B et 2.11C, nous pouvons voir qu'à ces temps les résidus E_{246} et K_{1271} sont en interaction de même que les résidus R_{279} et E_{1321}. Pour ce dernier couple de résidus, même si le profil d'énergie n'est pas parfaitement stable, la distance entre les deux résidus reste inchangée à 4 Å (voir figure 2.13) ; ce qui prouve que ces deux résidus sont toujours en interaction. Par contre, pour ces temps, les profils d'énergie montrent que les résidus D_{262} et K_{1326} n'interagissent quasiment plus. Si nous reprenons les résultats de mutations obtenus, nous constatons qu'en effet, les résidus E_{246} et R_{279} jouent un rôle dans l'interaction mais pas le résidu D_{262} (voir figure 2.15).

Ainsi, les solutions consensus ont ciblé une zone précise de la dynamique où les résidus en interaction correspondent aux résidus ayant un rôle lors des tests *in vitro*. Nous pouvons donc conclure quant à la capacité de ces serveurs de docking à obtenir une solution acceptable pour ce complexe, ceci avec ou sans informations pour contraindre les recherches.

2.4.4 Convergence des dynamiques : mise en évidence d'un *entonnoir* énergétique

Nous avons ensuite sélectionné une solution du tableau 2.2 pour chaque pas de temps trouvé. Nous avons sélectionné les solutions des serveurs de docking sans ajout d'information pour les temps $t = 11,8$ ns et $t = 13,1$ ns. Nous avons donc sélectionné la solution 24 de PatchDock (pour $t = 11,8$ ns) et la solution 3 de CLUSPRO (pour $t = 13,1$ ns). Nous avons réalisé une dynamique en solvant explicite de 10 ns pour ces deux solutions afin de tester leur stabilité respective. Nous avons également conduit une dynamique de 10 ns sur la solution 27 de Zdock. Cette dernière n'a pas été retenue comme solution consensus par les programmes de docking et elle pointe un autre pas de temps de la dynamique moléculaire ($t = 2,4$ ns). Pour $t = 2,4$ ns, toutes les interactions électrostatiques présentes dans les solutions consensus ne sont pas encore mises en place : en particulier, l'interaction R_{279}-E_{1321} (voir figures 2.10B et 2.13). Nous avons comparé l'évolution de cette structure par rapport à celles des deux autres solutions sélectionnées. Pour cela, nous avons calculé le RMSD du ligand entre tous les pas des dynamiques (figure 2.22). Nous avons aussi identifié les pas de chaque dynamique pour lesquels le RMSD du ligand été le plus bas (voir tableau 2.3). Pour calculer ce RMSD, nous avons éliminé les 25 premiers résidus du domaine PDZ formant le "bras" N-terminal. Les serveurs de docking n'ont pas placé ce fragment dans une position stable : cette portion était donc libre d'évoluer dans le solvant durant toute la durée des dynamiques. Ce mouvement aléatoire a engendré un bruit de fond important pour le calcul du RMSD ; c'est pourquoi nous avons omis cette partie dans le calcul du RMSD du ligand.

La figure 2.21 montre que les solutions consensus sont restées stables tout au long de la dynamique moléculaire puisque le centre de masse du domaine PDZ a oscillé autour d'un même point. Par contre, la solution isolée de Zdock a subi un fort changement conformationnel afin de replacer le domaine PDZ dans une position équivalente à celles trouvées pour les solutions consensus. Ceci est visible dans le tableau 2.3, où l'on peut constater que le RMSD du ligand

a fortement baissé, passant de 19 Å entre les solutions consensus et la solution isolée, avant dynamique, à ≈9 Å vers la fin de la dynamique moléculaire (pour ≈9,2 ns).

TAB. 2.3 – Comparaison des solutions sélectionnées avant et pendant la dynamique moléculaire.

Comp. serveur[a] (S_1-S_2)	Temps S_1	Temps S_2	L_{rmsd} avant Dyn.	L_{rmsd} pdt Dyn.
Cluspro(3)-Patchdock(24)	0,48	3,22	6,95	3,01
Cluspro(3)-Zdock(27)	3,52	9,2	19,73	9,32
Patchdock(24)-Zdock(27)	7,06	9,26	18,62	9,56

[a] Le chiffre entre parenthèses correspond au rang de la solution pour chaque serveur.
Le RMSD du ligand (en Å) est calculé en enlevant les 25 premiers résidus du domaine PDZ. Le RMSD du ligand pendant la dynamique correspond au RMSD le plus bas trouvé en comparant chaque pas des dynamiques.

Nous avons, ensuite, voulu savoir si ces minima mettaient en évidence une zone particulière au niveau de la dynamique conduite sur le modèle par analogie. Pour cela, nous avons repris le protocole précédent en calculant le RMSD du ligand entre chaque pas de la dynamique conduite sur le modèle par analogie et chaque pas des dynamiques réalisées sur les solutions consensus de PatchDock et Cluspro et la solution isolée de Zdock. Les résultats sont présentés figure 2.22. La comparaison des dynamiques nous a permis de mettre en évidence deux zones temporelles au niveau de la trajectoire de la dynamique moléculaire conduite à partir du modèle par analogie : une zone entre 5 et 6 ns et une autre entre 11,5 et 13,5 ns. Ces zones correspondent, encore une fois, à des temps de la dynamique pour lesquelles les interactions chargées R_{279}-E_{1321} et E_{246}-K_{1271} sont présentes (voir figures 2.10B et 2.11C). De plus, les minima entre solutions consensus et isolée présentés tableau 2.3 permettent eux aussi de cibler des zones temporelles. Le croisement de ces zones temporelles définissent des fenêtres au sein de la dynamique moléculaire du modèle par analogie (zone en jaunes sur la figure 2.22). Dans deux cas sur trois, ces fenêtres ciblent le L_{rmsd} le plus faible entre la dynamique du modèle par analogie et les dynamiques des solutions consensus et isolées. Ainsi, nous obtenons une convergence de la majorité des résultats vers un même minimum défini, au niveau de la dynamique moléculaire comme une structure transitoire entre les temps 5 et 6 ns et 11,5 et 13,5 ns. Nous retrouvons une structure pratiquement conservée entre ces deux laps de temps comme le montre le diagramme de RMSD figure 2.13. Si l'on superpose les trajectoires des dynamiques de la solution 27 de Zdock et du modèle par analogie (voir figure 2.21), nous remarquons que celles-ci se recouvrent complètement et ce même si les modèles ne partent pas de la même position. Ces trajectoires rejoignent le point d'oscillation des centres de masse des solutions consensus de CLUSPRO et PatchDock. Ceci peut donc définir un tunnel énergétique qui attire les solutions proches du minimum et bloque les solutions tombées dans celui-ci. Nous déterminons ce point d'ancrage au niveau de Smad3 comme étant le triplet d'arginines R_{279}, R_{287} et R_{288} (voir figure 2.12). Ainsi, le complexe Erbin PDZ/Smad3 MH2 est un bon exemple de l'importance des résidus chargés pour la reconnaissance protéique (Janin, 1997; Sheinerman *et al.*, 2000; Sheinerman et Honig, 2002; Dong et Zhou, 2006).

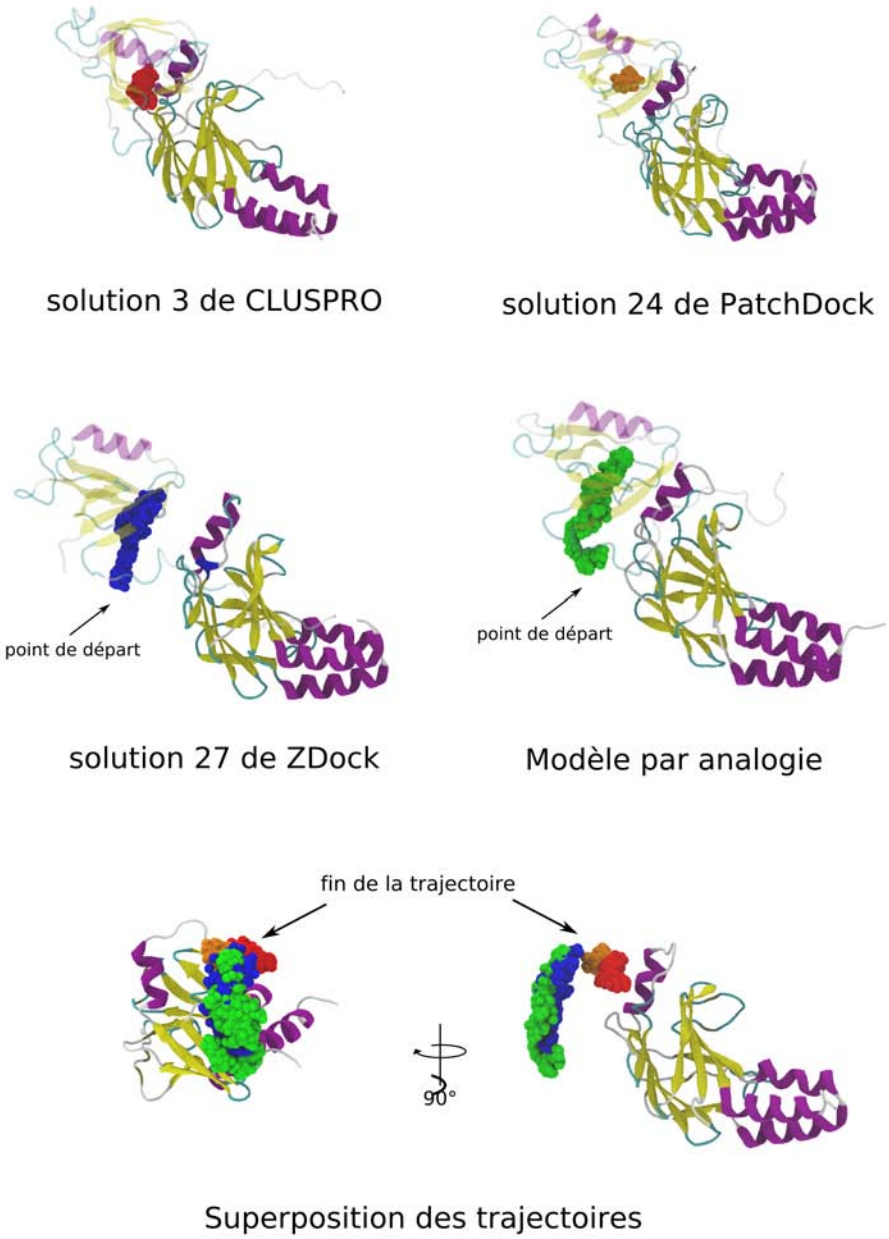


FIG. 2.21 – Visualisation des trajectoires du centre de masse du domaine PDZ pour les dynamiques conduites à partir des solutions consensus (solution 24 de PatchDock et solution 3 de CLUSPRO) et de la solution isolée (solution 27 de Zdock) ainsi que du modèle par analogie. Dernière image, superposition des différentes trajectoires.

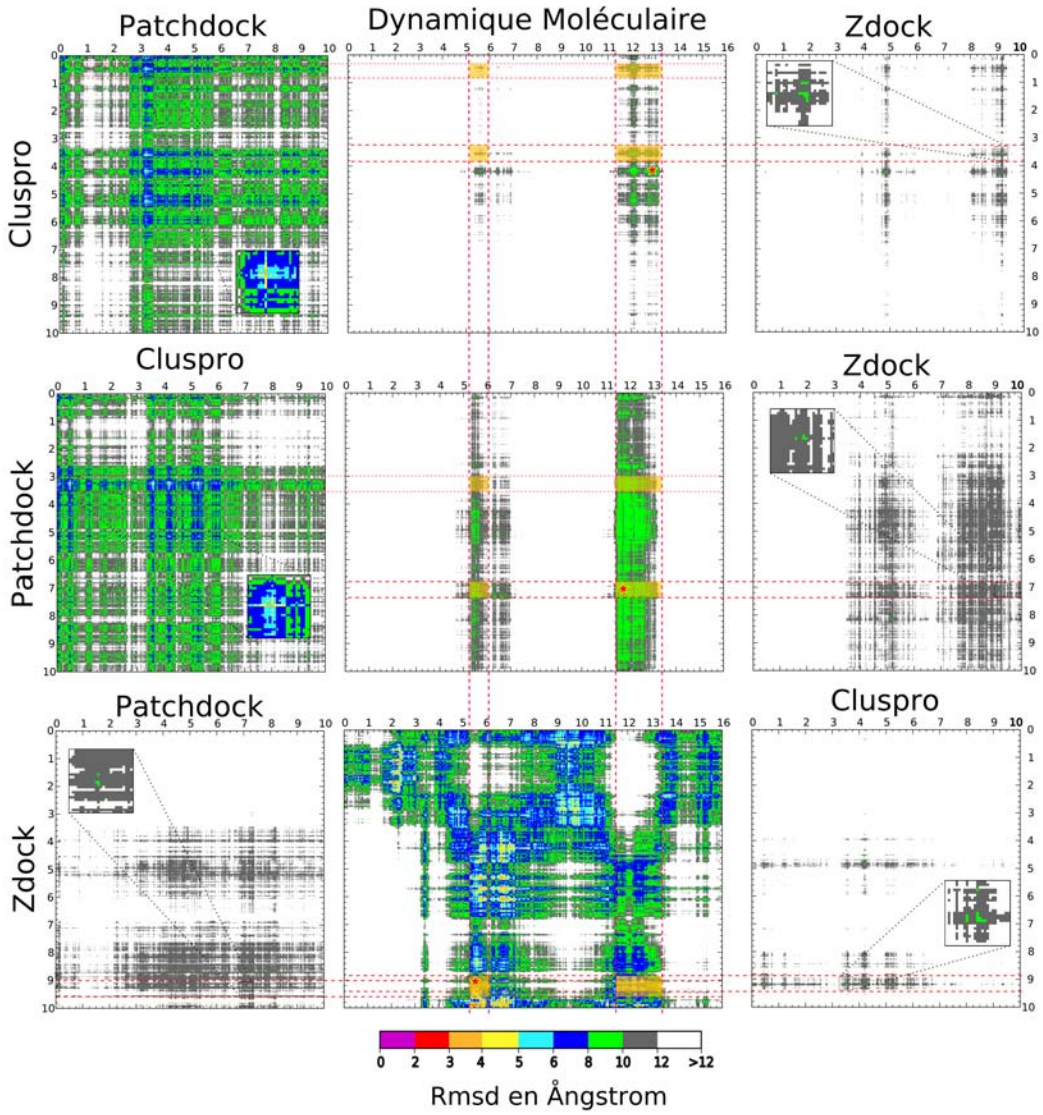


FIG. 2.22 – Comparaison des différentes dynamiques réalisées à partir des solutions consensus de PatchDock et CLUSPRO et de la solution isolée de Zdock ainsi qu' à partir du modèle par analogie. Les parties de dynamique encadrées correspondent aux minima présentés tableau 2.3. Au niveau de la dynamique moléculaire conduite à partir du modèle par analogie, les étoiles rouges correspondent aux L_{rmsd} minimum.

2.4.5 Conclusion

Ce travail nous permet d'aboutir aux conclusions suivantes :

- En utilisant divers serveurs de docking, nous avons consolidé notre modèle créé par analogie en retrouvant des structures proches de celui-ci.
- Cette étude nous a aussi permis de montrer l'utilité de la dynamique moléculaire pour raffiner les résultats. En effet, pour le modèle par analogie et pour la solution 27 du serveur Zdock, même si les points de départ n'étaient pas parfaitement positionnés, la dynamique moléculaire a permis un replacement des domaines PDZ.
- L'étude croisée des dynamiques a localisé certains minima. Ce protocole pourrait être étendu pour une recherche globale d'entonnoirs énergétiques. Dans ce cas précis, il ne faudrait pas se limiter à un seul assemblage mais regarder l'évolution de nombreux complexes bien définis. L'utilisation d'un *Benchmark* se révélerait alors très utile.
- Enfin, les dynamiques à partir du modèle par analogie et de la solution 27 de Zdock ont montré que le complexe Erbin PDZ/Smad3 MH2 pouvait subir de fort réarrangements structuraux (voir figures 2.13 et 2.21). Ceux-ci apparaissent après un laps de temps de 4-5 ns que nous qualifierons de temps de relaxation du système.

Ce travail n'est cependant pas encore terminé : la prochaine étape consistera à étudier les diverses solutions consensus ayant un L_{rmsd} de plus de 15 Å et voir si ces solutions peuvent converger elles aussi vers le même minimum que les modèles présentés au cours de cette étude. En effet, la mise en évidence d'un entonnoir énergétique nécessite l'analyse d'un grand nombre de trajectoires (Kozakov *et al.*, 2008; Hunjan *et al.*, 2008).

2.5 Extension de l'affinement de docking rigide par simulations courtes de dynamique moléculaire : exemple sur la cible 34 du challenge CAPRI

Les travaux réalisés sur le modèle par homologie et la solution 27 du serveur Zdock ont révélé que la durée moyenne de relaxation du complexe Erbin PDZ/Smad3 MH2 était de 4 à 5 ns. Cette durée est nécessaire pour voir des évolutions notables dans la structure du complexe. Cette évolution, relativement longue, d'un système pourrait être comparée à la formation d'un complexe de rencontre, ici guidée par des forces électrostatiques importantes (voir section Cinétique de l'association). Le calcul de simulations de dynamiques moléculaires de 5 nano-secondes, ou plus, est encore coûteux en temps et ne peut être appliqué dans des processus de routine.

A l'inverse, la comparaison des dynamiques moléculaires présentée précédemment, entre les solutions consensus de CLUSPRO et Patchdock, a mis en évidence un L_{rmsd} minimum atteint après un temps très court, ≈0,5 ns (voir tableau 2.3), ce qui laisse à penser que la solution de CLUSPRO était très proche d'un minimum local. Le calcul de simulations de quelques centaines de pico-secondes est maintenant possible en un temps relativement court.

Nous nous sommes alors posé la question suivante : est-ce que des simulations de dynamique moléculaire de courtes durées peuvent améliorer les résultats de docking rigide ? Pour répondre à cette problématique, nous avons testé l'utilisation de simulations courtes (0,5 ns) de dynamiques moléculaires pour raffiner les résultats de docking rigide. Nous avons comparé cette approche à des simulations courtes de Monte-Carlo.

La cible n° 34 du challenge CAPRI, complexe entre une protéine et un molécule d'ARN, a été choisie comme sujet d'étude. Ce complexe n'a pas encore été publié, c'est pourquoi nous ne nommerons pas explicitement la protéine mise en jeu ni ne donnerons les références des publications utilisées.

Ce travail a été réalisé en collaboration avec le chercheur écossais Dave Ritchie, développeur du programme de docking HEX. Nous avons, dans un premier temps, modélisé le complexe protéine/ARN à l'aide du programme HEX puis nous avons réalisé un affinement des meilleurs résultats de HEX par dynamique moléculaire en solvant explicite (DMSE) ou par Monte-Carlo (MC). Ces résultats ont ensuite été classés grâce à une fonction de score prenant en compte l'énergie potentielle d'interaction de l'assemblage.

2.5.1 La stratégie employée

Les informations disponibles pour modéliser la cible 34 : Lors de la 4ème édition de CAPRI, une grande variété de nouveaux complexes à assembler a été proposée. Cette édition suit l'impulsion de l'édition précédente en ne proposant pas seulement les "classiques" complexes anticorps-antigène et protéase-inhibiteur mais des complexes beaucoup plus variés reflétant ainsi

l'intérêt des cristallographes pour les nouveaux systèmes moléculaires (Janin, 2007). Le choix de la cible 34 ne fait pas exception puisqu'il s'agit du complexe entre une protéine et une molécule d'ARN : ce type de complexe n'avait encore jamais été présenté dans le challenge CAPRI.

Pour modéliser ce complexe, nous avions à notre disposition la structure cristallographique de la molécule d'ARN sous sa forme liée. Par contre, nous n'avions pas la structure de la protéine associée. Il a donc fallu, dans un premier temps, modéliser cette protéine à partir de la structure d'une protéine homologue fournie par les organisateurs du challenge CAPRI. Aucune étude structurale sur le complexe à modéliser n'avait été publiée, la seule référence que nous avions était celle de la protéine homologue. Cette publication présentait la protéine homologue associée à un modèle de molécule d'ARN. Cet assemblage a permis d'identifier le site de fixation de l'ARN au niveau de la protéine. En ce qui concerne la molécule d'ARN, des travaux expérimentaux sur l'interaction de celle-ci avec la protéine que nous avions à modéliser ont permis d'identifier certaines bases impliquées dans l'interaction. Ces informations ont donc ensuite été utilisées comme filtre pour choisir les solutions potentielles.

Modélisation par homologie de la protéine : Pour nous aider à modéliser la protéine réceptrice à partir de la structure de son homologue, un alignement structural était fourni. Nous avons pris le parti de faire notre propre alignement car l'alignement structural proposé coupait certaines structures secondaires, ce qui nous paraissait peu recommandé. Ce nouvel alignement avait un pourcentage d'identité et de similarité supérieur à celui de l'alignement structural (respectivement 27,3% d'identité et 42% de similarité contre 24,1% et 39,5%). Ces résultats ont été obtenus avec la matrice d'alignement BLOSUM62. A partir de cet alignement nous avons utilisé le programme MODELLER (Fiser et Sali, 2003) pour créer le modèle par homologie. Les résidus manquants ont été rajoutés à partir d'extraits de séquences issus de la PDB. Enfin, les parties ajoutées ont été minimisées grâce à 5000 pas de gradient conjugué (le reste de la protéine restant fixe). Puis l'ensemble de la structure a été relaxée par une minimisation de 6400 pas de gradient conjugué en utilisant le champ de forces CHARMM (Brooks *et al.*, 1983).

Docking des partenaires : Le programme utilisé pour créer notre assemblage est le programme HEX (Ritchie et Kemp, 2000). Celui-ci permet de représenter la surface des protéines en approximant cette dernière à partir d'harmoniques sphériques. Nous avons "docké" la structure de l'ARN et de notre modèle de protéine en prenant, dans un premier temps, le centre de gravité de chaque partenaire comme centre de rotation/translation. Pour cela, nous avons d'abord effectué une première recherche gros grains restreinte à $45^{\circ} \times 45^{\circ}$ pour les deux partenaires avec N=25. Nous avons ensuite sélectionné la $5^{\text{ème}}$ solution de ce premier filtrage par rapport aux informations obtenues dans la littérature. Cette solution 5 fut ensuite utilisée comme orientation de départ pour une recherche plus fine en utilisant la méthode 6D (Ritchie *et al.*, 2008) caractérisée par un pas de recherche beaucoup plus fin (0,5 Å) et des degrés de recherche de 45×180. Les 9 premiers résultats donnés par HEX ainsi que l'orientation de départ furent soumis comme les prédictions du groupe D. Ritchie à CAPRI. Les 14 meilleurs résultats obtenus avec HEX ont ensuite été raffinés par dynamique moléculaire ou Monte-Carlo.

Simulations par Monte-Carlo : La procédure de Monte-Carlo mise en place est constituée de deux étapes :

1. Une minimisation par échantillonnage des mouvements conformationnels a été réalisée. Celle-ci correspond à des mouvements aléatoires de rotation et de translation des deux partenaires, suivis par des mouvements des chaînes latérales des résidus et des acides nucléiques. Les mouvements locaux des chaînes latérales ont été modélisés en utilisant une bibliothèque de rotamères. Les mouvements globaux des deux partenaires ont été contraints à 0,2 Å pour la translation et de 5 à 10 degrés pour la rotation.
2. Chaque résultat a ensuite été reclassé grâce à une fonction de score.

L'équation de la fonction de score est :

$$\Delta G_{association} = \Delta E_{el+vdw} + \Delta G_{el,solv} + \Delta G_{np,solv}$$

où ΔE_{el+vdw} est l'énergie interne du système (van der Waals + électrostatique) calculée par CHARMM, $\Delta G_{el,solv}$ est la contribution électrostatique de l'énergie de solvatation calculée par APBS en résolvant l'équation de Poisson-Boltzmann et $\Delta G_{np,solv}$ est la variation de surface accessible au solvant calculée par CHARMM. Le champ de forces utilisé est CHARMM27.

Pour le challenge CAPRI, les simulations étaient conduites sur 200 itérations. Cependant, la configuration la mieux classée a souvent été obtenue après la première itération, à l'exception du résultat n°10 de HEX, car les mouvements conformationnels ont été assez faibles. Les 10 modèles reclassés par la procédure de Monte-Carlo ont été soumis comme les prédictions du groupe de Fabrice Leclerc.

Simulations par dynamique moléculaire : Pour améliorer la rapidité d'exécution, nous avons créé un "Plug'in" au logiciel VMD (Humphrey *et al.*, 1996) permettant de prendre directement un fichier solution de HEX 5.1 et de créer une boîte d'eau à partir de celui-ci. Les molécules d'eau utilisées sont du type TIP3P. Des ions Na^+ et Cl^- sont ajoutés automatiquement pour atteindre la neutralité de l'ensemble. Le protocole est ensuite le même que celui utilisé pour modéliser le complexe entre le domaine PDZ d'Erbin et le domaine MH2 de Smad3 ; le programme de dynamique moléculaire utilisé est NAMD (Phillips *et al.*, 2005) avec le champ de force CHARMM27 (MacKerell *et al.*, 2000). La température et la pression ont été maintenues respectivement à 300 Kelvins et 1 atmosphère (équations de Langevin). Les équations de mouvement furent intégrées avec un pas de 1 fs en utilisant r-RESPA (Tuckerman *et al.*, 1992). Les interactions électrostatiques à longue distance sont traitées en utilisant l'algorithme *Smooth Particle Mesh Ewald* avec une distance limite de 11 Å. De plus, nous avons contraint la structure de l'ARN à rester fixe puisqu'il s'agissait de la forme cristallisée. A l'inverse, la protéine et l'environnement (molécules d'eau et ions) n'ont subi aucune contrainte. Nous avons conduit une dynamique moléculaire de 0,5 ns pour les 14 meilleurs résultats de HEX et nous avons sélectionné le dernier pas de chaque dynamique comme solution. Nous avons ensuite classé ces solutions en fonction de leur énergie potentielle d'interaction.

Fonction de score utilisée : Il s'agit seulement de classer les solutions en fonction de l'énergie d'interaction entre la protéine et la molécule d'ARN. Cette énergie est calculée de la même

manière que celle calculée pour mettre en évidence les résidus en interaction pour le complexe Erbin/Smad3. Il s'agit de l'énergie potentielle d'interaction (*i.e.* énergie électrostatique+énergie de Van der Waals) calculée par le programme NAMD grâce au champ de forces CHARMM27. En pratique, nous avons utilisé le plug'in NAMD Energy[18] pour cibler la zone d'interaction et définir les paramètres d'entrée. Nous avons reclassé les 14 résultats de HEX et avons sélectionné les 10 modèles ayant les énergies d'interaction les plus basses comme les prédictions de notre groupe (groupe B. Maigret) à CAPRI.

2.5.2 Les résultats obtenus

La cible T34 en chiffres : Le challenge CAPRI se déroule en différents *rounds* : chaque équipe de recherche s'inscrit pour chaque round. Les équipes inscrites durant un round ne sont pas obligées de donner un résultat pour chaque complexe. Ainsi, la cible T34 fait partie du *round* 15 où il y avait 44 participants inscrits. Sur ces 44 participants, 27 seulement ont donné au moins un modèle pour cette cible. Ceci refléterait peut-être le caractère nouveau et spécifique de l'assemblage à modéliser. De plus, cet assemblage fut proposé préalablement dans ce même *round* mais sous la forme protéine à modéliser et ARN non lié. Aucune prédiction correcte n'a été obtenue. Tous ces paramètres ont peut-être découragé certains participants.

Néanmoins, sur les 27 participants engagés, 14 ont prédit au moins un résultat acceptable (voir figure 2.23). Parmi ceux-ci, 9 participants ont donné au moins un résultat medium et 14 un résultat acceptable. Notons que tous les participants (exceptée JC Mitchell) ayant obtenu un résultat medium ont aussi obtenu un ou plusieurs résultats acceptables ; preuve que leur ensemble de données se situait dans la bonne région d'association. Ces résultats sont honorables mais restent un peu en dessous de ce qui a déjà été obtenu pour les complexes du type lié/non lié. Par exemple, pour la cible 25 de l'édition précédente, il y eu 18 participants ayant donné des résultats au moins acceptables dont 1 groupe ayant donné un résultat haut et 13 autres ayant donné des résultats medium (Lensink *et al.*, 2007). Il faut néanmoins noter que le nombre de participants pour la cible 25 était plus élevé puisqu'il y avait 37 groupes prédicteurs.

Si l'on regarde le nombre de résultats acceptables et medium, la cible 34 donne de meilleurs résultats que la cible 25 avec 40 résultats acceptables (contre 20 pour T25) et 25 résultats medium (contre 13 pour la cible 25). Nous devons néanmoins retrancher les résultats redondants de l'équipe "Écosse-Lorraine", ce qui laisse 32 résultats acceptables et 21 résultats medium. Ceci est dans la moyenne haute des résultats de CAPRI, preuve que ce complexe devait être l'un des plus faciles de cette 4$^{\text{ème}}$ édition.

[18] `http://www.ks.uiuc.edu/Research/vmd/plugins/namdenergy/`

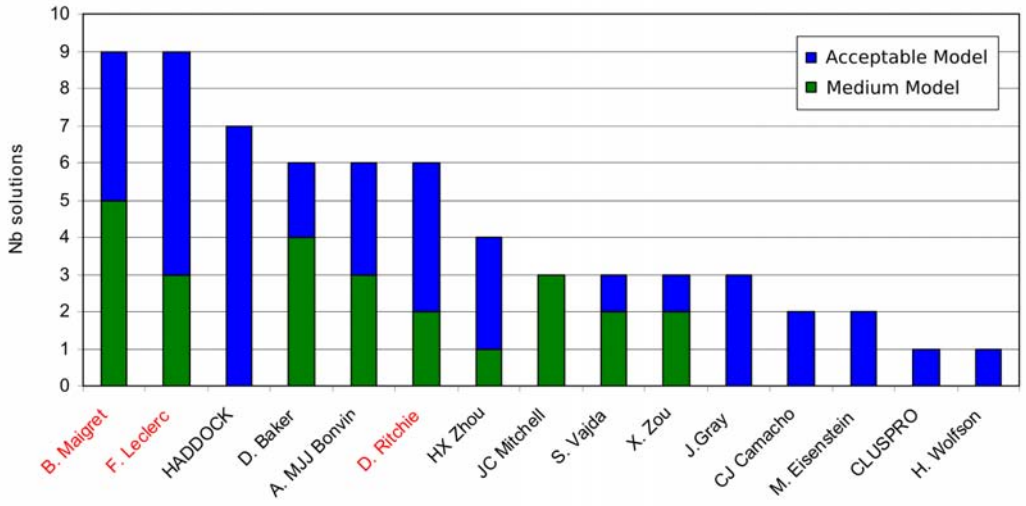


FIG. 2.23 – Résultats des participants ayant donné au moins un modèle acceptable. Résultats classés en fonction du nombre de modèles ayant une précision acceptable ou supérieure. En rouge, l'équipe "Écosse-Lorraine", constituée par 3 équipes : celle de Dave Ritchie chercheur à Aberdeen (au LORIA pendant la manche T34), une équipe du LORIA menée par Bernard Maigret et une équipe de la Faculté de sciences de Nancy menée par Fabrice Leclerc.

L'ensemble des résultats obtenus pour la cible 34 sont disponibles sur le site CAPRI[19]. Si l'on détaille les résultats obtenus (voir figure 2.23), l'équipe "Écosse-Lorraine" a obtenu de bons résultats en proposant un grand nombre de solutions correctes : 9/10 pour l'équipe de Bernard Maigret et de Fabrice Leclerc et 6/10 pour les résultats de HEX (Dave Ritchie). Il faut noter que 3 autres solutions du programme HEX étaient correctes mais ont été éliminées à cause d'un trop grand nombre d'interpénétrations (*clashes*). Au niveau du nombre de résultats medium, l'équipe reste encore en tête avec 5 résultats medium sur les 10 modèles proposés pour Bernard Maigret et deux résultats honorables pour les résultats de Fabrice Leclerc (3 résultats medium) et de HEX (2 résultats medium). D'autres équipes ont, elles aussi, obtenu de très bons résultats : l'équipe de David Baker (avec le programme RosettaDock) a donné 6 résultats corrects dont 4 medium. Parmi ces 4 résultats medium, leurs modèles n°1 et n°3 ont été les meilleurs de tous les modèles proposés. Nous pouvons aussi noter le très bon résultat en termes de nombre de solutions correctes du serveur HADDOCK avec 7 modèles acceptables. Les résultats de ce serveur semblent aller de pair avec ceux d'Alexandre Bonvin qui développe celui-ci. En effet, A. Bonvin a également obtenu de bons résultats avec 3 modèles médium et 3 modèles acceptables. Enfin, nous pouvons noter le cas de Julie C. Mitchell qui a obtenu 3 résultats medium mais aucun résultat acceptable, les autres modèles proposés par son équipe ayant été éliminés car ils contenaient trop de *clashes*. Peut-être faut-il y voir un protocole de l'affinement particulier employé sur ces 3 résultats médium mais pas sur les autres (par faute de temps par exemple) ?

[19]http ://www.ebi.ac.uk/msd-srv/capri/round15/R15_T34/

Affinement des résultats de docking rigide : Au delà des bons résultats obtenus, le véritable intérêt de cette étude est de pouvoir comparer, avec les mêmes outils d'analyse et sans *a priori*, les modèles obtenus après docking rigide (résultats de HEX) suivis d'une minimisation courte (résultats de Mont-Carlo) ou d'un affinement plus long et plus coûteux (dynamique moléculaire de 0,5 ns avec solvant explicite). De plus, la structure cristallographique du complexe n'ayant pas encore été divulguée, nous n'utiliserons que les paramètres donnés lors des résultats de CAPRI comme éléments d'évaluation. Ces paramètres sont : la fraction de contacts natifs et le RMSD du ligand (voir le paragraphe *Évaluation des résultats*, dans la section consacrée au challenge CAPRI, pour la définition de ces paramètres).

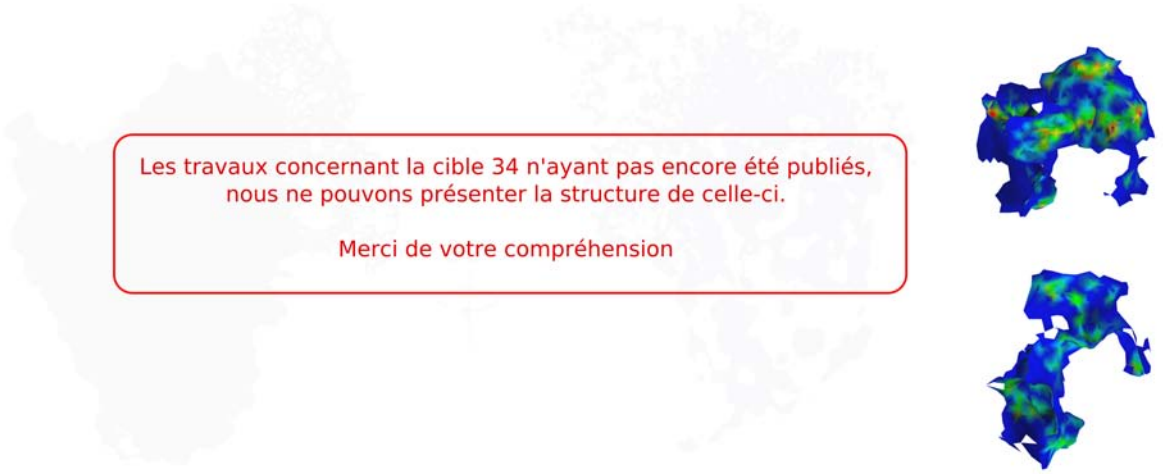


FIG. 2.24 – Comparaison du meilleur modèle de HEX (n°8), en rouge transparent, et du meilleur modèle de la dynamique moléculaire (n°3) en bleu. Notons que ce modèle n°3 correspond au modèle n°8 de HEX après affinement par DMSE. Il est possible de voir les réarrangements de surface entre ces deux modèles. A droite, en haut, interface du modèle de HEX et, en bas, interface du modèle après dynamique moléculaire. Les zones en rouge sur cette interface symbolise les *clashes*.

Nous sommes donc partis des 14 meilleurs résultats de HEX ; nous les avons raffinés puis nous les avons reclassés (voir figure 2.26). Le premier résultat intéressant est que les deux méthodes de l'affinement (DMSE et MC) ont permis de supprimer les interpénétrations des deux partenaires présentes après le docking rigide. En particulier, le modèle le plus précis fournit par HEX (n°8) a été éliminé car il contenait trop de *clashes* (156 pour une limite maximale de 136,3). Ce même modèle, après minimisation par MC ou DMSE, ne possédait plus que 15 *clashes* (voir figure 2.24). Cette tendance est confirmée au niveau de tous les modèles présentés puisque ceux-ci, après dynamique moléculaire, possédaient moins de 20 *clashes* (dont 5 ayant 10 *clashes* ou moins) de même que tous les modèles après Monte-Carlo possédaient moins de 23 *clashes* (dont 3 ayant 10 *clashes* ou moins) alors que tous les résultats présentés par HEX avaient

plus de 60 *clashes*. Ces résultats sont donc encourageants sachant que la moyenne des *clashes* de l'ensemble des solutions correctes présentées pour cette cible était de 30. Ainsi, nous avons prouvé l'intérêt de simulations courtes de Monte-Carlo et dynamique moléculaire pour la suppression des interpénétrations après docking rigide. Ces résultats confirment ce qui a d'ailleurs été trouvé dans d'autres travaux (Bonvin, 2006). Ainsi, les simulations courtes de dynamique moléculaire et de Monte-Carlo ont permis de raffiner la structure des modèles obtenus par docking rigide. La simulation par Monte-Carlo est plus adaptée à une élimination rapide des *clashes* car elle est plus rapide que la dynamique moléculaire.

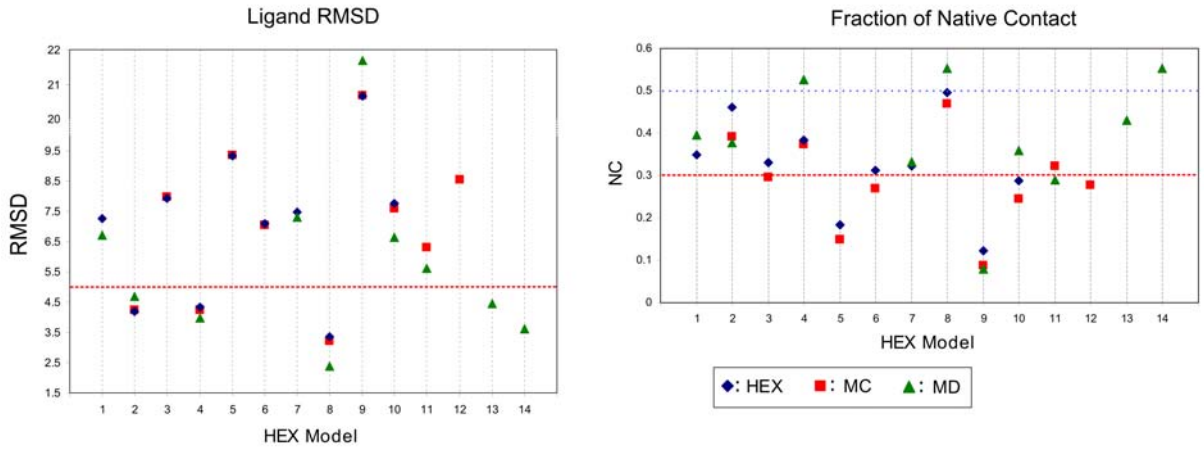


FIG. 2.25 – Valeurs de RMSD du ligand et de fraction de contacts natifs pour les 14 modèles présentés par l'équipe Écosse-Lorraine. La ligne en pointillés rouges représente la limite entre les résultats acceptables et medium : 5 Å pour le RMSD du ligand et 30% de contacts natifs (voir Méndez *et al.* (2005)). La limite de 50% de contacts natifs est indiquée par des pointillés bleus.

La dynamique moléculaire peut améliorer les résultats de docking rigide en augmentant la fraction de contacts natifs jusqu'à plus de 70% (Król *et al.*, 2007b). De même, la minimisation par Monte-Carlo a montré son efficacité dans ce domaine (Kozakov *et al.*, 2008). Nous avons voulu tester ces deux méthodes sur des laps de temps très courts (quelques centaines de pas pour les simulations de Monte-Carlo et quelques centaines de picosecondes pour la Dynamique Moléculaire). Les résultats sont présentés figure 2.25. Il apparaît que la simulation par Monte-Carlo n'a pas amélioré la position du ligand. A l'inverse, la dynamique moléculaire a amélioré les résultats de docking rigide, même si ce n'est que de quelques dizièmes d'angströms. Seuls deux cas ont montré une déterioration des résultats de dynamique moléculaire par rapport au docking rigide : les modèles n°2 et n°9. Pour le modèle n°9, nous pouvons expliquer ceci par le fait que la solution donnée par le programme HEX était déjà une mauvaise solution ; la dynamique moléculaire a donc accentué cette erreur. Nous n'avons pas d'explications pour le cas n°2. Nous pouvons quand même noter que l'un des meilleurs affinements se fait pour le modèle n°8 (-0,967 Å entre HEX et DMSE), correspondant au modèle le plus précis de HEX.

Les résultats par rapport à la fraction de contacts natifs sont plus explicites : la majorité des résultats de simulation par Monte-Carlo détériorent les contacts natifs trouvés par docking rigide. La dynamique moléculaire améliore les contacts natifs par rapport au docking rigide dans 5 cas sur 7, avec des améliorations pouvant aller jusqu'à ≈14% de contacts natifs en plus (pour le modèle n°4). De plus, seules les solutions après dynamique moléculaire ont un pourcentage de contacts natifs supérieur à 50% (voir figure 2.25). De même que pour le RMSD du ligand, la fraction en contact natif du meilleur résultat (n°8) est améliorée après DMSE (+ ≈6%) tandis que la fraction du modèle le moins bon est déteriorée après DMSE. Au regard de ces résultats, la dynamique moléculaire accentuerait les tendances en améliorant les résultats pour des complexes proches de la structure native et dégraderait les résultats déjà aberrants.

Ainsi, si l'on compare les résultats des affinements par DMSE et MC, la dynamique moléculaire semble, pour ce cas, plus efficace que la simulation par Monte-Carlo. Les simulations par MC étant meilleures seulement dans 2 cas dont un cas où le modèle de docking rigide était considéré comme une solution non correcte. Néanmoins, nous pouvons noter que les temps de simulation de Monte-Carlo étaient très courts et nous pensons donc refaire les mêmes tests sur des durées plus longues pour que la comparaison soit plus juste.

Une fonction de score simple mais efficace : Comme nous l'avons expliqué dans la partie précédente, la fonction de score utilisée pour classer les résultats après dynamique moléculaire est basée sur le champ de forces CHARMM27. Comme le montre la figure 2.26, les trois meilleurs résultats reclassés après dynamique moléculaire sont des modèles appartenant à la meilleure catégorie pour cette cible, la catégorie *medium*. Si l'on détaille, on peut voir que les 2 meilleurs résultats selon les critères CAPRI sont parmi les 3 premiers modèles classés par la fonction de score. Ces modèles étaient classés en $8^{\text{ème}}$, $13^{\text{ème}}$ et $14^{\text{ème}}$ position par le programme HEX. La fonction de score a donc reclassé correctement ces modèles. Il reste le problème des modèles 7 et 9 qui sont des modèles medium mais ont pourtant été classés après les modèles acceptables. Une explication pourrait être que des interactions chargées, présentes au niveau des autres modèles médium mais absentes au niveau de ces modèles, diminuent considérablement l'énergie potentielle de ces derniers (losanges notés 3 et 5 figure 2.27). De plus, même si cette fonction n'a pu éliminer le modèle incorrect, celui-ci est classé en dernière position.

Nous avons ensuite voulu savoir si cette fonction de score permettait de mieux classer *a posteriori* les résultats obtenus après docking rigide et Monte-Carlo. Le reclassement des résultats de docking rigide et de Monte-Carlo par notre fonction de score est présenté en figure 2.27. Les 3 meilleurs modèles de HEX ont été reclassés dans les premières positions. Il faut ici noter que, du fait des interpénétrations, la fonction de score utilisée n'a pris en compte que le paramètre électrostatique, celui de Van der Waals n'ayant pu être calculé. Néanmoins, dans le cas du complexe protéine-ARN, ce dernier avait une valeur très faible comparée à la valeur des interactions électrostatiques (pour les résultats de DMSE et de MC). Sur le diagramme d'énergie potentielle par rapport au RMSD du ligand ou aux contacts natifs nous pouvons voir une corrélation acceptable avec, pour l'ensemble des complexes, $R^2 = 0{,}71$. Cette corrélation est plus basse pour le RMSD avec $R^2 \approx 0{,}4$ mais, en enlèvant le résultat incorrect, nous obtenons $R^2 = 0{,}67$. Enfin,

nous notons que le résultat incorrect a été replacé en dernière position.

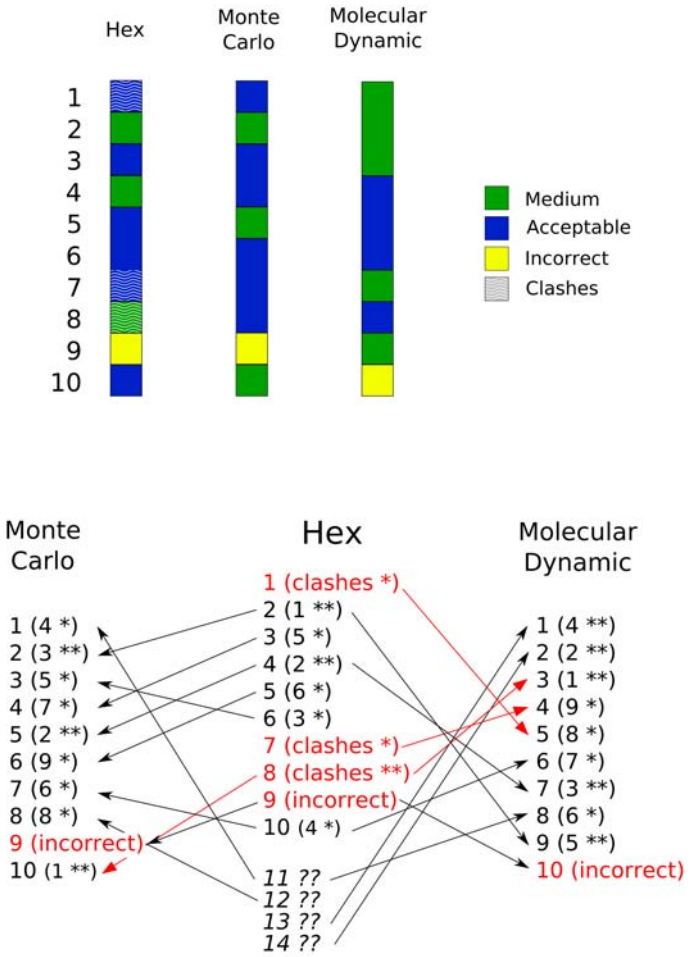


FIG. 2.26 – En haut, résultats après docking rigide (HEX) suivi d'une minimisation par Monte-Carlo (Monte Carlo) ou suivi d'un affinement par dynamique moléculaire en solvant explicite (Molecular dynamic). Ces résultats sont colorés en fonction de leur exactitude par rapport à la structure cristallographique : en vert, les résultats *medium* ; en bleu, les résultats acceptables et en jaune les résultats incorrects. Code couleur repris de Lensink *et al.* (2007). En tramé, résultats éliminés car ayant trop de *clashes*. En bas, reclassement des résultats de docking rigide. Entre parenthèses, le rang de chaque résultat ainsi que leur exactitude donnés par les experts de CAPRI : * correspond à un modèle accepatble, ** à un modèle *medium*. Les modèles incorrects ou ayant trop de *clashes* sont représentés en rouge.

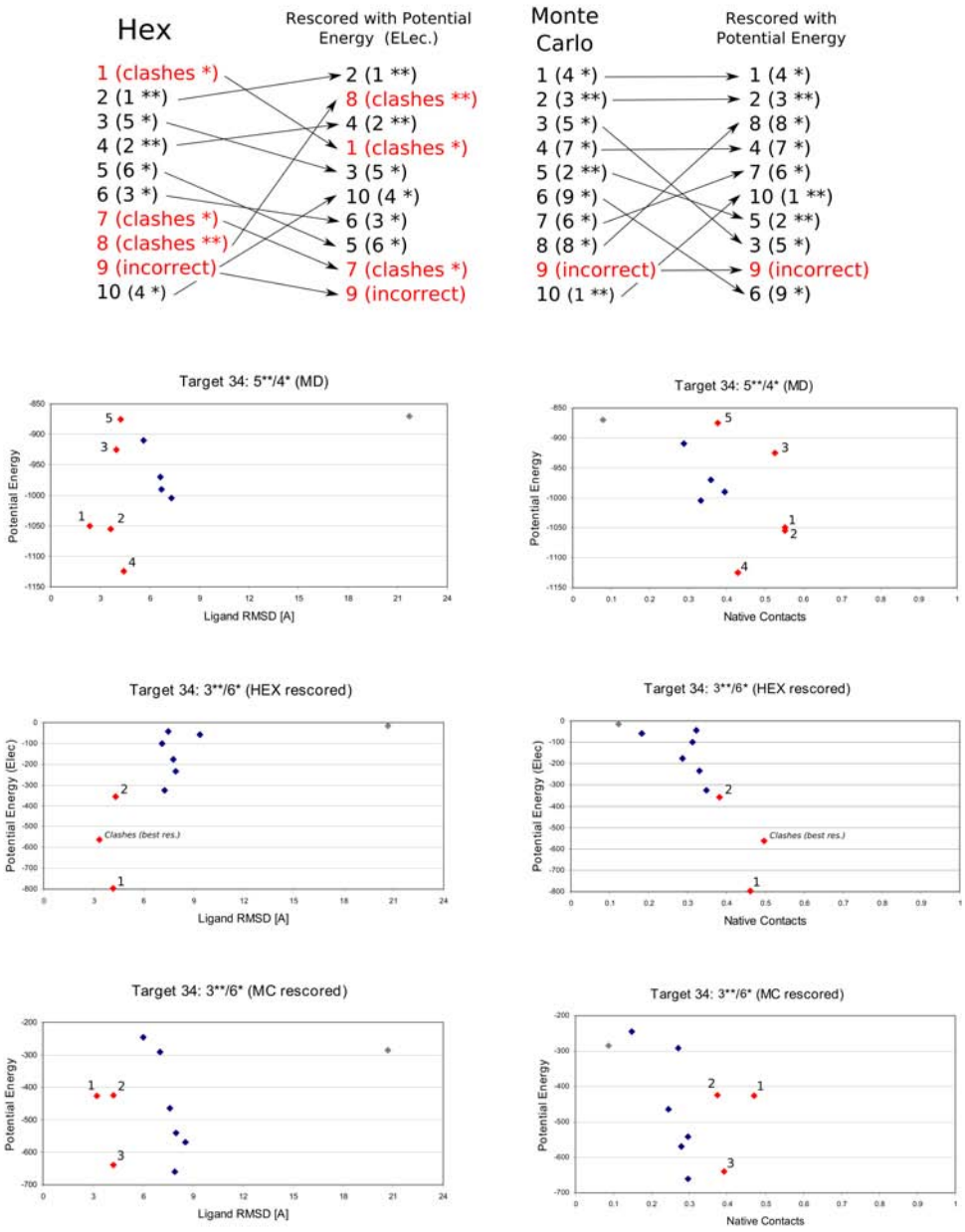


FIG. 2.27 – Reclassement des résultats de docking rigide et de MC par la fonction de score (en haut). Code couleur et classement identiques à la figure précédente. En dessous, l'énergie potentielle en fonction du RMSD du ligand ou de la fraction de contacts natifs. Résultats obtenus pour le classement après dynamique moléculaire et pour le reclassement des résultats de MC et de HEX par la fonction de score. En rouge, ensemble des résultats *medium* et en gris résultats incorrects. A côté de chaque symbole rouge est indiqué le rang du modèle donné par les experts de CAPRI.

Dans le cas des résultats de Monte-Carlo, la fonction de score utilisée n'a pas permis un

meilleur reclassement des modèles. En effet, les modèles médium ont été mélangés avec le reste des modèles acceptables plutôt que d'être reclassés dans les premières positions. Le seul résultat medium classé dans les premiers n'a pas évolué dans le classement. De même, le modèle incorrect n'a pu être classé en dernière position.

Ainsi, la fonction de score utilisée a, finalement, montré ses meilleures performances pour le classement des résultats de docking rigide. Néanmoins, le classement donné après DMSE reste très acceptable et permet de reclasser des résultats de manière convenable.

2.5.3 Discussion sur les résultats obtenus

La première explication de ces bons résultats est qu'un des deux partenaires (la molécule d'ARN) était sous forme liée. En effet, nous avons montré en première partie qu'il pouvait y avoir d'importants changements conformationnels pour ce type d'assemblage. Ceci devait être le cas pour ce complexe puisque l'interface entre la protéine et la molécule d'ARN a une taille d'au moins 3000 $Å^2$. Or ce genre de mouvements reste encore très difficile à modéliser (Bonvin, 2006; Lensink *et al.*, 2007; Andrusier *et al.*, 2008). La modélisation de ce complexe à partir des deux partenaires non-liés n'a d'ailleurs donné aucun résultat satisfaisant (voir résultats sur la cible 33[20]).

Ces résultats s'expliquent aussi par une analyse exhaustive des informations sur le site de fixation. Enfin, même si la dynamique moléculaire est plus coûteuse que les simulations de Monte-Carlo, elle permet une modélisation plus réaliste du système par la prise en compte des molécules d'eau et l'utilisation d'un champ de forces approprié.

La modélisation explicite des molécules d'eau a aussi permis une meilleure représentation des interactions potentielles au niveau de l'interface entre protéine et acide nucléique. En effet, nous avons expliqué dans notre première partie l'influence de ces molécules pour améliorer la complémentarité entre protéine et acide nucléique. Ces molécules d'eau peuvent aussi servir d'écran entre résidus de même charge ou permettent de ponter des interactions entre résidus trop éloignés. C'est pourquoi, il était important de modéliser explicitement celles-ci. Cette conclusion est confirmée par des travaux récents sur le sujet (Samsonov *et al.*, 2008).

Le choix d'un champ de forces approprié pour modéliser le complexe protéine-ARN fut aussi un facteur déterminant. En effet, les simulations de dynamique moléculaire d'un assemblage protéine-acide nucléique sont plus complexes que l'étude des composants séparés (Mackerell et Nilsson, 2008). Le champ de forces doit donc être bien équilibré et il faut faire très attention à traiter correctement le solvant et les interactions électrostatiques. Pour cela, nous avons utilisé le champ de forces CHARMM27. Celui-ci est reconnu pour être un bon champ de forces pour l'étude des interactions protéine-acide nucléique comme indiqué par divers travaux (MacKerell *et al.*, 2000; MacKerell et Banavali, 2000; Mackerell et Nilsson, 2008).

[20] voir `http://www.ebi.ac.uk/msd-srv/capri/round15/R15_T33/`

Enfin, la dynamique moléculaire en solvant explicite a permis une diminution du RMSD et une augmentation des contacts natifs dans la majorité des cas, même si les gains ne sont pas importants. Il faut aussi noter que les résultats obtenus avec HEX étaient déjà proches de la structure native et que les temps de simulation furent très courts. De plus, contrairement à ce qu'indique Król *et al.* (Król *et al.*, 2007a), la dynamique moléculaire en solvant explicite peut améliorer les modèles proches de la structure du complexe natif (à 3-4 Å de RMSD du complexe cristallin) : la preuve en est notre meilleur modèle (n°8 pour HEX et n°3 pour DMSE) avec un gain de ≈1 Å pour le RSMD du ligand (passant de 3,35 Å pour la solution de HEX à 2,38 Å pour le résultat après DMSE) et un gain de ≈5% de la fraction de contacts natifs (passant de ≈50% de contacts natifs pour la solution de HEX à ≈55% pour le résultat après DMSE). Enfin, ce modèle, avec un pourcentage de contacts natifs de 0,553 et un I_RMSD de 1,493 Å, était proche du classement en solution *high* (nécessitant une fraction de contacts natifs $\geq$ 0,5 et un I_RMSD $\leq$ 1,0 Å). Toutes ces données tendent à prouver l'intérêt de la DMSE dans ce cas particulier et plus généralement de l'intérêt de modéliser la flexibilité pour raffiner les résultats de docking rigide : ceci est clairement visible puisque, actuellement, la majorité des programmes de docking modélise cette flexibilité (voir tableau 1.3).

Nous attendons maintenant la structure cristallographique du complexe pour pouvoir finaliser notre étude. En effet, nous aimerions vérifier si les positions des molécules d'eau potentiellement présentes à l'interface du complexe cristallisé ont été conservées au cours de la dynamique moléculaire. Nous voulons aussi comparer les pas des trajectoires de la dynamique avec le complexe cristallisé afin de vérifier si une structure plus proche de la structure native a pu être échantillonnée.

2.6 Conclusion sur l'utilisation de la dynamique moléculaire en solvant explicite

Ainsi, nous avons montré, à travers ces quelques exemples, que la dynamique moléculaire en solvant explicite permettait de raffiner les structures de complexes après la phase de docking mais aussi permettait de mettre en évidence les résidus en interaction ou encore de visualiser les entonnoirs d'énergie. La prise en compte du solvant nous semble très importante, que cela soit pour la modélisation du repliement des protéines ou de l'assemblage macromoléculaire. De nombreuses études confirment cette hypothèse (Janin, 1999; Papoian *et al.*, 2004; Rodier *et al.*, 2005; Samsonov *et al.*, 2008). Le programme de docking HADDOCK, prenant en compte les molécules d'eau dans la phase de l'affinement (Dominguez *et al.*, 2003; van Dijk et Bonvin, 2006), a obtenu de très bons résultats sur cette cible et plus généralement sur l'ensemble du challenge CAPRI (de Vries *et al.*, 2007). Des méthodes prenant en compte le solvant dans la phase de l'affinement se mettent d'ailleurs en place (Qin et Zhou, 2007; Heifetz *et al.*, 2007).

De plus, la modélisation des molécules d'eau était particulièrement bien adaptée à nos systèmes (complexe erbin PDZ/Smad3 MH2 ou complexe protéine/ARN) car l'assemblage de ceux-

ci semble être guidé par les forces électrostatiques. Or, les molécules d'eau peuvent, dans ce cas, servir de relai pour les interactions longues distantes entre résidus polaires ou chargés (Hildebrandt *et al.*, 2007; Mackerell et Nilsson, 2008).

Le calcul de simulations de dynamiques moléculaires en solvant explicite reste encore une méthode très coûteuse. En effet, l'ajout du solvant augmente considérablement le nombre de particules du système : pour exemple, le complexe protéine/ARN ne compte que $\approx$ 3500 atomes ; l'ajout des molécules d'eau et des contre-ions fait passer le système à $\approx$ 94500 atomes. De ce fait, 80 à 90% du temps de calcul est dédié aux interactions entre les molécules constituant le solvant.

Néanmoins, nous avons maintenant accès à de plus en plus de puissance via les *clusters* de PC, grilles de calculs ou plus récemment les nouvelles utilisations des cartes graphiques. Des travaux sont d'ailleurs entrepris pour utiliser la pleine puissance des processeurs graphiques (*Graphic Processor Unit* ou GPU) et permettent déjà de réduire considérablement le temps de calculs (Stone *et al.*, 2007). Ceci nous permet de lancer des simulations plus poussées dans un laps de temps raisonnable ce qui n'aurait pu être le cas sur une seule machine (même très puissante). Les simulations de DMSE sur de gros systèmes seront bientôt calculables en temps réel ouvrant la voie à la dynamique moléculaire interactive (Grayson *et al.*, 2003).

Mais, si l'on arrive à modéliser la flexibilité des systèmes et à interagir avec ceux-ci en temps réel, les déformations de surfaces moléculaires restent encore très difficiles à visualiser de manière interactive. Nous allons voir, dans la partie suivante, comment l'emploie des GPU et de la méthode de lancer de rayons permet d'accélérer ce rendu de surface.

Chapitre 3

MetaMol : nouvelle approche pour la visualisation moléculaire interactive

Sommaire

Contexte

Comme nous l'avons vu dans les parties précédentes, les données issues de la littérature aident particulièrement à la construction d'un modèle de complexe. Il reste néanmoins encore très difficile d'inclure correctement ces données pour contraindre la recherche d'assemblages. Les solutions employées au cours des travaux présentés dans ce manuscrit restent très sommaires. En effet, nous avons sélectionné visuellement les assemblages les plus intéressants en fonction des connaissances que nous avions de l'interaction. Ce système a pourtant bien fonctionné sur les cas présentés précédemment (le complexe Erbin/Smad3 et la cible 34 du challenge CAPRI). Nous avons même montré que nous pouvions nous passer de programmes de docking dans le cas du complexe Erbin/Smad3 et qu'une étude *a posteriori* en utilisant divers serveurs nous permettait d'obtenir un modèle de complexe équivalent à celui créé "à la main".

Il est bien sûr impensable de se passer complètement des programmes de docking mais il serait intéressant d'ajouter une étape supplémentaire pour valider *de visu* les modèles obtenus et, à la manière de l'artisan finalisant son oeuvre, le bioinformaticien structural pourrait raffiner interactivement les partenaires constituant le complexe. Ainsi, il manipulerait une boucle flexible ou bougerait des chaînes latérales à l'interface, tout ceci dans le but de régler les derniers détails du complexe et d'adapter celui-ci à la vision qu'il se fait de l'assemblage (après avoir synthétisé les données bibliographiques sur le sujet). Bien sûr, l'homme reste limité par sa capacité à traiter un nombre d'actions (de calculs) très réduit par seconde, à l'inverse de l'ordinateur. C'est pourquoi, cette étape de raffinage manuel ne peut venir qu'au terme d'un processus permettant d'éliminer la majorité des complexes pour ne garder que les plus prometteurs.

Cette idée de manipulation interactive a déjà été développée dans différents travaux présentés ci-dessous.

3.1 Visualisation et interactivité au service de la bioinformatique structurale

3.1.1 Une brève histoire de la visualisation moléculaire

En quelques dizaines d'années la visualisation moléculaire a considérablement évolué. Les premières représentations de modèles moléculaires, à la fin des années 50, se rapprochaient plus du jeu de "mécano" géant que de la visualisation moléculaire (voir figure 3.1). L'un des plus ancien modèle moléculaire, celui de la myoglobine (Kendrew *et al.*, 1958) réalisé par Sir John Kendrew, était constitué d'un grand nombre de tubes, de fils de fer et de sphères [21].

A la fin des années 60, la première tentative de visualisation électronique de modèle moléculaire utilisait un oscilloscope contrôlé par ordinateur pour visualiser l'image de la structure d'une protéine (Levinthal, 1966). Il était possible de faire apparaître ou disparaître les chaînes latérales ou de mettre la molécule en rotation[22].

Une fois que Levinthal et ses collègues eurent démontré l'utilité de la visualisation électronique, d'autres groupes de recherche essayèrent de développer leurs propres systèmes de visualisation. Au milieu des années 70, une première société spécialisée dans le matériel informatique scientifique dédié à la visualisation fut créée : *Techtronics*. Puis sont venues d'autres sociétés proposant du matériel plus performant comme *Evans & Sutherland* ou *Megatek*. Milieu des années 80, l'entreprise *Silicon Graphics* fut créée et devint très rapidement le leader du domaine.

Cette suprématie a perduré jusqu'à la fin des années 90 où les PC, ou *Personal Computer*, ont fait leur apparition. Ceux-ci ont rapidement comblé le retard avec les puissantes stations de

[21] une image de ce modèle est visible à l'adresse : `http://en.wikipedia.org/wiki/John_Kendrew`

[22] une photographie du système est visible à l'adresse : `http://www.umass.edu/molvis/francoeur/levinthal/lev-index.html` et des vidéos sont disponibles sur le site : `http://www.umass.edu/molvis/francoeur/movgallery/moviegallery.html`

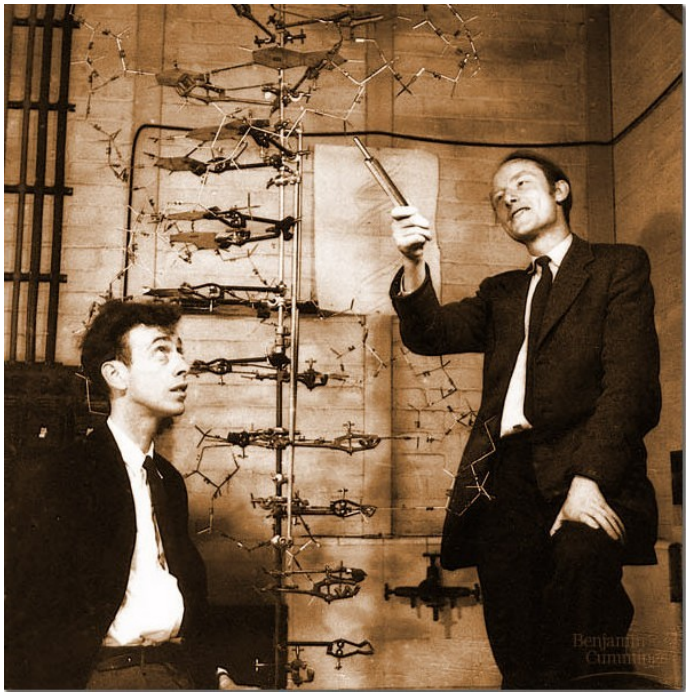


FIG. 3.1 – James Watson et Francis Crick posant devant leur modèle de molécule d'ADN.

travail développées par l'entreprise *Silicon Graphics* grâce, en particulier, aux développements de nouveaux matériels de calcul et de visualisation dédiés au jeu vidéo.

Enfin, même si ces dernières années la visualisation sur ordinateur a remplacé les modèles physiques (constitués de pièces en plastique ou en fer) dans le domaine de la recherche, ceux-ci restent largement utilisés dans l'enseignement. Par exemple, les olympiades de modélisation moléculaire[23] proposent aux étudiants participants de recréer la structure de protéines à partir de matériaux déformables[24]. Récemment, Gillet *et al.* ont assemblé un modèle physique, obtenu avec une imprimante 3D, avec son partenaire visualisé sur l'écran de l'ordinateur grâce à la technique de réalité augmentée (Gillet *et al.*, 2005). Ainsi, les avancées technologiques permettent maintenant de combiner réel et virtuel dans le but de rendre la visualisation moléculaire la plus conviviale possible.

3.1.2 Visualisation : de la molécule à la cellule

Il est possible de visualiser les molécules par diverses représentations. On peut citer, par exemple, la représentation des atomes par des sphères et les liaisons entre ceux-ci par des cylindres connue sous le terme *Ball and Stick*, la représentation des structures secondaires ou encore la visualisation de la surface moléculaire (Goodsell, 2005; Goddard et Ferrin, 2007). Ces types de représentation sont, en général, bien adaptées pour visualiser une partie ou la totalité d'une pro-

[23] `http://education.pdb.org/olympiad/index.html`
[24] `http://www.3dmoleculardesigns.com/news2.php#aminoacid`

téine mais se révèlent peu appropriés pour la visualisation de grands assemblages moléculaires. Il faut alors faire appel à des représentations simplifiées grâce à des méthodes approximant la forme des sous-unités comme présenté figure 3.2A (Goddard *et al.*, 2005).

Afin d'améliorer l'aspect des images de synthèse , une technique possible consiste à calculer une grandeur en chaque pixel, qui correspond à l'occlusion de l'éclairage "ambiant" (Kontkanen et Laine, 2005). Intuitivement, en supposant que la lumière provient uniformément de toutes les directions (éclairage "ambiant"), ceci revient à calculer en chaque point la quantité de lumière effectivement reçue par le point, à savoir la proportion de directions n'intersectant pas l'objet. Ceci permet de renforcer les reliefs de l'objet en rendant plus sombres les zones situées au fond des cavités. Récemment, cette technique a été implémentée dans le programme QuteMol (Tarini *et al.*, 2006) (voir figure 3.2B).

Une autre technique permettant de mieux appréhender les formes tridimensionnelles des objets est l'ajout des contours plus marqués lorsqu'une différence de profondeur apparaît dans l'image (Deussen et Strothotte, 2000). Ceci permet de faire une différence entre les parties située à l'avant de la scène de celles placées en arrière plan. Cette technique, en combinaison avec l'aplat de couleur, est employée par David S. Goodsell pour ses articles web mensuels : *Molecule of the Month*[25]. Cette représentation couplée à une vue *métaphorique* des machineries cellulaires permet de révéler l'intérieur des cellules dans toute leur complexité (voir figure 3.2C et Goodsell (2005)).

Pour plus d'informations sur la visualisation moléculaire, le lecteur peut consulter la section 7 du livre *Structural bioinformatics*(Bourne et Weissig, 2003) ou la publication Goddard et Ferrin (2007) ainsi que le site : *Molecular Surfaces : A Review*[26].

3.1.3 Mise en place d'outils interactifs

Nous avons montré, dans le chapitre précédent, l'utilité des simulations de dynamique moléculaire. Comme nous l'avons fait remarquer précédemment, un problème de la dynamique moléculaire est son échantillonnage sur une fenêtre de temps assez courte (de l'ordre de quelques nano-secondes). Or, certains mécanismes biologiques requièrent un laps de temps plus long. Pour répondre à cette problématique, de nouvelles techniques ont été mises en place comme la *steered molecular dynamic*. Cette méthode consiste à guider les simulations de dynamique moléculaire en accélérant artificiellement le processus biologique, ceci par l'ajout de forces extérieures (Isralewitz *et al.*, 2001a). Elle permet également d'étudier les forces mises en jeu, par exemple, lors du décrochage d'un ligand de son site actif ou lors de la déstructuration d'une protéine (Isralewitz *et al.*, 2001b). Dernièrement, la *steered molecular dynamic* a été utilisée pour raffiner

[25] `http://www.rcsb.org/pdb/static.do?p=education_discussion/molecule_of_the_month/alphabetical_list.html`

[26] `http://www.netsci.org/Science/Compchem/feature14.html`

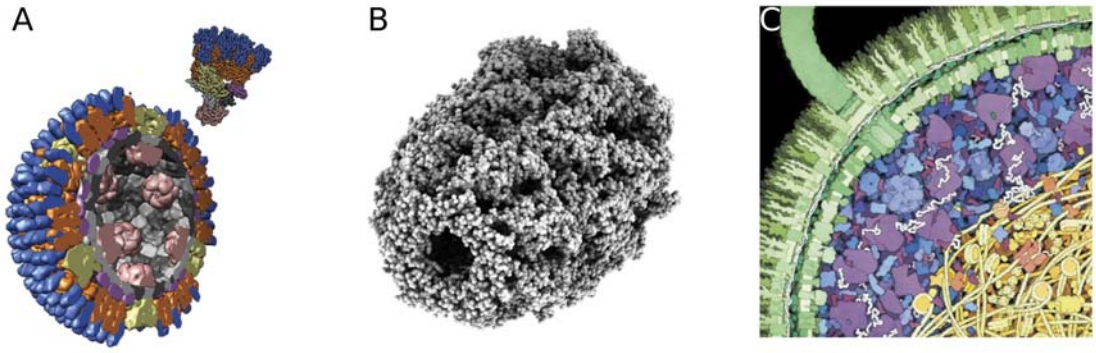


FIG. 3.2 – A- Surface basse résolution, pour représenter la capside d'un virus, implémentée dans le logiciel UCSF Chimera (Goddard *et al.*, 2005). B- *Ambient Occlusion*, pour visualiser la forme de la protéine GroEL, intégré au programme QuteMol (Tarini *et al.*, 2006). C- Représentation d'une coupe de la la bactérie *E. Coli* avec aplat de couleurs et mise en évidence des contours (Goodsell, 2005). *Figure modifiée de Goddard et Ferrin (2007)*

les solutions de docking (Heifetz *et al.*, 2007).

Les chercheurs peuvent également agir directement sur la dynamique grâce à des simulations d'*interactive molecular dynamic* (IMD). Dans ce cas, les forces extérieures sont ajustées de façon continue par l'utilisateur qui guide la dynamique grâce à un bras à retour de force (voir figure 3.3). Cette méthode, employée pour modéliser le passage d'une petite molécule à travers un canal membranaire, permet un gain de temps important en comparaison d'une dynamique moléculaire sans contraintes (Grayson *et al.*, 2003).

Une étape supplémentaire a été franchie avec le programme SAMSON réalisé par l'équipe de Stéphane Redon[27]. Cette équipe a développé une nouvelle méthode, *adaptative torsion-angle quasi-statics*, permettant de déterminer les zones de flexibilité de la molécule de manière automatique. Cette approche permet de limiter les calculs à certaines zones subissant des changements conformationnels tandis que les régions éloignées sont considérées comme rigides (Rossi *et al.*, 2007). Ceci permet de lancer des dynamiques interactives sur le processeur d'un seul ordinateur alors qu'il aurait fallu une puissance de calcul bien plus importante dans le cas d'une dynamique interactive classique. Cette méthode utilise, pour l'instant, les champs de forces CHARMM19 et CHARMM22.

Ces dernières années, la modélisation, et en particulier la visualisation, moléculaire a profité des avancées technologiques faites par l'industrie du jeu vidéo comme la puissance toujours accrue des cartes graphiques. Elle en reprend désormais les codes : en effet, le programme *Fold it*[28], développé par l'équipe de David Baker, laisse le soin aux joueurs du monde entier de trouver

[27] `http://nano-d.inrialpes.fr/?page_id=21`
[28] `http://www.fold.it/`

FIG. 3.3 – Démonstration d'IMD dans le *reality center* du LORIA. *Photo issue du rapport 2006 du CRVHP* : `http://crvhp.loria.fr/pages/`

la structure de plus basse énergie de plusieurs protéines (les résultats de CASP8 montrent le succès de cette méthode[29]). Les "scores" et les structures associées de chaque joueur sont ensuite stockés dans une base de données pouvant être ensuite réutilisées pour entraîner les programmes de repliements.

Ainsi, de nouveaux outils se développent pour rendre la biologie moléculaire structurale plus conviviale. Néanmoins, il reste encore une étape supplémentaire à réaliser pour que ceux-ci puissent vraiment servir dans le domaine du docking macromoléculaire : la visualisation des déformations de la surface moléculaire en temps réel.

3.2 Metamol : visualisation haute-qualité de la surface moléculaire

La création du programme MetaMol a été réalisée grâce à une collaboration entre les équipes Orpailleur et ALICE du LORIA et plus particulièrement avec Bruno Levy, Directeur de Recherche INRIA et responsable scientifique de l'équipe ALICE. Ce travail a fait l'objet d'une publication dans le journal : "*Journal of Molecular Graphics and Modelling*" (voir annexe E). MetaMol est un programme permettant de générer des images de haute qualité en temps réel. A l'inverse des programmes précédents approximant la surface moléculaire par des triangles ou une grille, MetaMol utilise et visualise l'équation de celle-ci. Ceci est rendu possible grâce à un algorithme de lancer de rayons et aux potentialités des nouvelles cartes graphiques. Cette nouvelle méthode permet d'avoir une qualité de rendu graphique et des performances supérieures à

[29] http ://fold.it/portal/node/729520

celles des techniques employées jusqu'à présent.

3.2.1 Définition des différents types de surfaces moléculaires

Visualiser la surface des molécules est très important pour comprendre leur fonctionnement ainsi que leurs interactions. Il existe plusieurs types de surfaces moléculaires définis ci-dessous (voir figure 3.4) :

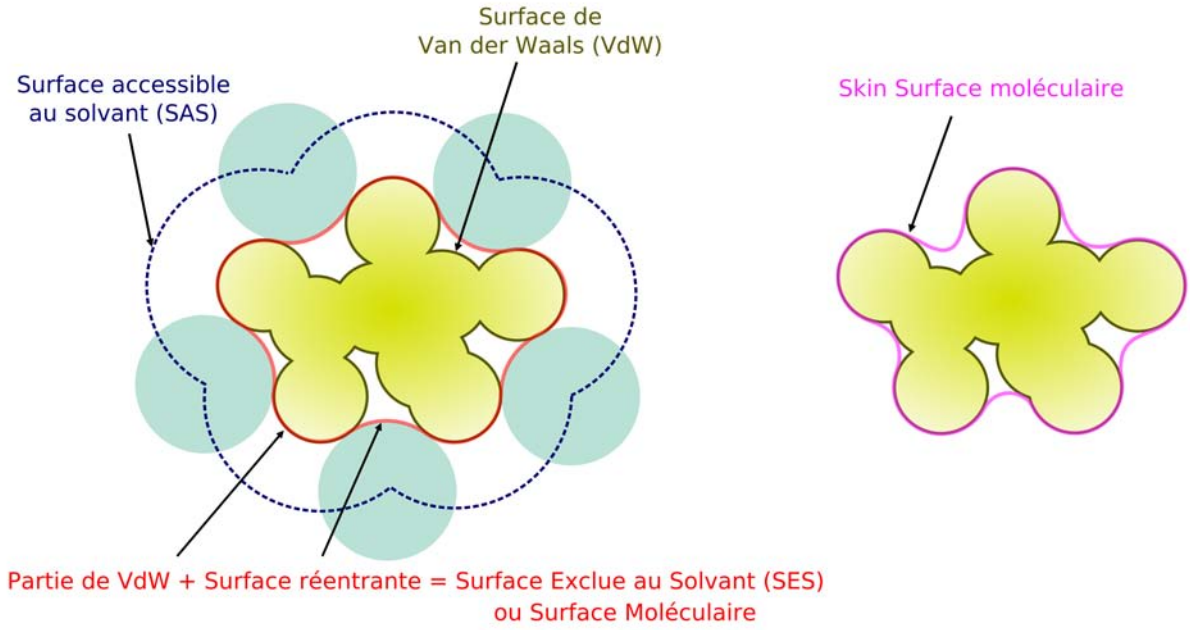


FIG. 3.4 – Représentation des différents types de surfaces moléculaires en deux dimensions.

- La première surface définie fut la surface de Van der Waals (VdW). Il s'agit de l'union des sphères atomiques. Ce type de surface donne une bonne approximation de la surface moléculaire pour les petites molécules mais devient rapidement trop complexe pour représenter de grosses molécules à cause d'un grand nombre d'interstices (Chapman et Connolly, 2001).
- Pour décrire la surface des macromolécules et leurs interactions potentielles avec les molécules d'eau, Lee et Richards définirent la surface accessible au solvant ou SAS (Lee et Richards, 1971). Elle est définie comme la trajectoire du centre d'une sphère (*probe sphere*) lorsque celle-ci roule sur la surface de Van der Waals. Cette sphère peut représenter une molécule d'eau, c'est pourquoi le rayon de celle-ci est le plus souvent égal à 1,4 Å (rayon de la molécule d'eau).
- Quelques années plus tard, Richards définit une surface moléculaire plus lisse (Richards, 1977). Cette surface est constituée de deux parties : la surface de Van des Waals directement en contact avec la *probe sphere* et la surface "réentrante". Cette dernière est constituée par

la surface de la *probe sphere* lorsque celle-ci roule entre deux atomes. Malheureusement, cette nouvelle surface, définie par Richards, souffrait d'auto-intersections. C'est pourquoi Greer et Bush définirent la surface exclue au solvant ou SES (Greer et Bush, 1978). Michael L. Connolly développa une méthode permettant de calculer analytiquement cette surface, plus connue sous le terme de *surface moléculaire* (Connolly, 1983).

– Plus récemment, dans le contexte plus général de la géométrie algorithmique, Herbert Edelsbrunner a défini une surface adoucie : la *Skin Surface* (Edelsbrunner, 1999). Cette surface - que nous définirons dans la section suivante - peut être calculée à partir de sphères, c'est pourquoi on peut l'utiliser pour représenter la surface moléculaire (en utilisant les sphères de Van der Waals). Elle est alors appelée "*Skin Surface* Moléculaire" (ou *Molecular Skin Surface*). Cette surface est proche de la surface moléculaire tout en possédant des propriétés plus intéressantes que cette dernière comme nous l'expliquerons ultérieurement.

De nombreux travaux, à commencer par l'algorithme fondateur de Connolly (Connolly, 1985), ont été dédiés à l'amélioration des méthodes de calcul et de représentation de la *surface moléculaire*. En 1994, Varshney *et al.* développèrent un programme facilement parallélisable (Varshney *et al.*, 1994). Quelques années plus tard, M. Sanner détermina une méthode basée sur les "surfaces réduites" (Sanner *et al.*, 1996) pour visualiser de gros assemblages (des molécules d'une taille allant jusqu'à 10000 atomes). Plus récemment, Can *et al.* proposèrent un algorithme utilisant les méthodes de Level-Set pour générer une surface moléculaire (Can *et al.*, 2006) et Bates *et al.* définirent une surface moléculaire minimale (Bates *et al.*, 2007). Ces deux dernières méthodes utilisent une grille et la technique de *Marching Front* pour représenter la *surface moléculaire* et les cavités de la molécule.
Toutes ces méthodes sont très efficaces mais souffrent d'un problème de précision : pour les algorithmes de Varshney et Sanner, la surface moléculaire est triangulée tandis que pour les algorithmes de Can et Bates, la surface est représentée comme l'union de cubes (sur une grille), d'où l'existence d'un niveau de zoom où les triangles (ou les cubes) apparaissent.

Quelques travaux représentent la *Skin Surface* en la triangulant (Kruithof et Vegter, 2004; Cheng et Shi, 2004, 2005). Cette méthode a deux inconvénients :

1. Il est nécessaire de s'assurer que la topologie de la surface est bien préservée, ce qui rend l'algorithme compliqué et lent.

2. Il existe, là encore, un niveau de zoom où les triangles générés provoquent des artefacts (voir figure 3.16B).

Pour dépasser ces problèmes de précision, nous proposons l'utilisation de la technique de lancer de rayons calculée sur processeurs graphiques (GPU). Cette méthode utilise directement l'équation de la *Skin Surface* afin d'obtenir des images d'une précision de l'ordre du pixel. La méthode de lancer de rayons sur GPU a déjà été utilisée pour représenter des modèles moléculaires simples comme la représentation "CPK" ou "*Balls and Sticks*" (Toledo et Levy, 2004; Sigg *et al.*, 2006). Le travail présenté dans cette thèse traite un cas bien plus complexe : la représentation de la *Skin Surface* Moléculaire.

3.2.2 Comparaison de la *Skin Surface* Moléculaire et la Surface Moléculaire

Nous aurions pu développer un programme de visualisation de la Surface Moléculaire. Nous avons plutôt pris le parti de développer un programme de visualisation de la *Skin Surface* Moléculaire. Nous expliquons, dans cette section, quelles en sont les raisons.

Ces deux types de surfaces sont définies par morceaux (voir figure 3.5A). Chacune peut être modélisée comme l'union de différentes formes définies par une équation simple. La Surface Moléculaire est définie par 3 types de morceaux : des morceaux de sphères ou de tores joints par des arcs circulaires (Connolly, 1983). De même, la *Skin Surface* Moléculaire peut être aussi définie par 3 types de régions : des bouts de sphères, d'hyperboloïdes de révolution à une nappe ou à deux nappes (voir figure 3.8).

Il est possible d'obtenir, dans ces deux cas, une surface ayant plus ou moins de détails, ceci en faisant varier un facteur. Il s'agit du rayon de la *probe sphere* pour la Surface Moléculaire et du facteur de réduction (ou *shrink factor* - que nous détaillerons par la suite) pour la *Skin Surface* Moléculaire.

Par contre, la manière de construire chaque surface diffère totalement. La Surface Moléculaire se construit en utilisant la *probe sphere* pour définir les morceaux (Connolly, 1983) tandis que les morceaux de la *Skin Surface* Moléculaire sont définis à partir d'une sous-structure (le *Mixed Complex*) composée par le diagramme de Voronoï et la tétraédrisation de Delaunay de la molécule (Edelsbrunner, 1999).

De plus, la *Skin Surface* Moléculaire possède des propriétés plus intéressantes que celles de la Surface Moléculaire (Edelsbrunner, 1999). D'un point de vue géométrique, la *Skin Surface* est une surface C1 : elle est continue et sa dérivée première l'est aussi. Ceci est clairement un avantage en comparaison de la Surface Moléculaire (voir figure 3.5B) qui possède de nombreuses singularités (comme des zones d'interpénétrations) et n'est donc pas une surface continue (Sanner *et al.*, 1996; Vorobjev et Hermans, 1997; Geng *et al.*, 2007; Bates *et al.*, 2007). Ces singularités ne permettent pas un calcul rapide de propriétés comme le potentiel électrostatique (en résolvant l'équation de Poisson-Boltzmann) ou le calcul d'effets de solvatation implicites (Zauhar, 1995; Vorobjev et Hermans, 1997; Bates *et al.*, 2007). Les propriétés de la *Skin Surface* évoquées précédemment font, potentiellement, de celle-ci une surface plus adaptée à de tels calculs. Les calculs de potentiels électrostatiques sur ce type de surface n'ont cependant pas encore été réalisés et une étude approfondie sera nécessaire pour valider cette hypothèse.

La *Skin Surface* Moléculaire permet aussi de générer simplement une Surface Accessible au Solvant lissée : il suffit d'ajouter le rayon de la *probe sphere* au rayon de Van der Waals de chaque atome.

Une autre propriété très intéressante de la *Skin Surface* est la possibilité de déformer celle-ci et de calculer relativement facilement ces déformations (Edelsbrunner, 1995, 1999; Cheng *et al.*, 2001).

Enfin, d'un point de vue calculatoire, la *Skin Surface* est définie uniquement par des équations de degré-2 (équation d'une sphère ou d'hyperboloïdes de révolution à une ou deux nappes) tandis que la Surface Moléculaire est définie par des équations de degré 2 (équation de la sphère) mais aussi par des équations de degré 4 (équation du tore). De ce fait, les calculs de la *Skin Surface*

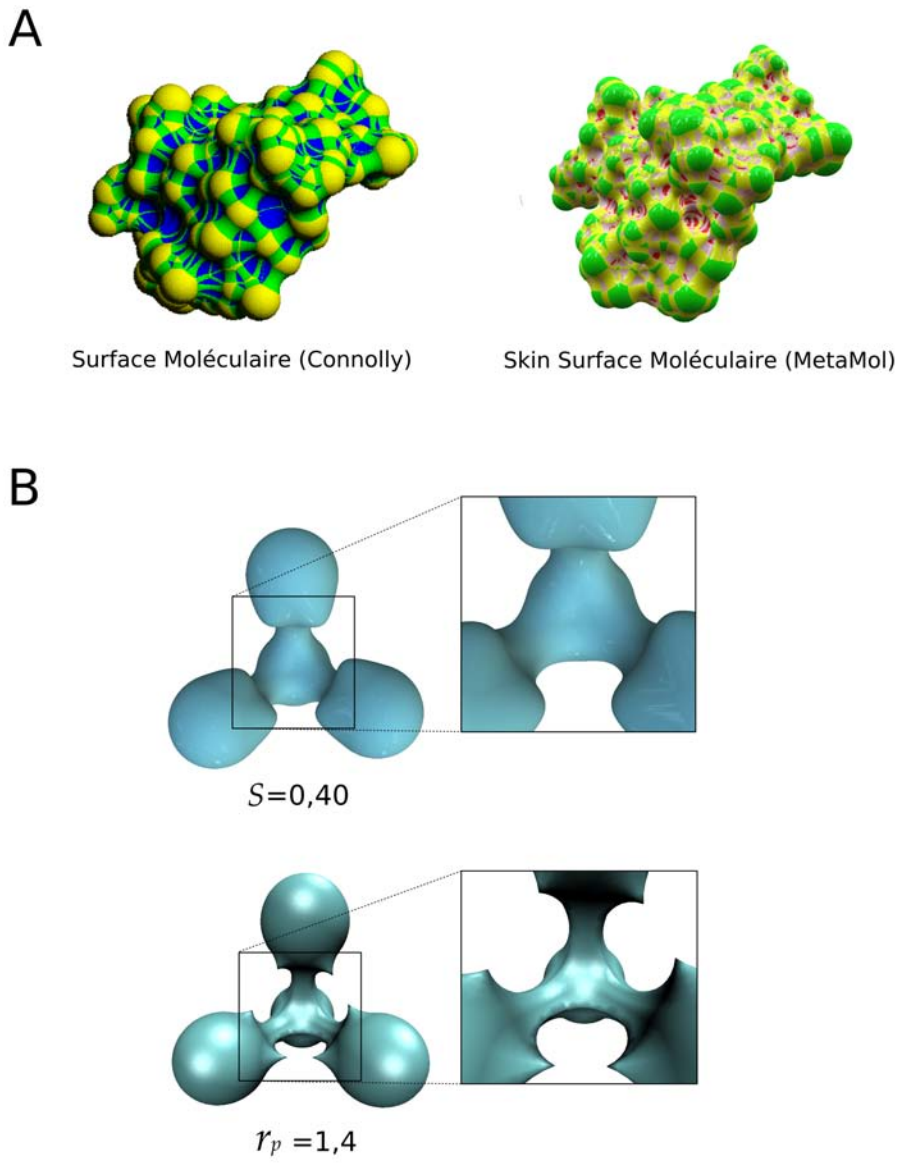


FIG. 3.5 – A- Visualisation des différents morceaux constituant la Surface Moléculaire et la *Skin Surface* Moléculaire (molécule de crambine). B- Comparaison de la *Skin Surface* Moléculaire (en haut) obtenue avec MetaMol et de la Surface Moléculaire (en bas) obtenue avec le programme MSMS (inclus dans le programme VMD (Humphrey *et al.*, 1996)). S représente le *Shrink Factor* et r_p le rayon de la *probe sphere*.

Moléculaire peuvent être plus rapides que ceux de la Surface Moléculaire.

Ainsi, la *Skin Surface* Moléculaire est potentiellement une surface très intéressante sur bien des points et pourrait, dans les prochaines années, être appelée à jouer un rôle aussi (voire plus) important que la Surface Moléculaire. Néanmoins, avant cela, une étude approfondie comparant clairement les deux surfaces est nécessaire.

3.2.3 Définition et construction de la *Skin Surface* Moléculaire

Dans cette partie, nous allons brièvement présenter la *Skin Surface* Moléculaire et la méthodologie pour construire celle-ci. Il est possible de se référer à la publication d'Herbert Edelsbrunner (Edelsbrunner, 1999) pour l'introduction originelle de cette surface ou à Cheng et Shi (2004); Kruithof et Vegter (2004, 2007) pour une explication plus détaillée de celle-ci.

Définition : La *Skin Surface* est définie par plusieurs paramètres :

- un ensemble de points pondérés (P)

$$P = \{p_i = (z_i, w_i) \text{ dans } \mathbb{R}^3 \times \mathbb{R} \mid i = 1, \ldots, n\}$$

  où p_i est un point pondéré avec z_i ses coordonnées cartésiennes et w_i son poids.
- un facteur de réduction (ou *shrink factor*) s, avec $0 \leq s \leq 1$.

La *Skin Surface* $skn^s(P)$ délimite le Corps $bdy^s(P)$ de l'ensemble de points pondérés P :

$$skn^s(P) = \partial bdy^s(P)$$

où $bdy^s(P)$ est défini comme une famille infinie de sphères réduites générée à partir de l'ensemble fini de points pondérés (P) (voir figure 3.6 et la section 4 de Edelsbrunner (1999) pour plus de détails sur le Corps). ∂ dénote la frontière reliant l'ensemble des sphères.

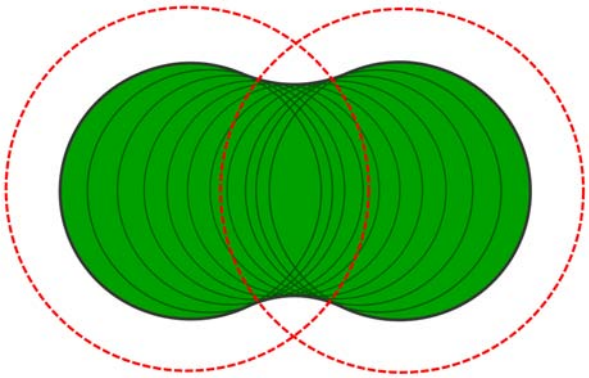


FIG. 3.6 – La *Skin Surface* de deux points pondérés en 2D. Les points pondérés sont représentés par des pointillés rouges. Le Corps (bdy^s) est représenté par un sous-ensemble fini de disques verts et la *Skin Surface* par une courbe noire reliant chaque disque.

Comme nous pouvons le voir figure 3.6, la *Skin Surface* n'enveloppe pas les points pondérés. Ceci est dû à la définition même du poids et à l'utilisation du facteur de réduction dans l'équation de chaque morceau (comme nous le verrons dans la partie suivante).

En fait, il faut voir chaque point pondéré $p(z, w)$ comme une sphère de centre de coordonnées $z \in \mathbb{R}^3$ et de rayon $\sqrt{w}$ avec $w \in \mathbb{R}$. Il faut donc redéfinir le poids pour que la *Skin Surface* puisse envelopper les sphères de Van der Waals (Kruithof et Vegter, 2004) :

$$w_p = \left(\frac{1}{s}\right) \times {r^2}_{vw} \tag{3.1}$$

où s est le facteur de réduction. Ce facteur de réduction se trouve en dénominateur ; il faut donc limiter le domaine d'existence de celui-ci par rapport à la définition générale de la *Skin Surface* : $0 < s \leq 1$. r_{vw} est le rayon de Van der Waals de chaque atome.

Le programme MetaMol peut donc prendre en entrée une liste de points pondérés ou un fichier PDB (Berman *et al.*, 2000). Dans ce dernier cas, les rayons des atomes sont convertis en poids en utilisant l'équation (3.1).

Construction de la *Skin Surface* : Comme nous l'avons expliqué précédemment, la *Skin Surface* est une surface définie par morceaux. Ces morceaux sont obtenus à partir d'une structure appelée le *Mixed Complex*. Ce *Mixed Complex*, associé à un facteur de réduction s, peut être considéré comme une structure intermédiaire entre la tétraédrisation de Delaunay pondérée et son dual le diagramme de Voronoï pondéré.

La tétraédrisation de Delaunay pondérée permet de relier différents points pondérés afin de former une structure en tétraèdres (voir Edelsbrunner et Shah (1992) pour une définition plus précise). Le diagramme de Voronoï pondéré est la structure duale de la tétraédrisation de Delaunay pondérée : chaque noeud du diagramme de Voronoï est le centre de la sphère circonscrite à chaque tétraèdre (voir figure 3.7B). Chaque cellule du *Mixed Complex* peut être vue comme une structure "mixée" des cellules de Voronoï et de Delaunay qui lui sont associées (voir figure 3.7B). Ce type de cellule s'obtient en faisant la somme de Minkowski des cellules de Delaunay et de Voronoï.

$${\mu^s}_X = s \cdot \upsilon_X \ \oplus \ (1 - s) \cdot \delta_X \tag{3.2}$$

où X est un sous-ensemble fini de sphères et ${\mu^s}_X$ représente une cellule "mixée", υ_X une cellule de Voronoï et δ_X une cellule de Delaunay. Le symbole [$\cdot$] dénote la multiplication par un scalaire et [$\oplus$] la somme de Minkowski. Pour une cellule du *Mixed Complex* donnée, la *Skin Surface* Moléculaire est complètement déterminée par, au plus, quatre points pondérés du sous-ensemble X.

Si s tend vers 0, la cellule mixée tend vers la cellule de Delaunay. Lorsque l'on augmente le facteur de réduction, cette cellule se déforme jusqu'à une cellule du diagramme de Voronoï (voir figure 3.7A).

En trois dimensions, le *Mixed Complex* est divisé en 4 parties différentes (voir figure 3.8), chaque partie est obtenue en mixant une cellule de la tétraédrisation de Delaunay avec sa cellule

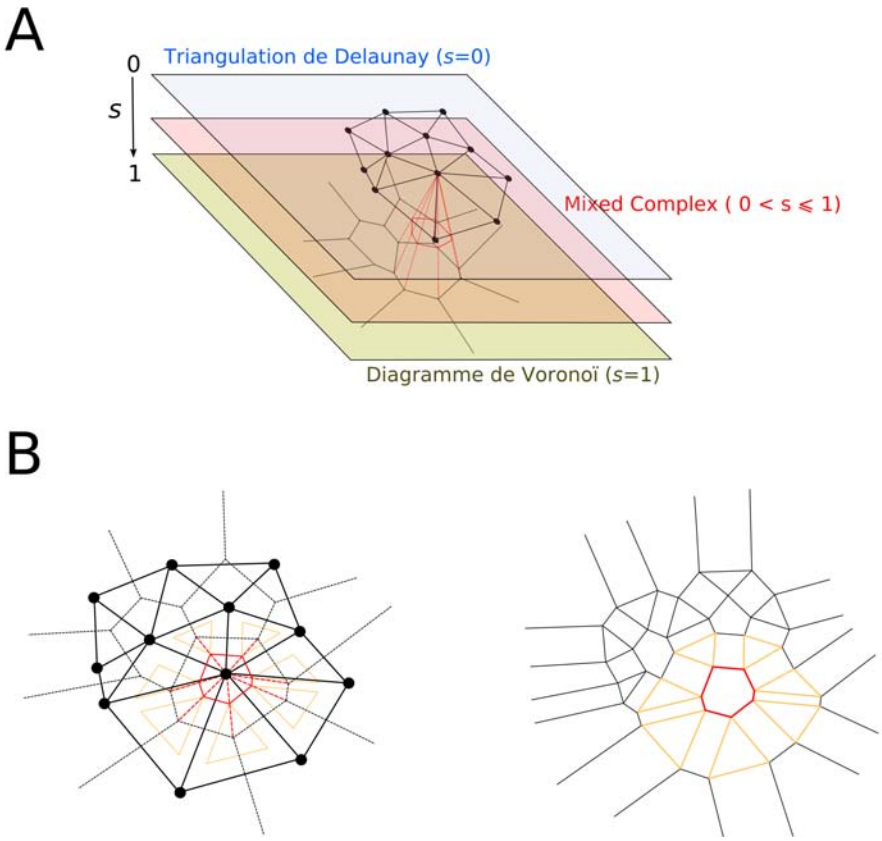


FIG. 3.7 – A- Création, en deux dimensions, d'une cellule du *Mixed Complex* à partir des cellules de Voronoï et de Delaunay associées : le sommet de la triangulation de Delaunay est relié aux noeuds duaux du diagramme de Voronoï. La cellule du *Mixed Complex* est obtenue en faisant varier un plan , le plan du *Mixed Complex*, entre les plans de Voronoï et de Delaunay. B- Vue de dessus du résultat : en rouge, la cellule de Voronoï réduite résultant du *mix* entre une celulle de Voronoï et un sommet de Delaunay. En orange, les triangle réduits résultant du *mix* entre des triangles de Delaunay et des noeuds du diagramme de Voronoï.

duale du diagramme de Voronoï :

- D'abord, les cellules de Voronoï réduites sont obtenues en mixant un sommet de la tétraédrisation de Delaunay avec une cellule du diagramme de Voronoï.
- Ensuite, les patchs H1 sont obtenus en mixant une arête d'un tétraèdre de Delaunay avec une face d'une cellule de Voronoï.
- Puis, les patchs H2 sont créés à partir d'une face d'un tétraèdre avec une arête d'une cellule de Voronoï.
- Enfin, les tétraèdres réduits sont obtenus à partir d'une cellule de Delaunay et d'un noeud du diagramme de Voronoï.

À chaque morceau du *Mixed Complex* est associée une surface quadratique dont l'équation

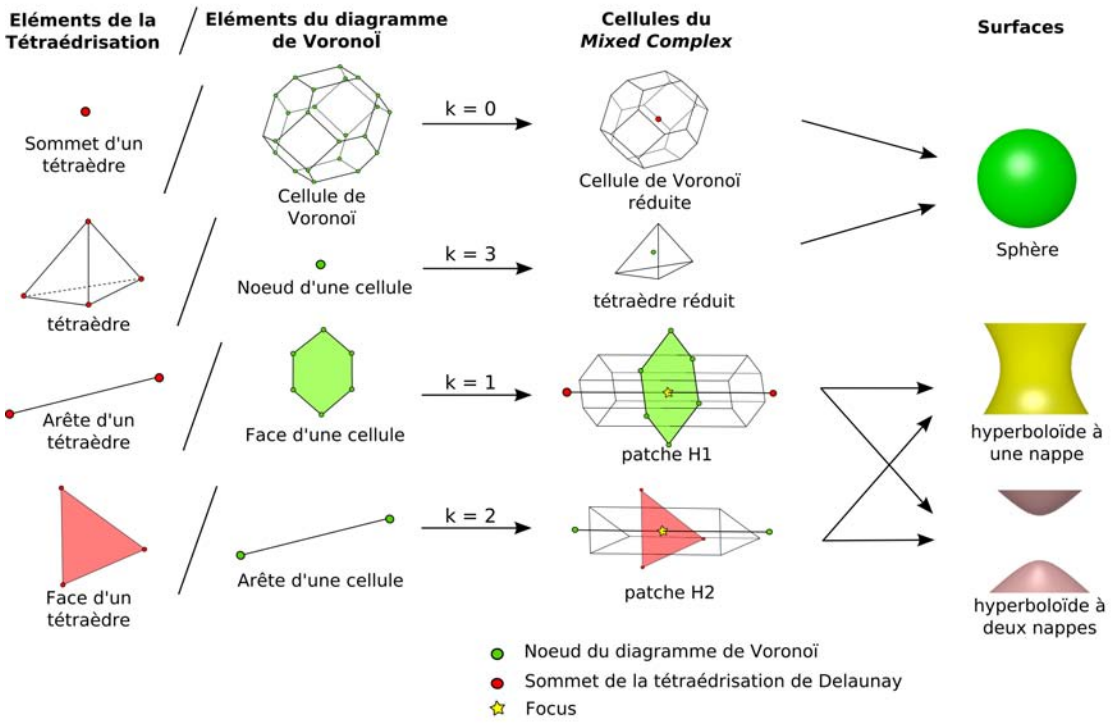


FIG. 3.8 – En mixant les éléments de la tétraédrisation de Delaunay et du diagramme de Voronoï, il est possible d'obtenir de nouvelles cellules : les cellules du *Mixed Complex*. Chaque type de cellule peut couper un morceau de surface, soit de sphère soit d'hyperboloïdes à une ou deux nappes.

générale est :

$$S_X(x) = -\frac{1}{1-s}\sum_{i=1}^{k} x_i^2 + \frac{1}{s}\sum_{i=k+1}^{3} x_i^2 - R^2 = 0 \tag{3.3}$$

avec $x = (x_1, x_2, x_3)$, k est un entier définissant le type de cellule ($k \in [0,3]$)et R^2 le rayon de la surface coupée.

Le rayon R^2 se définit comme le produit de puissance entre un sommet de la tétraédrisation de Delaunay ($p(z,w)$) et le foyer (*focus*), $f(X)$, associé à chaque cellule :

$$R^2 = w - ||z - f(X)||^2$$

où $||z - f(X)||$ est la distance euclidienne entre z et $f(X)$. Ce rayon peut prendre des valeurs positives, négatives ou nulles suivant les valeurs de la distance euclidienne.

Ainsi, à chaque type de cellule mixée sont associés une équation de surface, un type (k) et un foyer (voir figure 3.8).

Posons, $t_1 = -\frac{1}{1-s}$ et $t_2 = \frac{1}{s}$.

La cellule de Voronoï réduite est une cellule de type 0 dont le foyer se définit comme le sommet de la tétraédrisation de Delaunay associé à celle-ci (voir figure 3.8). L'équation générale de la surface (3.3) devient alors :

$$x_1^2 + x_2^2 + x_3^2 = \frac{R^2}{t_2} \tag{3.4}$$

Par définition, pour une cellule de type 0, R^2 est toujours supérieur ou égal à 0. L'équation obtenue est donc l'équation d'une sphère. Cette sphère est centrée sur le sommet de Delaunay associé à cette cellule de Voronoï réduite (voir figure 3.8).

Le patch H1 est une cellule de type 1 dont le foyer se définit comme le point d'intersection entre la face de la cellule de Voronoï et l'arête du tétraèdre de Delaunay "mixés" (voir figure 3.8). L'équation de la surface associée à cette cellule est alors :

$$t_1 x_1^2 + t_2 x_2^2 + t_2 x_3^2 = R^2 \tag{3.5}$$

R^2, pour cette cellule, peut prendre des valeurs positives, négatives ou nulles. L'équation obtenue peut donc être celle d'un hyperboloïde à une nappe ou à deux nappes suivant les valeurs de R^2. Cet hyperboloïde a pour axe de rotation l'arête du tétraèdre de Delaunay mixée pour obtenir ce patch (voir figure 3.8).

Le patch H2 est une cellule de type 2 dont le foyer se définit comme le point d'intersection entre l'arête de la cellule de Voronoï et la face du tétraèdre de Delaunay "mixés" (voir figure 3.8). L'équation de la surface associée à cette cellule est alors :

$$t_1 x_1^2 + t_1 x_2^2 + t_2 x_3^2 = R^2 \tag{3.6}$$

R^2, pour cette cellule, peut prendre des valeurs positives, négatives ou nulles. L'équation obtenue peut donc être celle d'un hyperboloïde à une nappe ou à deux nappes suivant les valeurs de R^2. Cet hyperboloïde a pour axe de rotation l'arête de la cellule de Voronoï mixée pour obtenir ce patch (voir figure 3.8).

Enfin, le tétraèdre réduit est une cellule de type 3 dont le foyer se définit comme le noeud du diagramme de Voronoï associé à celle-ci (voir figure 3.8). L'équation de la surface est alors :

$$x_1^2 + x_2^2 + x_3^2 = \frac{R^2}{t_1} \tag{3.7}$$

Par définition, pour une cellule de type 3, R^2 est toujours inférieur ou égal à 0. L'équation obtenue est donc l'équation d'une sphère. Cette sphère est centrée sur le noeud du diagramme de Voronoï associé à ce tétraèdre réduit (voir figure 3.8).

Ainsi, nous avons défini les différents morceaux du *Mixed Complex* et les équations associées à ceux-ci. Nous allons maintenant décrire les algorithmes permettant de visualiser la *Skin Surface* Moléculaire.

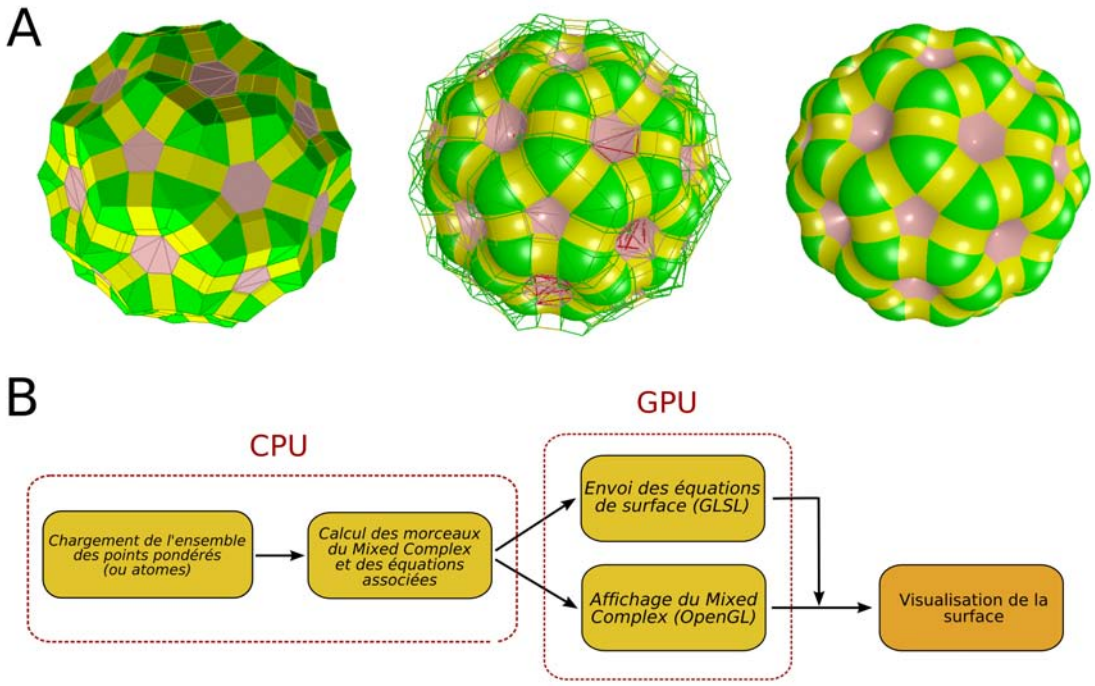


FIG. 3.9 – A- Visualisation de la molécule de fullerène C_{60}. A gauche, représentation de l'enveloppe du *Mixed Complex* ; au centre superposition du *Mixed Complex* et de la surface obtenue par lancer de rayons. A droite, la *Skin Surface* Moléculaire obtenue après lancer de rayons. B- Pipeline général de l'algorithme.

3.2.4 Visualisation de la *Skin Surface* Moléculaire

A la différence des programmes existants, qui n'utilisent en général que le processeur de l'ordinateur (*Central Processor Unit* ou CPU) pour calculer la surface, notre approche divise les calculs (voir figure 3.9 pour une vue d'ensemble) :
– Le *Mixed Complex* est calculé sur le processeur de l'ordinateur (CPU). Le *Mixed Complex* est indépendant du point de vue, il n'est donc calculé qu'une seule fois lors du chargement des données.
– La technique de lancer de rayons est mise en oeuvre sur les processeurs de la carte graphique (*Graphic Processor Unit* ou GPU) pour visualiser la *Skin Surface* Moléculaire.

Calculs sur CPU : Nous allons, dans cette partie, décrire l'algorithme permettant d'obtenir le *Mixed Complex* et les équations associées à chaque morceau du *Mixed Complex*. Il faut souligner que les équations présentées précédemment ne sont valables que dans le cas général où chaque élément de surface (sphère ou hyperboloïdes) est centré à l'origine : O(0,0,0) et orienté suivant l'axe $\vec{Ox_3}$ (dans le cas des hyperboloïdes). De plus, dans le cas général, les cellules de Voronoï périphériques sont des cellules infinies, *i.e.* elles ne sont pas closes (voir figure 3.7). Or l'utilisation

de la technique de lancer de rayons impose de clore ces cellules (voir dessin de l'enveloppe du *Mixed Complex* figure 3.9). Il faut donc, dans un premier temps, fermer ces cellules.

Tous les calculs réalisés sur CPU ont été faits à l'aide de la librairie mathématique CGAL[30].

Dans un premier temps, il est nécessaire de créer une tétraédrisation de Delaunay pondérée à partir d'une liste de points pondérés. La création de cette tétraédrisation de Delaunay pondérée se fait grâce à la classe *Regular_triangulation_3* de CGAL. De plus, chaque point ajouté durant cette phase sera marqué comme étant un atome.

Une fois la tétraédrisation obtenue, il faut clore les cellules de Voronoï infinies.

Fermeture des cellules de Voronoï infinies : Pour fermer les cellules de Voronoï infinies, nous avons développer notre propre méthodologie : nous n'avons pas travaillé sur le diagramme de Voronoï directement mais sur son dual, la tétraédrisation de Delaunay. En effet, clore les cellules de Voronoï revient à créer des cellules de Delaunay supplémentaires (*i.e.* à clore des cellules de Delaunay infinies). Il faut donc ajouter des points à la tétraédrisation de Delaunay. De plus, il nécessaire d'imposer certaines contraintes à ces cellules pour pouvoir ensuite appliquer la technique de lancer de rayons.

La première contrainte est d'obtenir une cellule convexe. Cette contrainte est, par définition, remplie par la cellule de Voronoï réduite. La deuxième contrainte est que cette cellule soit suffisamment grande pour envelopper l'atome à l'intérieur. Si cette contrainte n'est pas remplie, des problèmes de fenêtrage (*clipping*) peuvent apparaître lors de la visualisation de la surface.

Il faut donc, en fait, calculer la position potentielle du centre de la sphère circonscrite à la cellule de Delaunay infinie pour que celui-ci remplisse les contraintes, puis, à partir de ce point, calculer le point à ajouter à la tétraédrisation de Delaunay pour clore la cellule de Delaunay infinie. Le protocole utilisé pour créer un point est présenté figure 3.10 :

1. On cherche une cellule de Delaunay infinie. Une cellule de Delaunay est considérée comme infinie si elle ne possède que 3 points (au lieu de 4). Lorsque l'on trouve une cellule infinie, on récupère la face finie de celle-ci (*i.e.* la face commune avec une cellule finie).
2. On calcule ensuite le dual de cette cellule de Delaunay infinie. Il s'agit d'un rayon dont l'origine est le centre de la sphère circonscrite à la cellule de Delaunay voisine.
3. On calcule le barycentre de la face stockée en 1 et on calcule le vecteur normé à partir du rayon. En positionnant ce vecteur au niveau du barycentre de la face, on obtient un triangle rectangle formé par le barycentre, un des sommets de la face et le centre potentiel de la sphère circonscrite. On calcule ensuite la distance entre le barycentre et le sommet de la face ainsi que la distance entre ce sommet et la position potentielle du centre circonscrit. Ceci permet ensuite de calculer la distance entre le barycentre et la position du centre circonscrit en utilisant le théorème de Pythagore.
4. Les coordonnées du centre circonscrit (*i.e.* le noeud du diagramme de Voronoï) sont ensuite obtenues à partir des coordonnées du barycentre, du vecteur normé et de la distance calculée en 3 :

$$cc = \text{Barycentre } + \overrightarrow{\text{Vecteur Normé}} \times |Barycentre - cc|$$

[30] http ://www.cgal.org/

où cc représente les coordonnées du centre circonscrit.
On calcule ensuite le sommet de Delaunay manquant à partir des sommets de la face stockée en 1 et du centre de la sphère circonscrite calculé en 3. Ce point est ensuite ajouté à la tétraédrisation de Delaunay. Ce point n'est pas marqué comme étant un atome.

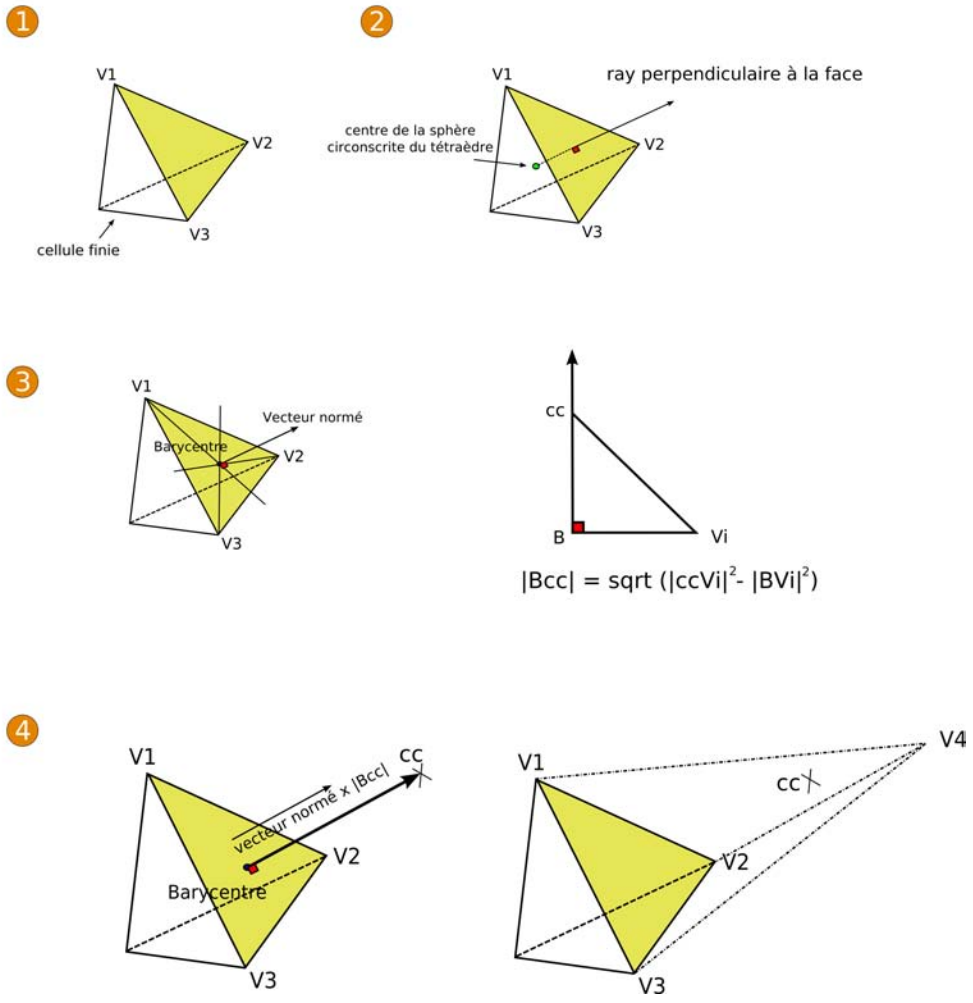


FIG. 3.10 – Visualisation des différentes étapes pour ajouter un point à la tétraédrisation de Delaunay. En jaune, face d'une cellule finie adjacente à une cellule infinie.*Les chiffres correspondent aux différentes étapes énoncées dans le paragraphe précédent.*

Ce processus est réitéré tant qu'il reste des cellules de Delaunay infinies composées d'au moins un sommet marqué comme atome.

Algorithme 1 : Fermeture des cellules

tant que ($Nb.cellule\_inf(atome) \neq 0$) {
Parcourir les cellules
  si (cellule est infinie) {
    Ajouter un point
    Marquer ce point comme non_atome
  }
}

Stockage des différents morceaux du *Mixed Complex* et des équations associées : En remarque préalable à cette partie, il est intéressant de noter que l'équation d'une surface du second degré peut se mettre sous la forme matricielle :

$$\begin{aligned} f(x,y,z) =& Ax^2 + 2Bxy + 2Cxz + 2Dx + Ey^2 + \\ & 2Fyz + 2Gy + Hz^2 + 2Iz + J = 0 \end{aligned} \tag{3.8}$$

Les 10 coefficients peuvent être disposés dans une matrice 4×4 symétrique notée $\mathbf{Q}$ et l'équation (3.8) peut être réécrite comme :

$$\begin{bmatrix} x & y & z & 1 \end{bmatrix} \begin{bmatrix} A & B & C & D \\ B & E & F & G \\ C & F & H & I \\ D & G & I & J \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = 0$$

Ainsi, chaque équation définissant une partie de la *Skin Surface* Moléculaire pourra être stockée sous la forme d'une matrice 4×4.

L'algorithme de création et de stockage des différentes parties du *Mixed Complex* et de leurs équations associées peut être divisé en 4 fonctions (voir figure 3.11).

- D'abord, le calcul de tous les points constituant le *Mixed Complex* ainsi que de tous les points constituant le diagramme de Voronoï : ceci se fait lors d'une première passe sur l'ensemble des cellules finies de la tétraédrisation de Delaunay. Ce calcul est couplé au stockage des points de chaque tétraèdre réduit (cellule de type 3) ainsi que de leur équation associée. Cette étape doit être réalisée en premier puisqu'elle permet de calculer tous les points du *Mixed Complex*. Les 3 étapes suivantes sont indépendantes et peuvent être réalisées dans n'importe quel ordre.
- Stockage des points des cellules de Voronoï réduites (cellule de type 0) et de leur équation associée.
- Stockage des points des patchs H1 (cellule de type 1) et de leur équation associée.
- Stockage des points des patchs H2 (cellule de type 2) et de leur équation associée.

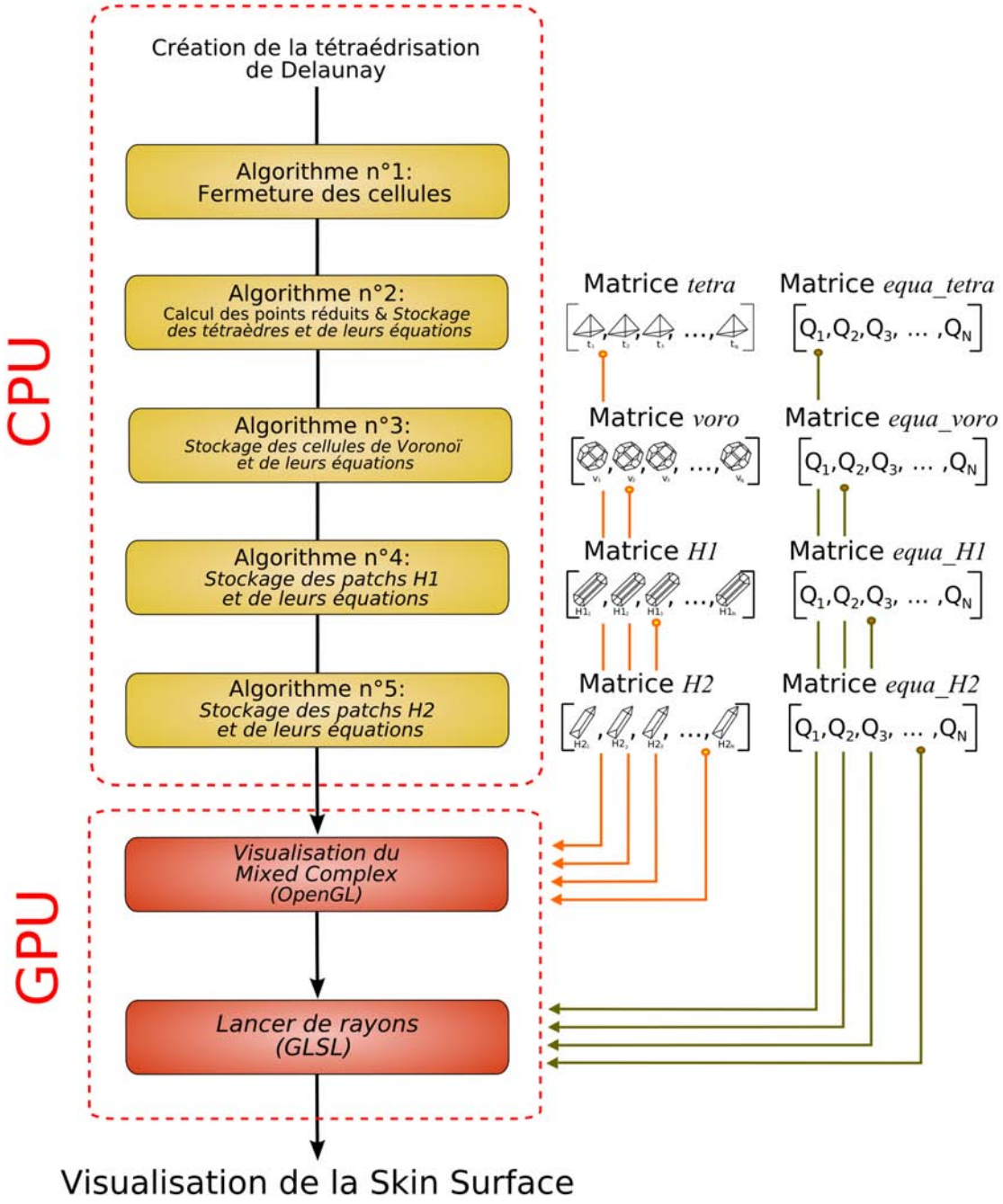


FIG. 3.11 – Présentation détaillée de l'algorithme général. A droite de l'algorithme, les différentes matrices créées à chaque étape. Les matrices de points (M_{tetra}, M_{voro}, M_{H1} et M_{H2}) seront utilisées pour visualiser l'enveloppe du *Mixed Complex* avec OpenGL. Les matrices d'équations (M_{equa_tetra}, M_{equa_voro}, M_{equa_H1} et M_{equa_H2}) serviront au lancer de rayons pour afficher la *Skin Surface* Moléculaire à partir de l'enveloppe du *Mixed Complex*.

Le premier algorithme pourrait se résumer comme suit :

Algorithme 2 : Calcul des points réduits & stockage des tétraèdres réduits et de leurs équations

pour (chaque cellule finie) {
Calcul du centre circonscrit
Calcul du tétraèdre réduit

Stockage du centre circonscrit dans une extension de la cellule
Stockage de chaque sommet "réduit" dans une extension de la cellule

si (cellule ne contient que des atomes) {
Calcul de l'équation
Stockage des sommets du tétraèdre réduit dans une matrice M_{tetra}
Stockage de l'équation associée dans une matrice M_{equa_tetra}
}
}

Détaillons maintenant les calculs effectués pour une cellule de Delaunay donnée. Le calcul des centres circonscrits se fait en utilisant la fonction *dual* de CGAL qui prend une cellule de Delaunay en entrée. Pour tout sommet V_{Di} (avec $i \in [0,3]$) de la cellule, on peut calculer son point réduit en fonction du centre (cc) de la sphère circonscrite au tétraèdre et du facteur de réduction s (voir exemple figure 3.12) :

$$V'_{Di} = V_{Di} + s(cc - V_{Di})$$

où V'_{Di} est le point réduit associé à V_{Di}.

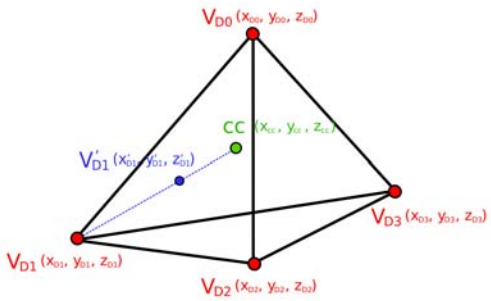


FIG. 3.12 – Cellule de Delaunay (tétraèdre). En rouge, les points pondérés, sommets du tétraèdre. En vert, centre de la sphère circonscrite au tétraèdre, noeud du diagramme de Voronoï. En bleu, un sommet réduit.

Posons $x = x_1$, $y = x_2$ et $z = x_3$, l'équation (3.7) s'écrit alors :

$$x^2 + y^2 + z^2 = \frac{R^2}{t_1}$$

Cette équation peut donc s'écrire sous forme matricielle :

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -\frac{R^2}{t_1} \end{bmatrix}$$

Comme précisé précédemment, cette équation est valable dans le cas général pour une sphère centrée en (0,0,0). Pour les tétraèdres réduits, cette sphère est centrée sur le noeud du diagramme de Voronoï (équivalent au centre de la sphère circonscrite au tétraèdre). Il faut donc appliquer une translation. La matrice de translation est donc :

$$T = \begin{bmatrix} 1 & 0 & 0 & x_{cc} \\ 0 & 1 & 0 & y_{cc} \\ 0 & 0 & 1 & z_{cc} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

avec x_{cc}, y_{cc} et z_{cc} les coordonnées du centre de la sphère circonscrite au tétraèdre.

Appliquer la translation revient à calculer la matrice $\mathbf{Q}$:

$$\mathbf{Q} = (T^{-1})^t \cdot M \cdot T^{-1}$$

$$\Leftrightarrow \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -x_{cc} & -y_{cc} & -z_{cc} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -\frac{R^2}{t_1} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & -x_{cc} \\ 0 & 1 & 0 & -y_{cc} \\ 0 & 0 & 1 & -z_{cc} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Leftrightarrow \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & -x_{cc} \\ 0 & 1 & 0 & -y_{cc} \\ 0 & 0 & 1 & -z_{cc} \\ -x_{cc} & -y_{cc} & -z_{cc} & tot \end{bmatrix}$$

Avec $tot = -\frac{R^2}{t_1} + x_{cc}^2 + y_{cc}^2 + z_{cc}^2$

Pour chaque cellule, l'ensemble des points du tétraèdre réduit sera stocké dans un élément de la matrice globale M_{tetra} et l'équation associée dans un élément de la matrice globale M_{equa_tetra}. De plus, les quatre sommets réduits et le centre circonscrit sont stockés dans une extension de chaque cellule de Delaunay. Tous les points du *Mixed Complex* sont donc calculés après ce premier passage. Il faut maintenant parcourir les cellules dans le bon sens pour reconstituer les autres morceaux du *Mixed Complex*.

Pour calculer les cellules de Voronoï réduites, l'algorithme est le suivant :

Algorithme 3 : Stockage des cellules de Voronoï réduites et de leurs équations

pour (chaque atome) {
Calcul de l'équation

Parcours de toutes les cellules adjacentes
  pour (chaque cellule) {
    Récupération du sommet réduit associé à cet atome
    Stockage de celui-ci dans une matrice temporaire M_{temp}
  }

Stockage de l'ensemble des sommets réduits (M_{temp}*) dans une matrice* $M_{voronoi}$
Stockage de l'équation associée dans une matrice $M_{equa\_voronoi}$
}

Pour reconstituer les cellules de Voronoï réduites, il suffit de parcourir, pour chaque point de la tétraédrisation de Delaunay marqué comme atome, toutes les cellules adjacentes à cet atome de Delaunay ; c'est à dire, toutes les cellules ayant ce point comme sommet. Chaque sommet réduit associé à l'atome est récupéré dans l'extension de chaque cellule puis est stocké dans une matrice temporaire. A la fin du parcours de toutes les cellules adjacentes, l'ensemble des sommets réduits (*i.e.* la cellule de Voronoï réduite) sont stockés dans la matrice temporaire. Celle-ci est ensuite stockée dans un élément de la matrice globale $M_{voronoi}$ et l'équation associée dans un élément de la matrice globale $M_{equa\_voronoi}$. Pour calculer cette équation, il faut reprendre l'équation (3.4). Comme dans le cas des tétraèdres réduits, cette équation est centrée à l'origine. Il faut donc translater celle-ci au niveau du noeud du diagramme de Voronoï. Comme dans le cas du tétraèdre réduit, il est possible de mettre cette équation sous forme matricielle :

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{R^2}{t_2} \end{bmatrix}$$

La matrice de translation est cette fois :

$$T = \begin{bmatrix} 1 & 0 & 0 & x_{VD} \\ 0 & 1 & 0 & y_{VD} \\ 0 & 0 & 1 & z_{VD} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

avec x_{VD}, y_{VD} et z_{VD} les coordonnées du sommet du tétraèdre de Delaunay associé à cette cellule.

$$\mathbf{Q} = (T^{-1})^t \cdot M \cdot T^{-1}$$

$$\Leftrightarrow \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -x_{VD} & -y_{VD} & -z_{VD} & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{R^2}{t_2} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & -x_{VD} \\ 0 & 1 & 0 & -y_{VD} \\ 0 & 0 & 1 & -z_{VD} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Leftrightarrow \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & -x_{VD} \\ 0 & 1 & 0 & -y_{VD} \\ 0 & 0 & 1 & -z_{VD} \\ -x_{VD} & -y_{VD} & -z_{VD} & tot \end{bmatrix}$$

Avec $tot = \frac{R^2}{t_2} + x_{VD}^2 + y_{VD}^2 + z_{VD}^2$

Pour calculer les équations des patchs H1 et H2, une étape supplémentaire est nécessaire : appliquer une rotation. En effet, les équations (3.5) et (3.6) présentées précédemment décrivent des hyperboloïdes dont l'axe de révolution est orienté suivant l'axe $\vec{Ox_3}$. Or, dans le cas de la *Skin Surface* Moléculaire, l'axe de rotation de l'hyperboloïde est orienté suivant l'axe du patch (ce qui est équivalent à une arête d'un tétraèdre de Delaunay pour les patchs H1 et à une arête d'une cellule de Voronoï pour les patchs H2 - voir figure 3.8).

Posons, $x = x_1$, $y = x_2$ et $z = x_3$. Cette nouvelle matrice de rotation peut aussi être vue comme une matrice de changement de bases orthonormées permettant de passer du système (O,$\vec{Ox}$,$\vec{Oy}$,$\vec{Oz}$) au système ($f(X)$,$\vec{e_1}$,$\vec{e_2}$,$\vec{e_3}$) - voir figure 3.13.

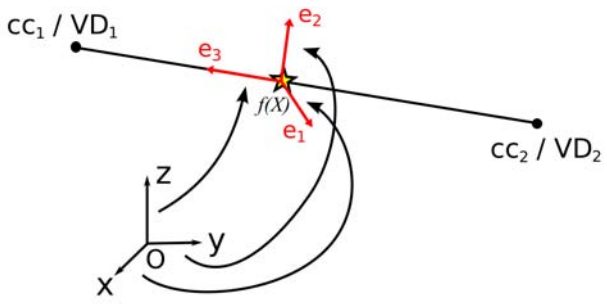


FIG. 3.13 – Changement de bases entre le système (O,$\vec{Ox}$,$\vec{Oy}$,$\vec{Oz}$) et le système (f(X),$\vec{e_1}$,$\vec{e_2}$,$\vec{e_3}$). Le point O(0,0,0) représente l'origine du système et $f(X)$ le foyer.

Cette matrice peut alors s'écrire :

$$R = \begin{bmatrix} x_{e_1} & x_{e_2} & x_{e_3} & 0 \\ y_{e_1} & y_{e_2} & y_{e_3} & 0 \\ z_{e_1} & z_{e_2} & z_{e_3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

où $\vec{e_1}$, $\vec{e_2}$ et $\vec{e_3}$ dépendent de 3 points : le foyer et les sommets de l'arête étudiée. Ainsi, dans le cas des patchs H1, il faudra considérer les sommets d'une arête d'un tétraèdre de Delaunay et le foyer associé tandis que dans le cas des patchs H2, il faudra considérer les sommets d'une arête d'une cellule de Voronoï et le foyer associé (voir figures 3.13 et 3.8).

Pour les patchs H1, l'équation (3.5) peut donc s'écrire sous forme matricielle :

$$M = \begin{bmatrix} t_1 & 0 & 0 & 0 \\ 0 & t_2 & 0 & 0 \\ 0 & 0 & t_2 & 0 \\ 0 & 0 & 0 & -R^2 \end{bmatrix}$$

Comme pour les cellules de Voronoï réduites et les tétraèdres réduits, il est nécessaire d'appliquer une translation pour positionner le centre de l'hyperboloïde au niveau du foyer. La matrice de translation s'écrit :

$$T = \begin{bmatrix} 1 & 0 & 0 & x_{f(X)} \\ 0 & 1 & 0 & y_{f(X)} \\ 0 & 0 & 1 & z_{f(X)} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Ensuite il faut composer les deux transformations (rotation puis translation) pour obtenir la matrice de l'équation $\mathbf{Q}$. Notons C, la matrice composée de la rotation et de la translation : $C = T \cdot R$.

$$\begin{aligned} \mathbf{Q} &= (C^{-1})^t \cdot M \cdot C^{-1} \\ \Leftrightarrow \mathbf{Q} &= ((T \cdot R)^{-1})^t \cdot M \cdot (T \cdot R)^{-1} \\ \Leftrightarrow \mathbf{Q} &= ((R^{-1} \cdot T^{-1})^t \cdot M \cdot R^{-1} \cdot T^{-1} \\ \Leftrightarrow \mathbf{Q} &= (T^{-1})^t \cdot (R^{-1})^t \cdot M \cdot R^{-1} \cdot T^{-1} \end{aligned}$$

Nous ne détaillerons pas plus les calculs. CGAL permet de calculer la forme développée de cette matrice à partir des matrices M, T et R.

L'algorithme permettant de stocker les patchs H1 et les équations associées peut s'écrire sous la forme :

Algorithme 4 : Stockage des patchs H1 et de leurs équations

pour (chaque arête de tétraèdre de Delaunay pour laquelle les 2 sommets sont des atomes) {
Calcul de l'équation

Parcours de toutes les cellules adjacentes
  pour (chaque cellule) {
    récupération de la paire de sommets réduits associés à cette arête
    Stockage de cette paire dans une matrice temporaire M_{temp}
  }

Stockage de l'ensemble des paires (M_{temp}*) dans une matrice* M_{H1}
Stockage de l'équation associée dans une matrice M_{equa_H1}
}

Pour reconstituer les patchs H1, il suffit de parcourir, pour chaque arête de la tétraédrisation de Delaunay pour laquelle les deux sommets sont marqués comme atomes, toutes les cellules adjacentes à celle-ci, c'est à dire, toutes les cellules possédant cette arête. La paire de sommets réduits associée à chaque arête est récupérée dans l'extension de chaque cellule puis est stockée dans une matrice temporaire. À la fin du parcours de toutes les cellules adjacentes, l'ensemble des sommets réduits (*i.e.* le patch H1) est stocké dans la matrice temporaire. Celle-ci est ensuite stockée dans un élément de la matrice globale M_{H1} et l'équation associée dans un élément de la matrice globale M_{equa_H1}.

Pour calculer la forme matricielle de l'équation (3.5) associée aux patchs H2, il suffit de reprendre exactement le même protocole que pour les patchs H1 en changeant le foyer et les sommets de l'arête (ici remplacés par des noeuds du diagramme de Voronoï). L'algorithme de stockage des patchs H2 et de leurs équations est :

Algorithme 5 : Stockage des patchs H2 et de leurs équations

pour (chaque face des tétraèdres de Delaunay constituée uniquement d'atomes) {
Calcul de l'équation

pour (chaque cellule voisine) {
récupération du triplet de sommets réduits associé à cette face
Stockage de ce triplet dans une matrice temporaire M_{temp}
}

Stockage de l'ensemble des triplets (M_{temp}*) dans une matrice* M_{H2}
Stockage de l'équation associée dans une matrice M_{equa_H2}
}

Les patchs H2 sont des cellules mixées entre une face de tétraèdre de Delaunay et une arête de cellule de Voronoï. Pour reconstituer les patchs H2, il suffit de parcourir l'ensemble des faces des cellules de Delaunay finies. Chaque cellule de Delaunay partage une face avec une cellule voisine. Pour une face donnée, il suffit de visiter les deux cellules voisines et de récupérer les triplets de points réduits associés à ces faces. On obtient alors 6 points constituant les patchs H2. Ces points sont ensuite stockés dans un élément de la matrice globale M_{H2} et l'équation associée dans un élément de la matrice globale M_{equa_H2}.

Ainsi, ces algorithmes permettent de créer et de stocker chaque élément du *Mixed Complex* ainsi que les équations associées à ceux-ci. Chaque morceau du *Mixed Complex* va ensuite être triangulé et envoyé au pipeline graphique OpenGL (voir annexe D).

Calculs sur GPU : L'enveloppe du *Mixed Complex* est ensuite visualisée grâce à l'API OpenGL (Segal et Akeley, 2004; Shreiner *et al.*, 2005). Une étape supplémentaire est nécessaire pour visualiser la *Skin Surface* Moléculaire : l'utilisation du lancer de rayons (ou *Ray Casting*) sur GPU. Cette étape n'est réalisable, sur GPU, que depuis le début des années 2000, lorsque le pipeline graphique a été ouvert à l'exécution de codes utilisateurs permettant aux unités parallèles de traitement de sommets et de fragments de devenir programmables (voir annexe D). Cette programmation se fait au niveau des processeurs de sommets ou de fragments (respectivement *Vertex Shader units* et *Fragment Shader units*). Ces processeurs sont programmés à l'aide de langages spécifiques comme, par exemple, l'*OpenGL Shading Language* (ou *GLSL*) utilisé pour ce travail (Kessenich *et al.*, 2003; Rost, 2006) ou encore le langage Nvidia Cg (Fernando et Kilgard, 2003). Il est donc nécessaire de posséder une carte graphique supportant le "*Shader Model 3.0*", ce qui revient à avoir une carte graphique Nividia Geforce série 6 (ou une carte graphique ATI Radeon X1300) ou supérieure.

Le lancer de rayons est une technique couramment utilisée pour simuler les phénomènes de réflexion ou de réfraction par le parcours inverse d'un rayon lumineux. Elle consiste, pour chaque pixel de l'image générée, à lancer un rayon depuis le point de vue (la caméra) dans la scène 3D.

Le point d'impact du rayon sur un objet permet de définir l'objet concerné par le pixel correspondant. D'autres rayons sont ensuite lancés de ce point d'impact vers les sources lumineuses pour déterminer le taux d'éclairage. Cette technique est utilisée par divers programmes comme, par exemple, POV-ray (pov, 2003). L'intérêt de cette technique permet de définir mathématiquement les objets à représenter (*i.e.* par une équation) et non plus seulement par une multitude de facettes. Ceci est très intéressant pour représenter les objets lisses (*i.e.* sans arête) comme une sphère. En effet, une représentation de celle-ci uniquement par des triangles oblige à créer une grande quantité de facettes pour approximer correctement la surface (voir figure 3.14)

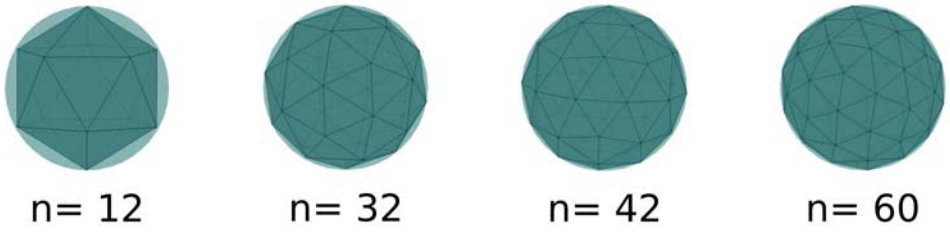


Fig. 3.14 – Approximation d'une sphère par triangulation. n représente le nombre de triangles visualisés.

En utilisant la méthode de lancer de rayons, il est possible de diminuer considérablement le nombre de triangles tout en améliorant la qualité du rendu. Reprenons l'exemple de la sphère. Prenons, un cube : cet objet peut être triangulé simplement par 12 triangles, deux par face. Ainsi, nous avons autant de triangles que l'approximation de la sphère par un dodécaèdre (figure 3.14). Ce cube est représenté sur l'écran par un certain nombre de pixels (voir figure 3.15A).

À l'intérieur de ce cube, définissons une sphère par son équation puis lançons un rayon à partir de chaque pixel de l'écran représentant le cube (voir figure 3.15B).

Ces rayons vont, soit intersecter la sphère, soit passer de part et d'autre de celle-ci (voir figure 3.15C). Il faut noter, à ce niveau que, contrairement au lancer de rayons pour calculer le taux d'éclairage, la technique utilisée dans ce travail ignore les rayons réfléchis : seule l'intersection entre la surface implicite et le premier rayon partant du pixel est prise en compte.

Les pixels dont les rayons n'ont pas intersecté la sphère seront enlevés tandis que les pixels dont les rayons ont intersecté la sphère seront déplacés de leur position à la surface du cube à la position d'intersection du ray et de la surface (voir figure 3.15C). On peut ainsi obtenir une image de sphère précise au pixel près avec, au départ, seulement 12 triangles.

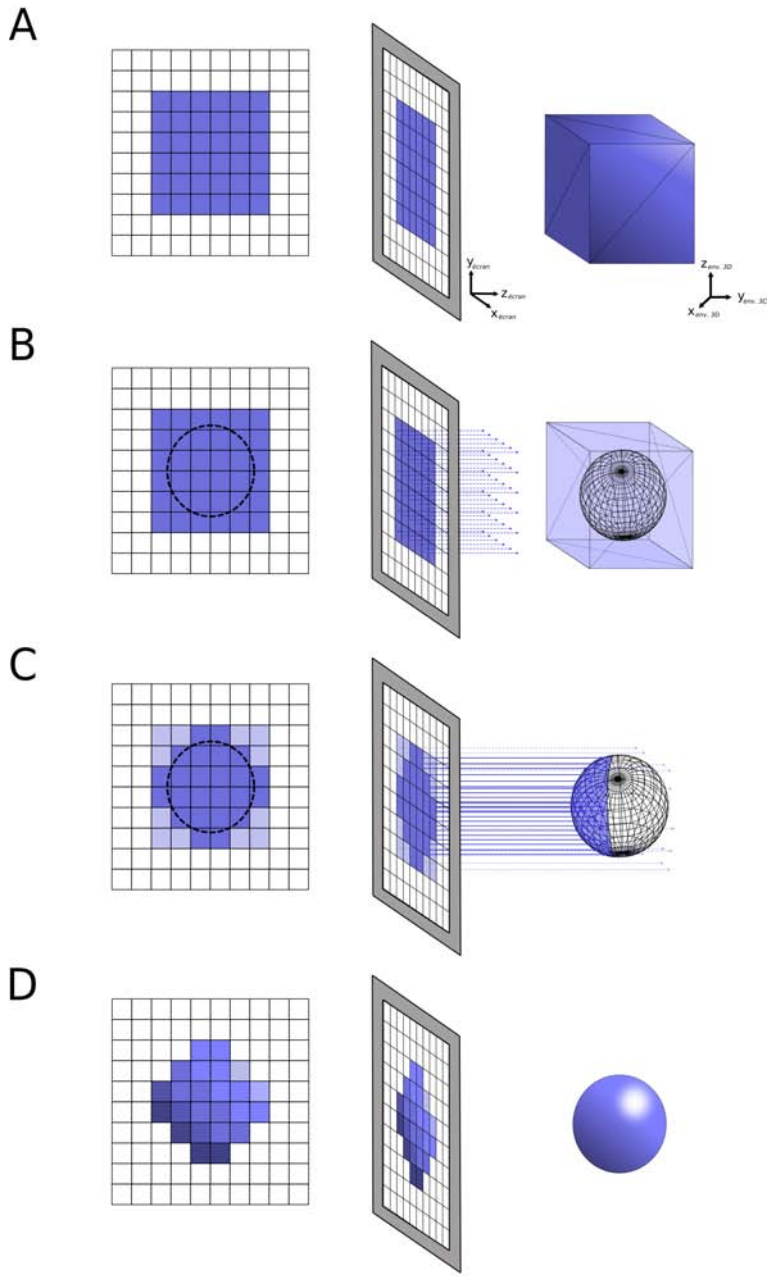


FIG. 3.15 – A- Visualisation d'un cube avec OpenGL. B- Lancer de rayons à partir des pixels représentant le cube. C- Intersection de la sphère par certains rayons. D- Destruction des pixels dont le rayon n'a pas intersecté la sphère et déplacement des pixels restant à la surface de celle-ci. À gauche, représentation simplifiée de l'écran ; chaque carré représentant un pixel. A droite, représentation des objets en trois dimensions ; en A : enveloppe de départ et en B : surface après le lancer de rayon. Il faut aussi noter que l'environnement 3D et l'environnement de l'écran possèdent leur propre repère. Il faut donc faire des changements de repère pour passer des coordonnées d'un pixel aux coordonnées de la portion de surface qu'il représente en 3D.

Le calcul de l'intersection d'un rayon avec une surface implicite de degré-2 peut se résoudre en calculant un discriminant, comme expliqué dans (Toledo et Levy, 2004) :

Soit R, un rayon :

$$R : (x, y, z) = o + t\vec{v} \tag{3.9}$$

avec o, l'origine, $\vec{v}$ le vecteur direction du rayon et t le pas.

Le changement de variable (3.9) dans l'équation (3.8) nous permet d'obtenir l'équation de l'intersection entre un rayon est une surface implicite de degré-2 :

$$V\mathbf{Q}V\ t^2 + 2O\mathbf{Q}V\ t + O\mathbf{Q}O = 0 \tag{3.10}$$

avec V le vecteur direction du rayon : $[\vec{v_x}, \vec{v_y}, \vec{v_z}, 0]$ et O l'origine du rayon : $[O_x, O_y, O_z, 1]$.

La position de l'origine et de la direction de chaque rayon sont connues, il est donc nécessaire de calculer le pas t. Pour cela, il suffit de calculer le discriminant de l'équation (3.10). Posons, $a = V\mathbf{Q}V$; $b = O\mathbf{Q}V$; et $c = O\mathbf{Q}O$.

$$\Delta = b^2 - ac$$

$$t = \frac{-b - \sqrt{\Delta}}{a}$$

Si $\Delta < 0$, il n'y a pas d'intersection et le pixel est éliminé. De plus, on ne considère que la première intersection du rayon. On ne prend donc pas en compte la deuxième solution du discriminant : $(-b + \sqrt{\Delta})/a$.

Sur la carte graphique, ce calcul se déroule en 2 étapes :
– Lors de l'affichage d'un morceau du *Mixed Complex*, un rayon est calculé à partir de chaque pixel de l'enveloppe. Ce calcul est divisé entre le processeur de sommets et le processeur de fragments.
– Puis on calcule l'intersection entre le rayon et l'équation associée au morceau du *Mixed Complex* sur le processeur de fragments.

Nous allons maintenant décrire le code GLSL utilisé pour réaliser ces deux étapes.

Le **Code Shaders 1** permet, pour chaque pixel de l'écran, de calculer sa position dans l'espace 3D. En effet, le système de coordonnées n'est pas le même entre l'écran et l'environnement 3D (voir figure 3.15A et Annexe D). Ce calcul se fait grâce à la matrice de projection *gl_ModelViewProjectionMatrixInverse*. Le point d'origine du rayon est calculé en prenant les coordonnées x et y du pixel à l'écran et en mettant son z à 0 pour se trouver dans le plan de l'écran. Ces coordonnées sont ensuite transformées pour passer dans l'environnement 3D grâce à la matrice citée ci-dessus. Le point infini est calculé à partir du point origine et de la matrice *gl_ModelViewProjectionMatrixInverse*. Ces points seront ensuite envoyés au processeur de fragments.

Code Shaders 1 : Calcul de l'origine du rayon et du point à l'infini (*Vertex Shader*)

```
varying vec4 i_near;
varying vec4 i_far;

void main() {
   vec4 p = ftransform();
   gl_Position = p;

   vec4 near = p; near.z = 0.0;
   near = gl_ModelViewProjectionMatrixInverse * near;

   i_near = near;
   i_far = near + p.w*gl_ModelViewProjectionMatrixInverse[2];
}
```

Dans le **Code Shaders 2** nous ne détaillons pas tout le code du processeur de fragment mais seulement les fonctions permettant de calculer l'intersection entre l'équation de la surface quadrique et du rayon. La structure *Ray* est définie comme ayant, en premier élément, un vecteur *origin* et, en deuxième élément, un vecteur *direction*. La fonction *primary_ray()* construit le rayon à partir des points origine et infini envoyés par le *Vertex Shader*. Ce rayon et la matrice équation vont ensuite être passés en arguments de la fonction *isect_surf*. Celle-ci va calculer le discriminant comme expliqué précédemment et éliminera le pixel ou retournera les nouvelles coordonnées de celui-ci en fonction de la valeur du discriminant.

Code Shaders 2 : Calcul du rayon et de son intersection avec la surface (*Fragment Shader*)

```
varying vec4 i_near;
varying vec4 i_far;

struct Ray {
   vec3 origin;
   vec3 direction;
};

Ray primary_ray() {
   vec3 near = i_near.xyz / i_near.w;
   vec3 far = i_far.xyz / i_far.w;

   return Ray(near,far-near);
}

vec3 isect_surf (Ray r, mat4 Q) {;

   vec4 direction = vec4(r.direction, 0.0);
   vec4 origin = vec4(r.origin, 1.0);

   float a = dot(direction,Q*direction);
   float b = dot(origin,Q*direction);
   float c = dot(origin,Q*origin);
   float Delta = b*b - a*c;

   if(Delta < 0.0) {discard; }
   else {
      float t = (-b-sqrt(Delta))/a;
      return r.origin = t*r.direction;
   }
}
```

Les codes Shaders 1 et 2, décrits brièvement, permettent de passer de la structure grossièrement triangulée du *Mixed Complex* à la *Skin Surface* Moléculaire.

3.2.5 Intérêt de notre approche pour la visualisation moléculaire

La technique de lancer de rayons permet d'obtenir un rendu d'une qualité supérieure et l'utilisation des processeurs de la carte graphique permet d'accélérer considérablement les calculs.

Précision du rendu : L'utilisation du lancer de rayons sur GPU nous permet de recalculer la surface pour tout mouvement de caméra. Nous pouvons donc faire des zooms sur la surface

moléculaire d'une macromolécule sans perte de qualité visuelle (voir figure 3.16A). Ceci peut être très appréciable si l'on veut étudier l'interaction d'une petite molécule au sein d'un assemblage moléculaire de grande taille (comme au niveau du ribosome par exemple). A notre connaissance, MetaMol est actuellement le seul programme qui puisse visualiser de manière interactive une surface moléculaire d'une telle qualité.

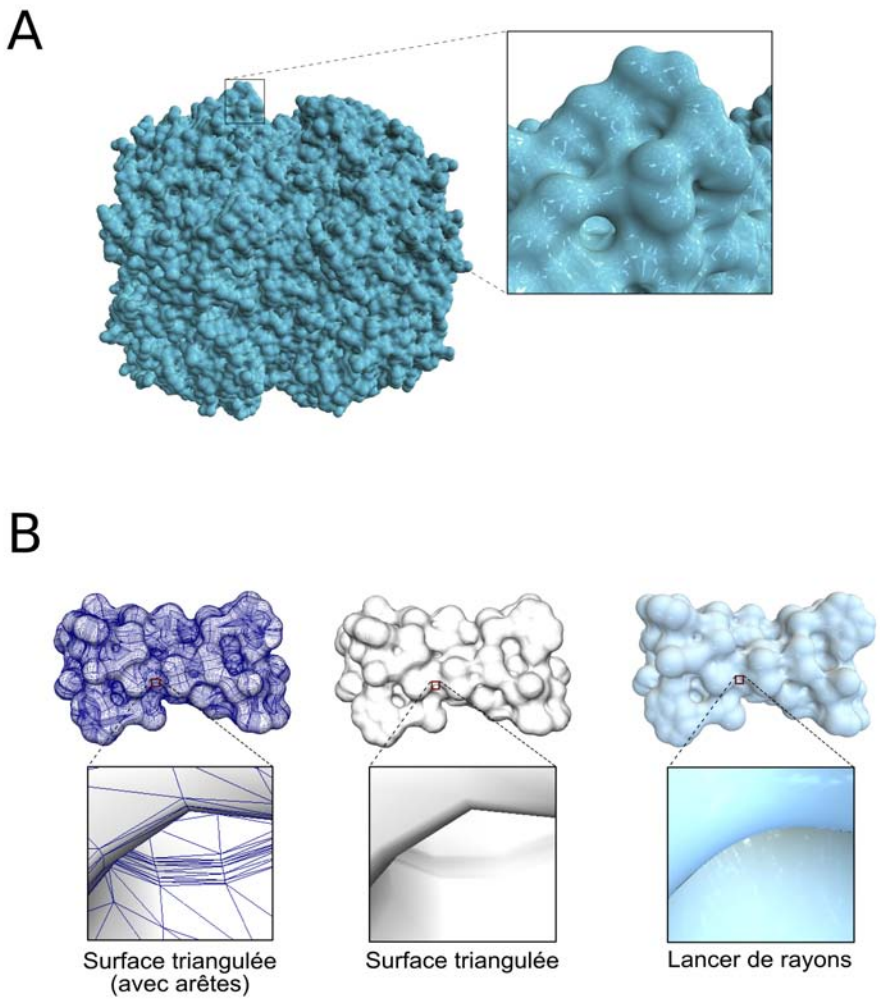


FIG. 3.16 – A- Visualisation de la molécule de *E. Coli* ClpP (identifiant PDB : 2FZS) contenant 20620 atomes. Un zoom sur cette macromolécule présente une surface parfaitement lisse sans perte d'information. B- Visualisation de la molécule de Gramicidine A (identifiant PDB : 1GRM) avec l'approche développée par Nico Kruithof (deux premières images en partant de la gauche) et avec notre méthode.

Performances du lancer de rayons par rapport à la triangulation : Pour avoir une référence, nous avons comparé, du point de vue du temps de calcul, notre méthode avec celle développée par Nico Kruithof et Gert Vegter (Kruithof et Vegter, 2004, 2007). Cette méthode

permet de représenter la *Skin Surface* Moléculaire en la triangulant. Les résultats sont présentés dans le tableau 3.1 et la figure 3.16B.

TAB. 3.1 – Comparaison de notre approche avec celle développée par Nico Kruithof.

Code PDB	Nb. d'atomes	Triangulation			Lancer de rayons		
		Nb. de triangles	Temps de calcul[a]	FPS[b]	Nb. de triangles	Temps de calcul[a]	FPS[b]
7TMN	33	23424	1,1	800	7116	0,02	200
1GRM	272	310488	16,1	130	73416	1,7	50
1G6X	509	481856	28,7	95	146476	3,6	25
1CBS	1091	1664184	93,1	30	325076	8,2	12
1J4N	1852	2165268	137,4	25	558372	15,4	7

Nous avons utilisé un processeur INTEL 2,4 GHz couplé à une carte graphique Nvidia Geforec 8800 GTX.
[a] temps de calcul en secondes.
[b] FPS : *Frame per Second* : nombre d'images par seconde, pour une résolution de 1024 × 1024.

Nous avons distingué le temps de calcul, pour créer les morceaux et générer les équations de surface, du temps d'affichage (mesuré en images par secondes ou *Frames Per Second, FPS*). Le temps de calcul obtenu avec notre approche est meilleur que celui obtenu avec la méthode de Nico Kruithof et ce pour une qualité de visualisation meilleure (voir l'exemple de la gramidicine A figure 3.16B). La raison principale en est que nous créons beaucoup moins de triangles que la triangulation (voir tableau 3.1). En effet, les seuls triangles que nous avons besoin de générer sont ceux des faces du *Mixed Complex*. Cette enveloppe n'a pas besoin d'être finement triangulée (comme le montre la figure). A l'inverse, la méthode de Kruithof nécessite une triangulation très fine pour approximer la *Skin Surface* Moléculaire. De plus, avec cette méthode, il est nécessaire de vérifier que la triangulation préserve la topologie de la surface (Kruithof et Vegter, 2007). L'utilisation du lancer de rayons permet de dépasser ce problème en calculant la valeur de la surface pour chaque pixel de la fenêtre d'affichage. La puissance des processeurs graphiques permet de mettre à jour la surface "à la volée" lorsque le point de vue ou le facteur de réduction est changé. Ceci peut être utilisé pour visualiser la déformation de surface (voir paragraphe suivant).

Cependant, après la phase de calcul du *Mixed Complex*, les performances de la triangulation sont meilleures que celles de notre approche (voir tableau 3.1) : le nombre de FPS (permettant de mesurer l'interactivité du programme, comme la réponse de l'image à une action de la souris) est meilleure pour la triangulation. Le résultat de notre approche s'explique par un grand nombre d'accès en lecture-écriture au *Z-Buffer* (Foley *et al.*, 1995) ce qui diminue considérablement la fréquence d'affichage. Néanmoins, si l'on veut obtenir avec la triangulation le même niveau de détails que notre approche, il faudrait créer une très grande quantité de petits triangles-idéalement ayant au maximum la taille d'un seul pixel affiché à l'écran - ce qui démultiplierait le temps de calcul et ferait s'effondrer les performances d'affichage.

Il reste que notre approche est, pour l'instant, dédiée à la visualisation de la *Skin Surface*

Moléculaire et ne permet pas le calcul de propriétés physico-chimiques à partir de celle-ci. Dans ce cas, la triangulation développée par Nico Kruithof[31] apparaît donc comme complémentaire à notre approche.

Visualisation de la déformation de surface : La *Skin Surface* est aussi bien adaptée à la visualisation de déformation de surface. En effet, cette surface peut se déformer librement grâce à des transitions douces (Edelsbrunner, 1995, 1999). Contrairement aux programmes qui triangulent la *Skin Surface* (Kruithof et Vegter, 2007; Cheng et Shi, 2004), la technique de lancer de rayons sur GPU permet de visualiser des déformations en temps réel[32].

Déformation lors du changement du facteur de réduction : Comme on peut le voir figure 3.17, il est possible de faire varier le facteur de réduction et de voir évoluer la surface en fonction de ce facteur. Il est possible de passer d'une représentation proche de la surface de Van der Waals (où tous les atomes sont représentés par des sphères), via une représentation proche de la Surface Moléculaire (pour s autour de 0,5) à une représentation simplifiée de la surface (si l'on continue à diminuer davantage le facteur de réduction). Cette surface simplifiée peut être intéressante pour visualiser la forme globale de la molécule afin de comparer rapidement différentes silhouettes de protéines (Cipriano et Gleicher, 2007). Cette représentation peut aussi être utile dans un processus de docking protéique haut-débit pour lequel les modèles basse résolution sont utiles pour un calcul rapide des assemblages lors des premières étapes du docking (Ritchie et Kemp, 1999; Tovchigrechko *et al.*, 2002). Nous pouvons enfin noter que le passage d'une représentation à une autre se fait de manière continue.

[31] disponible avec la libraire mathématique CGAL

[32] Une vidéo exemple est disponible : `http://www.loria.fr/~chavent/video/video2.avi`
Pour visualiser la vidéo il est peut être nécessaire d'installer ffdshow : `http://www.clubic.com/telecharger-fiche11020-ffdshow.html`

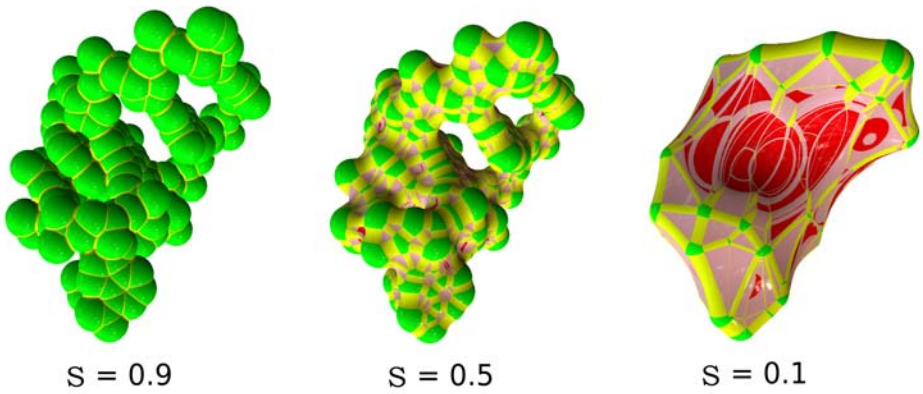


FIG. 3.17 – Représentation d'un fragment d'ADN (identifiant PDB : 200D). Évolution de la *Skin Surface* Moléculaire en fonction du facteur de réduction (s). Les couleurs permettent de visualiser les différents morceaux de la *Skin Surface* Moléculaire : en vert, la surface à l'intérieur les cellules de Voronoï réduites ; en rouge, la surface à l'intérieur des tétraèdres réduits ; en jaune, la surface à l'intérieur des patchs H1 et en rose, la surface à l'intérieur des patchs H2.

Déformation lors de mouvements moléculaires : Une visualisation claire et efficace des mouvements moléculaires est un outil précieux pour mieux comprendre les mécanismes moléculaires. Au cours de ces 10 dernières années, des efforts ont été faits pour atteindre cet objectif.

Nous pouvons citer Eyal et Halperin qui ont développé un algorithme pour mettre à jour, de manière dynamique, la surface de Van der Waals et la surface accessible au solvant lors de changements conformationnels (Eyal et Halperin, 2005). Hao et Varshney ont développé un programme permettant de visualiser de larges mouvements de protéines en temps réel en utilisant la technique d'*occlusion culling* (Hao et Varshney, 2004). Plus récemment, Lampe *et al.* ont défini une approche à deux niveaux de précision pour visualiser la dynamique des protéines (Lampe *et al.*, 2007). Le problème de ces approches est de ne traiter que des types de surface simples : comme la surface de Van der Waals ou une représentation "*Balls and Sticks*", ce qui a un intérêt limité pour l'étude des interactions moléculaires. Il n'existe, à l'heure actuelle, que peu de travaux dédiés à la visualisation des déformations de surfaces plus complexes telle la Surface Moléculaire (Sanner et Olson, 1997; Bajaj *et al.*, 2003). Une explication est que mettre à jour une surface triangulée pour de larges mouvements est très difficile et coûteux en temps de calcul.

Avec l'utilisation du lancer de rayons sur GPU, il n'y a pas de triangulation à maintenir et le temps de calcul du *Mixed Complex* est relativement court, ce qui nous permet déjà de visualiser, en temps réel, des déformations de surface pour de petites molécules (voir figure 3.18). Des optimisations ultérieures nous laissent espérer une visualisation en temps réel de déformation de surface pour des macromolécules de grande taille.

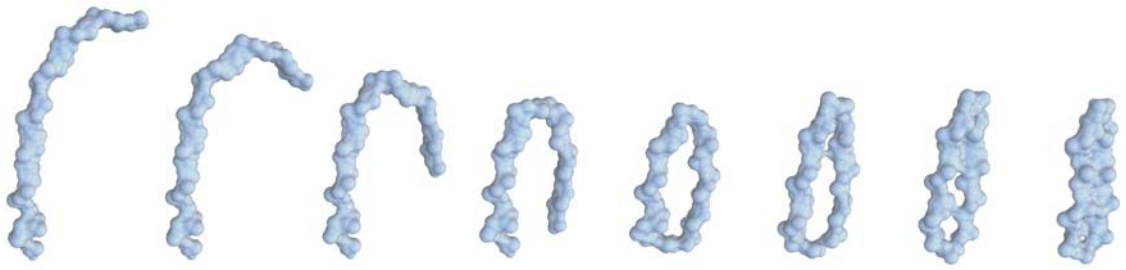


FIG. 3.18 – Déformation d'une chaîne de poly-alanine lors d'un passage d'une forme désorganisée à une forme en feuillet-β.

3.2.6 Discussion et futures optimisations

Ainsi, nous avons développé un nouvel outil de visualisation qui semble avoir un grand potentiel. L'utilisation de la *Skin Surface* Moléculaire associée à la technique de lancer de rayons sur GPU permet un rendu en temps réel jusqu'à présent jamais atteint. Ce travail démontre la faisabilité et l'intérêt d'une telle approche. Néanmoins, notre programme n'en est qu'à ces débuts et des optimisations sont encore possibles :

- D'abord, la création de matrices globales côté CPU et l'envoi de celles-ci à la carte graphique n'est pas la façon la plus adaptée de transmettre des données aux GPU. Il serait préférable d'envoyer directement l'information (points et équations) à la carte graphique lors de l'extraction de celle-ci.
- Ensuite, nous envoyons chaque équation de surface sous la forme d'une matrice 4×4. Or la simplification de ces équations pourrait permettre de passer beaucoup moins d'arguments à la carte graphique, limitant les temps de transfert et par conséquent la fréquence d'affichage.
- De même, il serait nécessaire de créer un *shader* spécifique à chaque type de surface (sphère, hyperboloïdes à une ou deux nappes) pour remplacer notre *shader* générique.
- Enfin, des techniques de classement d'objets dans la scène 3D (comme l'*occlusion culling*) permettraient de limiter les calculs aux seuls pixels visibles.

Conscients de ces problèmes, nous continuons à travailler sur le programme MetaMol afin de le rendre utilisable, le plus rapidement possible, par l'ensemble de la communauté scientifique.

Une avancée significative a d'ailleurs été réalisée dernièrement : il s'agit de l'ajout de *Ambient Occlusion* au programme MetaMol. Cet ajout a été réalisé par Cécile Poisot durant son stage de 3$^{\text{ème}}$ année d'ingénieur sous la direction de Bruno Levy.

L'utilisation de l'*Ambient Occlusion* avec le programme MetaMol permet de mettre en valeur le rendu de celui-ci et de mieux appréhender la forme des macromolécules (voir figure 3.19).

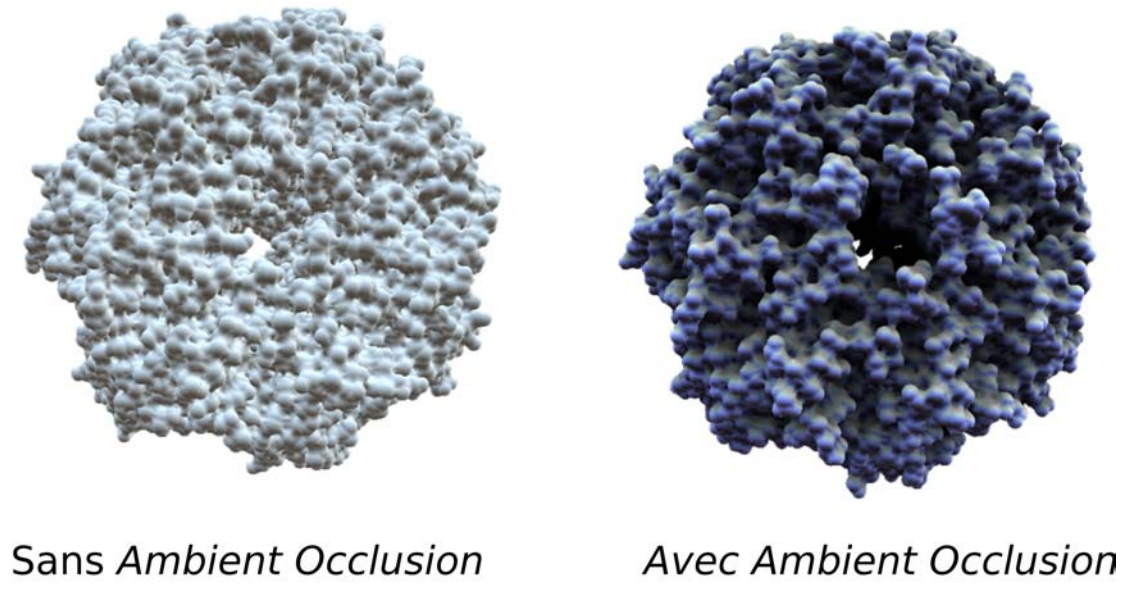


FIG. 3.19 – Visualisation de la molécule de ClpP d'*E. Coli* (identifiant PDB 2FZS) avec et sans ajout d'*Ambient Occlusion*.

3.3 Conclusion : Vers un outil multi-résolution et interactif

En conclusion, le programme MetaMol propose un rendu graphique de qualité de la *Skin Surface* Moléculaire et l'ajout de nouveaux effets de lumière va permettre de mieux appréhender la forme complexe des protéines. Ce programme reste cependant un outil de visualisation et ne permet pas une analyse structurale poussée. C'est pourquoi, il serait intéressant d'inclure celui-ci dans des logiciels d'analyse plus complets comme VMD, Chimera ou Antheprot3D.

De plus, la *Skin Surface* moléculaire peut être utilisée pour représenter d'autres types de surfaces comme les enveloppes SAXS ou celles de cryo-microscopie électronique. MetaMol pourrait donc être utilisé pour développer un logiciel multi-résolution. La *Skin Surface* moléculaire pourrait aussi être utilisée pour représenter la surface de représentations simplifiées, comme le modèles réduits du programme ATTRACT ou ceux utilisés pour étudier la flexibilité des protéines grâce aux modes normaux.

Il est possible d'envisager d'utiliser MetaMol avec de nouveaux outils afin d'améliorer l'interactivité avec l'utilisateur (voir figure 3.20). Nous pouvons déjà utiliser la commande de la console Wii (la wiimote) afin de remplacer la souris et le clavier de l'ordinateur. Nous envisageons d'utiliser la webcam pour manipuler des représentations 3D réelles des molécules et de voir les mouvements de celle-ci reproduits à l'écran, comme l'avait déjà réaliser l'équipe de Michek Sanner (Gillet *et al.*, 2005). Enfin, nous réfléchissons à l'utilisation d'un bras à retour de force pour manipuler les molécules. Le LORIA offre aussi du matériel de visualisation haute définition

comme le *Reality Center* ou un projet de *CAVE* (voir figure 3.20) qui pourraient être utilisés pour visualiser la surface moléculaire obtenue avec MetaMol.

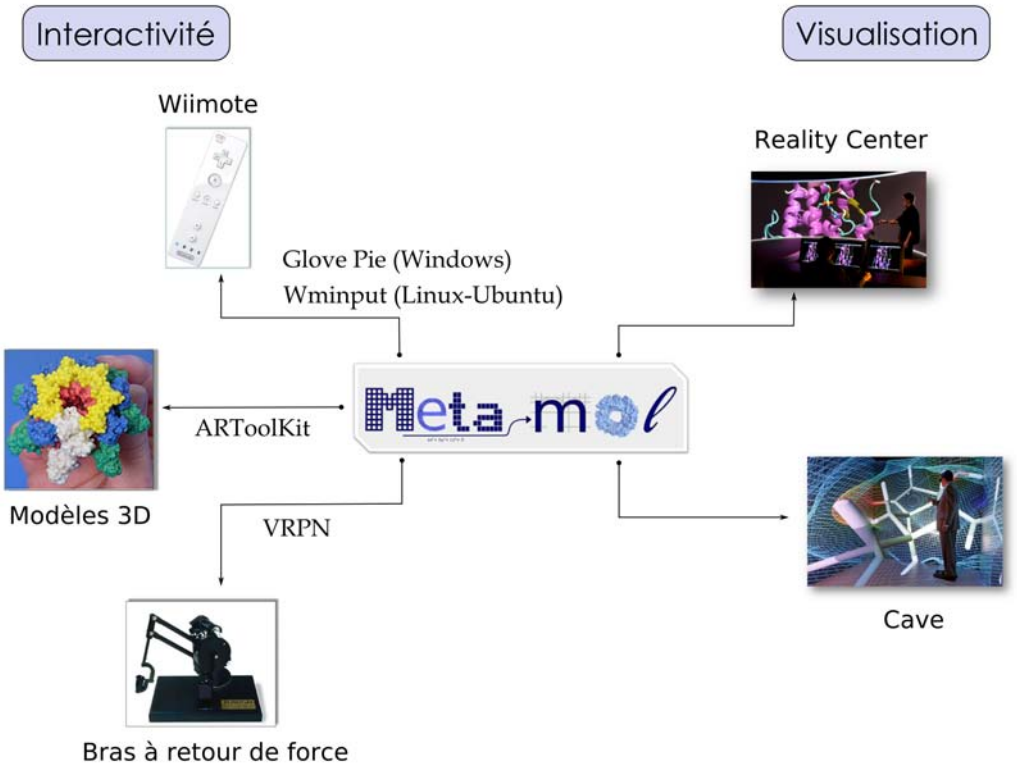


FIG. 3.20 – Les différentes extensions envisagées à MetaMol. Pour augmenter l'interactivité, l'utilisation des commandes de la console Wii, d'une représentation 3D d'une molécule ou d'un bras à retour de force. Pour la visualisation haute définition, l'utilisation d'un CAVE ou du reality center pourrait accentuer le rendu de MetaMol.

Enfin, une autre voie, qui nous semble très prometteuse, serait d'inclure MetaMol comme un plug'in du logiciel SAMSON afin de pouvoir visualiser les déformations de surface lors de simulations de dynamique moléculaire interactive.

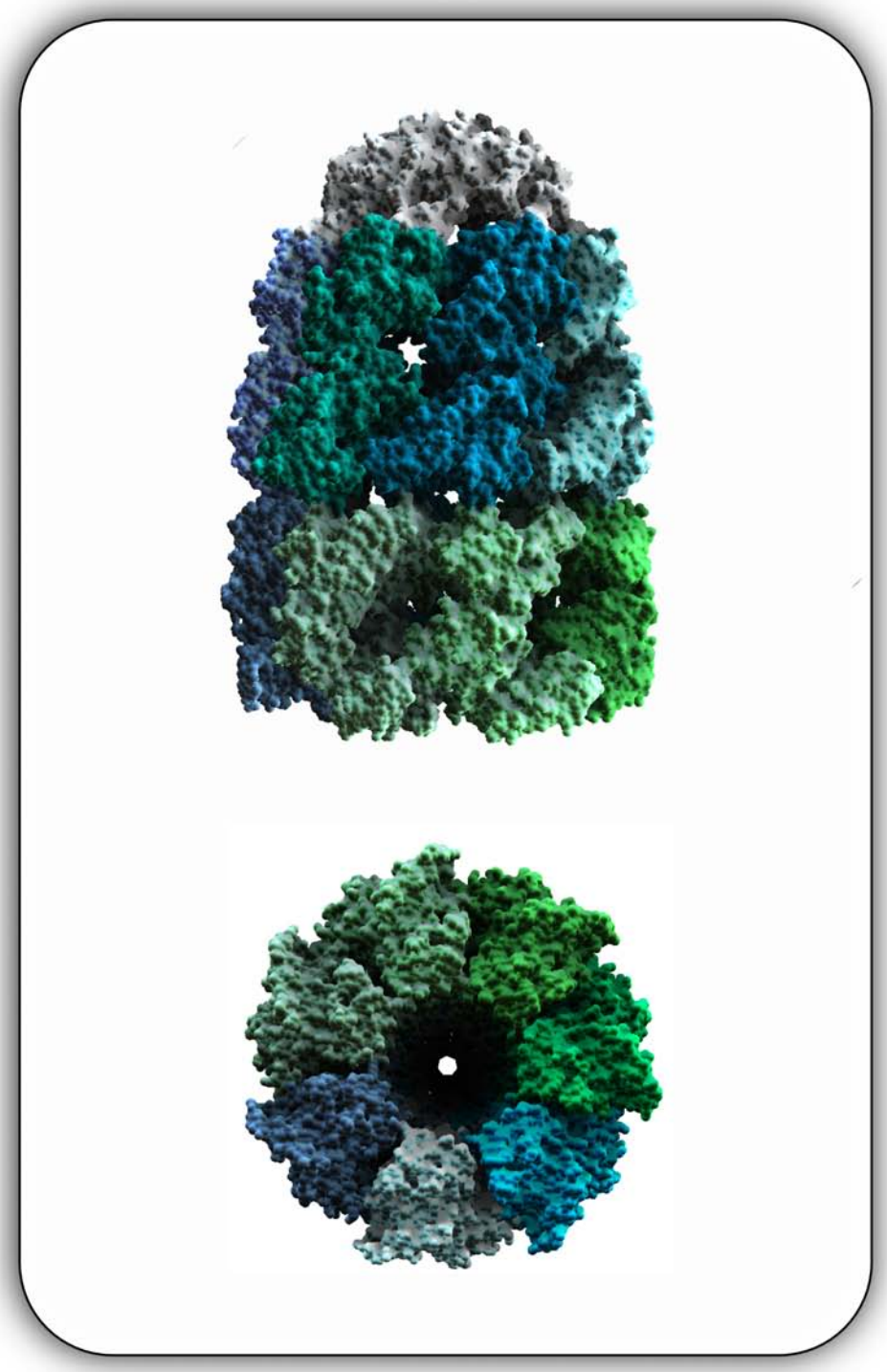


FIG. 3.21 – Visualisation de l'assemblage GroEL-GroES (identifiant PDB : 1AON) par le programme MetaMol avec l'*Ambient Occlusion*.

Conclusion

Montaigne écrivait dans *Les Essais* : « Notre vie n'est que mouvement », cette phrase reste valable au niveau moléculaire. Nous avons constaté que les macromolécules étaient des objets flexibles et qu'il était nécessaire de prendre cette flexibilité en compte pour modéliser convenablement les systèmes biologiques. Pour cela, nous avons utilisé la dynamique moléculaire en solvant explicite.

La dynamique moléculaire nous a servi à relaxer les modèles de dimères créés, soit par analogie par rapport à un complexe existant, soit grâce à des programmes de docking. Suivant le temps de simulation, nous avons pu mettre en évidence des résidus clés pour l'interaction entre les domaines PDZ d'Erbin et MH2 de Smad3, et voir certains modèles converger vers une structure unique ou raffiner des solutions de docking rigide.

Le choix du temps de simulation est donc un élément important à déterminer en fonction des phénomènes biologiques que l'on veut observer. Dans le cas du complexe PDZ d'Erbin/MH2 de Smad3, un temps de 5 nano-secondes était nécessaire pour commencer à voir le système évoluer. Nous considérons cette durée de 5 nano-secondes comme une valeur minimale pour étudier la stabilité des complexes protéiques de taille standard. Néanmoins, dans le contexte de la post-génomique, pour raffiner un modèle de docking rigide, ce temps peut être réduit à quelques centaines de pico-secondes. Ce laps de temps permet de laisser les chaînes latérales et le squelette peptidique se réarranger localement mais elle n'autorise pas d'important changements conformationnels.

Une possibilité pour modéliser des mouvements de grande amplitude sur une durée réduite est de guider *de visu* la simulation. Des logiciels comme SAMSON peuvent être, dans ce cas, d'une aide précieuse. Il reste que la représentation des protéines proposée par ce programme n'est pas adaptée au raffinage de solution de docking. En effet, pour cela, l'important est de visualiser l'interface entre les protéines partenaires. Il faut alors représenter la surface des partenaires et étudier les déformations de celle-ci lors de la manipulation. Le programme MetaMol, développé durant cette thèse, peut répondre à ce besoin.

Nous pouvons combiner les différentes approches abordées au cours de cette thèse afin de créer une stratégie globale de docking (voir figure 2) :

– Après une courte dynamique moléculaire sur les structures des partenaires, celles-ci se-

ront transmises à un meta-programme qui lancera différents programmes de docking. Une analyse des solutions consensus limitera le nombre de solutions.

- Les solutions les plus intéressantes seront transmises ensuite à un programme de dynamique moléculaire interactive. L'expert réajustera, si besoin est, certaines parties en fonction de ses connaissances sur le sujet.
- Ces résultats pourront ensuite subir une dynamique moléculaire plus longue pour relaxer le système et mettre en évidence les résidus en interaction.
- Enfin, ces modèles serviront à guider les expérimentations biologiques.

Cette stratégie se veut interactive. En effet, l'élément central de celle-ci est le programme de dynamique moléculaire interactive et les informations obtenues à chaque étape peuvent être réexploitées dans les étapes précédentes. Par exemple, les informations expérimentales pourront servir à guider la dynamique interactive ou pourront être traduites sous forme de contraintes pour les programmes de docking, etc... Chaque étape de la stratégie est donc en interaction avec les autres.

Cette stratégie a été proposée sur un cas concret : la protéine FAK (pour *Focal Adhesion Kinase*) sujet central d'un projet ANR. Cette protéine, importante pour la propagation des signaux émanant de divers récepteurs membranaires, en particulier ceux des intégrines (Parsons, 2003), est composée de 4 domaines dont 3 ont été déterminés pas cristallographie (voir figure 1). Le problème posé est alors l'agencement de ces domaines les uns par rapport aux autres. De plus, les cristallographes du projet ANR ont déterminé l'enveloppe SAXS (Márquez *et al.*, 2003) de cette protéine. Ceci constitue donc un problème de docking protéine-protéine avec contraintes. Pour aborder ce problème nous avons procédé comme indiqué auparavant : d'abord une étape de docking rigide utilisant divers serveurs, puis nous avons utilisé le programme SITUS (Wriggers et Chacón, 2001) pour positionner les assemblages des différents domaines dans l'enveloppe SAXS. Nous avons également commencé à utiliser le programme SAMSON (Rossi *et al.*, 2007) pour déplacer certaines boucles flexibles dans l'enveloppe. Enfin, nous avons réalisé une dynamique en solvant explicite de 70 ns pour tester la stabilité de l'ensemble et identifier les résidus importants pour l'assemblage. Les résultats préliminaires sont très encourageants.

Le nombre de nouveaux types complexes macromoléculaires cristallisés est de plus en plus important. Or, si l'on suppose que le nombre de types d'interactions est limité à 10000 (Aloy et Russell, 2004), il arrivera un jour où la majorité des types d'interactions existants seront répertoriés. Nous en sommes encore loin puisque nous ne recensons actuellement qu'approximativement 2500 types différents. Néanmoins, nous utilisons déjà, dans certains cas, ces informations structurales pour modéliser des complexes non cristallisés. Dans cette thèse, nous avons développé une approche par analogie, basée sur l'utilisation de la structure d'un complexe déjà existant comme patron pour modéliser un nouveau complexe. Nous avons été plus loin pour la dernière cible en date de CAPRI (Janin *et al.*, 2003), la cible 37. En effet, nous avons modélisé les partenaires de cette cible par homologie en nous servant des partenaires complexés d'un assemblage déjà existant. Avec le nombre grandissant de structures disponibles, nous sommes persuadés que cette approche, que nous nommons *structural homology docking* et qu'il est possible de rappro-

cher du "*comparative modelling*" (Aloy et Russell, 2003; Aloy *et al.*, 2004), se développera dans les prochaines années.

La prochaine étape sera de modéliser la cinétique d'association des complexes. Les nouveaux développements à la fois théoriques, autour de l'analyse des modes normaux et de la simulation gros grains, et expérimentaux, comme de la spectroscopie par RMN ou le laser femto-seconde, vont apporter les informations nécessaires pour mieux étalonner les paramètres de dynamique moléculaire. Et, peut-être un jour, au côté des expériences CASP (Dunbrack *et al.*, 1997) et CAPRI un nouveau challenge apparaîtra : *CAMDs, a Critical Assessment for Molecular Dynamic simulations*.

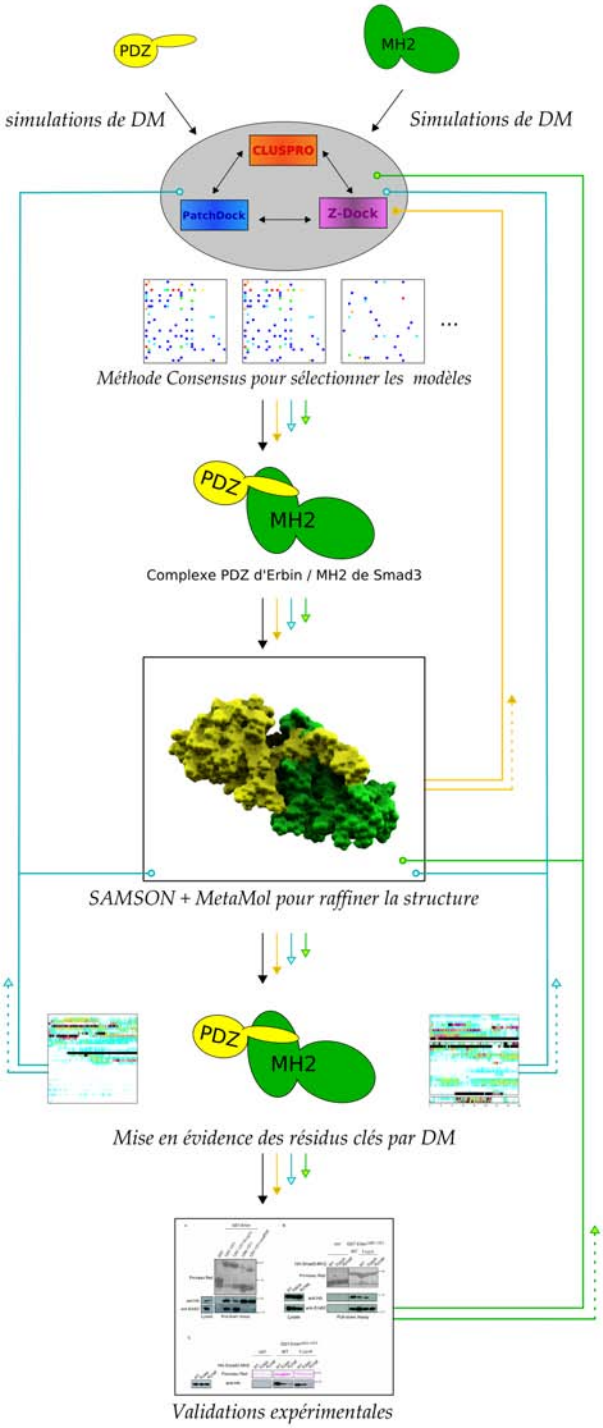


FIG. 1 – Visualisation de la stratégie interactive de docking. Exemple pour le complexe PDZ d'Erbin/MH2 de Smad3. Les lignes de différentes couleurs représentent les informations transmises par chaque étape : en orange, la structure du complexe raffiné "à la main"; en bleu, les résidus clés mis en évidence durant la dynamique moléculaire ; en vert, les données expérimentales.

Annexe A

Structure des macromolécules

Une très bonne revue sur les structures des macromolécules est disponible en sections I.2 et I.3 du livre : *Structural Bioinformatics* (Bourne et Weissig, 2003).

A.0.1 Structure des protéines

Les protéines représentent environ 50% de la masse sèche de la plupart des cellules et jouent un rôle dans presque toutes les fonctions cellulaires, fonctions sont souvent liées à la structure des protéines.

Les protéines sont des polymères élaborés à partir d'acides aminés. Les cellules élaborent leurs protéines à partir de 20 acides aminés. Ces acides aminés possèdent une structure commune composée d'un carbone alpha asymétrique sur lequel se fixe un groupement carboxyle (COO^-), une fonction amine (NH_3^+) et un atome d'hydrogène. Ils ne se différencient que par la partie attachée au carbone alpha au moyen de la quatrième liaison. Cette partie est appelée la chaine latérale. Les propriétés physiques et chimiques de la chaîne latérale déterminent les caractéristiques particulières d'un acide aminé (voir figure A.1).

Lorsque deux acides aminés sont placés de telle sorte que le groupement carboxyle de l'un se trouve à côté du groupement amine de l'autre, une réaction de condensation peut les unir. A l'aide d'une enzyme, cette réaction produit une liaison covalente appelée liaison peptidique et libère une molécule d'eau (voir figure A.2). Cette réaction permet de former des chaînes d'acides aminés appelées chaînes polypeptidiques.

Une protéine se compose d'une ou plusieurs chaînes polypeptidiques adoptant une forme tridimensionnelle définie, c'est à dire une certaine conformation. La conformation d'une protéine comporte quatre niveaux d'organisation structurale : primaire, secondaire, tertiaire et quaternaire.

La structure primaire : L'ordre de succession des acides aminés constitue la structure primaire ou séquence de la protéine. Cette séquence est donnée, par convention, dans le sens allant de l'extrémité N-terminale à l'extrémité C-terminale. A ces extrémités, la protéine présente des

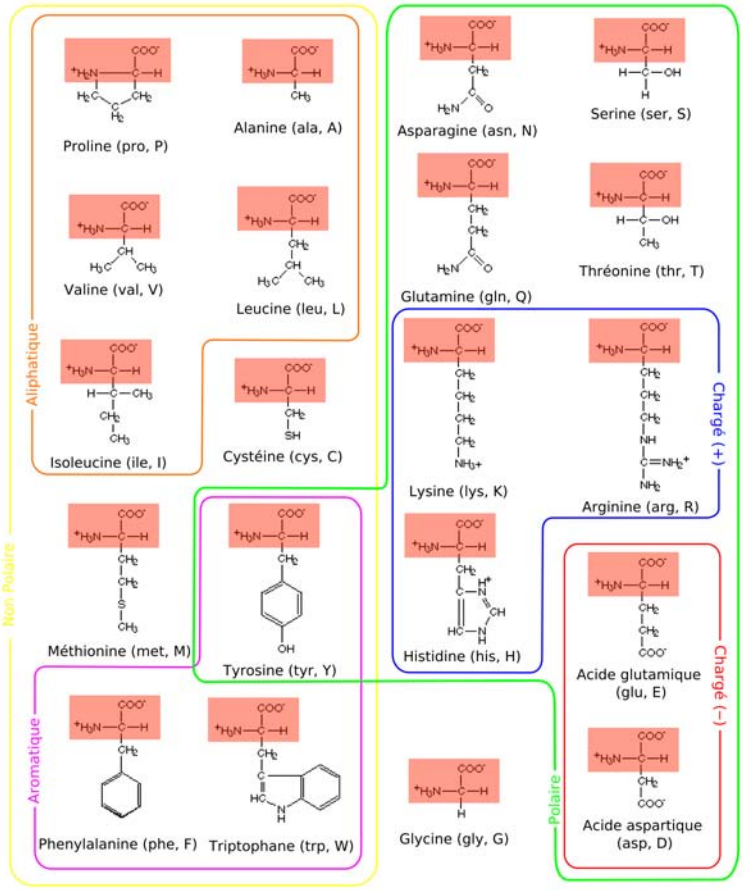


FIG. A.1 – Liste des 20 acides aminés. Ces acides aminés sont, ici, classés en différents groupes suivant les propriétés de leur chaîne latérale. Les acides aminés apparaissent ici dans leur forme ionique dominante au pH intracellulaire de 7 environ. En rouge, le squelette protéique formé par un carbone α reliant un groupement carboxyle et un groupement hydroxyle.

groupements chargés NH_3^+ (partie N-terminale) et COO^- (partie C-terminale).

La structure secondaire : Dans la plupart des protéines, certains segments de la chaîne polypeptidique sont enroulés ou repliés de façon répétitive et forment ainsi des motifs qui contribuent à la conformation globale de la protéine. L'ensemble de ces motifs constitue la structure secondaire et provient de liaisons hydrogène situées à intervalles réguliers le long de la chaîne polypeptidique. Les éléments de structure secondaire les plus courants dans la protéine sont les hélices α et les feuillets β.

Les hélices α : Une hélice est créée par une courbure au niveau du squelette polypeptidique jusqu'à ce qu'une une forme en spirale soit produite. Cette hélice peut être, en théorie, enroulée dans deux directions possibles (droite ou gauche). En pratique, la majorité des hélices sont orientées vers la droite. Parmi celles-ci orientées à droite, l'hélice α est, de loin, la plus

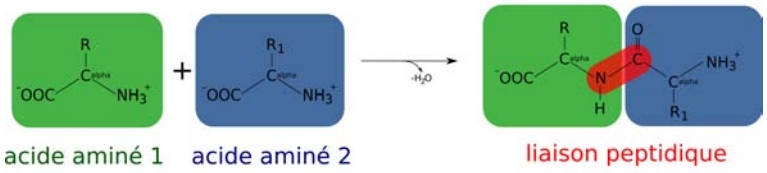


FIG. A.2 – Formation d'une liaison peptidique.

représentative et contient 3,6 résidus par tour. Cette structure est stabilisée par des liaisons hydrogène entre le groupement C=O de l'acide aminé en position i et le groupement N-H de l'acide aminé en position $i + 4$. Les chaînes latérales de tous les acides aminés pointent vers l'extérieur de l'hélice (voir figure A.3).

D'autres types d'hélice ont aussi été observés, plus ou moins fréquemment suivant leurs configurations (plus ou moins stables). Les hélices 3_{10} ont une période de 3 résidus par tour avec une liaison hydrogène entre l'acide aminé en position i et celui en position $i+3$. Ce type d'hélice est habituellement de petite taille et se trouve à la fin des hélices α. Il existe aussi des hélices π, celle-ci sont très rares. Elles ont une période de 4,4 résidus par tour avec des liaisons hydrogène formées entre les résidus i et $i + 5$. Cette forme a seulement été observée à la fin des hélices α.

Les feuillets β : Contrairement aux hélices α, les feuillets β sont formés par des liaisons hydrogène entre des acides aminés de polypeptides adjacents plutôt que par des résidus d'une même chaîne. Ces arrangements produisent une structure plane en accordéon où les chaînes latérales des résidus se trouvent alternativement de part et d'autre du plan formé par le feuillet-β (voir figure A.3). Il existe deux configurations possibles de feuillets-β : parallèle et anti-parallèle. Dans le premier cas, les chaînes polypeptidiques sont orientées dans le même sens tandis qu'elles sont orientées en tête bêche dans le deuxième cas.

Les autres structures secondaires : Les hélices α et les feuillets β sont reliés par des régions moins structurées appelées *boucles* ou *coudes*. Ces régions sont souvent des zones de transition entre les zones plus structurées. Néanmoins, elles peuvent avoir un rôle non négligeable dans les mécanismes cellulaires si elles sont placées près d'un site actif. Un parfait exemple est la boucle d'activation des kinases nécessaire pour la phosphorylation de ces dernières. Il faut noter que ce type de régions, très flexible, est très difficile à modéliser par les programmes de docking.

La structure tertiaire : La structure tertiaire d'une protéine correspond à l'ensemble des contorsions irrégulières dues aux liaisons entre les chaînes latérales des acides aminés. Ces contorsions permettent à la protéine d'adopter une structure tridimensionnelle définie (voir figure A.4). Les interactions hydrophobes contribuent en grande partie à la structure tertiaire. Lorsqu'un polypeptide adopte sa conformation native, les acides aminés portant une chaîne latérale hydrophobe se rassemblent au coeur de la protéine, s'éloignant ainsi de l'eau. Les liaisons hydrogène entre certaines chaînes latérales ainsi que les liaisons ioniques entre les chaînes latérales chargées

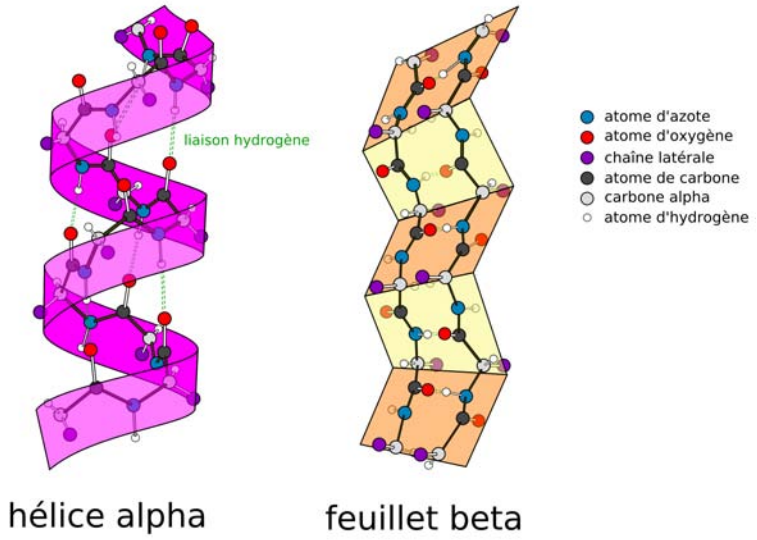


FIG. A.3 – Visualisation d'une hélice-α et d'un feuillet-β.

positivement et négativement sont également importantes . Enfin, les liaisons covalentes fortes entre deux résidus cystéines permettent de stabiliser fortement la structure de la protéine. Ces liaisons covalentes sont appelées des ponts disulfure.

La structure quaternaire : Certaines protéines se composent de deux ou plusieurs chaînes polypeptidiques assemblées pour former une macromolécule fonctionnelle. Chaque chaîne polypeptidique constitue une sous-unité de la protéine.

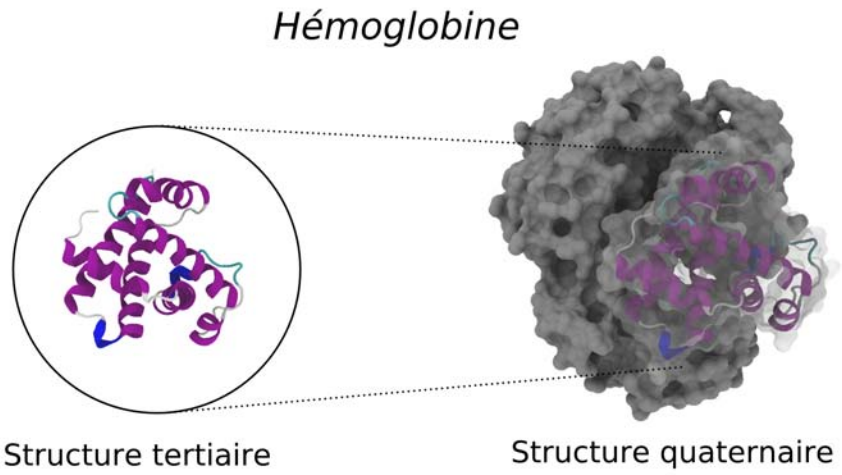


FIG. A.4 – Représentation de la structure tertiaire et quaternaire de la protéine d'hémoglobine. En cyan, les parties désorganisées appelées boucles. En violet, les hélices-α. En bleu, les hélices 3_{10}. En gris et en transparent, représentation de la surface de la protéine d'hémoglobine.

A.0.2 Structure des acides nucléiques

Il existe deux types d'acides nucléiques : l'acide désoxyribonucléique (ou ADN) et l'acide ribonucléique (ou ARN).

L'ADN constitue le matériel héréditaire tandis que l'ARN sert surtout d'intermédiaire dans la circulation de l'information génétique de l'ADN à la protéine. Ces deux molécules sont constituées par une succession d'unités comprenant toutes un groupement phosphate, un sucre et une base nucléique. Il existe 5 types de bases regroupés en deux familles : les purines (R) contenant l'adénine et la guanine (bases à deux cycles aromatiques) et les pyrimidines (Y) contenant la cytosine, la thymine et l'uracile (base à un cycle aromatique). Ces bases sont présentées figure A.5.

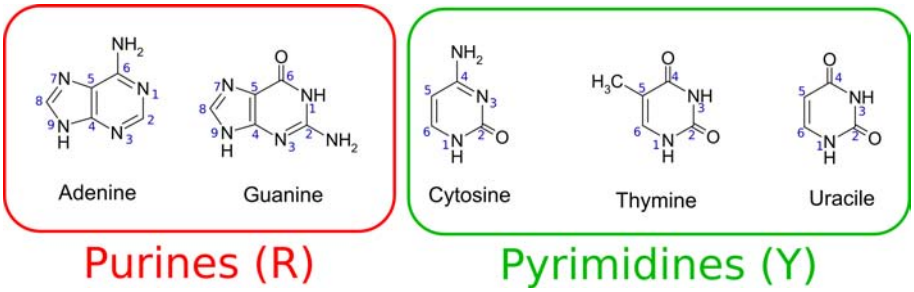


FIG. A.5 – Les cinq bases nucléiques. Les atomes sont numérotés selon la nomenclature standard.

Les molécules d'ADN et d'ARN diffèrent par deux caractères :
– La base thymine présente dans l'ADN est remplacée par la base uracile dans l'ARN.
– l'ARN possède un groupement hydroxyle en position 2' sur la molécule de sucre. C'est la différence au niveau de ce groupement qui explique le nom respectif de chaque molécule : le sucre présent au niveau de l'ARN est un ribose tandis que le sucre présent au niveau de l'ADN est un désoxyribose.

La structure de la molécule d'ADN : L'ADN, dans sa forme native, est une molécule en double hélice (voir figure A.6). Les bases jouent un rôle déterminant dans la stabilité de cette molécule :
– Chaque base purique d'une hélice est associée à une base pyrimidique. Il existe, plus exactement, deux types de paires. La paire adénine-thymine (A-T) est reliée par deux liaisons hydrogène tandis que trois liaisons hydrogène sont nécessaires pour la paire cytosine-guanine (C-G) (voir figure A.6).
– Les liaisons hydrogène sont des liaisons très flexibles et ne suffisent pas à elles seules à expliquer la stabilité de la double hélice d'ADN. Ces liaisons hydrogène sont complétées par des arrangements verticaux des cycles aromatiques des bases. Cet empilement, dit en *stacking*, permet de restreindre la flexibilité des bases.

La structure d'ADN la plus commune est la forme B : c'est cette structure qui a été décrite par Watson et Crick (voir figure A.7). Cette molécule est orientée vers la droite et compte, en moyenne, 10 paires de bases par tour représentant un pas d'hélice de 34 Å. Ces paires de bases

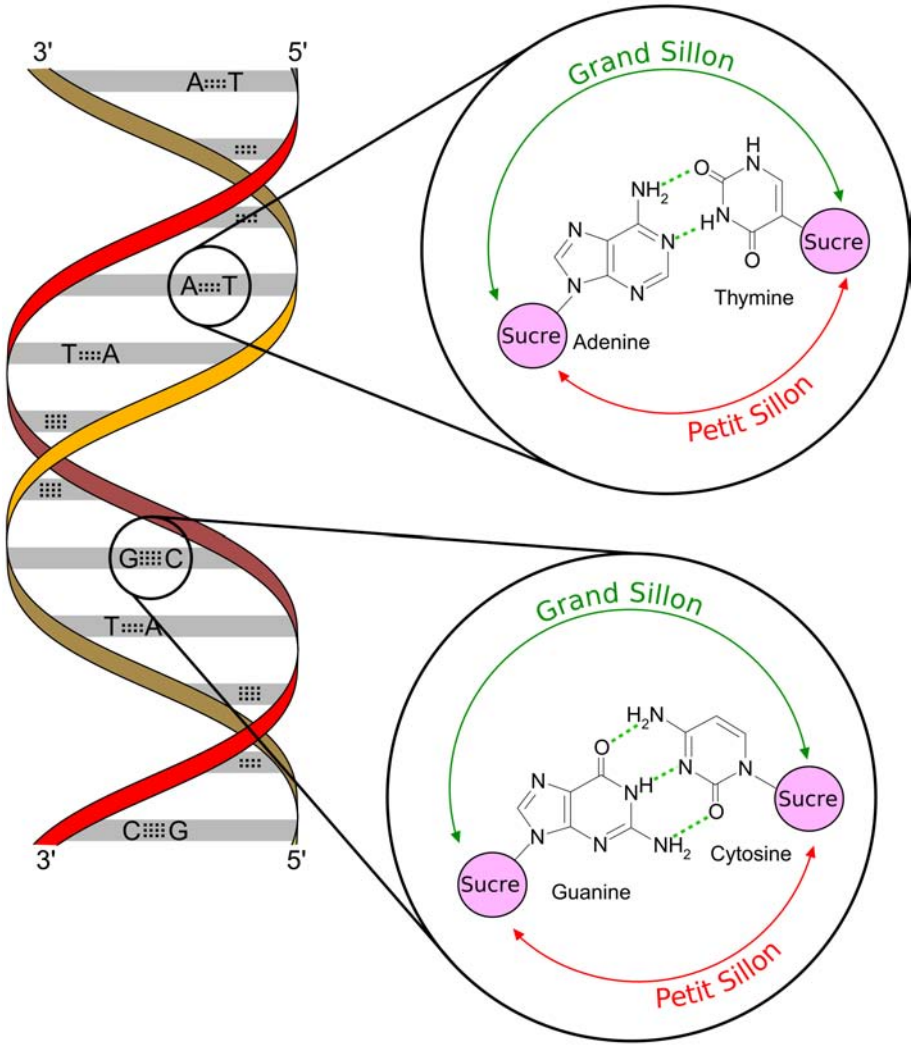


FIG. A.6 – Structure de l'ADN en double hélice. La structure de l'ADN est stabilisée par les paires de bases complémentaires : A-T et C-G. les liaisons hydrogène forment des appariements Watson-Crick. La position des sucres permet de définir un grand et un petit sillon représentant des zones d'accessibilité aux bases.

sont majoritairement perpendiculaires à l'axe de l'hélice.

Il existe aussi une forme A de la molécule d'ADN. Celle-ci se caractérise par des groupes phosphate plus rapprochés que dans le cas de la forme B (voir figure A.7). Ceci entraîne un éloignement des paires de bases par rapport à l'axe de l'hélice. Le pas de l'hélice est alors de 28 Å équivalent à 11 paires de bases par tour. Le diamètre de l'hélice est alors agrandi par rapport à la forme B. Cette conformation entraîne des changements conformationnels au niveau des sillons : le grand sillon devient profond et étroit tandis que le petit sillon s'élargit et devient superficiel.

Enfin, il existe une dernière forme *canonique* de molécule d'ADN : la forme Z (voir figure A.7). Ce type de molécule est orienté vers la gauche. Le pas de l'hélice est de 45 Å et il faut 12 bases pour faire un tour complet. Cette forme présente un grand sillon complètement convexe tandis que le petit sillon central s'approfondit et se remplit de molécules d'eau.

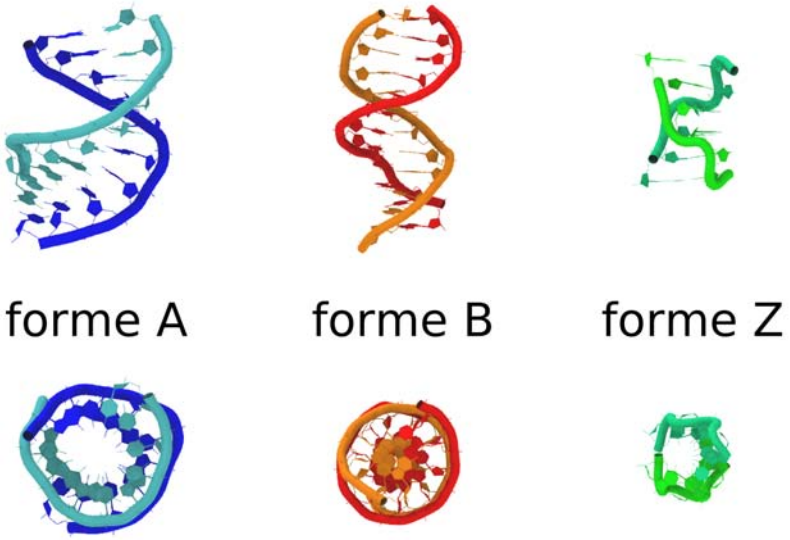


FIG. A.7 – Représentation des formes les plus classiques de la molécule d'ADN (vues de côté et vues de dessus) : la molécule d'ADN-A (PDB id : 414B), la molécule d'ADN-B (PDB id : 428D) et la molécule d'ADN-Z (PDB id : 1DNF).

Il existe d'autres assemblages que les doubles hélices d'ADN tels que les formes trimériques servant à la lecture des séquences d'ADN ou les formes tétramériques apparaissant dans les télomères.

La structure de la molécule d'ARN : Bien qu'en règle générale la molécule d'ARN soit simple brin, il existe des formes double brins. Dans ce cas, la paire thymine-adénine est remplacée par la paire uracile-adénine. Contrairement à la molécule d'ADN qui présente différentes formes de dimères, la molécule d'ARN se caractérise par une forme unique : la forme A. Cette forme a un pas d'hélice de 11 paires de bases et présente un grand sillon profond et étroit et un petit sillon large et superficiel. Cette forme d'ARN partage donc de nombreuses caractéristiques

avec la molécule d'ADN-A c'est pourquoi elle porte le même nom.

Les molécules d'ARN peuvent aussi former de nombreux mésappariements transformant la structure de celle-ci ; des protubérances apparaissent alors et la forme des sillons est transformée (voir figure A.8).

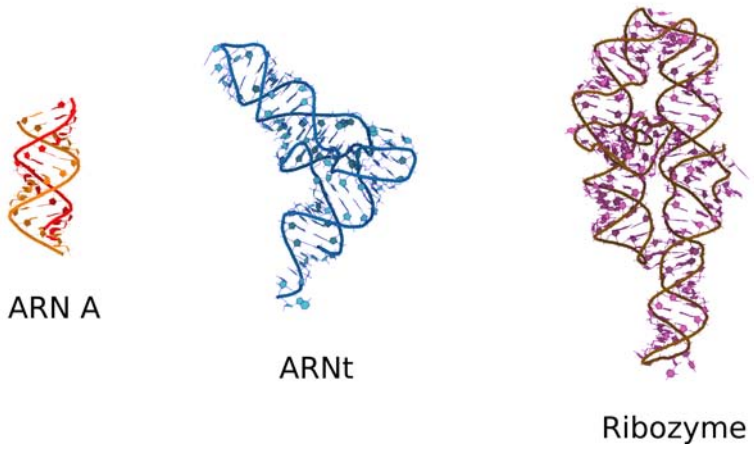


Fig. A.8 – Visualisation de 3 structures formées par des brins d'ARN : un dimère d'ARN (PDB id : 405D), un ARN de transfert (PDB id : 1EVV) et un ribozyme (PDB id : 1HR2).

Outre la fonction de messager de l'information génétique, les molécules d'ARN ont également d'autres fonctions comme les ARN de transfert ou les ribozymes (voir figure A.8). Les ribozymes (contraction de ribonucléique et enzyme) sont des molécules d'ARN capables de cliver d'autres molécules d'ARN. Il peuvent intervenir dans les phénomènes d'épissage en catalysant l'excision d'introns. Les ARN de transfert jouent un rôle primordial dans la phase de traduction qui permet de passer d'un ARN messager à une protéine. Les ARN de transfert prélèvent des acides aminés dans le cytoplasme et transfèrent ceux-ci au niveau du ribosome. Il ont une forme caractéristique en L avec deux bras en angle droit. Un bras est appelé bras accepteur : c'est sur celui-ci que se fixera l'acide aminé. L'autre bras est appelé bras anticodon et porte la boucle anticodon formée de 3 nucléotides permettant l'appariement à la molécule d'ARN messager.

Le ribosome est un assemblage d'un grand nombre de protéines et de molécules d'ARN ribosomique (ARNr). Sa fonction est de synthétiser les chaînes polypeptidiques à partir de molécules d'ARN messager et d'ARN de transfert. Le ribosome est constitué de deux sous-unités : la sous-unité 30S et la sous-unité 50S (voir figure A.9). De nombreuses études structurales ont été réalisées et se poursuivent pour comprendre les mécanismes exacts mis en jeu au niveau du ribosome.

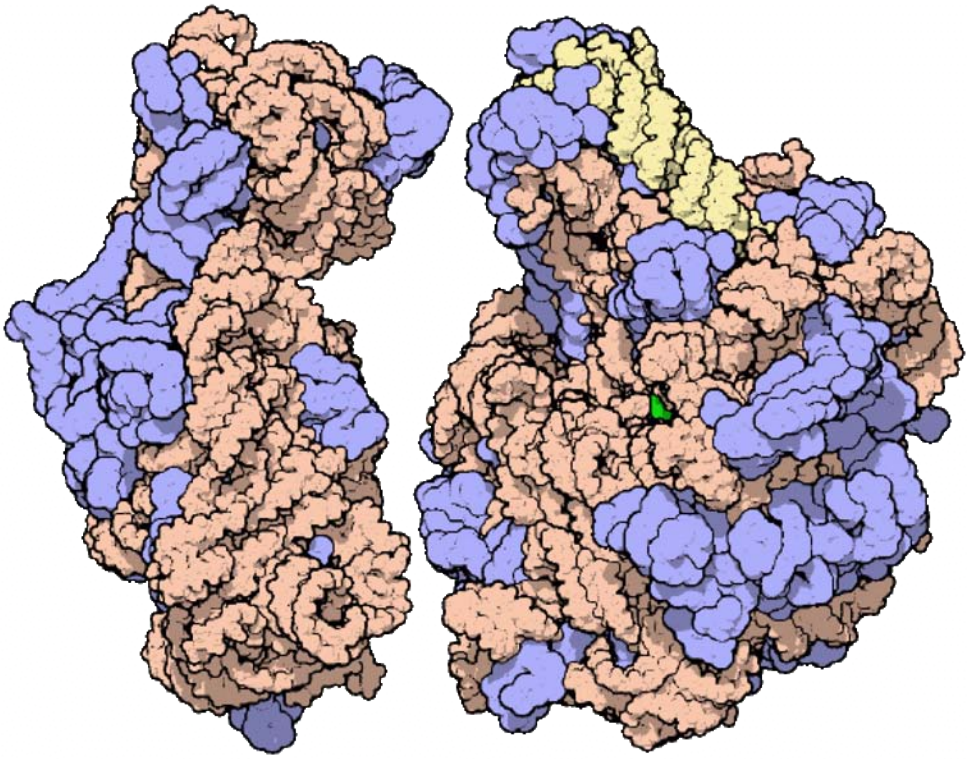


FIG. A.9 – Le ribosome : sous-unités 30S à gauche et 50S à droite. *Image créée par David Goodsell pour la série : Molecule of the Month.*

Annexe B

Calcul de la taille de l'interface à l'aide d'Intersurf

B.0.3 Définition de l'aire de l'interface

La taille de l'interface entre protéines est généralement mesurée par la différence entre l'aire de la Surface Accessible au Solvant (notée SAS : voir la section 3.2.1 pour la définition de cette surface) du complexe et celle des composés séparés. Cette mesure est appelée l'aire de la surface enfouie et est notée B :

$$B = Asas_{prot.1} + Asas_{prot.2} - Asas_{complexe}$$

Certaines études utilisent la valeur de la surface enfouie par sous-unité (Jones et Thornton, 1995, 1996). Dans ce cas, il est nécessaire de diviser B par deux.

Il existe divers programmes permettant de calculer cette aire d'interface. Nous citons, à titre d'exemple, NACCESS[33] (Hubbard et Thornton, 1992), programme à installer sur ordinateur. Il existe aussi divers serveurs comme PROTORP[34] (Reynolds *et al.*, 2008), PROFACE[35] (Saha *et al.*, 2006) ou INTERVOR[36] (Cazals *et al.*, 2006).

Enfin, H. Edelsbrunner *et al.* ont aussi développé un programme qui calcule certaines caractéristiques de l'interface macromoléculaire comme son aire ou sa forme (Ban *et al.*, 2004).

B.0.4 Mesure de l'aire de l'interface avec Intersurf

Nous avons modifié le programme Intersurf (Ray *et al.*, 2005). Ce programme visualisait l'interface de complexes macromoléculaires grâce à l'utilisation de la tétraédrisation de Delaunay mais ne calculait pas l'aire de cette interface (pour la version du Plugin fourni par VMD). Nous

[33] `http://www.bioinf.manchester.ac.uk/naccess/`
[34] `http://www.bioinformatics.sussex.ac.uk/protorp/`
[35] `http://202.141.148.29/resources/bioinfo/interface/`
[36] `http://cgal.inria.fr/Intervor/`

avons ajouté cette fonction en calculant la somme des aires de chaque triangle composant cette interface. Nous avons comparé les calculs de l'aire de l'interface de 70 complexes avec la valeur de l'aire de la surface enfouie calculée par Chakrabarti et Janin (Chakrabarti et Janin, 2002). Les valeurs respectives de l'aire de l'interface et de la surface enfouie sont présentées table B.1. Nous avons aussi comparé les valeurs obtenues avec celles calculées par le serveur PROTORP (Reynolds *et al.*, 2008). Les résultats de corrélation sont présentés figure B.1.

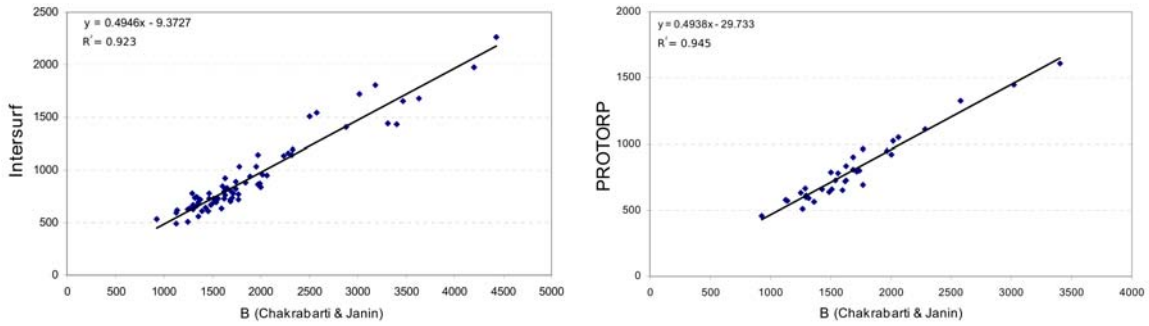


FIG. B.1 – Corrélation entre l'aire de l'interface calculée par Intersurf ou PROTORP et l'aire de la surface enfouies calculée par Chakrabarti et Janin.

Les valeurs de corrélation entre la surface calculée par Intersurf et la surface enfouie sont en accord : le facteur de corrélation R^2 étant égal à 0.923. Ce facteur est un peu moins élevé que celui de PROTORP ($R^2 = 0.945$) mais, contrairement à PROTORP, Intersurf permet de calculer l'aire de l'interface pour des oligomères et non pas seulement pour des dimères. Le facteur de corrélation reste en tout cas bien plus élevé que celui du programme développé par H. Edelsbrunner *et al.* ($R^2 = 0.742$) en utilisant les valeurs présentées dans la publication de *RECOMB '04* (Ban *et al.*, 2004). Toutes ces valeurs sont encore en deçà du facteur de corrélation de 0,98 obtenu avec INTERVOR (Cazals *et al.*, 2006).

Ces travaux restent, pour l'instant, préliminaires et des tests supplémentaires seront réalisés pour choisir au mieux certaines valeurs comme la distance entre les atomes considérés à l'interface : celle-ci est fixée actuellement à 9 Å mais pourrait être ajustée.

Les images et les calculs d'interfaces présentés dans les chapitres 1 et 2 de cette thèse ont été réalisés grâce à Intersurf.

Tab. B.1 – Comparaison des résultats d'Intersurf avec ceux de PROTORP pour les complexes présentés dans la publication (Chakrabarti et Janin, 2002)

Code PDB	Complexes	Aire de l'interface (en Å^2)		
		Chakrabarti et Janin	PROTROP[a]	Intersurf
Protease-inhibiteur (18)				
2ptc (E-I)	Trypsin-PTI	1430	660	648
1mct (A-I)	Trypsin-bitter gourd inhibitor	1510	663	725
1avw (A-B)	Trypsin-soybean inhibitor	1740	803	889
3tpi (Z-I)	Trypsinogen-PTI	1600	650	643
1tgs (Z-I)	Trypsinogen-PSTI	1720	795	778
1cho (FG-I)	Chymotrypsin-ovomucoid	1470	—	733
1acb (E-I)	Chymotrypsin-eglinC	1540	728	696
1cbw (BC-D)	Chymotrypsin-PTI	1460	—	613
1ppf (E-I)	Elastase-ovomucoid	1320	592	735
1fle (E-I)	Elastase-elafin	1770	692	717
2kai (AB-I)	Kallikrein-PTI	1340	—	745
1hia (AB-I)	Kallikrein-hirustatin	1740	—	824
3sgb (E-I)	*S. griseus* protease B-ovomucoid	1270	513	632
1cse (E-I)	Subtilisin-eglinC	1490	640	671
2sic (E-I)	Subtilisin-SSI	1620	723	731
2sni (E-I)	Subtilisin-CI2	1630	727	762
1stf (E-I)	Papain-stefin	1690	805	803
4cpa (A-I)	Carboxypeptidase A-inhibitor	1360	569	561
Large protease complexes (5)				
1bth (LH-P)	ThrombinE192Q-PTI	2240	—	1134
4htc (LH-I)	Thrombin-hirudin	3310	—	1438
1tbq (LH-R)	Thrombin-rhodniin	3470	—	1656
1dan (LH-TU)	Factor VIIA-soluble tissue factor	3180	—	1802
Antibody-antigen (18)				
1jhl (HL-A)	Fv D11.15-lysozyme	1250	—	630
1vfb (AB-C)	Fv D1.3-lysozyme	1380	—	723
1mlc (AB-E)	Fab D44.1-lysozyme	1390	—	607
1yqv (HL-Y)	Fab HyHEL5-lysozyme	1710	—	739
3hfm (HL-Y)	Fab HyHEL10-lysozyme	1610	—	851
1fbi (HL-X)	Fab 9.13.7-lysozyme	1690	—	723
1mel (A-L)	Camel H chain-lysozyme	1690		702
1dvf (AB-CD)	Fv D1.3-Fv E5.2	1630	—	820
1nfd (AB-EF)	Fab H57-N15 T cell receptor	1620	—	797
1ao7 (DE-A)	T cell receptor-HLA A2	1990	—	872
1jel (HL-P)	Fab Jel42-HPR	1360	—	694
1nca (HL-N)	Fab NC41-flu neuraminidase	1950	—	1037
1nmb (HL-N)	Fab NC10-flu neuraminidase	1290	—	783
1nsn (HL-S)	Fab N10-Staph. nuclease	1780	—	1038
1osp (HL-O)	Fab-*Borrelia* OSP A	1470	—	776
1qfu (HL-AB)	Fab BH151-flu H1X31	1840	—	882
1iai (HL-MI)	Fab 730.1.4-Fab 409.5.3	1890	—	943
1kb5 (HL-AB)	Fab Désiré-1-TCR Fv	2320	—	1142

TAB. B.2 – Comparaison des résultats d'Intersurf avec ceux de PROTORP *(suite)*.

Code PDB	Complexes	Aire de l'interface (en Å^2)		
		Chakrabarti et Janin	PROTROP[a]	Intersurf
Enzyme complexes (8)				
2pcc (A-B)	Peroxidase-Cytochrome c	1140	570	615
1gla (G-F)	Glycerol kinase-Factor IIIGlc	1300	615	671
1brs (A-D)	Barnase-Barstar	1560	781	731
1udi (E-I)	Uracil DNA glycosylase-inhibitor	2020	1025	956
1dhk (A-B)	A-Amylase-bean inhibitor	3020	1446	1724
1fss (A-B)	Acethylcholinesterase-fasciculin	1970	952	865
1ydr (E-I)	Protein kinase A-inhibitor	2000	921	843
1dfj (E-I)	RNase A-RNase inhibitor	2580	1328	1546
G-proteins, cell cycle, signal transduction (11)				
1a0o (A-B)	CheA-CheY	1130	579	495
1gua (A-B)	Rap1A-cRaf1	1290	669	632
1a2k (CD-B)	Ran-NFT2	1650	—	828
1agr (A-E)	$G_{1\alpha}$-RGS4	1630	832	920
1tx4 (B-A)	Rho-Rho GAP	2280	1110	1158
1gg2 (A-BG)	$G_{1\alpha}$- $G_{1\beta1\gamma2}$	2330	—	1196
1got (A-BG)	Transducin $G_{t\alpha}$- $G_{t\beta\gamma}$	2500	—	1507
2trc (BG-P)	$G_{t\beta\gamma}$-phosducin	4430	—	2259
1fin (A-B)	CDK2-cyclin A	3400	1609	1435
1aip (A-CD)	EFtu-EFts *T. thermophilus*	2880	—	1410
1efu (A-B)	EFtu-EFts *E. coli*	3630	—	1677
Miscellaneous (10)				
1ak4 (A-C)	Cyclophilin-HIV capsid	930	461	537
1igc (A-LH)	Protein G-Fab MOPC21	1130	—	595
1efn (A-B)	Fyn SH3 domain-HIV Nef	1250	630	510
1fc2 (D-C)	Protein A-FC fragment	1300	602	630
1seb (AB-D)	HLA DR1-enterotoxin B	1340	—	661
1atn (A-D)	Actin-DNase I	1770	960	772
1ycs (A-B)	p53 core-53BP2	1500	786	688
2btf (A-P)	Actin-Profilin	2060	1051	953
1hwg (BC-A)	HGH receptor-human growth hormone	4200	—	1976
1dkg (AB-D)	Grep E-DNA K	1970	—	1143
All interfaces	Average	1883	—	922

Annexe C

Calcul de la valeur de propension

Le terme de propension (*propensity*) est généralement défini comme le ratio entre la contribution d'un résidu à l'interface d'un complexe et la contribution de celui-ci au niveau de la surface des protéines.

Jones et Thornton ont défini ce terme comme :

$$\text{propension d'un acide aminé } AA_j = \frac{\sum_{i=1}^{N_i} ASA_{AA_j(i)} / \sum_{i=1}^{N_i} ASA_{(i)}}{\sum_{i=1}^{N_s} ASA_{AA_j(s)} / \sum_{i=1}^{N_s} ASA_{(s)}}$$

où $\sum_{i=1}^{N_i} ASA_{AA_j(i)}$ représente la somme des surfaces accessibles au solvant (ASA : voir annexe précédente) des résidus de type j à l'interface. $\sum_{i=1}^{N_i} ASA_{(i)}$ la somme des surfaces accessibles au solvant de tous les types de résidus à l'interface. $\sum_{i=1}^{N_s} ASA_{AA_j(s)}$ représente la somme des surfaces accessibles au solvant des résidus de type j à la surface d'une protéine. $\sum_{i=1}^{N_s} ASA_{(s)}$ la somme des surfaces accessibles au solvant de tous les types de résidus à la surface d'une protéine. N_i est le nombre de résidus à l'interface et N_s est le nombre de résidus présent à la surface d'une protéine (Jones et Thornton, 1996).

Cette définition a été utilisée dans plusieurs travaux comme ceux de Jones *et al.*, Nadassy *et al.* ou Ellis *et al.* (Jones et Thornton, 1997; Jones *et al.*, 1999; Nadassy *et al.*, 1999; Jones *et al.*, 2001; Ellis *et al.*, 2007).

Ce terme de propension a aussi été défini par Kim *et al.* comme un ratio de fréquences (Kim *et al.*, 2006). Soit f_i la fréquence d'un acide aminé AA_i à la surface d'une protéine et $\overline{f}_i$ la fréquence d'un acide aminé AA_i à l'interface d'un complexe.

$$f_i = \frac{n_i}{\sum_{i=1}^{20} n_i}, \overline{f}_i = \frac{\overline{n}_i}{\sum_{i=1}^{20} \overline{n}_i}$$

où n_i est le nombre d'acides aminés de type i à la surface de la protéine et $\overline{n}_i$ est le nombre d'acides aminés de type i à l'interface du complexe.

La propension P_i pour un acide aminé AA_i s'écrit donc :

$$P_i = \frac{\overline{f}_i}{f_i}$$

Cette définition a été reprise par Ponstingl *et al.* ou par Lejeune *et al.* (Lejeune *et al.*, 2005; Ponstingl *et al.*, 2005).

Enfin, certains travaux présentent aussi les deux types de calculs basés sur la surface accessible au solvant ou sur la fréquence des résidus à l'interface (Bahadur *et al.*, 2003, 2008).

Annexe D

Le pipeline graphique

Au milieu des années 90, les cartes graphiques, qui jusque là se limitaient à afficher l'image construite sur les processeurs de l'ordinateur (CPU), furent dotées de processeurs dédiés : les GPU. Ces derniers ont acquis au fur et à mesure de leur évolution une mémoire propre dédiée, la capacité à dessiner des objets 2D puis 3D. L'ensemble des étapes de traitement intervenant dans un GPU est regroupé sous l'appellation : *pipeline graphique*.

Comme le montre la figure D.1, le pipeline prend en entrée des sommets qui peuvent être assemblés en primitives (points, lignes, triangles ou polygones). Ces primitives sont ensuite discrétisées en fragments durant l'étape de *rasterisation*. Les fragments peuvent être considérés comme des "pré-pixels" en trois dimensions. Ces fragments vont ensuite subir des transformations pour être transformés en pixels. Ces pixels seront stockés dans la mémoire écran (*framebuffer*) pour donner l'image finale. Deux unités sont essentielles dans le pipeline graphique : l'unité de traitement des sommets et l'unité de traitement des fragments. Vers le début des années 2000, ce pipeline a été ouvert à l'exécution de code utilisateur en permettant aux unités de sommets et de fragments de devenir programmables. Ces unités ont alors été renommées "processeur de fragments" et "processeur de sommets" (ou respectivement *Vertex Shader* et *Fragment Shader*).

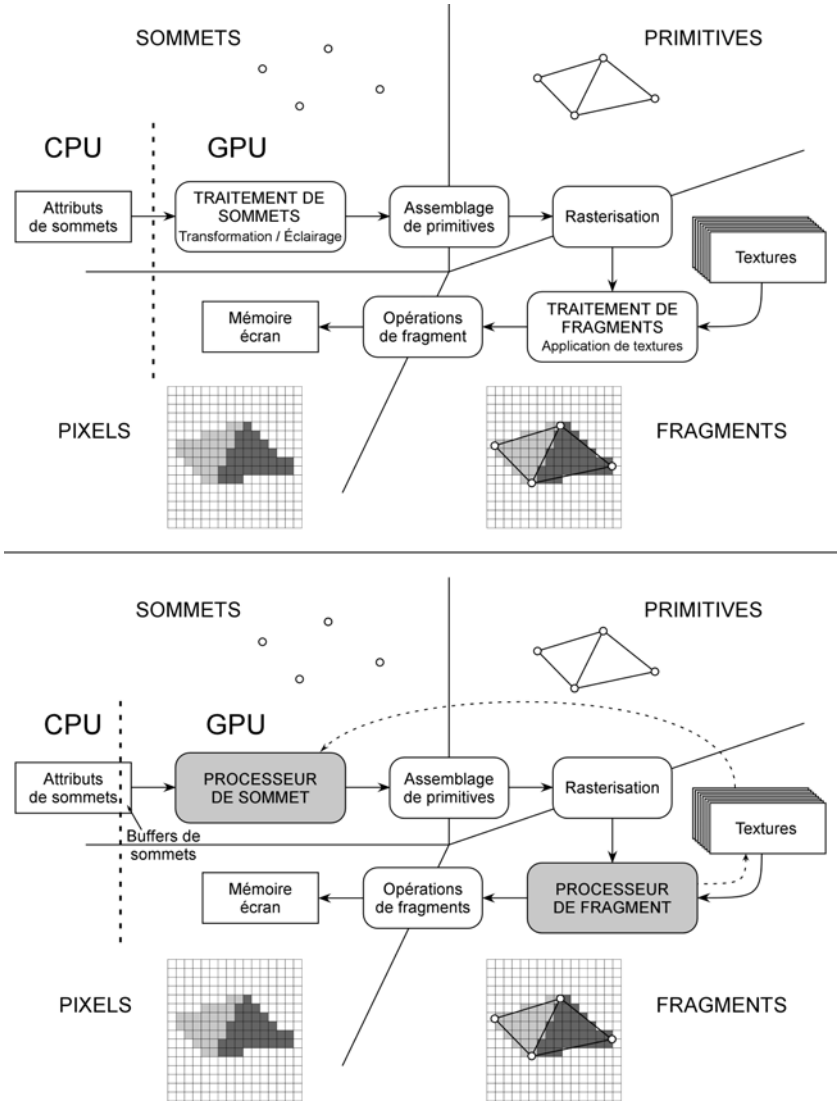


FIG. D.1 – Principe du pipeline graphique standard (haut) et son évolution vers plus de flexibilité depuis le début des années 2000 (bas). *Figure issue de (Castanié, 2006)*

Annexe E

Articles publiés

Biochemical and Biophysical Research Communications xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Biochemical studies and molecular dynamics simulations of Smad3–Erbin interaction identify a non-classical Erbin PDZ binding

Nadine Déliot [a,b,c,1], Matthieu Chavent [d,1], Claire Nourry [a,b,c,1], Patrick Lécine [a,b,c], Camille Arnaud [a,b,c], Aurélie Hermant [a,b,c], Bernard Maigret [d], Jean-Paul Borg [a,b,c,*]

[a] Inserm, U891, Centre de Recherche en Cancérologie de Marseille, Pharmacologie Moléculaire, F-13009 Marseille, France
[b] Institut Paoli-Calmettes, F-13009 Marseille, France
[c] Univ Méditerranée, F-13007 Marseille, France
[d] Equipe Orpailleur, LORIA Campus Scientifique, 54506 Vandoeuvre-lès-Nancy, France

ARTICLE INFO

Article history:
Received 29 October 2008
Available online xxxx

Keywords:
Erbin PDZ
Smad3 MH2
Proteins interaction
Non-canonical PDZ binding
Electrostatic interactions

ABSTRACT

In this work, we describe how the Erbin PDZ domain interacts with Smad3, a transductor of the Transforming Growth Factor-beta (TGFβ) pathway, via its MH2 domain. This interaction was described as important for TGFβ signaling as it could potentially repress the transcriptional activity of the growth factor. In order to clarify our preliminary experimental observations pointing this interaction, we built a 3D model of the Erbin PDZ/Smad3 MH2 complex and checked its stability using molecular dynamics simulations. This model pointed out charged residues in Smad3 and Erbin which could be important for the interaction. By introducing point mutations of these residues within the proposed binding domains, we experimentally confirmed that arginine 279, glutamic acid 246 in Smad3 and glutamic acid 1321 in Erbin are important for the binding. These data suggest a possible novel interface of binding in the Erbin PDZ domain and reveal an unconventional mode of interaction for a PDZ domain and its ligand.

PDZ (*P*SD-95, *D*iscs-Large and *Z*O-1) domains are ~90 residues in length and adopt a common fold consisting of a β-barrel comprising 6 β-strands (βA to βF) capped by two α-helices (αA and αB) [1]. Most frequently, PDZ domains recognize the very C-terminal peptide of proteins, thereby bringing signaling pathway components into proximity [2]. Erbin is a scaffold cytoplasmic protein originally identified as an interactor for the receptor tyrosine kinase ErbB2 [3]. Erbin belongs to the LAP (*L*RR *A*nd *P*DZ) protein family which is composed of 16 LRRs (*L*eucine-*R*ich *R*epeats) in the amino-terminal position and PDZ domains in the carboxy-terminal region [4]. Erbin contains a single PDZ domain able to bind to the VPV carboxy-terminal sequence of ErbB2 (class II motif) or with class I interacting proteins [5,6]. This versatility can be explained in part by the structure of the Erbin PDZ domain that deviates from the canonical PDZ fold in that it contains a single α-helix [7,8].

The Smad proteins are transducers of TGFβ receptors and are categorized in three distinct groups based on function and sequence homologies [9]. The receptor-regulated Smads (R-smad) are substrates for TGFβ type I receptors (Smad 2, 3, 5, 8). Smad4 is a "common" Smad that heterodimerizes with phosphorylated R-Smads. Finally, the inhibitory Smads (Smad 6 and 7) antagonize R-Smad functions and therefore TGFβ signalling. Smad3 shares a common domain configuration with other R-smads and Smad4 consisting of an amino-terminal DNA-binding domain (MH1 domain) and a carboxy-terminal effector domain (MH2 domain) separated by a linker region. At the plasma membrane, the MH2 domain of Smad3 interacts with SARA [10] and with the TGFβ receptor [11,12]. After dissociation from the receptor, the MH2 domain of Smad3 interacts with Smad4. Following dimerization or trimerization, Smad3 and Smad4 translocate into the nucleus to turn on a transcriptional program [9].

Using biochemical approaches, we have shown that the Erbin C-terminal region comprising the PDZ domain is required to interact with the Smad3 MH2 domain. To unravel the molecular mechanism of this association, we used a protein docking strategy to model this interaction. According to the results obtained from the simulations, residues within the Smad3 MH2 and the Erbin PDZ domains were pinpointed for their role in the interaction. We validated their role by introducing point mutations in Smad3 and Erbin, and by experimentally confirming their contribution. According to our model, the PDZ domain of Erbin is engaged in a novel interaction that does not require the classical binding pocket.

* Corresponding author. Address: Inserm, U891, Centre de Recherche en Cancérologie de Marseille, Pharmacologie Moléculaire, F-13009 Marseille, France. Fax: +33 (0)4 91 26 03 64.
E-mail address: jean-paul.borg@inserm.fr (J.-P. Borg).
[1] These authors contributed equally to this work.

0006-291X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2008.10.175

Materials and methods

Plasmids and antibodies. Antibodies directed against Smad3 (FL425, I20), Smad1/2/3 Myc (9E10), and Scrib (C20) were obtained from Santa Cruz. The anti-β-catenin antibody was purchased from BD Transduction Laboratories. Anti-HA (4F10) and anti-Flag (M2) antibodies were purchased from Roche and Sigma, respectively. The rabbit polyclonal anti-Erbin antibody was described previously [3].

The human Erbin constructs were described previously [3]. The human Smad3 cDNA (generous gift from Dr. T. Wang) was cloned into pcDNA-HA and pCMV6-Flag. The pcDNA-HA or pDEST15 Smad3 and pDEST15-Erbin (GST fused proteins) constructs were obtained by PCR using the Gateway cloning system (Invitrogen). HA and GST-Smad3-MH2 (residues 229–425) and GST-Erbin$^{1257-1371}$ were mutated using the site-directed mutagenesis Quick Change kit according to manufacturer's instructions (Stratagene) to obtain the following mutants: $E_{246}L$, $D_{262}A$, $R_{279}M$, $R_{279}E$, $E_{284}A$, $E_{1321}R$ and $E_{1321}L$.

Cell culture. Non-transformed human mammary epithelial MCF10-2A cells were cultured in Ham'sF12/DMEM Glutamax (Gibco) supplemented with 5% horse serum, 10 μg/mL insulin, 20 ng/mL EGF, 100 ng/mL cholera toxin, 500 ng/mL hydrocortisone, 100 U/mL penicillin and 100 μg/mL streptomycin. COS cells were grown in Dulbecco's Modified Eagle's Medium-glutamax (DMEM, Gibco) containing 10% fetal calf serum and transfected using FuGENE6 reagent (Roche) according to manufacturer's instructions. Two days after transfection, cells were lysed in lysis buffer (50 mM HEPES pH 7.5, 10% glycerol, 150 mM NaCl, 1% Triton X100 1.5 mM $MgCl_2$, 1 mM EDTA).

Biochemical experiments. For GST pull-down assays, cell lysates were incubated with 4 μg of GST-fusion proteins for 2 h at 4 °C. Beads were then washed once with lysis buffer and twice with HNTG buffer (50 mM HEPES pH 7.5, 10% glycerol, 150 mM NaCl, 0.1% Triton X-100). For immunoprecipitation, 300 μg of cell lysates were incubated with 2 μg of the indicated antibodies overnight at 4 °C followed by additional 1 h incubation with protein A or protein G agarose beads. Beads were then washed once with lysis buffer and twice with HNTG buffer. Cell extracts and immunoprecipitates were analyzed by immunoblotting using the indicated antibodies.

Modeling interacting domains, docking and molecular dynamics calculations. To model the Erbin interacting region of interest (residues 1257–1371), we used as a starting point the crystal structure of the human Erbin PDZ (PDB Id: 1MFG [8]) that included residues A_{1277} to S_{1371}. In order to have the structure of the whole C-terminal interacting domain, it was necessary to model the part between V_{1257} and L_{1276}. For that purpose, we used the PDZ domain of Par6 (PDB Id: 1NF3 [13]). We used the crystal structure of human Smad3 MH2 domain (PDB Id: 1MJS [14]). As a template to model the complex between the Erbin PDZ and Smad3 MH2 domains, we used the published Par6/Cdc42 complex [13]. In order to check the stability of the obtained 3D model of the complex, an energy minimization procedure was used followed by 16 ns molecular dynamics (MD) simulation recorded for the system immerged in a explicit solvent box (details in Supplementary material).

Results and discussions

Smad3 interacts with the C-terminal region of Erbin

In an effort to identify partners for Erbin, we and other groups performed yeast two hybrid screens and, among the obtained clones, proteins of the Smad family were pulled out [15]. By this

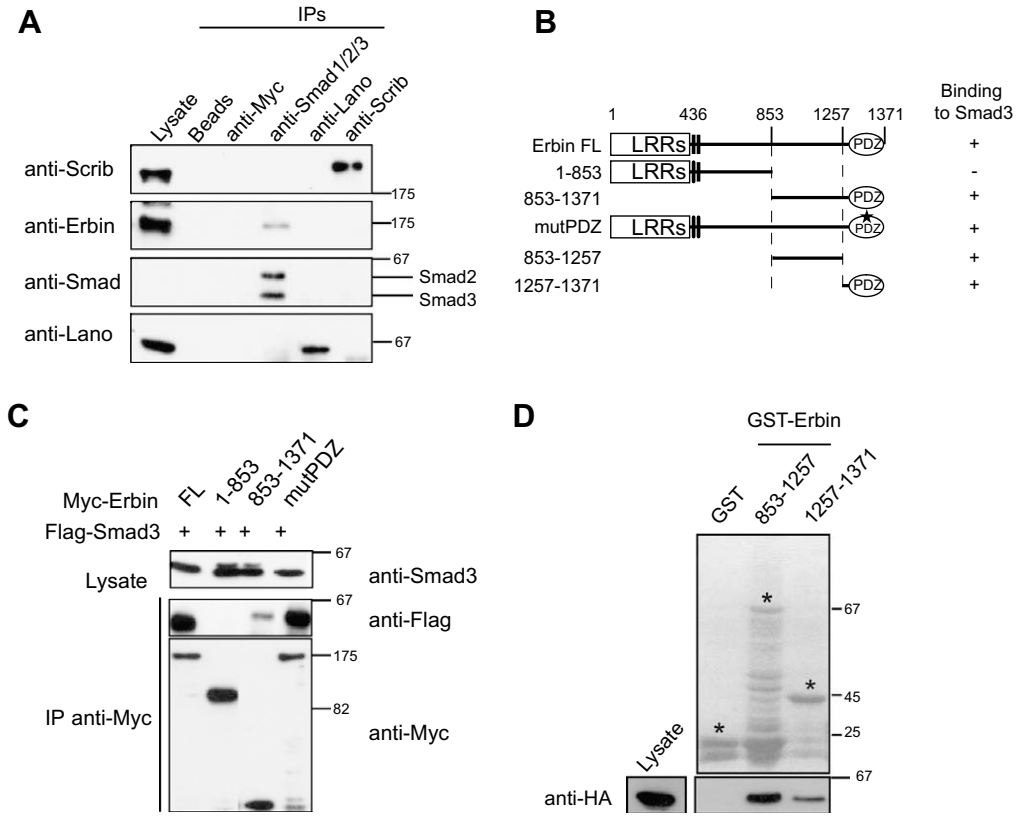


Fig. 1. Smad3 interacts with the carboxy-terminal region of Erbin. (A) MCF10.2A cell extracts were subjected to immunoprecipitation with anti-Myc and anti-Smad1/2/3 antibodies. The anti-Smad3 antibody recognizes Smad2 and Smad3. Expression levels of endogenous Smad2 and Smad3 are too weak to be detected in total cell lysates. (B) Scheme of Erbin constructs. (C) COS cells were transfected with Flag-Smad3 and Myc-Erbin constructs Cell lysates were subjected to immunoprecipitation with anti-Myc antibody. (D) Pull-down assay on COS cell extracts expressing HA-Smad3s. Asterisks indicate the position of the GST proteins revealed by Ponceau Red staining.

Please cite this article in press as: N. Déliot et al., Biochemical studies and molecular dynamics simulations of Smad3–Erbin ..., Biochem. Biophys. Res. Commun. (2008), doi:10.1016/j.bbrc.2008.10.175

method, a direct interaction between Smad3 and Erbin was characterized. A detailed picture of this novel interaction remains elusive so that we first tried to confirm it by biochemical means. Proteins extracted from MCF10.2A, a mammary epithelial cell line, were subjected to immunoprecipitation using a pan anti-Smad antibody recognizing Smad1, Smad2 and Smad3 (Smad1/2/3), or anti-Myc antibodies. After western blot with appropriate antibodies, Erbin was present in the anti-Smad immunoprecipitation. Lano and Scrib, two homologs of Erbin, did not interact with Smad proteins (Fig. 1A). Identical results were obtained using specific anti-Smad3 antibody for immunoprecipitation (data not shown). We then mapped the Smad3 binding domain in Erbin. First, we cotransfected HA-Smad3 with Myc-Erbin full length (FL), Myc-Erbin$^{1-853}$ (1–853) and Myc-Erbin$^{853-1371}$ (853–1371) constructs in COS cells (Fig. 1B). Immunoprecipitation with anti-Myc antibody showed that Smad3 interacts with the carboxy-terminal region of Erbin (853–1371) that contains the PDZ domain (Fig. 1C). In order to evaluate the role of the PDZ domain in this interaction, we expressed a mutated version of Erbin unable to bind PDZ ligands (Myc-Erbin mutPDZ). Two point mutations (residues $H_{1347}G_{1348}$ changed to $Y_{1347}D_{1348}$) introduced into the helix αB of the Erbin PDZ domain were shown to provoke a poor binding of the Erbin PDZ domain to class I and II ligands [3]. This mutation does not abolish Smad3 binding, suggesting that either the PDZ domain is not involved in this interaction or that the classical binding pocket is not required for the binding (Fig. 1C).

To map the Erbin region involved in the interaction with Smad3, we also used different constructs of Erbin fused to the GST. We showed that Erbin$^{853-1257}$ (853–1257) and Erbin$^{1257-1371}$ (1257–1371) interact with Smad3 (Fig. 1D). Interestingly, the fragment 1257–1371 of Erbin also contains the PDZ domain while Smad3 does not contain an obvious PDZ binding motif in its C-terminal position. As demonstrated hereafter, the PDZ alone fused to the GST protein (GST-Erbin$^{1280-1371}$) is able to bind to Smad3 (Fig. 4C).

Our data suggest therefore that the interaction between Smad3 and the C-terminal region of Erbin implies several protein moieties and is specific since Erbin homologs can not bind Smad3 (Fig. 1A). We identified the Smad3 MH2 domain binding sites in the Erbin C-terminal part where two regions (residues 853–1257 and 1257–1371, Fig. 1D) were shown to bind to Smad3. These data are in partial agreement with two recent papers describing a direct interaction between the Smad3 MH2 domain and Erbin: Warner et al. identified the Erbin 1004–1280 sequence as important for the binding whereas Dai et al. narrowed down the interacting region to residues 1172–1282 [15,16]. We have also shown that the Erbin PDZ alone (residues 1280–1371) can interact with Smad3. Warner et al. proposed, by two-hybrid assay in yeast, that this domain is not sufficient for Erbin to bind to Smad3. Nevertheless, a deletion of the PDZ domain in Erbin was shown to diminish the Smad3/Erbin interaction suggesting a potential contribution of this domain [15].

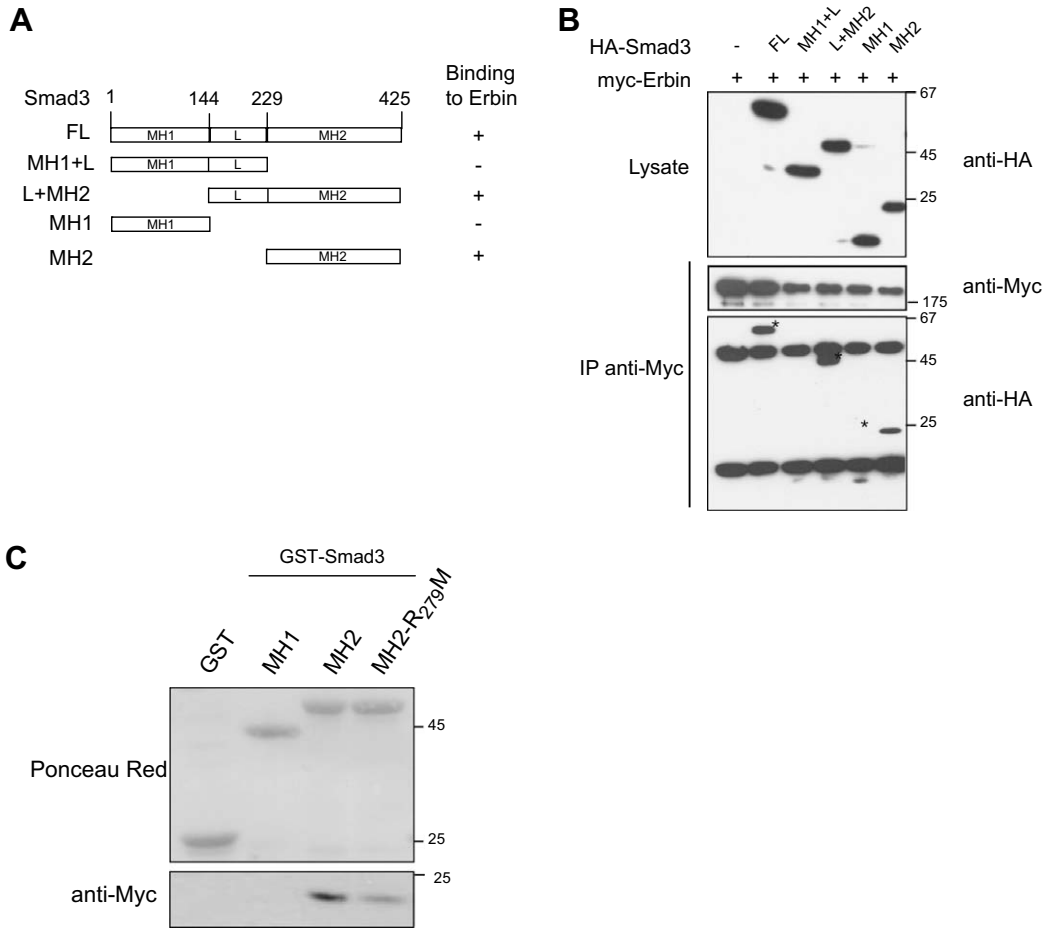


Fig. 2. The Smad3 MH2 domain interacts with the carboxy-terminal region of Erbin. (A) Scheme of Smad3 constructs. (B) COS cells were transfected with the indicated HA-Smad3 constructs and Myc-Erbin. Cell lysates were subjected to immunoprecipitation with anti-Myc antibody. Asterisks indicate the HA-Smad3 constructs coimmunoprecipitated with Erbin (FL, L+MH2 and MH2). (C) The mentioned GST proteins were used to pulldown Myc-Erbin$^{1257-1371}$ revealed with anti-Myc antibody (bottom panel). The MH2 domain binds to Erbin. The $R_{279}M$ mutation decreases the binding to Erbin.

Please cite this article in press as: N. Déliot et al., Biochemical studies and molecular dynamics simulations of Smad3–Erbin ..., Biochem. Biophys. Res. Commun. (2008), doi:10.1016/j.bbrc.2008.10.175

The MH2 domain of Smad3 is required for the Erbin–Smad3 interaction

We next characterize the region of Smad3 involved in binding to Erbin. HA-tagged Smad3 constructs corresponding to the different regions of the protein (Fig. 2A) were transfected into COS cells together with a Myc-tagged version of Erbin. Immunoprecipitation using anti-Myc antibody confirmed that the Smad3 MH2 domain interacts with Erbin (Fig. 2B), a result in agreement with previous reports [15,16]. This interaction was confirmed by GST pull-down using GST-MH1 and MH2 fused proteins in presence of Myc-Erbin (Fig. 2C). Using different truncation MH2 constructs, we could show that the carboxyl-terminal SSVS residues of Smad3 are not required for the interaction (data not shown). We thus conclude that Erbin and Smad3 are engaged in an interaction involving the MH2 domain of Smad3.

Molecular dynamics analysis highlights strong charged interactions

To better understand this interaction involving a PDZ domain, we build a 3 D model of the complex and try to identify particular residues involved in the interaction between the C-terminal region of Erbin (residues 1257–1371) and the Smad3 MH2 domain.

The interface between the two proteins has a classical size for a protein–protein heterodimer: ~1600 Å^2 [17]. This contact area is divided in two parts: one is between the Erbin residues 1257–1280 and the MH2 domain (~1200 Å^2); the other is between the PDZ itself and the MH2 domain (~400 Å^2) (Supplementary Fig. 4A). The analysis of the interface properties along the 16 ns Molecular Dynamics (MD) trajectory revealed interesting features concerning these two interacting parts. Firstly, at the 1200 Å^2 interface, there are mostly hydrophobic and polar interactions between the Erbin peptide chain (residues 1257–1280) and the MH2 domain (Supplementary Fig. 4B). These interactions are numerous and stable but not really strong according to the obtained energy profiles (Fig. 3). On the contrary, we identified stronger interactions between charged residues which are located in the second smaller area of ~400 Å^2 (Fig. 3). Here, three couples of interacting residues are found (presenting similar energy profiles): Smad3-E_{246} interacts with Erbin-R_{1271}, Smad3-D_{262} interacts with Erbin-K_{1326} and Smad3-R_{279} in Smad3 interacts with Erbin-E_{1321}. Furthermore, close to Smad3-R_{279}, we observed that R_{287} and R_{288} also interact with Erbin-E_{1321} with less strength (Fig. 3 and Supplementary Fig. 2). Thus, from our MD simulations, we highlighted two possible sets of anchoring: one composed of numerous weak interactions (hydrophilic and hydrophobic) and another driven by strong electrostatic interactions. Furthermore, our model predicts that these interactions are situated at the opposite side of the canonical PDZ binding site (Fig. 3 and Supplementary Fig. 4A).

Glutamic acid 246 and arginine 279 in the MH2 domain are important for the Smad3–Erbin interaction

To confirm our predicted model, we experimentally introduced point mutations in the MH2 domain of Smad3 that target charged residues important for the interaction with Erbin (E_{246}L, D_{262}A and R_{279}M). A MH2 E_{284}A mutant was produced as a control. The HA-tagged Smad3 MH2 constructs were transfected into COS cells

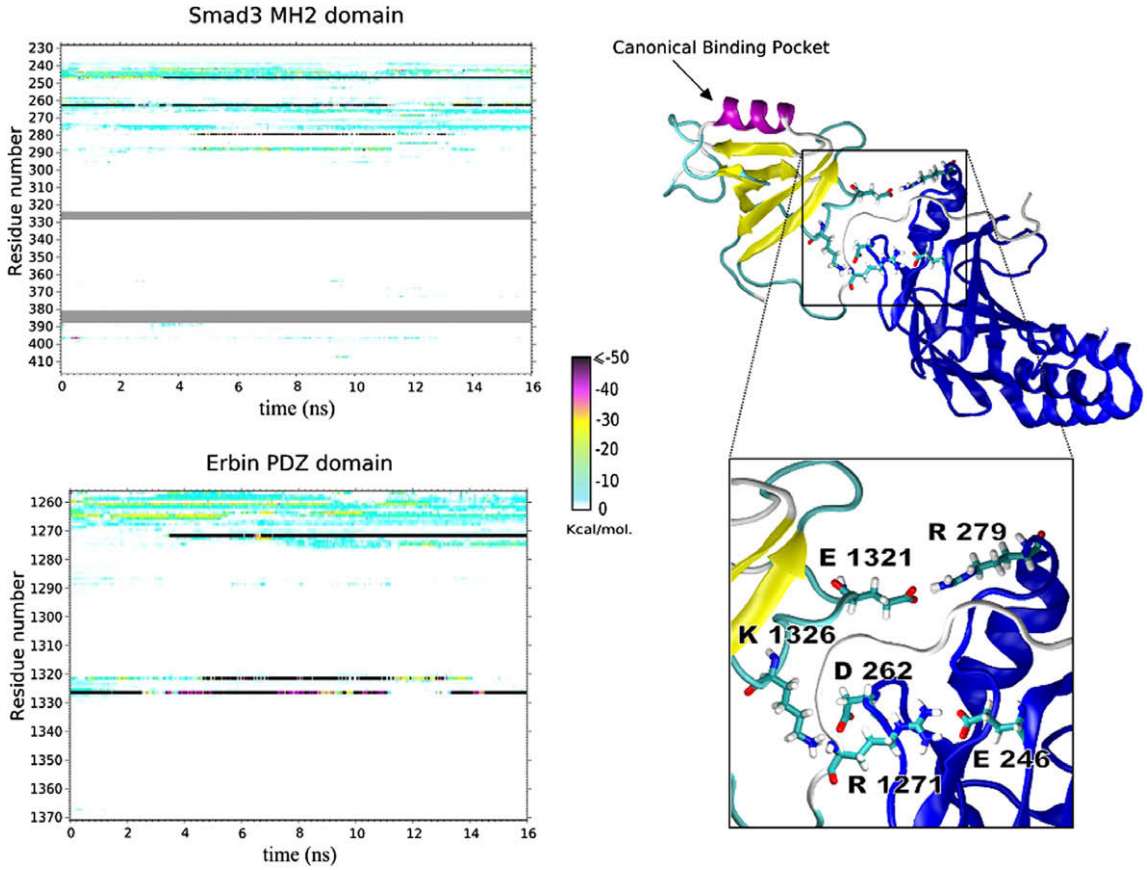


Fig. 3. Model of the interaction between the PDZ and MH2 domains. Interaction energy profiles for the PDZ and MH2 domains, and representation of the complex with the crucial interaction residues. Colored lines represent interaction energy from light blue (weak interactions) to black (strong interactions). The gray areas in the MH2 domain diagram represent the sequences that are too flexible to be represented in the structure. The snapshot was taken at 11ns where all charged interactions were present. The Smad3 MH2 domain is painted in blue. The Erbin PDZ domain is colored in function of the structure. α-Helices are in purple and β-strands are in yellow. An arrow points out the canonical binding pocket of the Erbin PDZ domain.

Please cite this article in press as: N. Déliot et al., Biochemical studies and molecular dynamics simulations of Smad3–Erbin ..., Biochem. Biophys. Res. Commun. (2008), doi:10.1016/j.bbrc.2008.10.175

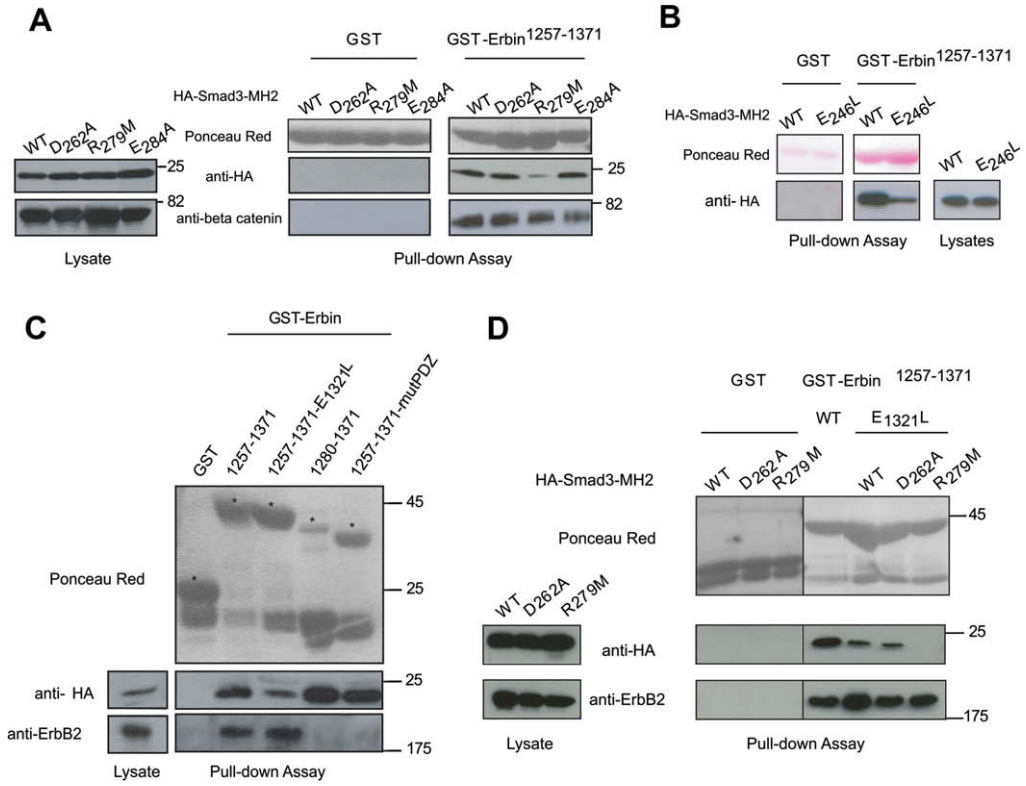


Fig. 4. R_{279}, E_{246} in the MH2 domain and E_{1321} in PDZ domain are important for the Smad3–Erbin interaction. (A) Pull-down assays with recombinant GST fused to Erbin$^{1257-1371}$. Wild type (WT) or single mutants ($D_{262}A$, $R_{279}M$ $E_{284}A$) HA-Smad3 MH2 were expressed into COS cells. The pull-down was analyzed using anti-HA antibody and β-catenin antibody as a control. (B) Same experiments using the HA-Smad3-MH2 $E_{246}L$ mutant. (C) COS cells were transfected with HA-Smad3-MH2 construct and lysate was incubated with different versions of GST-Erbin$^{1257-1371}$ (wild type, $E_{1321}L$ or mutPDZ mutants) or with GST-Erbin$^{1280-1371}$. Asterisks indicate the position of the GST proteins revealed by Ponceau Red staining (D) Same experiments using the previous described HA-Smad3-MH2 mutants precipitated by the $E_{1321}L$ mutant of GST-Erbin$^{1257-1371}$.

and expressed proteins were pulled-down using the GST-Erbin $^{1257-1371}$. GST-Erbin$^{1257-1371}$ precipitated wild type, $D_{262}A$ and $E_{284}A$ Smad3 MH2 mutants but had less affinity for the Smad3 MH2 $R_{279}M$ mutant and $E_{246}L$ mutants (Fig. 4A and B). β-catenin, a known Erbin PDZ partner was used as a control [18]. Conversely, a GST-MH2 $R_{279}M$ precipitated weakly Myc-Erbin$^{1257-1371}$ compared to GST-MH2 (Fig. 2C). Similar results were obtained with the full version of Smad3 (data not show).

Taken together, the *in vitro* experiments confirmed the predicted model of the interaction between Smad3 MH2 and PDZ of Erbin, and pointed out residues E_{246} and R_{279} of Smad3 as residues implicated in the Erbin–Smad3 interaction. The arginine 279 of MH2 domain when mutated leads to a decrease of binding to Erbin (Supplementary Fig. 5). R_{279} is located in the L2 loop of the Smad3 MH2 domain which is involved in Smad heteromerization and regulation [19,20]. As R_{279} was shown to be important for Smad proteins heteromerization and regulation [19], we focused on its interacting residue in Erbin (E_{1321}) revealed by the 3D model.

Glutamic acid 1321 in the Erbin PDZ domain is involved in the interaction with Smad3

To further validate our model, we mutated glutamic acid 1321 (E_{1321}) in the Erbin PDZ domain. This residue is predicted to interact with arginine 279 (R_{279}) of Smad3 (Fig. 3 and Supplementary Fig. 2). Glutamic acid was changed to leucine ($E_{1321}L$) in the GST-Erbin$^{1257-1371}$ protein. A pull-down was performed using COS cell extracts containing HA-Smad3-MH2. The $E_{1321}L$ mutation reduced the binding to HA-Smad3-MH2 but not to ErbB2 (Fig. 4C). Like already described in the Fig. 1B, the specific mutation in PDZ domain of Erbin (GST-Erbin$^{1257-1371}$ mutPDZ) abrogated the interaction with ErbB2 but not with Smad3 (Fig. 4C). We also produced a GST protein fused to the PDZ alone (GST-Erbin$^{1280-1371}$). This recombinant protein is difficult to express in bacteria. Absence of the residues amino-terminal to the PDZ domain (1257–1276) may impair the folding and/or the stability of the domain and so this domain is no longer able to bind ErbB2 (Fig. 4C). As it was shown that the (1280–1371) region non-fused to the GST is able to bind ErbB2 in solution [13], the presence of the GST may also impinge on the ErbB2 PDZ interaction. Nevertheless, GST-Erbin $^{1280-1371}$ still binds to Smad3 (Fig. 4C) suggesting that the PDZ domain possesses different protein interfaces to bind Smad3 and ErbB2. Both mutations R_{279} in the Smad3 MH2 domain and E_{1321} in the Erbin PDZ domain partially impair the interaction.

Potential interaction between R_{279} of Smad3 and E_{1321} of Erbin

To evidence the interaction between R_{279} and E_{1321}, we pulled-down Smad3-MH2 $R_{279}M$ with GST-Erbin$^{1257-1371}$ $E_{1321}L$. As shown in Fig. 4D, the interaction was completely abrogated in the presence of the combined mutations. It must be noted that the $D_{262}A$ Smad3 MH2 mutant has no effect on the interaction like previously demonstrated (Fig. 4A). Again, binding to ErbB2 was not affected by the $E_{1321}L$ mutation (Fig. 4D). These results suggest that arginine 279 and glutamic acid 1321 potentially interact according to our model (Fig. 3 and Supplementary Fig. 2).

Thus, taken together, these data are in agreement with our predicted model suggesting that (1) the Erbin PDZ and Smad3 MH2 domains interact via an interface independent of the classical PDZ binding pocket, (2) that charged residues R_{279}, E_{246} in Smad3 and E_{1321} in Erbin are important for this interaction. R_{279} is conserved in the Smad family and was previously predicted to play a role in the heteromerization between Smad3 and Smad4 [19]. It contributes to the strong electrostatic interaction at the heterotri-

Please cite this article in press as: N. Déliot et al., Biochemical studies and molecular dynamics simulations of Smad3–Erbin ..., Biochem. Biophys. Res. Commun. (2008), doi:10.1016/j.bbrc.2008.10.175

meric interface. While electrostatic interactions are predicted between the PDZ and MH2 domains, hydrophobic-hydrophilic interactions allow engagement between the MH2 domain and the peptide (1257–1276) of Erbin. Dai et al. identified a region called SID (Smad-Interaction Domain, residues 1172 to 1282) in Erbin that interacts with the Smad3 MH2 domain and partially overlaps with this peptide, which is in agreement with our model of interaction [16]. Furthermore, the energy profile of the model (Fig. 3) shows that E_{1321} in Erbin interacts with the Smad3 "arginine patch" formed by R_{279}, R_{287} and R_{288} similarly to the negatively charged D_{493} in Smad4 (Supplementary Fig. 2) [19]. A recent paper shows that mutation of R_{287} in Smad3 inhibits the TGFβ signaling [19]. Thus, we postulate that engagement of Erbin with the Smad3 MH2 domain may compete with the Smad3–Smad4 interaction, and impair the ability of Smad3 to heteromerize with Smad4 following TGFβ activation. This hypothesis is in agreement with recent results showing that over expression of Erbin inhibits the nuclear translocation and the transcriptional activity of Smad3 by reducing the dimerization of Smad3/Smad4 [16]. Interaction between the Erbin PDZ domain and the Smad3 MH2 domain leaves free the PDZ domain for canonical PDZ interactions.

Conclusions

Thus, combined experimental and theoretical results highlight a possible new type of interaction for a PDZ domain and its ligand provided by electrostatic interactions. On the Smad3 side, PDZ/MH2 interaction involves residues important for the heteromerization of Smads: this interaction can therefore compete with the Smad3/Smad4 interaction in provoking sterical constraints. Via this interaction, Erbin could act as a repressor in the TGFβ pathway. Erbin also negatively regulates the MAPK pathway downstream of tyrosine kinases including ErbB2 [13]. Erbin is thus at the crossroads of two signaling pathways of crucial importance in physiology and pathology.

Acknowledgments

We thank Brynn Taylor for her comments, Dr. T. Wang for the Smad3. This study was supported by La Ligue Contre le Cancer (Label Ligue 2007 to J.P.B.), Institut National du Cancer, Institut Paoli-Calmettes, ARC and CNRS/Région Lorraine.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2008.10.175.

References

[1] C. Nourry, S.G. Grant, J.P. Borg, PDZ domain proteins: plug and play!, Sci STKE 2003 (2003) RE7
[2] B.Z. Harris, W.A. Lim, Mechanism and role of PDZ domains in signaling complex assembly, J. Cell Sci. 114 (2001) 3219–3231.
[3] J.P. Borg, S. Marchetto, A.L. Bivic, V. Ollendorff, F. Jaulin-Bastard, H. Saito, E. Fournier, J. Adéla, B. Margolis, D. Birnbaum, ERBIN: a basolateral PDZ protein that interacts with the mammalian ERBB2/HER2 receptor, Nat. Cell Biol. 2 (2000) 407–414.
[4] P.J. Bryant, A. Huwe, LAP proteins: what's up with epithelia?, Nat Cell Biol. 2 (2000) E141–E143.
[5] B.A. Appleton, Y. Zhang, P. Wu, J.P. Yin, W. Hunziker, N.J. Skelton, S.S. Sidhu, C. Wiesmann, Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1, Insights into determinants of PDZ domain specificity, J. Biol. Chem. 281 (2006) 22312–22320.
[6] F. Jaulin-Bastard, J.P. Arsanto, A. Le Bivic, C. Navarro, F. Vely, H. Saito, S. Marchetto, M. Hatzfeld, M.J. Santoni, D. Birnbaum, J.P. Borg, Interaction between Erbin and a Catenin-related protein in epithelial cells, J. Biol. Chem. 277 (2002) 2869–2875.
[7] N.J. Skelton, M.F.T. Koehler, K. Zobel, W.L. Wong, S. Yeh, M.T. Pisabarro, J.P. Yin, L.A. Lasky, S.S. Sidhu, Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain, J. Biol. Chem. 278 (2003) 7645–7654.
[8] G. Birrane, J. Chung, J.A.A. Ladias, Novel mode of ligand recognition by the Erbin PDZ domain, J. Biol. Chem. 278 (2003) 1399–1402.
[9] P. ten Dijke, C.S. Hill, New insights into TGF-beta-Smad signalling, Trends Biochem. Sci. 29 (2004) 265–273.
[10] G. Wu, Y.G. Chen, B. Ozdamar, C.A. Gyuricza, P.A. Chong, J.L. Wrana, J. Massagué, Y. Shi, Structural basis of Smad2 recognition by the Smad anchor for receptor activation, Science 287 (2000) 92–97.
[11] R.S. Lo, Y.G. Chen, Y. Shi, N.P. Pavletich, J. Massagué, The L3 loop: a structural motif determining specific interactions between SMAD proteins and TGF-beta receptors, EMBO J. 17 (1998) 996–1005.
[12] U. Persson, H. Izumi, S. Souchelnytskyi, S. Itoh, S. Grimsby, U. Engström, C.H. Heldin, K. Funa, P.t. Dijke, The L45 loop in type I receptors for TGF-beta family members is a critical determinant in specifying Smad isoform activation, FEBS Lett. 434 (1998) 83–87.
[13] S.M. Garrard, C.T. Capaldo, L. Gao, M.K. Rosen, I.G. Macara, D.R. Tomchick, Structure of Cdc42 in a complex with the GTPase-binding domain of the cell polarity protein, Par6, EMBO J. 22 (2003) 1125–1133.
[14] B.Y. Qin, S.S. Lam, J.J. Correia, K. Lin, Smad3 allostery links TGF-beta receptor kinase activation to transcriptional control, Genes Dev. 16 (2002) 1950–1963.
[15] D.R. Warner, M.M. Pisano, E.A. Roberts, R.M. Greene, Identification of three novel Smad binding proteins involved in cell polarity, FEBS Lett. 539 (2003) 167–173.
[16] F. Dai, C. Chang, X. Lin, P. Dai, L. Mei, X.H. Feng, Erbin inhibits transforming growth factor beta signaling through a novel Smad-interacting domain, Mol. Cell. Biol. 27 (2007) 6183–6194.
[17] L. Lo Conte, C. Chothia, J. Janin, The atomic structure of protein–protein recognition sites, J. Mol. Biol. 285 (1999) 2177–2198.
[18] A. Ress, K. Moelling, Interaction partners of the PDZ domain of erbin, Protein Pept. Lett. 13 (2006) 877–881.
[19] B.M. Chacko, B.Y. Qin, A. Tiwari, G. Shi, S. Lam, L.J. Hayward, M. De Caestecker, K. Lin, Structural basis of heteromeric Smad protein assembly in TGF-beta signaling, Mol. Cell 15 (2004) 813–823.
[20] V. Prokova, S. Mavridou, P. Papakosta, K. Petratos, D. Kardassis, Novel mutations in Smad proteins that inhibit signaling by the transforming growth factor beta in mammalian cells, Biochemistry 46 (2007) 13775–13786.

Journal of Molecular Graphics and Modelling 27 (2008) 209–216

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

MetaMol: High-quality visualization of molecular skin surface

Matthieu Chavent [a,*], Bruno Levy [b], Bernard Maigret [a]

[a] Orpailleur Group, Nancy University, Loria, BP 239, 54506 Vandoeuvre les Nancy Cedex, France
[b] ALICE Group, Nancy University, Loria, France

ARTICLE INFO

Article history:
Received 11 February 2008
Received in revised form 17 April 2008
Accepted 22 April 2008
Available online 29 April 2008

Keywords:
Molecular skin surface
Mixed complex
Ray-casting
Smooth molecular surface
GPU

ABSTRACT

Modeling and visualizing molecular surfaces is an important and challenging task in bioinformatics. Such surfaces play an essential role in better understanding the chemical and physical properties of molecules. However, constructing and displaying molecular surfaces requires complex algorithms.

In this article we introduce MetaMol, a new program that generates high-quality images in interactive time. In contrast with existing software that discretizes the surface with triangles or grids, our program is based on a GPU accelerated ray-casting algorithm that directly uses the piecewise-defined algebraic equation of the molecular skin surface. As a result, both better performances and higher quality are obtained.

1. Introduction

Visualizing molecular surfaces is important since such surfaces play a central role in molecular interactions: depicting the surfaces is essential in understanding, for example, protein–protein or protein–ligand assemblies.

Several surface definitions exist (see Fig. 1):

- The Van der Waals (VdW) surface was firstly defined as the topological boundary of atom spheres. This surface gives a good approximation of molecular surface for small molecules but becomes rapidly complicated due to gaps and crevices. Therefore, it is not suitable for large molecular systems [1].
- To describe surfaces of macromolecules and their possible interactions with water molecules, Lee and Richards introduced the solvent accessible surface (SAS) [2]; this surface is obtained by rolling a probe sphere depicting a water molecule over the Van der Waals surface. The trajectory of the rolling sphere center traces out the SAS.
- A few years later, Richards defined a smooth molecular surface [3] constructed as the union of two parts: the contact surface and the re-entrant surface. The contact surface is the part of the Van der Waals surface that is accessible to the probe sphere. Re-entrant regions are constructed from the probe surface as it rolls between pairs of atom. This surface representation suffers from self-intersections. For this reason, Greer and Bush [4] decided to define the solvent excluded surface (SES). Connolly has given an analytical method to compute this surface [5] commonly referred to as the molecular surface (MS).
- More recently, in the more general context of computational geometry, Edelsbrunner proposed to define a smoother surface: the skin surface [6]. This surface can be used to represent a molecular surface and is named molecular skin surface (MSS). It is close to the MS representation but has extra-properties such as smoothness and decomposability. The MSS has therefore several advantages as compared to other molecular surface definitions [6]: (1) the surface does not self-intersect and (2) is everywhere tangent continuous. Moreover, MSS is composed of quadrics – whereas MS comprises torus slices – which simplify calculations.

Starting from the pioneering algorithm proposed by Connolly [7], numerous works have been devoted to the improvement of methods to calculate and represent molecular surface. The goals were still the same: providing fast and robust methods that allow generating high-quality pictures of MS in real time. In 1994, Varshney et al. developed a program that was easily parallelizable [8]. The year after, Sanner proposed a method based on reduced surfaces [9] to visualize large molecules (more than 10,000 atoms). More recently, Can et al. proposed an algorithm to generate molecular surface using level-set methods [10] and Bates et al.

* Corresponding author. Tel.: +33 3 54 95 85 92.
E-mail address: chavent@loria.fr (M. Chavent).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.04.007

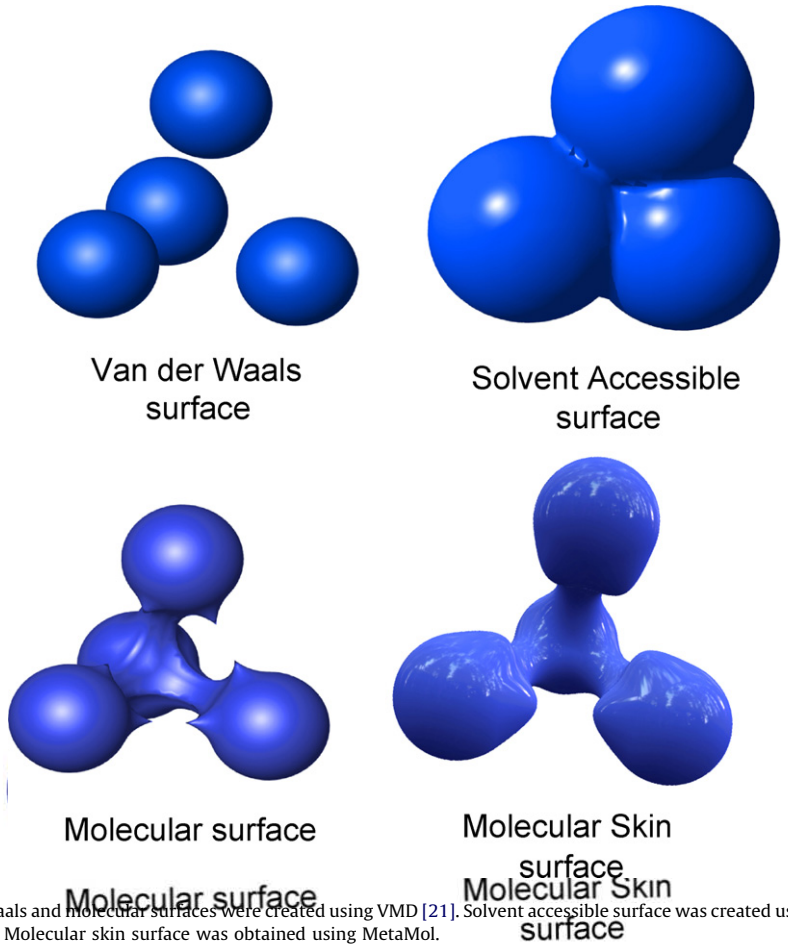


Fig. 1. Surface definitions. Van der Waals and molecular surfaces were created using VMD [21]. Solvent accessible surface was created using YASARA [38]. These three images were post-processed using POV-ray. Molecular skin surface was obtained using MetaMol.

defined a minimal molecular surface [11]. These latter two methods use a grid and marching front to display molecular surface and cavities. All these approaches are efficient but suffer from precision problems: in the Varshney and Sanner algorithms, the molecular surface is triangulated while for the Can and Bates algorithms, the surface is represented as the union of cubes so that a level of zoom is always found where triangles or cubes appear. Some works already triangulate skin surface [12–14]. This has two drawbacks: (1) they need to make sure that the surface topology is preserved. This makes the algorithm complicated (and slow). (2) At a certain level of zoom, the triangles generate display artifacts (see Fig. 9).

To overcome these limitations, we propose to use a ray-casting method performed on graphics processing unit (GPU). This has two advantages: (1) the ray-casting algorithm directly uses the equation of the MSS and does not need to resample it and (2) pixel-accurate images are generated in interactive time. GPU ray-casting has been already used to represent simple molecular models as "CPK" or "Balls and Sticks" [15,16]. In this paper, we deal with the much more complicated case of MSS. We present the MetaMol program devoted to display in interactive time MSS with the best visualizing quality.

2. Methods

In Ref. [6], Edelsbrunner defines the skin surface as the boundary of an infinite family of spheres defined from a set of weighted points (explained further below). The skin surface has the interesting property of being piecewise defined as a set of quadratic patches. It is this latter property that is beneficial in our case: using the ray-casting approach, we can display the surface from its equation, without triangulating it nor using a grid.

In this section, we will present briefly the molecular skin surface description and properties, and how we can exploit them with the ray-casting approach. The reader who wishes more background material about skin surface is referred to Ref. [6] for the original introduction of skin surface and to Refs. [17,12,13] for detailed explanations.

2.1. Weighted points

The input data of MSS are weighted points with atomic coordinates of a 3D molecular structure and a weight associated to each point. A weighted point is a pair $p = (z, w_p)$ of an atom position $z \in \mathbf{R}^3$ and a weight $w_p \in \mathbf{R}$. To make the molecular skin surface wrap around atom spheres (as explained in Ref. [12]), it is necessary to define the weight as

$$w_p = \left(\frac{1}{s}\right) \times r_{vw}^2 \tag{1}$$

where s is named the *shrink factor*, with $0 < s \leq 1$ and r_{vw} is the Van der Walls radius of the sphere centered on each atom. A weighted points list or PDB file [18] can be used as input for MetaMol: in this last case, atom radii are converted into weights using Eq. (1).

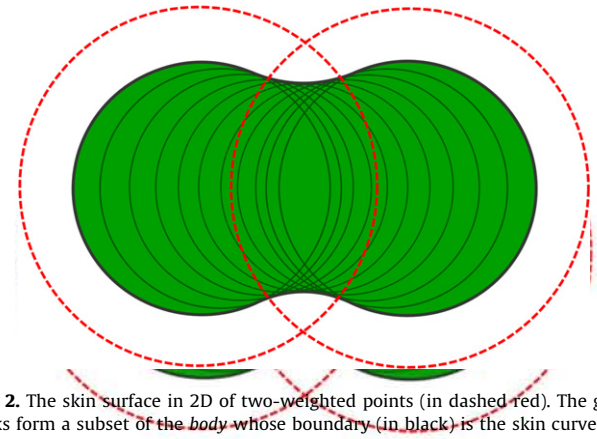


Fig. 2. The skin surface in 2D of two-weighted points (in dashed red). The green disks form a subset of the *body* whose boundary (in black) is the skin curve.

2.2. Molecular skin surface

2.2.1. Definition

Molecular skin surface is defined by:

- a set of weighted points (P):

$$P = \{ p_i = (z_i, w_i) \text{ in } \mathbf{R}^3 \times R | i = 1, \ \ldots \ , n\}$$

where p_i is a weighted point with z_i as the position of ith atom (among n atoms) and w_i as its weight.

- a shrink factor s, with $0 < s \leq 1$.

The skin surface $\text{skn}^s(P)$ is its boundary of $\text{bdy}^s(P)$ (see Fig. 2):$\text{skn}^s(P) = \partial \text{bdy}^s(P)$where $\text{bdy}^s(P)$ is an infinite family of shrunk spheres generated from the finite set of weighted points (P) and is named the *body* (see Section 4 of Ref. [6] for details). ∂ denotes the boundary of the union of the set of balls.

2.2.2. Equation of the MSS

Using this definition, Edelsbrunner proved that the MSS is a piecewise quadratic surface. This property is interesting for us since quadratic surfaces can be easily ray-casted. The intersection between a ray and a quadratic surface can be obtained in a closed form. The pieces of the MSS correspond to the cells of a geometric structure called the *mixed complex*.

The mixed complex, associated with a shrink factor s, may be thought of as an intermediate structure between the weighted Delaunay tetrahedralization and the weighted Voronoï diagram. The weighted Delaunay tetrahedralization, or regular tetrahedralization, is a collection of points that form tetrahedra. The weighted Voronoï Diagram is the dual of the weighted Delaunay tetrahedralization: basically, each node of Voronoï diagram is the center of the circumscribing sphere to each tetrahedron (see Fig. 3A). Each mixed cell, in the mixed complex, is obtained by taking the Minkowski sum of Delaunay and Voronoï cells (see Fig. 3B):

$$\mu_X^s = s \cdot \upsilon_X \oplus (1 - s) \cdot \delta_X \tag{2}$$

where X is a subset of a finite set of spheres and μ_X^s represents the mixed complex cell, υ_X the Voronoï cell and δ_X the Delaunay cell. The symbol $\cdot$ denotes the multiplication of a set by a scalar and $\oplus$denotes the Minkowski sum. Within the mixed cell μ_X^s, the MSS is completely determined by, at most, four weighted points in X. As s tends towards 0, the mixed cell tends towards the Delaunay cell. When s increases it deforms affinely into the Voronoï cell until $s = 1$.

The mixed complex is divided into four different parts (see Fig. 4). First, shrunk Voronoï cells stem from Delaunay vertices. Then, shrunk tetrahedra stem from Delaunay cells. These two cells clip pieces of the sphere. Between these two objects "species", created from the Voronoï Diagram and the Delaunay tetrahedralization, there appear two other polyhedra: prisms with shrunk Voronoï facets at their base (that we called H1 patches) and prisms with shrunk Delaunay triangles at their base (that we called H2 patches). For H1 (respectively H2) patch, the cell clips a one-sheeted or a two-sheeted hyperboloid with its symmetry axis aligned along the Delaunay (respectively Voronoï) edge.

2.3. Calculation pipeline

In contrast with programs that create molecular surfaces by using the central processing unit (CPU) alone for the calculations, our approach splits the calculations:

- The mixed complex envelope is calculated on the CPU. Note that the mixed complex computation is view-independent and is therefore computed once only at the loading time.
- The ray-casting to display the MSS is performed on GPU.

The use of both CPU and GPU reduces computation times and improves the quality of the surface (see Fig. 5B for a global view).

2.3.1. CPU calculations

Firstly, we construct the Delaunay tetrahedralization using atoms centers as vertices. Then, we calculate the Voronoï diagram from the Delaunay cells. From these two structures, we create mixed cells as a function of s using Eq. (2). Finally, we calculate the quadratic equation for each mixed complex cell using the computational geometry library CGAL [19].

2.3.2. GPU calculations

Then, we display the mixed complex envelope using the OpenGL API (Fig. 5A, left picture). We obtain the surface by using the ray-casting method (Fig. 5A, center and right pictures). In classical programs that perform ray-tracing, such as POV-ray, each ray is traced from each pixel of the screen to the object and, when a ray intersects the object surface, a reflecting ray is calculated to define the lighting. In our case, we perform ray-casting, which is simpler and ignores secondary reflections: for each mixed complex cell, rays are launched from the screen, as in the ray-tracing method, but when they touch the mixed cell, the intersection between the implicit surface and ray is calculated. If there is an intersection, the color of the pixel is computed by a simple shading model, and the Z-buffer coordinates are updated at the actual ray–surface intersection. If there is no intersection, the pixel is discarded [20]. This part is performed on the GPU, using the OpenGL shading language (GLSL). More precisely, the graphic card must support "Shader Model 3.0". In practice, it is necessary to have a Geforce 6 series (or equivalently an ATI Radeon $\times$1300) or greater.

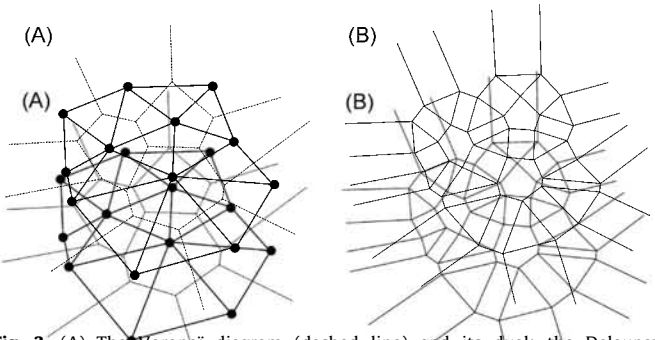


Fig. 3. (A) The Voronoï diagram (dashed line) and its dual: the Delaunay triangulation (solid line) in 2D and (B) the mixed complex (for $s = 0.5$) based on Delaunay Triangulation and Voronoï Diagram: here, the Voronoï and Delaunay cells are shrunk and we can notice the appearance of new patches between cells.

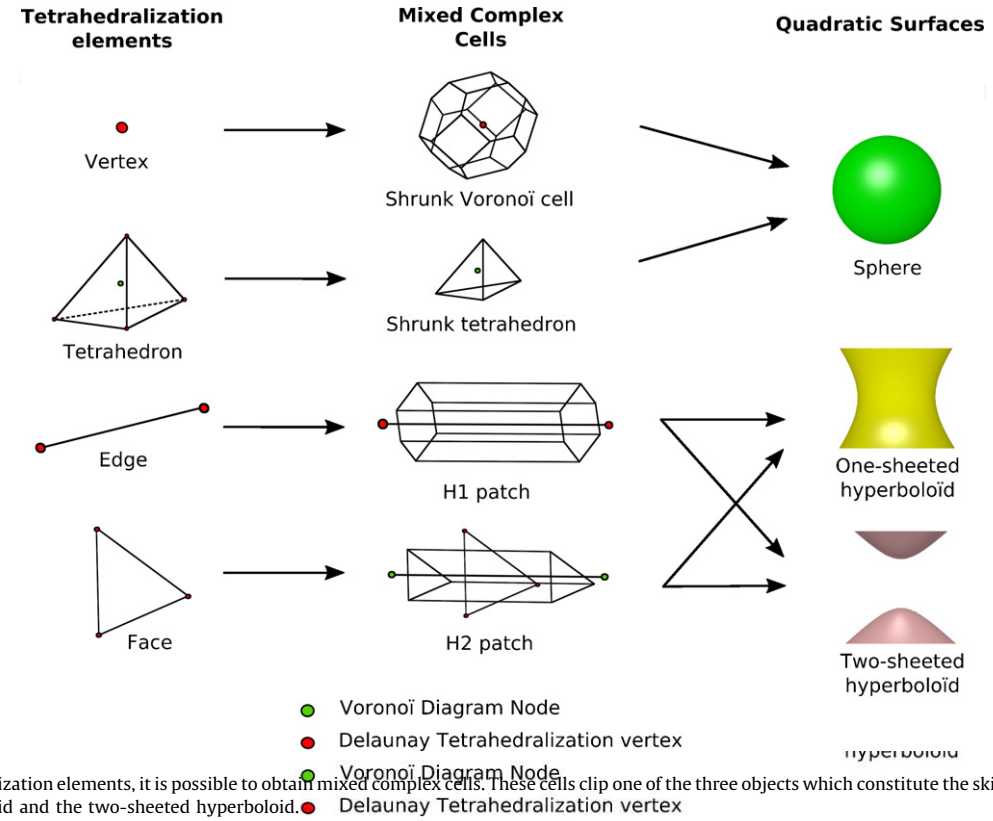


Fig. 4. From tetrahedralization elements, it is possible to obtain mixed complex cells. These cells clip one of the three objects which constitute the skin surface: the sphere, the one-sheeted hyperboloid and the two-sheeted hyperboloid.

3. Results and discussion

The main benefit of our method is the quality and the smoothness of the surface at each level of zoom thanks to the ray-casting method. The generated images are pixel accurate. Note that thanks to the high efficiency and parallel power of the GPU, the viewpoint and the shrink factor can be changed interactively.

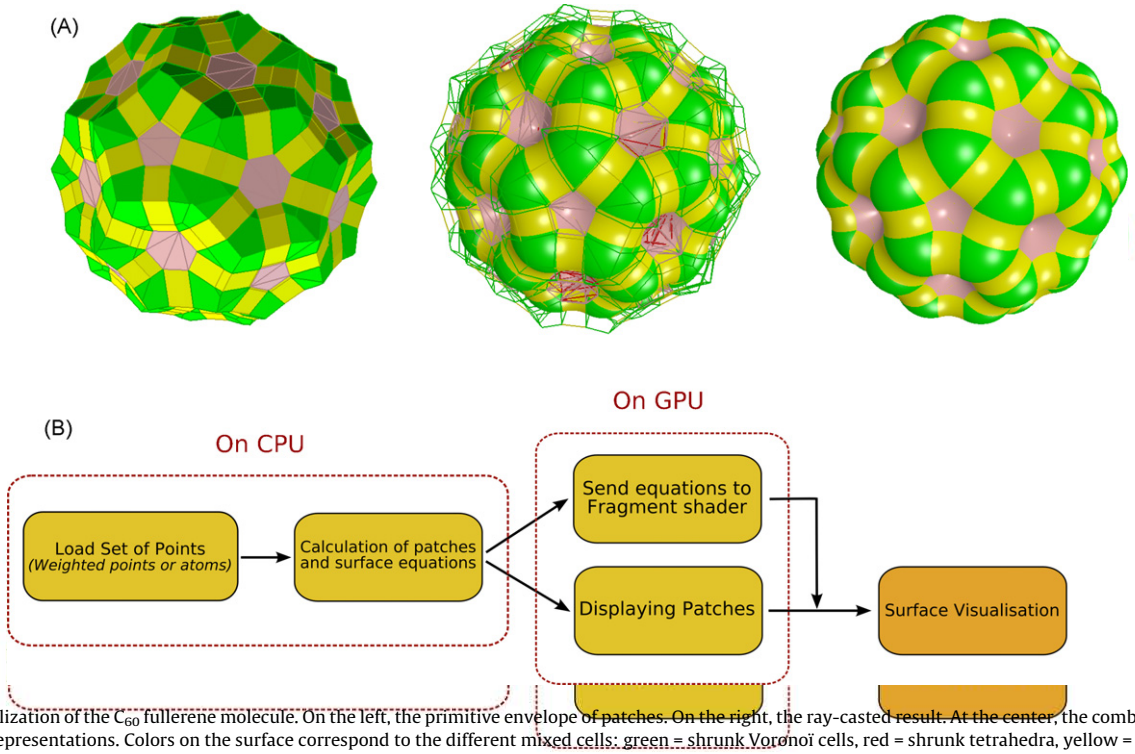


Fig. 5. (A) Visualization of the C_{60} fullerene molecule. On the left, the primitive envelope of patches. On the right, the ray-casted result. At the center, the combined ray-casted and primitive representations. Colors on the surface correspond to the different mixed cells: green = shrunk Voronoï cells, red = shrunk tetrahedra, yellow = H1 patches and pink = H2 patches. (B) Pipeline of our algorithm.

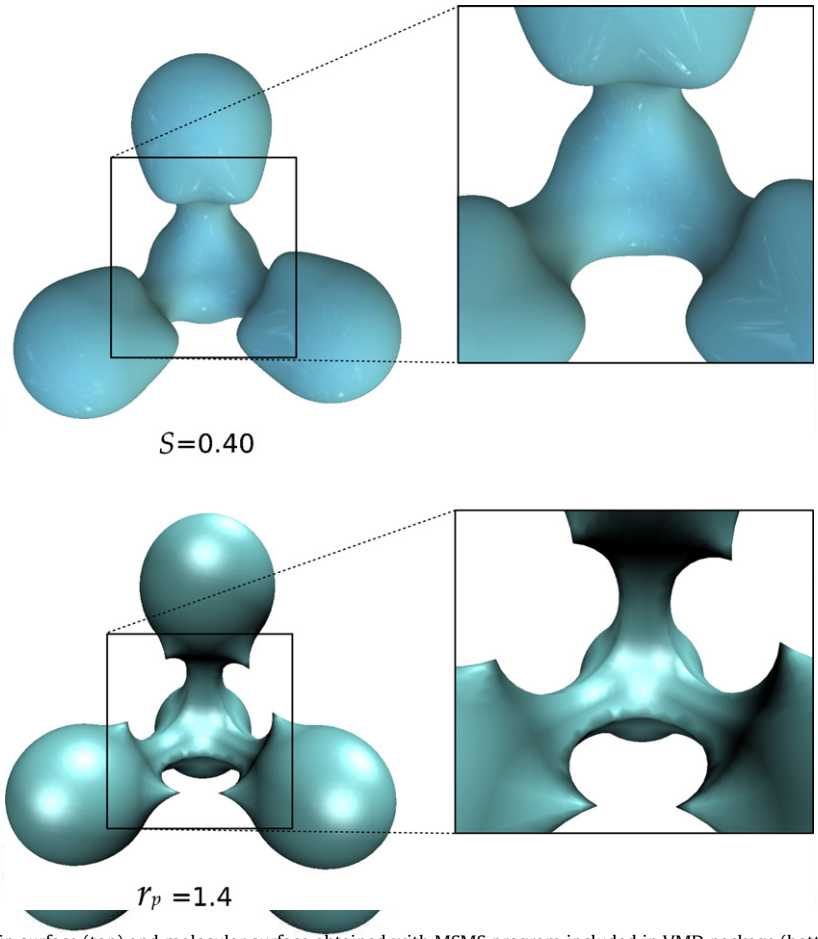


Fig. 6. Comparison of molecular skin surface (top) and molecular surface obtained with MSMS program included in VMD package (bottom) with the maximum density of points. s represents the shrink factor and r_p represents the radius of the rolling probe sphere.

3.1. Molecular skin surface versus molecular surface

To highlight our choice of surface, we compared the smoothness of the molecular surface (obtained with MSMS program included in VMD [21] package) with the molecular skin surface. As it is presented in Fig. 6, the cusps observed on the MS surface are replaced, on the skin surface, by two-sheeted hyperboloids so that the surface is clearly continuous. Beyond the unquestionable display quality, the skin surface smoothness is suitable to compute and display physical and chemical properties of molecules. In contrast with the MS, where singularities are encountered [22,23,9,11], the molecular skin surface does not self intersect, there are no cusps and the surface is C^1 (continuous and its first order derivatives are also continuous).

Another advantage of the MSS is the ability to easily generate a smoothed solvent accessible surface (SAS) by adding the probe radius to the Van der Waals radius when defining the atom weight.

3.2. Deforming molecular skin surface

The molecular skin surface is also well adapted to support molecular deformations as this surface can be freely deformed with smooth transitions [24,6]. Contrary to programs which compute the molecular skin surface as a mesh [25,13], using GPU ray-casting allows deformations to be displayed in real time.

3.2.1. In changing the shrink factor

As shown in Fig. 7, it is possible to interactively change the shrink factor, going from a Van der Waals surface (s = 1.0) where all atoms are represented by spheres, via the molecular skin surface (s = 0.5) which is close to the MS, ending in a simplified surface if we continue to decrease the shrink factor. When the shrink factor is increased after 0.5, minor details (as re-entrant regions) tend to be removed from the scene. This is interesting to have a global perception of the structure and to focus on the global shape in order to compare, for example, several protein figures [26]. Furthermore, this "simplified representation" can be useful for docking in an high-throughput approach where low-resolution models are used to perform fast calculations [27,28] and can serve as starting points for further structural refinements [29]. When going from one representation to another, it is noticeable that the surface changes continuously.

3.2.2. During molecular movement

Efficient visualization of molecular movements is an important tool to understand the structural behavior of biological samples. Several efforts have been made this last decade to reach this goal: Eyal and Halperin dynamically maintained the Van der Waals and solvent accessible surfaces [30]. Hao et al. developed a linear scalable program to visualize large time-varying molecules using occlusion culling [31]. Recently, Lampe et al. defined a two-level approach to visualize protein dynamics [32]. Unfortunately, these works used quite simple representation as "ball and stick" or union of spheres and few works are dedicated to represent deformations on more complicated surface such as MS [33,34]. One explanation is that maintaining a mesh during molecular movements is very costly in computation time. With ray-casting, there is no mesh and the precomputation time is reduced so that it is possible to

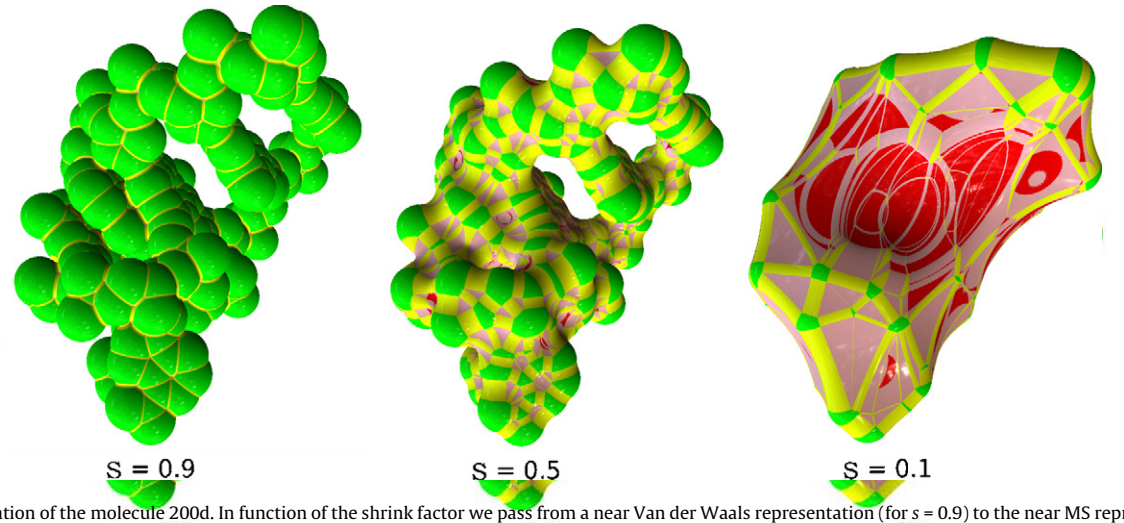


Fig. 7. Representation of the molecule 200d. In function of the shrink factor we pass from a near Van der Waals representation (for $s = 0.9$) to the near MS representation (for $s = 0.5$) to a simplified form (for $s = 0.1$). Colors on the surface correspond to the different mixed cells: green = shrunk Voronoi cells, red = shrunk tetrahedra, yellow = H1 patches and pink = H2 patches.

Table 1
Performance of our approach compared with Kruithof's approach

PDB code	No. of atoms	Kruithof's approach			MetaMol		
		Nb of triangles	Computing time (s)	FPS (1024 × 1024)	Nb of triangles	Computing time (s)	FPS (1024 × 1024)
7tmn	33	23,424	1.1	800	7,116	0.02	200
1grm (*Gramicidin A*)	272	310,488	16.1	130	73,416	1.7	50
1g6x	509	481,856	28.7	95	146,476	3.6	25
1cbs	1091	1,664,184	93.1	30	325,076	8.2	12
1j4n	1852	2,165,268	137.4	25	558,372	15.4	7

We use an Intel CPU at 2.40 GHz with a Nvidia GeForce 8800 GTX graphic card. Note that heteroatoms in PDB files were removed before measurement.

visualize real-time deformations on a surface. For the moment, it is possible to visualize movements of small proteins in concatenated PDB file formats. But, in the near future, MetaMol will be able to read molecular dynamics trajectory files and to display the associated moving surface in real time.

3.3. Discussion

3.3.1. Precision

Using ray-casting allows us to visualize the surface of large objects (see Fig. 8) with high rendering quality. Any zoom on this representation will provide a surface that is still very precise (pixel accurate), without any loss of information (see Figs. 8 and 9). To our knowledge, MetaMol is the first program that interactively visualizes molecular surface with such a quality.

3.3.2. Performances

To have a reference, we first measured the computing time of a program that computes the same type of skin surface but by tessellating it [25,12]. We then compared its performance with ours (see Table 1 and Fig. 9). We have to separate the computing time, to create the patches and generate the surfaces and the

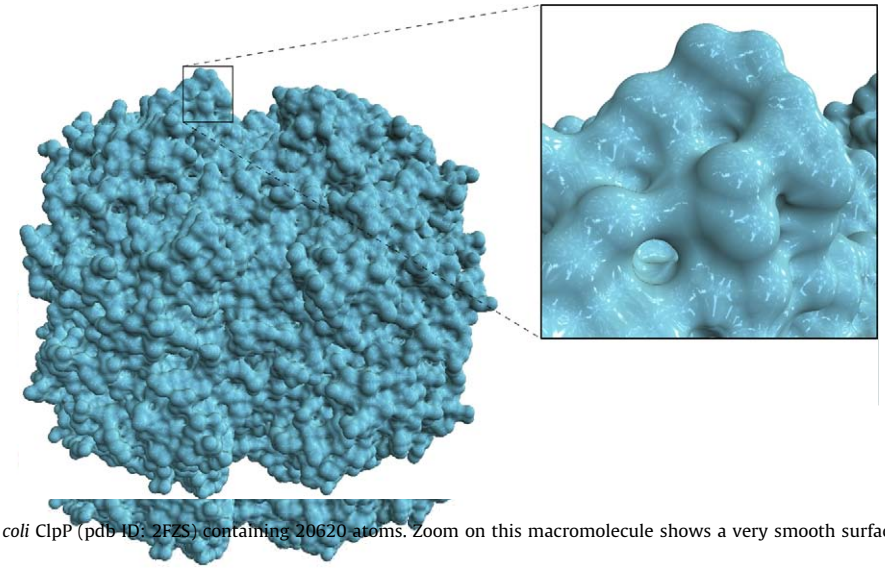


Fig. 8. Visualization of the *E. coli* ClpP (pdb ID: 2FZS) containing 20620 atoms. Zoom on this macromolecule shows a very smooth surface without loss of information.

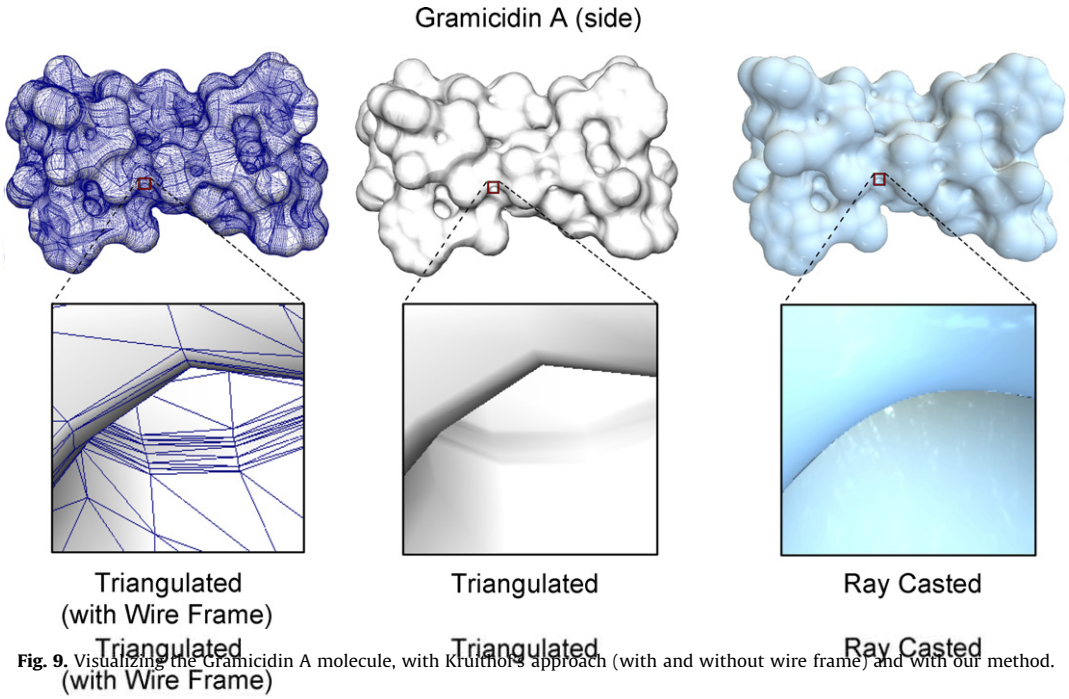


Fig. 9. Visualizing the Gramicidin A molecule, with Kruithof's approach (with and without wire frame) and with our method.

display time (measured in frames per second, fps). The computing times that we obtained are clearly better for a superior display quality than Kruithof's algorithm. The main reason is that the only triangles we need to generate are the faces of the mixed complex whereas Kruithof needs a very fine triangulation that approximates the MSS. In addition, Kruithof's method needs to ensure that the generated triangulation preserves the topology of the MSS, which is time consuming [25]. The use of ray-casting overcomes this limitation by calculating the position on the surface for each pixel of the screen. The efficiency of the GPU allows updating the surface "on-the-fly" when the viewpoint and/or the shrink factor are changed. It is used to visualize the surface deformations (as presented in Section 3.2). However, after the pre-processing step, *i.e.* once the surface is computed, the performance of Kruithof's algorithm is better than ours: this is due to multiple read–write accesses to the *Z*-buffer [35]. Note that this can be improved by occlusion culling techniques, which we will implement in future work. Nevertheless, to have a surface quality equivalent to ours with Kruithof's algorithm, it would be necessary to create huge number of very small triangles, so that computing and display times would increase dramatically. However, Kruithof's method provides other benefits: it is necessary to perform calculations only once for a structure (mesh is stored in an .off file) in comparison to our algorithm in which the surface is ray-casted at each frame. Unfortunately, high-resolution .off files can be memory consuming: the .off file size of the 2FZS molecule (presented in Fig. 8) is 439 MB (~12 millions of triangles). So, decreasing the level of detail could be required to visualize and manipulate large structures interactively. With Kruithof's algorithm, the surface is meshed so that it is possible to have information about the neighbors of any point on the surface. Therefore, the surface can rapidly be crossed in a scanning perspective: for example, to

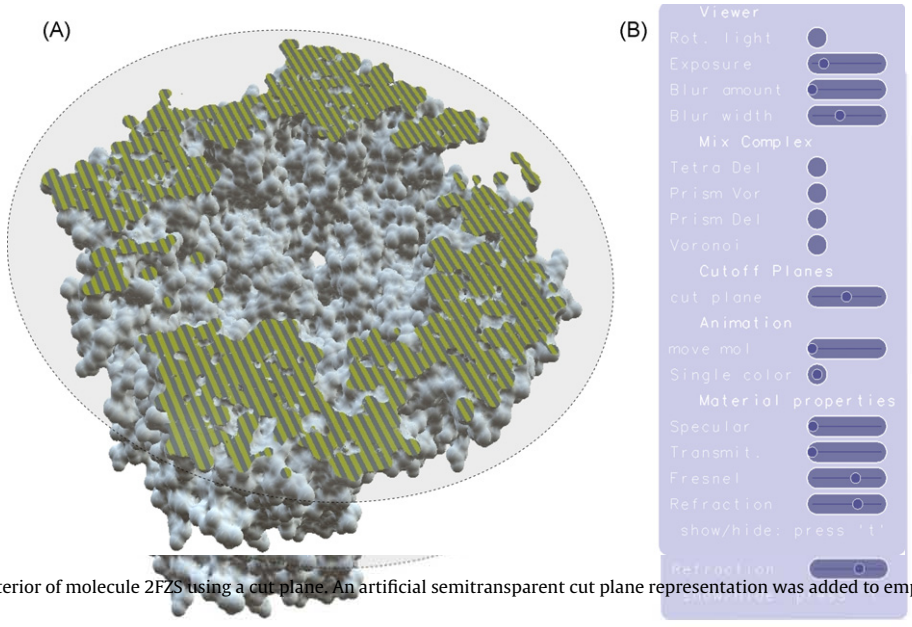


Fig. 10. (A) Visualizing the interior of molecule 2FZS using a cut plane. An artificial semitransparent cut plane representation was added to emphasize the image. (B) Viewer panel.

perform physical calculations or to get information about solvent accessible areas. In its current status our algorithm is dedicated to visualize molecular surface and is not intended to perform calculation on the surface so that our approach and Kruithof's approach are quite complementary.

3.4. User interface

The user interface is depicted in Fig. 10B. The user can easily change the visualization parameters: in particular, lighting functions (exposure, light rotation or specular effect) or to scroll the clipping plane to visualize the internal structure of the molecule (see Fig. 10A).

4. Conclusion

In this article, we introduce a method to accurately and interactively visualize the MSS. With our program, several types of molecular surfaces can be visualized (in changing the shrink factor): from Van der Waals surface to coarse-grained surface. It is clear that, for the moment, our algorithm is less efficient for displaying huge assemblies than algorithms cited above (e.g. MSMS or Varshney's program). To deal with this issue, we are developing a multi-resolution program to visualize surfaces ranging from atomic description to cryo-electron microscopy level with the same precision. We will optimize the algorithm to visualize interactively larger structures with real-time surface deformations. We plan to add new effects such as ambient occlusion (already used in Tachyon [36] and Qutemol [37]) in order to improve the perception of the global shape of large molecules.

Acknowledgments

M. Chavent is supported by CNRS/Région Lorraine. B. Levy is supported by Microsoft Research (“Geometric Intelligence” Grant). Authors would like to thank L. Provot and R. Toledo for their technical supports and B. Vallet for his fruitful discussions.

References

[1] M. Chapman, M.L. Connolly, Molecular surfaces: calculations, uses and representations, Int. Tables Crystallogr. (2001) 539–545.
[2] B. Lee, F.M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–400.
[3] F.M. Richards, Areas, volumes, packing and protein structure, Annu. Rev. Biophys. Bioeng. 6 (1977) 151–176.
[4] J. Greer, B.L. Bush, Macromolecular shape and surface maps by solvent exclusion, Proc. Natl. Acad. Sci. U.S.A. 75 (1978) 303–307.
[5] M.L. Connolly, Analytical molecular surface calculation, J. Appl. Crystallogr. 16 (1983) 548–558.
[6] H. Edelsbrunner, Deformable smooth surface design, Discrete Comput. Geom. 21 (1999) 87–115.
[7] M.L. Connolly, Molecular surface triangulation, J. Appl. Crystallogr. 18 (1985) 499–505.
[8] A. Varshney, F.P.J. Brooks, W.V. Wright, Linearly scalable computation of smooth molecular, surfaces, IEEE Comput. Graph. Appl., 14 (1994) 19–25.
[9] M.F. Sanner, A.J. Olson, J.C. Spehner, Reduced surface: an efficient way to compute molecular surfaces, Biopolymers 38 (1996) 305–320.
[10] T. Can, C.-I. Chen, Y.-F. Wang, Efficient molecular surface generation using level-set methods, J. Mol. Graph. Model. 25 (2006) 442–454.
[11] P.W. Bates, G.W. Wei, S. Zhao, Minimal molecular surfaces and their applications, J. Comput. Chem. (2007).
[12] N. Kruithof, G. Vegter, Approximation by skin surfaces, Comput-Aid. Des. 36 (2004) 1075–1088.
[13] H.-L. Cheng, X. Shi, Guaranteed quality triangulation of molecular skin surfaces, in: Proceedings of the Conference on Visualization '04, IEEE Computer Society, (2004), pp. 481–488.
[14] H.-L. Cheng, X. Shi, Quality mesh generation for molecular skin surfaces using restricted union of balls, IEEE Visualization, 2005, 399–405.
[15] R. Toledo, B. Levy, Extending the graphic pipeline with new GPU-accelerated primitives, Tech report, 2004.
[16] C. Sigg, T. Weyrich, M. Botsch, M. Gross, GPU-based ray-casting of quadratic surfaces, in: Proceedings of the Symposium on Point-Based Graphics, 2006, pp. 56–65.
[17] N.G.H. Kruithof, Envelope Surfaces, Surface Design and Meshing, University of Groningen, 2006 125 pp..
[18] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242. http://www.rcsb.org/pdb/home/home.do.
[19] Cgal, Computational Geometry Algorithms Library, 1999, http://www.cgal.org.
[20] C. Loop, J. Blinn, Real-time GPU rendering of piecewise algebraic surfaces, ACM Trans. Graph. 25 (2006) 664–670.
[21] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, J. Mol. Graph. 14 (1996), 33–38, 27–28..
[22] Y.N. Vorobjev, J. Hermans, SIMS: computation of a smooth invariant molecular surface, Biophys. J. 73 (1997) 722–732.
[23] W. Geng, S. Yu, G. Wei, Treatment of charge singularities in implicit solvent models, J. Chem. Phys. 127 (2007), 114106 pp..
[24] H. Edelsbrunner, Smooth surfaces for multi-scale shape representation, in: Proceedings of the 15th Conference on Foundations of Software Technology and Theoretical Computer Science, Springer–Verlag, 1995, pp. 391–412.
[25] N. Kruithof, G. Vegter, Meshing skin surfaces with certified topology, Comput. Geom. Theor. Appl. 36 (2007) 166–182.
[26] G. Cipriano, M. Gleicher, Molecular surface abstraction, IEEE Trans. Vis Comput. Graph. 13 (2007) 1608–1615.
[27] A. Tovchigrechko, C.A. Wells, I.A. Vakser, Docking of protein models, Protein Sci. 11 (2002) 1888–1896.
[28] D.W. Ritchie, G.J.L. Kemp, Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces, J. Comput. Chem. 20 (1999) 383–395.
[29] J.J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, D. Baker, Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations, J. Mol. Biol. 331 (2003) 281–299.
[30] E. Eyal, D. Halperin, Dynamic maintenance of molecular surfaces under conformational changes, in: Proceedings of the twenty-first annual symposium on Computational geometry, ACM, Pisa, Italy, (2005), pp. 45–54.
[31] X. Hao, A. Varshney, S. Sukharev, Real-time visualization of large time-varying molecules, in: Proceedings of the High-Performance Computing Symposium '04, (2004) 109–114.
[32] O.D. Lampe, I. Viola, N. Reuterm, H. Hauser, Two-level approach to efficient visualization of protein dynamics, in: Proceedings of the IEEE Transactions on Visualization and Computer Graphics, vol. 13, November–December, (2007), pp. 1616–1623.
[33] M.F. Sanner, A.J. Olson, Real time surface reconstruction for moving molecular fragments, Pac. Symp. Biocomput. (1997) 385–396.
[34] C.L. Bajaj, V. Pascucci, A. Shamir, R.J. Holt, A.N. Netravali, Dynamic maintenance and visualization of molecular surfaces, Discrete Appl. Math. 127 (2003) 23–51.
[35] J. Foley, A. Van Dam, S. Feiner, J. Hughes, Computer Graphics: Principles and Practice, Second edition in C, Addison-Wesley Professional, 1995.
[36] J. Stone, An efficient Library for Parallele Ray Tracing and Animation, Computer Science Department, University of Missouri-Rolla, 1998.
[37] M. Tarini, P. Cignoni, C. Montani, Ambient occlusion and edge cueing for enhancing real time molecular visualization, IEEE Trans. Vis. Comput. Graph. 12 (2006) 1237–1244.
[38] E. Krieger, YASARA, 2003, www.yasara.org.

Bibliographie

(2003). Pov-ray, persistence of vision ray-tracer, http ://www.povray.org.

Abagyan, R. et Totrov, M. (1994). Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol*, 235(3):983–1002.

Abagyan, R., Totrov, M. et Kuznetsov, D. (2004). Icm : A new method for protein modeling and design : Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, 15(5):488–506.

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.-C., Bork, P., Superti-Furga, G., Serrano, L. et Russell, R. B. (2004). Structure-based assembly of protein complexes in yeast. *Science*, 303(5666):2026–2029.

Aloy, P., Pichaud, M. et Russell, R. B. (2005). Protein complexes : structure prediction challenges for the 21st century. *Curr. Op. Struct. Biol.*, 15:15–22.

Aloy, P., Querol, E., Aviles, F. X. et Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins : applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol*, 311(2):395–408.

Aloy, P. et Russell, R. B. (2002). The third dimension for protein interactions and complexes. *Trends Biochem Sci*, 27(12):633–638.

Aloy, P. et Russell, R. B. (2003). Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci.*, 99:5896–5901.

Aloy, P. et Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, 22:1317–1321.

Aloy, P. et Russell, R. B. (2005). Structure-based systems biology : a zoom lens for the cell. *FEBS Lett*, 579(8):1854–1858.

André, I., Bradley, P., Wang, C. et Baker, D. (2007). Prediction of the structure of symmetrical protein assemblies. *Proc Natl Acad Sci U S A*, 104(45):17656–17661.

Andrusier, N., Mashiach, E., Nussinov, R. et Wolfson, H. J. (2008). Principles of flexible protein-protein docking. *Proteins*, 73(2):271–289.

Andrusier, N., Nussinov, R. et Wolfson, H. J. (2007). FireDock : fast interaction refinement in molecular docking. *Proteins*, 69(1):139–159.

Appleton, B. A., Zhang, Y., Wu, P., Yin, J. P., Hunziker, W., Skelton, N. J., Sidhu, S. S. et Wiesmann, C. (2006). Comparative structural analysis of the erbin pdz domain and the first pdz domain of zo-1. insights into determinants of pdz domain specificity. *J Biol Chem*, 281(31):22312–22320.

Bahadur, R. P., Chakrabarti, P., Rodier, F. et Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins : Struct. Func. Genet.*, 53:708–917.

Bahadur, R. P., Chakrabarti, P., Rodier, F. et Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4):943–955.

Bahadur, R. P., Zacharias, M. et Janin, J. (2008). Dissecting protein-RNA recognition sites. *Nucleic Acids Res*, 36(8):2705–2716.

Bajaj, C. L., Pascucci, V., Shamir, A., Holt, R. J. et Netravali, A. N. (2003). Dynamic maintenance and visualization of molecular surfaces. *Discrete Appl. Math.*, 127(1):23–51.

Baker, C. M. et Grant, G. H. (2007). Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers*, 85(5-6):456–470.

Ban, Y.-E. A., Edelsbrunner, H. et Rudolph, J. (2004). Interface surfaces for protein-protein complexes. *In RECOMB '04 : Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 205–212, New York, NY, USA. ACM.

Barbey-Martin, C., Gigant, B., Bizebard, T., Calder, L. J., Wharton, S. A., Skehel, J. J. et Knossow, M. (2002). An antibody that prevents the hemagglutinin low pH fusogenic transition. *Virology*, 294(1):70–74.

Basdevant, N., Weinstein, H. et Ceruso, M. (2006). Thermodynamic basis for promiscuity and selectivity in protein-protein interactions : Pdz domains, a case study. *J Am Chem Soc*, 128(39):12766–12777. 0002-7863 (Print)Journal article.

Bastard, K., Prévost, C. et Zacharias, M. (2006). Accounting for loop flexibility during protein-protein docking. *Proteins : Struct. Func. Bioinf.*, 62:956–969.

Bates, P. W., Wei, G. W. et Zhao, S. (2007). Minimal molecular surfaces and their applications. *J Comput Chem.* 0192-8651 (Print)Journal article.

Baumeister, W. (2005). From proteomic inventory to architecture. *FEBS Lett*, 579(4):933–937.

Beeman, D. (1976). Some multistep methods for use in molecular dynamics calculations. *Journal of Computational Physics*, 20:130–139.

Berchanski, A. et Eisenstein, M. (2003). Construction of molecular assemblies via docking : modeling of tetramers with D_2 symmetry. *Proteins : Struct. Func. Genet.*, 53:817–829.

Berchanski, A., Segal, D. et Eisenstein, M. (2005). Modeling oligomers with Cn or Dn symmetry : application to CAPRI target 10. *Proteins*, 60(2):202–206.

Berchanski, A., Shapira, B. et Eisenstein, M. (2004). Hydrophobic complementarity in protein-protein docking. *Proteins : Struct. Func. Bioinf.*, 56:130–142.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. et Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 6 No 1):899–907.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. et Schneider, B. (1992). The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J*, 63(3):751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. et Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–242.

Betts, M. J. et Sternberg, M. J. (1999). An analysis of conformational changes on protein-protein association : implications for predictive docking. *Protein Eng*, 12(4):271–283.

Bezprozvanny, I. et Maximov, A. (2001). Classification of pdz domains. *FEBS Lett*, 509(3): 457–62. 0014-5793 (Print)Journal ArticleResearch Support, Non-U.S. Gov'tResearch Support, U.S. Gov't, P.H.S.

Birrane, G., Chung, J. et Ladias, J. A. A. (2003). Novel mode of ligand recognition by the erbin pdz domain. *J Biol Chem*, 278(3):1399–1402.

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

Bon, M., Vernizzi, G., Orland, H. et Zee, A. (2008). Topological classification of RNA structures. *J Mol Biol*, 379(4):900–911.

Bonsor, D. A., Grishkovskaya, I., Dodson, E. J. et Kleanthous, C. (2007). Molecular mimicry enables competitive recruitment by a natively disordered protein. *J Am Chem Soc*, 129(15):4800–4807.

Bonvin, A. M. (2006). Flexible protein-protein docking. *Curr Opin Struct Biol*, 16(2):194–200. 0959-440X (Print)Journal ArticleReview.

Bordner, A. J. et Abagyan, R. (2005). Statistical analysis and prediction of protein-protein interfaces. *Proteins : Struct. Func. Bioinf.*, 60:353–366.

Borg, J. P., Marchetto, S., Bivic, A. L., Ollendorff, V., Jaulin-Bastard, F., Saito, H., Fournier, E., Adélaïde, J., Margolis, B. et Birnbaum, D. (2000). Erbin : a basolateral pdz protein that interacts with the mammalian erbb2/her2 receptor. *Nat Cell Biol*, 2(7):407–414.

Boudon, S., Wipff, G. et Maigret, B. (1990). Monte carlo simulations on the like-charged guanidinium-guanidinium ion pair in water. *Journal of Physical Chemistry*, 94(15):6056–6061.

Bourne, P. et Weissig, H., éditeurs (2003). *Structural Bioninformatics*. Wiley-Liss.

Bradford, J. R. et Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machine. *Bioinformatics*, 21:1487–1494.

Bressanelli, S., Stiasny, K., Allison, S. L., Stura, E. A., Duquerroy, S., Lescar, J., Heinz, F. X. et Rey, F. A. (2004). Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO J*, 23(4):728–738.

Brooks, B. et Karplus, M. (1983). Harmonic dynamics of proteins : normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A*, 80(21):6571–6575.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. et Karplus, M. (1983). Charmm : A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*, 4:187–217.

Camacho, C. J. (2005). Modeling side-chains using molecular dynamics improves recognition of binding region in CAPRI targets. *Proteins : Struct. Func. Bioinf.*, 60:245–251.

Camacho, C. J. et Gatchell, D. W. (2003). Successful discrimination of protein interactions. *Proteins : Struct. Func. Bioinf.*, 52:92–97.

Camacho, C. J., Kimura, S. R., DeLisi, C. et Vajda, S. (2000). Kinetics of desolvation-mediated protein-protein binding. *Biophys J*, 78(3):1094–1105.

Camacho, C. J., Ma, H. et Champ, C. (2006). Scoring a diverse set of high quality docked conformations : A metascore based on electrostatic and desolvation interactions. *Proteins : Struct. Func. Bioinf.*, 63:868–877.

Camacho, C. J. et Vajda, S. (2001). Protein docking along smooth association pathways. *Proc. Natl. Acad. Sci.*, 98:10636–10641.

Camacho, C. J., Weng, Z., Vajda, S. et DeLisi, C. (1999). Free energy landscapes of encounter complexes in protein-protein association. *Biophys J*, 76(3):1166–1178.

Can, T., Chen, C.-I. et Wang, Y.-F. (2006). Efficient molecular surface generation using level-set methods. *J Mol Graph Model*, 25(4):442–454.

Carvalho, A. L., Dias, F. M. V., Prates, J. A. M., Nagy, T., Gilbert, H. J., Davies, G. J., Ferreira, L. M. A., Romão, M. J. et Fontes, C. M. G. A. (2003). Cellulosome assembly

revealed by the crystal structure of the cohesin-dockerin complex. *Proc Natl Acad Sci U S A*, 100(24):13809–13814.

Castanié, L. (2006). *Visualisation de Donnéees Volumiques Massives, Applications aux Données sismiques*. Thèse de doctorat, Institut National Polytechnique de Lorraine.

Cazals, F., Proust, F., Bahadur, R. P. et Janin, J. (2006). Revisiting the Voronoi description of protein-protein interfaces. *Protein Sci*, 15(9):2082–2092.

Chacko, B. M., Qin, B. Y., Tiwari, A., Shi, G., Lam, S., Hayward, L. J., De Caestecker, M. et Lin, K. (2004). Structural basis of heteromeric smad protein assembly in tgf-beta signaling. *Mol Cell*, 15(5):813–23. 1097-2765 (Print)Journal Article.

Chakrabarti, P. et Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins : Struct. Func. Genet.*, 47:334–343.

Chapman et Hall-CRC, éditeurs (2006). *Normal Mode Analysis : theory and applications to biological and chemical systems*. Q. Cui and I. Bahar.

Chapman, M. et Connolly, M. L. (2001). Molecular surfaces : calculations, uses and representations. *In International Tables for Crystallography*, volume F, pages 539–545.

Chaudhury, S., Sircar, A., Sivasubramanian, A., Berrondo, M. et Gray, J. J. (2007). Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6-12. *Proteins*, 69(4):793–800.

Chen, R., Li, L. et Weng, Z. (2003a). ZDOCK : an initial-stage protein-docking algorithm. *Proteins : Struct. Func. Genet.*, 52:80–87.

Chen, R., Mintseris, J., Janin, J. et Weng, Z. (2003b). A protein-protein docking benchmark. *Proteins : Struct. Func. Genet.*, 52:88–91.

Chen, R. et Weng, Z. (2003). A novel shape complementarity scoring function for protein-protein docking. *Proteins : Struct. Func. Genet.*, 51:397–408.

Chen, X., Rubock, M. J. et Whitman, M. (1996). A transcriptional partner for MAD proteins in TGF-beta signalling. *Nature*, 383(6602):691–696.

Chen, X., Weisberg, E., Fridmacher, V., Watanabe, M., Naco, G. et Whitman, M. (1997). Smad4 and FAST-1 in the assembly of activin-responsive factor. *Nature*, 389(6646):85–89.

Cheng, H.-L., Edelsbrunner, H. et Fu, P. (2001). Shape space from deformation. *Computational Geometry*, 19(2-3):191 – 204.

Cheng, H.-L. et Shi, X. (2004). Guaranteed quality triangulation of molecular skin surfaces. *In Proceedings of the conference on Visualization '04*, pages 481–488. IEEE Computer Society.

Cheng, H.-L. et Shi, X. (2005). Quality mesh generation for molecular skin surfaces using restricted union of balls. *vis*, 00:51–57.

Christen, M., Hünenberger, P. H., Bakowies, D., Baron, R., Bürgi, R., Geerke, D. P., Heinz, T. N., Kastenholz, M. A., Kräutler, V., Oostenbrink, C., Peter, C., Trzesniak, D. et van Gunsteren, W. F. (2005). The GROMOS software for biomolecular simulation : GROMOS05. *J Comput Chem*, 26(16):1719–1751.

Cipriano, G. et Gleicher, M. (2007). Molecular surface abstraction. *IEEE Trans Vis Comput Graph*, 13(6):1608–15. 1077-2626 (Print)Journal ArticleResearch Support, N.I.H., Extramural.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. et Walters, L. (1998). New goals for the U.S. Human Genome Project : 1998-2003. *Science*, 282(5389):682–689.

Comeau, S. R. et Camacho, C. J. (2005). Predicting oligomeric assemblies : N-mers a primer. *J. Struct. Biol.*, 150:233–244.

Comeau, S. R., Gatchell, D. W., Vajda, S. et Camacho, C. J. (2004a). ClusPro : a fully automated algorithm for protein-protein docking. *Nucleic Acids Res.*, 32:W96–W99.

Comeau, S. R., Gatchell, D. W., Vajda, S. et Camacho, C. J. (2004b). ClusPro : an automated docking and discriminiation method for the prediction of protein complexes. *Bioinformatics*, 20:45–50.

Comeau, S. R., Vajda, S. et Camacho, C. J. (2005). Performance of the first protein docking server *ClusPro* in CAPRI rounds 3–5. *Proteins : Struct. Func. Bioinf.*, 60:239–244.

Connolly, M. L. (1983). Analytical molecular surface calculation. *Journal of Applied Crystallography*, 16(5):548–558.

Connolly, M. L. (1985). molecular surface triangulation. *Journal of Applied Crystallography*, 18(6):499–505.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. et Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197.

Crowley, P. B. et Golovin, A. (2005). Cation-pi interactions in protein-protein interfaces. *Proteins*, 59(2):231–239.

Dai, F., Chang, C., Lin, X., Dai, P., Mei, L. et Feng, X. H. (2007). Erbin inhibits transforming growth factor beta signaling through a novel smad-interacting domain. *Mol Cell Biol*, 27(17):6183–94. 0270-7306 (Print)Journal ArticleResearch Support, N.I.H., ExtramuralResearch Support, Non-U.S. Gov't.

Dai, P., Xiong, W. C. et Mei, L. (2006). Erbin inhibits RAF activation by disrupting the sur-8-Ras-Raf complex. *J Biol Chem*, 281(2):927–933.

Das, K., Acton, T., Chiang, Y., Shih, L., Arnold, E. et Montelione, G. T. (2004). Crystal structure of RlmAI : implications for understanding the 23S rRNA G745/G748-methylation at the macrolide antibiotic-binding site. *Proc Natl Acad Sci U S A*, 101(12):4041–4046.

Daura, X., Mark, A. E. et Gunsteren, W. F. V. (1998). Parametrization of aliphatic chn united atoms of GROMOS96 force field. *Journal of Computational Chemistry*, 19(5):535–547.

de Vries, S. J. et Bonvin, A. M. J. J. (2008). How proteins get in touch : interface prediction in the study of biomolecular complexes. *Curr Protein Pept Sci*, 9(4):394–406.

de Vries, S. J., van Dijk, A. D. J. et Bonvin, A. M. J. J. (2006). WHISCY : what information does surface conservation yield ? Application to data-driven docking. *Proteins*, 63(3):479–489.

de Vries, S. J., van Dijk, A. D. J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T. et Bonvin, A. M. J. J. (2007). HADDOCK versus HADDOCK : new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, 69(4):726–733.

Dennler, S., Huet, S. et Gauthier, J. M. (1999). A short amino-acid sequence in MH1 domain is responsible for functional differences between Smad2 and Smad3. *Oncogene*, 18(8):1643–1648.

Desmyter, A., Spinelli, S., Payan, F., Lauwereys, M., Wyns, L., Muyldermans, S. et Cambillau, C. (2002). Three camelid VHH domains in complex with porcine pancreatic alpha-amylase. Inhibition and versatility of binding topology. *J Biol Chem*, 277(26):23645–23650.

Deussen, O. et Strothotte, T. (2000). Computer-generated pen-and-ink illustration of trees. *In SIGGRAPH '00 : Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 13–18, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Dickerson, R. E. (1998). DNA bending : the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res*, 26(8):1906–1926.

Dillon, C., Creer, A., Kerr, K., Kümin, A. et Dickson, C. (2002). Basolateral targeting of ERBB2 is dependent on a novel bipartite juxtamembrane sorting signal but independent of the C-terminal ERBIN-binding domain. *Mol Cell Biol*, 22(18):6553–6563.

Dobbins, S. E., Lesk, V. I. et Sternberg, M. J. E. (2008). Insights into protein flexibility : The relationship between normal modes and conformational change upon protein-protein docking. *Proc Natl Acad Sci U S A*, 105(30):10390–10395.

Dominguez, C., Boelens, R. et Bonvin, A. M. J. J. (2003). HADDOCK : a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, 125: 1731–1737.

Dong, F. et Zhou, H.-X. (2006). Electrostatic contribution to the binding stability of protein-protein complexes. *Proteins*, 65(1):87–102.

Doudna, J. A. (2000). Structural genomics of RNA. *Nat Struct Biol*, 7 Suppl:954–956.

Douguet, D., Chen, H.-C., Tovchigrechko, A. et Vakser, I. A. (2006). DOCKGROUND resource for studying protein-protein interfaces. *Bioinformatics*, 22(21):2612–2618.

Draper, D. E. (1999). Themes in RNA-protein recognition. *J Mol Biol*, 293(2):255–270.

Duhovny, D., Nussinov, R. et Wolfson, H. (2002). Efficient unbound docking of rigid molecules. *In Proceedings of the 2nd Workshop on algorithms in bioinformatics*.

Dunbrack, R. L., Gerloff, D. L., Bower, M., Chen, X., Lichtarge, O. et Cohen, F. E. (1997). Meeting review : the Second meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2), Asilomar, California, December 13-16, 1996. *Fold Des*, 2(2):R27–R42.

Dunbrack, R. L. et Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol*, 230(2):543–574.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C. et Obradovic, Z. (2001). Intrinsically disordered protein. *J Mol Graph Model*, 19(1):26–59.

Dyson, H. J. et Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol*, 12(1):54–60.

Echols, N., Milburn, D. et Gerstein, M. (2003). MolMovDB : analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res*, 31(1):478–482.

Eddy, S. R. (2004). Where did the blosum62 alignment score matrix come from ? *Nat Biotechnol*, 22(8):1035–6. 1087-0156 (Print)Evaluation StudiesJournal ArticleReview.

Edelsbrunner, H. (1995). Smooth surfaces for multi-scale shape representation. *In Proceedings of the 15th Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 391–412. Springer-Verlag.

Edelsbrunner, H. (1999). Deformable smooth surface design. *Discrete & Computational Geometry*, 21(1):87–115.

Edelsbrunner, H. et Shah, N. R. (1992). Incremental topological flipping works for regular triangulations. *In SCG '92 : Proceedings of the eighth annual symposium on Computational geometry*, pages 43–52, New York, NY, USA. ACM.

Eghiaian, F., Grosclaude, J., Lesceu, S., Debey, P., Doublet, B., Tréguer, E., Rezaei, H. et Knossow, M. (2004). Insight into the PrPC–>PrPSc conversion from the structures of antibody-bound ovine prion scrapie-susceptibility variants. *Proc Natl Acad Sci U S A*, 101(28):10254–10259.

Eisen, M. B., Spellman, P. T., Brown, P. O. et Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868.

Eisenstein, M. et Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *Comptes Rendus Biologies*, 327:409–420.

Elcock, A. H. et McCammon, J. A. (2001). Identification of protein oligomerization states by analysis of interface conservation. *Proc Natl Acad Sci U S A*, 98(6):2990–2994.

Elcock, A. H., Sept, D. et McCammon, J. A. (2001). Computer simulation of protein-protein interactions. *J. Phys. Chem.*, 105:1504–1518.

Ellis, J. J., Broom, M. et Jones, S. (2007). Protein-RNA interactions : structural analysis and functional classes. *Proteins*, 66(4):903–911.

Ellis, J. J. et Jones, S. (2008). Evaluating conformational changes in protein structures binding RNA. *Proteins*, 70(4):1518–1526.

Emekli, U., Schneidman-Duhovny, D., Wolfson, H. J., Nussinov, R. et Haliloglu, T. (2008). HingeProt : automated prediction of hinges in protein structures. *Proteins*, 70(4):1219–1227.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. et Pedersen, L. G. (1995). A smooth particle mesh ewald method. *The Journal of Chemical Physics*, 103(19): 8577–8593.

Ewald, P. P. (1921). Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369:253–287.

Eyal, E. et Halperin, D. (2005). Dynamic maintenance of molecular surfaces under conformational changes. *In Proceedings of the twenty-first annual symposium on Computational geometry*, pages 45–54. ACM, Pisa, Italy.

Favre, B., Fontao, L., Koster, J., Shafaatian, R., Jaunin, F., Saurat, J. H., Sonnenberg, A. et Borradori, L. (2001). The hemidesmosomal protein bullous pemphigoid antigen 1 and the integrin beta 4 subunit bind to ERBIN. Molecular cloning of multiple alternative splice variants of ERBIN and analysis of their tissue expression. *J Biol Chem*, 276(35):32427–32436.

Fernández-Recio, J., Totrov, M. et Abagyan, R. (2002). Soft protein-protein docking in internal coordinates. *Protein Sci.*, 11:280–291.

Fernández-Recio, J., Totrov, M. et Abagyan, R. (2003). ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins : Struct. Func. Bioinf.*, 52:113–117.

Fernández-Recio, J., Totrov, M. et Abagyan, R. (2004). Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.*, 335:843–865.

Fernández-Recio, J., Totrov, M., Skorodumov, C. et Abagyan, R. (2005). Optimal docking area : a new method for predicting protein-protein interaction sites. *Proteins : Struct. Func. Bioinf.*, 58:134–143.

Fernando, R. et Kilgard, M. J. (2003). *The Cg Tutorial : The Definitive Guide to Programmable Real-Time Graphics*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Fieulaine, S., Morera, S., Poncet, S., Mijakovic, I., Galinier, A., Janin, J., Deutscher, J. et Nessler, S. (2002). X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr. *Proc Natl Acad Sci U S A*, 99(21):13437–13441.

Fischer, D., Lin, S. L., Wolfson, H. L. et Nussinov, R. (1995). A geometry-based suite of molecular docking processes. *J. Mol. Biol.*, 248:459–477.

Fiser, A. et Sali, A. (2003). Modeller : generation and refinement of homology-based protein structure models. *Methods Enzymol*, 374:461–91. 0076-6879 (Print)Journal ArticleResearch Support, Non-U.S. Gov'tResearch Support, U.S. Gov't, P.H.S.

Fitzjohn, P. W. et Bates, P. A. (2003). Guided docking : first step to locate potential binding sites. *Proteins*, 52(1):28–32.

Foley, J., Van Dam, A., Feiner, S. et Hughes, J. (1995). *Computer Graphics : Principles and Practice, second edition in C*. Addison-Wesley Professional.

Frankenstein, Z., Sperling, J., Sperling, R. et Eisenstein, M. (2008). FitEM2EM–tools for low resolution study of macromolecular assembly and dynamics. *PLoS ONE*, 3(10):e3594.

Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A. et Schulten, K. (2006). Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449.

Gabb, H. A., Jackson, R. M. et Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, 272(1):106–120.

Gabdoulline, R. R. et Wade, R. C. (1999). On the protein-protein diffusional encounter complex. *J Mol Recognit*, 12(4):226–234.

Gabdoulline, R. R. et Wade, R. C. (2001). Protein-protein association : investigation of factors influencing association rates by brownian dynamics simulations. *J Mol Biol*, 306(5):1139–1155.

Gallivan, J. P. et Dougherty, D. A. (1999). Cation-pi interactions in structural biology. *Proc Natl Acad Sci U S A*, 96(17):9459–9464.

Gao, Y., Douguet, D., Tovchigrechko, A. et Vakser, I. A. (2007). DOCKGROUND system of databases for protein recognition studies : unbound structures for docking. *Proteins*, 69(4):845–851.

Garrard, S. M., Capaldo, C. T., Gao, L., Rosen, M. K., Macara, I. G. et Tomchick, D. R. (2003). Structure of cdc42 in a complex with the gtpase-binding domain of the cell polarity protein, par6. *EMBO J*, 22(5):1125–1133.

Geng, W., Yu, S. et Wei, G. (2007). Treatment of charge singularities in implicit solvent models. *J Chem Phys*, 127(11):114106. 0021-9606 (Print)Journal Article.

Gerstein, M. et Echols, N. (2004). Exploring the range of protein flexibility, from a structural proteomics perspective. *Curr Opin Chem Biol*, 8(1):14–19.

Gherardi, E., Sandin, S., Petoukhov, M. V., Finch, J., Youles, M. E., Ofverstedt, L.-G., Miguel, R. N., Blundell, T. L., Woude, G. F. V., Skoglund, U. et Svergun, D. I. (2006). Structural basis of hepatocyte growth factor/scatter factor and MET signalling. *Proc Natl Acad Sci U S A*, 103(11):4046–4051.

Ghosh, A., Praefcke, G. J. K., Renault, L., Wittinghofer, A. et Herrmann, C. (2006). How guanylate-binding proteins achieve assembly-stimulated processive cleavage of GTP to GMP. *Nature*, 440(7080):101–104.

Gillet, A., Sanner, M., Stoffler, D. et Olson, A. (2005). Tangible interfaces for structural molecular biology. *Structure*, 13(3):483–491.

Glaser, F., Steinberg, D. M., Vakser, I. A. et Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43(2):89–102.

Günther, S., May, P., Hoppe, A., Frömmel, C. et Preissner, R. (2007). Docking without docking : ISEARCH–prediction of interactions using known interfaces. *Proteins*, 69(4):839–844.

Goddard, T. D. et Ferrin, T. E. (2007). Visualization software for molecular assemblies. *Curr Opin Struct Biol*, 17(5):587–595.

Goddard, T. D., Huang, C. C. et Ferrin, T. E. (2005). Software extensions to UCSF chimera for interactive visualization of large molecular assemblies. *Structure*, 13(3):473–482.

Goodsell, D. S. (2005). Visual methods from atoms to cells. *Structure*, 13(3):347–354.

Goto, D., Nakajima, H., Mori, Y., Kurasawa, K., Kitamura, N. et Iwamoto, I. (2001). Interaction between Smad anchor for receptor activation and Smad3 is not essential for TGF-beta/Smad3-mediated signaling. *Biochem Biophys Res Commun*, 281(5):1100–1105.

Graille, M., Heurgué-Hamard, V., Champ, S., Mora, L., Scrima, N., Ulryck, N., van Tilbeurgh, H. et Buckingham, R. H. (2005a). Molecular basis for bacterial class I release factor methylation by PrmC. *Mol Cell*, 20(6):917–927.

Graille, M., Mora, L., Buckingham, R. H., van Tilbeurgh, H. et de Zamaroczy, M. (2004). Structural inhibition of the colicin D tRNase by the tRNA-mimicking immunity protein. *EMBO J*, 23(7):1474–1482.

Graille, M., Stura, E. A., Bossus, M., Muller, B. H., Letourneur, O., Battail-Poirot, N., Sibaï, G., Gauthier, M., Rolland, D., Du, M.-H. L. et Ducancel, F. (2005b). Crystal structure of the complex between the monomeric form of Toxoplasma gondii surface antigen 1 (SAG1) and a monoclonal antibody that mimics the human immune response. *J Mol Biol*, 354(2):447–458.

Graille, M., Zhou, C.-Z., Receveur-Bréchot, V., Collinet, B., Declerck, N. et van Tilbeurgh, H. (2005c). Activation of the LicT transcriptional antiterminator involves a domain swing/lock mechanism provoking massive structural changes. *J Biol Chem*, 280(15):14780–14789.

Gray, J. J. (2006). High-resolution protein-protein docking. *Curr. Op. Struct. Biol.*, 16:183–193.

Gray, J. J., Moughan, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A. et Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.*, 331:281–299.

Grayson, P., Tajkhorshid, E. et Schulten, K. (2003). Mechanisms of selectivity in channels and enzymes studied with interactive molecular dynamics. *Biophys J*, 85(1):36–48.

Greer, J. et Bush, B. L. (1978). Macromolecular shape and surface maps by solvent exclusion. *Proc Natl Acad Sci U S A*, 75(1):303–7. 0027-8424 (Print)Journal ArticleResearch Support, U.S. Gov't, P.H.S.

Grünewald, K., Desai, P., Winkler, D. C., Heymann, J. B., Belnap, D. M., Baumeister, W. et Steven, A. C. (2003). Three-dimensional structure of herpes simplex virus from cryo-electron tomography. *Science*, 302(5649):1396–1398.

Grubmüller, H., Heller, H., Windemuth, A. et Schulten, K. (1991). Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions. *Molecular Simulation*, 6(1):121–142.

Grünberg, R., Leckner, J. et Nilges, M. (2004). Complementarity of structure ensembles in protein-protein docking. *Structure*, 12:2125–2136.

Halperin, I., Ma, B., Wolfson, H. et Nussinov, R. (2002). Principles of docking : An overview of search algorithms and a guide to scoring functions. *Proteins : Struct. Func. Genet.*, 47:409–443.

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

Hao, X. et Varshney, A. (2004). Real-time visualization of large time-varying molecules. *In Proceedings of the High-Performance Computing Symposium '04*.

Harris, B. Z. et Lim, W. A. (2001). Mechanism and role of pdz domains in signaling complex assembly. *J Cell Sci*, 114(Pt 18):3219–3231.

Harrison, R. W., Kourinov, I. V. et Andrews, L. C. (1994). The Fourier-Green's function and the rapid evaluation of molecular potentials. *Protein Eng*, 7(3):359–369.

Heifetz, A. et Eisenstein, M. (2003). Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking. *Protein Eng*, 16(3):179–185.

Heifetz, A., Katchalski-Katzir, E. et Eisenstein, M. (2002). Electrostatics in protein-protein docking. *Protein Sci.*, 11:571–587.

Heifetz, A., Pal, S. et Smith, G. R. (2007). Protein-protein docking : progress in CAPRI rounds 6-12 using a combination of methods : the introduction of steered solvated molecular dynamics. *Proteins*, 69(4):816–822.

Henikoff, S. et Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9. 0027-8424 (Print)Journal Article.

Hildebrandt, A., Blossey, R., Rjasanow, S., Kohlbacher, O. et Lenhof, H.-P. (2007). Electrostatic potentials of proteins in water : a structured continuum approach. *Bioinformatics*, 23(2):e99–103.

Hillier, B. J., Christopherson, K. S., Prehoda, K. E., Bredt, D. S. et Lim, W. A. (1999). Unexpected modes of pdz domain scaffolding revealed by structure of nnos-syntrophin complex. *Science*, 284(5415):812–5. 0036-8075 (Print)Journal Article.

Hockney, R. W. (1970). The potential calculation and some applications. *Methods in Computational Physics*, 9:136–211.

Hou, Z., Bernstein, D. A., Fox, C. A. et Keck, J. L. (2005). Structural basis of the Sir1-origin recognition complex interaction in transcriptional silencing. *Proc Natl Acad Sci U S A*, 102(24):8489–8494.

Hu, Z., Ma, B., Wolfson, H. et Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins*, 39(4):331–342.

Huang, S.-Y. et Zou, X. (2008). An iterative knowledge-based scoring function for protein-protein recognition. *Proteins*, 72(2):557–579.

Huang, Y. Z., Zang, M., Xiong, W. C., Luo, Z. et Mei, L. (2003). Erbin suppresses the MAP kinase pathway. *J Biol Chem*, 278(2):1108–1114.

Hubbard, S. et Thornton, J. (1992). Naccess v2.1.1 - solvent accessible area calculations.

Humphrey, W., Dalke, A. et Schulten, K. (1996). Vmd : visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8.

Humphreys, D. D., Freisner, R. A. et Berne, B. J. (1994). a multiple-time-step molecular dynamics algorithm for macromolecules. *Journal of Physical Chemistry*, 98:6885–6892.

Hunjan, J., Tovchigrechko, A., Gao, Y. et Vakser, I. A. (2008). The size of the intermolecular energy funnel in protein-protein interactions. *Proteins*, 72(1):344–352.

Hwang, H., Pierce, B., Mintseris, J., Janin, J. et Weng, Z. (2008). Protein-protein docking benchmark version 3.0. *Proteins*, 73(3):705–709.

Inman, G. J., Nicolás, F. J. et Hill, C. S. (2002). Nucleocytoplasmic shuttling of Smads 2, 3, and 4 permits sensing of TGF-beta receptor activity. *Mol Cell*, 10(2):283–294.

Isralewitz, B., Baudry, J., Gullingsrud, J., Kosztin, D. et Schulten, K. (2001a). Steered molecular dynamics investigations of protein function. *J Mol Graph Model*, 19(1):13–25.

Isralewitz, B., Gao, M. et Schulten, K. (2001b). Steered molecular dynamics and mechanical functions of proteins. *Curr Opin Struct Biol*, 11(2):224–230.

Janin, J. (1997). The kinetics of protein-protein recognition. *Proteins*, 28(2):153–161.

Janin, J. (1999). Wet and dry interfaces : the role of solvent in protein-protein and protein-DNA recognition. *Structure*, 7(12):R277–R279.

Janin, J. (2005a). Assessing predictions of protein-protein interactions : The CAPRI experiment. *Protein Sci.*, 14:278–283.

Janin, J. (2005b). The targets of CAPRI rounds 3-5. *Proteins*, 60(2):170–175.

Janin, J. (2007). The targets of CAPRI rounds 6-12. *Proteins*, 69(4):699–703.

Janin, J. et Chothia, C. (1990). The structure of protein-protein recognition sites. *J Biol Chem*, 265(27):16027–16030.

Janin, J., Henrick, K., Moult, J., Ten Eyck, L., Sternberg, M. J. E., Vajda, S., Vakser, I. et Wodak, S. J. (2003). CAPRI : a critical assessment of predicted interactions. *Proteins : Struct. Func. Genet.*, 52:2–9.

Janin, J., Rodier, F., Chakrabarti, P. et Bahadur, R. P. (2007). Macromolecular recognition in the Protein Data Bank. *Acta Cryst.*, D63:1–8.

Janin, J. et Séraphin, B. (2003). Genome-wide studies of protein-protein interaction. *Curr. Op. Struct. Biol.*, 13:383–388.

Janin, J. et Wodak, S. (2007). The third CAPRI assessment meeting Toronto, Canada, April 20-21, 2007. *Structure*, 15(7):755–759.

Jaulin-Bastard, F., Arsanto, J. P., Le Bivic, A., Navarro, C., Vely, F., Saito, H., Marchetto, S., Hatzfeld, M., Santoni, M. J., Birnbaum, D. et Borg, J. P. (2002). Interaction between erbin and a catenin-related protein in epithelial cells. *J Biol Chem*, 277(4):2869–75. 0021-9258 (Print)Journal ArticleResearch Support, Non-U.S. Gov't.

Jaulin-Bastard, F., Nola, S. et Borg, J. (2005). Protéines lap : de nouvelles clés de voûte de l'architecture épithéliale. *Medecine Sciences*, 21:267–272.

Jaulin-Bastard, F., Saito, H., Le Bivic, A., Ollendorff, V., Marchetto, S., Birnbaum, D. et Borg, J. P. (2001). The erbb2/her2 receptor differentially interacts with erbin and pick1 psd-95/dlg/zo-1 domain proteins. *J Biol Chem*, 276(18):15256–63. 0021-9258 (Print)Journal Article.

Jayaram, B. et Jain, T. (2004). The role of water in protein-DNA recognition. *Annu Rev Biophys Biomol Struct*, 33:343–361.

Jiang, F. et Kim, S. H. (1991). "Soft docking" : matching of molecular surface cubes. *J Mol Biol*, 219(1):79–102.

Jiang, L. et Lai, L. (2002). CH...O hydrogen bonds at protein-protein interfaces. *J Biol Chem*, 277(40):37732–37740.

Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. et Thornton, J. M. (2001). Protein-RNA interactions : a structural analysis. *Nucleic Acids Res*, 29(4):943–954.

Jones, S. et Thornton, J. M. (1995). Protein-protein interactions : a review of protein dimer structures. *Prog Biophys Mol Biol*, 63(1):31–65.

Jones, S. et Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci.*, 93(1):13–20.

Jones, S. et Thornton, J. M. (1997). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, 272:121–132.

Jones, S., van Heyningen, P., Berman, H. M. et Thornton, J. M. (1999). Protein-DNA interactions : A structural analysis. *J Mol Biol*, 287(5):877–896.

Jorgensen, W., Maxwell, D. et Tirado-Rives, J. (1996). Development and testing of the opls all-atom force-field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118 (45):11225–11236.

Jorgensen, W. et Tirado-Rives, J. (1988). The opls potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110 (6):1657–1666.

Kale, L. et Skeel, R. (1999). Namd2 : Greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, 151(1):283–312.

Karplus, M. et Kuriyan, J. (2005). Molecular dynamics and protein function. *Proc Natl Acad Sci U S A*, 102(19):6679–6685.

Karplus, M. et McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat Struct Biol*, 9(9):646–52. 1072-8368 (Print)Journal ArticleReview.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A. et Aflalo, C. (1992). Molecular surface recognition : Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci.*, 89:2195–2199.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. et Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666.

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

Keskin, O. et Nussinov, R. (2007). Similar binding sites and different partners : implications to shared proteins in cellular pathways. *Structure*, 15(3):341–354.

Kessenich, J., Baldwin, D. et Rost, R. (2003). *The opengl shading language*.

Kim, E. et Sheng, M. (2004). Pdz domain proteins of synapses. *Nat Rev Neurosci*, 5(10):771–81. 1471-003X (Print)Journal ArticleReview.

Kim, O. T. P., Yura, K. et Go, N. (2006). Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res*, 34(22):6450–6460.

Knight, J. L., Mekler, V., Mukhopadhyay, J., Ebright, R. H. et Levy, R. M. (2005). Distance-restrained docking of rifampicin and rifamycin SV to RNA polymerase using systematic FRET measurements : developing benchmarks of model quality and reliability. *Biophys J*, 88(2):925–938.

Koehl, P. (2006). Electrostatics calculations : latest methodological advances. *Curr. Op. Struct. Biol.*, 16:142–151.

Kolch, W. (2003). Erbin : sorting out ErbB2 receptors or giving Ras a break ? *Sci STKE*, 2003(199):pe37.

Kontkanen, J. et Laine, S. (2005). Ambient occlusion fields. *In Proceedings of ACM SIGGRAPH 2005 Symposium on Interactive 3D Graphics and Games*, pages 41–48. ACM Press.

Kornau, H. C., Schenker, L. T., Kennedy, M. B. et Seeburg, P. H. (1995). Domain interaction between nmda receptor subunits and the postsynaptic density protein psd-95. *Science*, 269(5231):1737–1740.

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc Natl Acad Sci U S A*, 44(2):98–104.

Kozakov, D., Schueler-Furman, O. et Vajda, S. (2008). Discrimination of near-native structures in protein-protein docking by testing the stability of local minima. *Proteins*, 72(3): 993–1004.

Krissinel, E. et Henrick, K. (2004). Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*, 60(Pt 12 Pt 1):2256–2268.

Król, M., Chaleil, R. A. G., Tournier, A. L. et Bates, P. A. (2007a). Implicit flexibility in protein docking : cross-docking and local refinement. *Proteins*, 69(4):750–757.

Król, M., Tournier, A. L. et Bates, P. A. (2007b). Flexible relaxation of rigid-body docking solutions. *Proteins*, 68(1):159–169.

Kruithof, N. et Vegter, G. (2004). Approximation by skin surfaces. *Computer-Aided Design*, 36:1075–1088.

Kruithof, N. et Vegter, G. (2007). Meshing skin surfaces with certified topology. *Comput. Geom. Theory Appl.*, 36(3):166–182.

Kräutler, V., F., v. W. et H., H. P. (2001). A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry*, 22(5):501–508.

Kufareva, I., Budagyan, L., Raush, E., Totrov, M. et Abagyan, R. (2007). PIER : protein interface recognition for structural proteomics. *Proteins*, 67(2):400–417.

Lampe, O., Viola, I., Reuter, N. et Hauser, H. (2007). Two-level approach to efficient visualization of protein dynamics. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1616–1623.

Lander, E. S. et the International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Laura, R. P., Witt, A. S., Held, H. A., Gerstner, R., Deshayes, K., Koehler, M. F., Kosik, K. S., Sidhu, S. S. et Lasky, L. A. (2002). The erbin pdz domain binds with high affinity and specificity to the carboxyl termini of delta-catenin and arvcf. *J Biol Chem*, 277(15):12906–14. 0021-9258 (Print)Journal Article.

Lawson, C. L., Dutta, S., Westbrook, J. D., Henrick, K. et Berman, H. M. (2008). Representation of viruses in the remediated PDB archive. *Acta Crystallogr D Biol Crystallogr*, D64(Pt 8):874–882.

Leach, A. R. (2001). *Molecular modelling : principles and applications (2nd edition)*. Prentice Hall.

Lee, B. et Richards, F. M. (1971). The interpretation of protein structures : estimation of static accessibility. *J Mol Biol*, 55(3):379–400. 0022-2836 (Print)Journal Article.

Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A. et Brasseur, R. (2005). Protein-nucleic acid recognition : statistical analysis of atomic interactions and influence of DNA structure. *Proteins*, 61(2):258–271.

Lensink, M. F., Méndez, R. et Wodak, S. J. (2007). Docking and scoring protein complexes : CAPRI 3rd Edition. *Proteins*, 69(4):704–718.

Leulliot, N. et Varani, G. (2001). Current topics in RNA-protein recognition : control of specificity and biological function through induced fit and conformational capture. *Biochemistry*, 40(27):7947–7956.

Levinthal, C. (1966). Molecular model-building by computer. *Sci Am*, 214(6):42–52.

Levitt, M. et Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, 261(5561): 552–558.

Levitt, M. et Sharon, R. (1988). Accurate simulation of protein dynamics in solution. *Proc Natl Acad Sci U S A*, 85(20):7557–7561.

Levitt, M. et Warshel, A. (1975). Computer simulation of protein folding. *Nature*, 253(5494): 694–698.

Levy, Y. et Onuchic, J. N. (2006). Water mediation in protein folding and molecular recognition. *Annu Rev Biophys Biomol Struct*, 35:389–415.

Li, C. H., Ma, X. H., Chen, W. Z. et Wang, C. X. (2003a). A protein-protein docking algorithm dependent on the type of complexes. *Protein Eng*, 16(4):265–269.

Li, L., Chen, R. et Weng, Z. (2003b). RDOCK : refinement of rigid-body protein docking predictions. *Proteins : Struct. Func. Genet.*, 53:693–707.

Lo Conte, L., Chothia, C. et Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–2198.

Loew, L. M. et Schaff, J. C. (2001). The Virtual Cell : a software environment for computational cell biology. *Trends Biotechnol*, 19(10):401–406.

Luscombe, N. M., Austin, S. E., Berman, H. M. et Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biol*, 1(1):REVIEWS001.

Lyskov, S. et Gray, J. J. (2008). The RosettaDock server for local protein-protein docking. *Nucleic Acids Res*, 36(Web Server issue):W233–W238.

Ma, B., Elkayam, T., Wolfson, H. et Nussinov, R. (2003). Protein-protein interactions : structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A*, 100(10):5772–5777.

Ma, J., Sigler, P. B., Xu, Z. et Karplus, M. (2000). A dynamic model for the allosteric mechanism of GroEL. *J Mol Biol*, 302(2):303–313.

Mackerell, A. D. (2004). Empirical force fields for biological macromolecules : overview and issues. *J Comput Chem*, 25(13):1584–1604.

MacKerell, A. D., Banavali, N. et Foloppe, N. (2000). Development and current status of the CHARMM force field for nucleic acids. *Biopolymers*, 56(4):257–265.

MacKerell, A. D. et Banavali, N. K. (2000). All-atom empirical force field for nucleic acids : II. Application to molecular dynamics simulations of DNA and RNA in solution. *Journal of Computational Chemistry*, 21:105–120.

MacKerell, A. D., Brooks, J. B., III, C. L. B., Nilsson, L., Roux, B., Won, Y. et Karplus, M. (1998). Charmm : The energy function and its parameterization with an overview of the program. *In The Encyclopedia of Computational Chemistry*, pages 271–277. John Wiley & Sons.

Mackerell, A. D., Feig, M. et Brooks, C. L. (2004). Extending the treatment of backbone energetics in protein force fields : limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem*, 25(11):1400–1415.

Mackerell, A. D. et Nilsson, L. (2008). Molecular dynamics simulations of nucleic acid-protein complexes. *Curr Opin Struct Biol*, 18(2):194–199.

Magalhaes, A., Maigret, B., Hoflack, J., Gomes, J. N. F. et Scheraga, H. A. (1994). Contribution of unusual arginine-arginine short-range interactions to stabilization and recognition in proteins. *Journal of Protein Chemistry*, 13:195–215.

Mandel-Gutfreund, Y., Margalit, H., Jernigan, R. L. et Zhurkin, V. B. (1998). A role for CH...O interactions in protein-DNA recognition. *J Mol Biol*, 277(5):1129–1140.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I. et Ten Eyck, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng.*, 14(2):105–113.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O. et Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753.

Martin, O. et Schomburg, D. (2008). Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. *Proteins*, 70(4):1367–1378.

Masuyama, N., Hanafusa, H., Kusakabe, M., Shibuya, H. et Nishida, E. (1999). Identification of two Smad4 proteins in Xenopus. Their common and distinct properties. *J Biol Chem*, 274(17):12163–12170.

May, A. et Zacharias, M. (2008). Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. *Proteins*, 70(3):794–809.

McCammon, J. A., Gelin, B. R. et Karplus, M. (1977). Dynamics of folded proteins. *Nature*, 267(5612):585–590.

Méndez, R., Leplae, R., De Maria, L. et Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions : current status of docking methods. *Proteins : Struct. Func. Genet.*, 52:51–67.

Méndez, R., Leplae, R., Lensink, M. F. et Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. *Proteins : Struct. Func. Bioinf.*, 60:150–169.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. et Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. et Weng, Z. (2005). Protein-protein docking benchmark 2.0 : An update. *Proteins : Struct. Func. Genet.*, 60:214–216.

Mittermaier, A. et Kay, L. E. (2006). New tools provide new insights in NMR studies of protein dynamics. *Science*, 312(5771):224–228.

Ménétrey, J., Perderiset, M., Cicolari, J., Dubois, T., Elkhatib, N., Khadali, F. E., Franco, M., Chavrier, P. et Houdusse, A. (2007). Structural basis for ARF1-mediated recruitment of ARHGAP21 to Golgi membranes. *EMBO J*, 26(7):1953–1962.

Moore, W. H. et Krimm, S. (1976). Vibrational analysis of peptides, polypeptides, and proteins. II. beta-poly-2l-alanine and beta-poly-2-alanylglycine). *Biopolymers*, 15(12):2465–2483.

Moreira, I. S., Fernandes, P. A. et Ramos, M. J. (2007). Hot spots–a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–812.

Morozova, N., Allers, J., Myers, J. et Shamoo, Y. (2006). Protein-RNA interactions : exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, 22(22):2746–2752.

Motiejunas, D., Gabdoulline, R., Wang, T., Feldman-Salit, A., Johann, T., Winn, P. J. et Wade, R. C. (2008). Protein-protein docking by simulating the process of association subject to biochemical constraints. *Proteins*, 71(4):1955–1969.

Moustakas, A., Souchelnytskyi, S. et Heldin, C. H. (2001). Smad regulation in TGF-beta signal transduction. *J Cell Sci*, 114(Pt 24):4359–4369.

Márquez, J. A., Smith, C. I. E., Petoukhov, M. V., Surdo, P. L., Mattsson, P. T., Knekt, M., Westlund, A., Scheffzek, K., Saraste, M. et Svergun, D. I. (2003). Conformation of full-length Bruton tyrosine kinase (Btk) from synchrotron X-ray solution scattering. *EMBO J*, 22(18):4616–4624.

Mustard, D. et Ritchie, D. W. (2005). Docking essential dynamics eigenstructures. *Proteins : Struct. Func. Bioinf.*, 60:269–274.

Nadassy, K., Wodak, S. J. et Janin, J. (1999). Structural features of protein-nucleic acid recognition sites. *Biochemistry*, 38(7):1999–2017.

Natarajan, P., Lander, G. C., Shepherd, C. M., Reddy, V. S., Brooks, C. L. et Johnson, J. E. (2005). Exploring icosahedral virus structures with VIPER. *Nat Rev Microbiol*, 3(10): 809–817.

Neumaier, A. (1997). Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Review*, 39:407–460.

Nielsen, T. K., Liu, S., Lührmann, R. et Ficner, R. (2007). Structural basis for the bifunctionality of the U5 snRNP 52K protein (CD2BP2). *J Mol Biol*, 369(4):902–908.

Nooren, I. M. A. et Thornton, J. M. (2003a). Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–3492.

Nooren, I. M. A. et Thornton, J. M. (2003b). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol*, 325(5):991–1018.

Nourry, C., Grant, S. G. et Borg, J. P. (2003). Pdz domain proteins : plug and play ! *Sci STKE*, 2003(179):RE7. 1525-8882 (Electronic)Journal ArticleReview.

Ofran, Y. et Rost, B. (2007). Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol*, 3(7):e119.

Oostenbrink, C., Villa, A., Mark, A. E. et van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation : the GROMOS force-field parameter sets 53A5 and 53A6. *J Comput Chem*, 25(13):1656–1676.

Owen, D., Mott, H. R., Laue, E. D. et Lowe, P. N. (2000). Residues in Cdc42 that specify binding to individual CRIB effector proteins. *Biochemistry*, 39(6):1243–1250.

Pabo, C. O. et Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces : why is there no simple code for recognition ? *J Mol Biol*, 301(3):597–624.

Palma, P. N., Krippahl, L., Wampler, J. E. et Moura, J. J. G. (2000). BiGGER : a new (soft) docking algorithm for predicting protein interactions. *Proteins : Struct. Func. Genet.*, 39:372–384.

Pandey, A. et Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, 405(6788): 837–846.

Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z. et Wolynes, P. G. (2004). Water in protein structure prediction. *Proc Natl Acad Sci U S A*, 101(10):3352–3357.

Parsons, J. T. (2003). Focal adhesion kinase : the first ten years. *J Cell Sci*, 116(Pt 8):1409–1416.

Parsons, M. R., Convery, M. A., Wilmot, C. M., Yadav, K. D., Blakeley, V., Corner, A. S., Phillips, S. E., McPherson, M. J. et Knowles, P. F. (1995). Crystal structure of a quinoenzyme : copper amine oxidase of Escherichia coli at 2 A resolution. *Structure*, 3(11):1171–1184.

Petrone, P. et Pande, V. S. (2006). Can conformational change be described by only a few normal modes ? *Biophys J*, 90(5):1583–1593.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. et Schulten, K. (2005). Scalable molecular dynamics with namd. *J Comput Chem*, 26(16):1781–1802.

Phipps, K. R. et Li, H. (2007). Protein-RNA contacts at crystal packing surfaces. *Proteins*, 67(1):121–127.

Pierce, B., Tong, W. et Weng, Z. (2005). M-ZDOCK : a grid-based approach for C_n symmetric multimer docking. *Bioinformatics*, 21:1472–1478.

Pierce, B. et Weng, Z. (2007). ZRANK : reranking protein docking predictions with an optimized energy function. *Proteins*, 67(4):1078–1086.

Pierce, B. et Weng, Z. (2008). A combination of rescoring and refinement significantly improves protein docking performance. *Proteins*, 72(1):270–279.

Ponder, J. W. et Case, D. A. (2003). Force fields for protein simulations. *Adv Protein Chem*, 66:27–85.

Ponomarenko, J. V. et Bourne, P. E. (2007). Antibody-protein interactions : benchmark datasets and prediction tools evaluation. *BMC Struct Biol*, 7:64.

Ponstingl, H., Kabir, T., Gorse, D. et Thornton, J. M. (2005). Morphological aspects of oligomeric protein structures. *Prog Biophys Mol Biol*, 89(1):9–35.

Porollo, A. et Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. *Proteins : Struct. Func. Bioinf.*, 66:630–645.

Prabakaran, P., Siebers, J. G., Ahmad, S., Gromiha, M. M., Singarayan, M. G. et Sarai, A. (2006). Classification of protein-DNA complexes based on structural descriptors. *Structure*, 14(9):1355–1367.

Prokova, V., Mavridou, S., Papakosta, P., Petratos, K. et Kardassis, D. (2007). Novel mutations in smad proteins that inhibit signaling by the transforming growth factor beta in mammalian cells. *Biochemistry*, 46(48):13775–13786. 0006-2960 (Print)Journal article.

Qin, B. Y., Lam, S. S., Correia, J. J. et Lin, K. (2002). Smad3 allostery links tgf-beta receptor kinase activation to transcriptional control. *Genes Dev*, 16(15):1950–1963.

Qin, S. et Zhou, H.-X. (2007). A holistic approach to protein docking. *Proteins*, 69(4):743–749.

Rajamani, D., Thiel, S., Vajda, S. et Camacho, C. J. (2004). Anchor residues in protein-protein interactions. *Proc. Natl. Acad. Sci.*, 101:11287–11292.

Ramesh, V., Mayer, C., Dyson, M. R., Gite, S. et RajBhandary, U. L. (1999). Induced fit of a peptide loop of methionyl-tRNA formyltransferase triggered by the initiator tRNA substrate. *Proc Natl Acad Sci U S A*, 96(3):875–880.

Ray, N., Cavin, X., Paul, J. C. et Maigret, B. (2005). Intersurf : dynamic interface between proteins. *J Mol Graph Model*, 23(4):347–54. 1093-3263 (Print)Journal Article.

Ress, A. et Moelling, K. (2006). Interaction partners of the pdz domain of erbin. *Protein Pept Lett*, 13(9):877–81. 0929-8665 (Print)Journal Article.

Reynolds, C., Damerell, D. et Jones, S. (2008). ProtorP : A Protein-Protein Interaction Analysis Server. *Bioinformatics*.

Rice, P. A., Yang, S., Mizuuchi, K. et Nash, H. A. (1996). Crystal structure of an IHF-DNA complex : a protein-induced DNA U-turn. *Cell*, 87(7):1295–1306.

Richards, F. M. (1977). Areas, volumes, packing, and protein structure. *Ann. Rev. Biophys. Bioeng.*, 6:151–176.

Ritchie, D. W. (2003). Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins*, 52(1):98–106.

Ritchie, D. W. (2008). Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci*, 9(1):1–15.

Ritchie, D. W. et Kemp, G. J. L. (1999). Fast computation, rotation, and comparison of low resolution spherical harmonic molecular surfaces. *Journal of Computational Chemistry*, 20(4):383–395.

Ritchie, D. W. et Kemp, G. J. L. (2000). Protein docking using spherical polar Fourier correlations. *Proteins : Struct. Func. Genet.*, 39(2):178–194.

Ritchie, D. W., Kozakov, D. et Vajda, S. (2008). Accelerating and focusing protein-protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics*, 24(17):1865–1873.

Rodier, F., Bahadur, R. P., Chakrabarti, P. et Janin, J. (2005). Hydration of protein-protein interfaces. *Proteins : Struct. Func. Bioinf.*, 60:36–45.

Rossi, R., Isorce, M., Morin, S., Flocard, J., Arumugam, K., Crouzy, S., Vivaudou, M. et Redon, S. (2007). Adaptive torsion-angle quasi-statics : a general simulation method with applications to protein structure analysis and design. *Bioinformatics*, 23(13):i408–i417.

Rost, R. J. (2006). *OpenGL(R) Shading Language (2nd Edition)*. Addison-Wesley Professional.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, P. et Sali, A. (2004). A structural perspective on protein-protein interactions. *Curr. Op. Struct. Biol.*, 14:313–324.

Russell, R. B. et Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison : assignment of global and residue confidence levels. *Proteins*, 14(2):309–323.

Ryckaert, J.-P., Ciccotti, G. et Berendsen, H. J. (1977). Numerical integration of the cartesian equations of motion of a system with constraints : molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341.

Sacquin-Mora, S., Carbone, A. et Lavery, R. (2008). Identification of protein interaction partners and protein-protein interaction sites. *J Mol Biol*, 382(5):1276–1289.

Saha, R. P., Bahadur, R. P., Pal, A., Mandal, S. et Chakrabarti, P. (2006). ProFace : a server for the analysis of the physicochemical features of protein-protein interfaces. *BMC Struct Biol*, 6:11.

Saito, M. (1994). Molecular dynamics simulations of proteins in solution : Artifacts caused by the cutoff approximation. *The Journal of Chemical Physics*, 101(5):4055–4061.

Samsonov, S., Teyra, J. et Pisabarro, M. T. (2008). A molecular dynamics approach to study the importance of solvent in protein interactions. *Proteins*, 73(2):515–525.

Sanbonmatsu, K. Y., Joseph, S. et Tung, C.-S. (2005). Simulating movement of tRNA into the ribosome during decoding. *Proc Natl Acad Sci U S A*, 102(44):15854–15859.

Sanner, M. F. et Olson, A. J. (1997). Real time surface reconstruction for moving molecular fragments. *Pac Symp Biocomput*, pages 385–96. 1793-5091 (Print)Journal ArticleResearch Support, U.S. Gov't, P.H.S.

Sanner, M. F., Olson, A. J. et Spehner, J. C. (1996). Reduced surface : an efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320.

Sansen, S., Ranter, C. J. D., Gebruers, K., Brijs, K., Courtin, C. M., Delcour, J. A. et Rabijns, A. (2004). Structural basis for inhibition of Aspergillus niger xylanase by triticum aestivum xylanase inhibitor-I. *J Biol Chem*, 279(34):36022–36028.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. et Wolfson, H. J. (2005a). Geometry-based flexible and symmetric protein docking. *Proteins : Struct. Func. Bioinf.*, 60:224–231.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. et Wolfson, H. J. (2005b). PatchDock and SymmDock : servers for rigid and symmetric docking. *Nucleic Acids Res.*, 33:W363–W367.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. et Wolfson, H. J. (2005). PatchDock and SymmDock : servers for rigid and symmetric docking. *Nucleic Acids Res*, 33(Web Server issue):W363–W367.

Schreiber, G. (2002). Kinetic studies of protein-protein interactions. *Curr Opin Struct Biol*, 12(1):41–47.

Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. et Bork, P. (2000). Smart : a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res*, 28(1):231–234.

Seeman, N. C., Rosenberg, J. M. et Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A*, 73(3):804–808.

Segal, D. et Eisenstein, M. (2005). The effect of resolution-dependent global shape modifications on rigid-body protein-protein docking. *Proteins : Struct. Func. Bioinf.*, 59:580–591.

Segal, M. et Akeley, K. (2004). *The OpenGL graphics system : a specification, version 2.0.*

Selvaraj, S., Kono, H. et Sarai, A. (2002). Specificity of protein-DNA recognition revealed by structure-based potentials : symmetric/asymmetric and cognate/non-cognate binding. *J Mol Biol*, 322(5):907–915.

Shajani, Z., Deka, P. et Varani, G. (2006). Decoding RNA motional codes. *Trends Biochem Sci*, 31(8):421–424.

Sheinerman, F. B. et Honig, B. (2002). On the role of electrostatic interactions in the design of protein-protein interfaces. *J Mol Biol*, 318(1):161–177.

Sheinerman, F. B., Norel, R. et Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr. Op. Struct. Biol.*, 10:153–159.

Shi, Y. et Massagué, J. (2003). Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell*, 113(6):685–700.

Shi, Y., Wang, Y. F., Jayaraman, L., Yang, H., Massagué, J. et Pavletich, N. P. (1998). Crystal structure of a Smad MH1 domain bound to DNA : insights on DNA binding in TGF-beta signaling. *Cell*, 94(5):585–594.

Shreiner, D., Woo, M., Neider, J. et Davis, T. (2005). *OpenGL(R) Programming Guide : The Official Guide to Learning OpenGL(R), Version 2 (5th Edition)*. Addison-Wesley Professional.

Sigg, C., Weyrich, T., Botsch, M. et Gross, M. (2006). Gpu-based ray-casting of quadratic surfaces. *Symposium on Point-Based Graphics*, pages 56–65.

Skelton, N. J., Koehler, M. F. T., Zobel, K., Wong, W. L., Yeh, S., Pisabarro, M. T., Yin, J. P., Lasky, L. A. et Sidhu, S. S. (2003). Origins of pdz domain ligand specificity. structure determination and mutagenesis of the erbin pdz domain. *J Biol Chem*, 278(9):7645–7654.

Slepchenko, B. M., Schaff, J. C., Macara, I. et Loew, L. M. (2003). Quantitative cell biology with the Virtual Cell. *Trends Cell Biol*, 13(11):570–576.

Smith, G. R., Fitzjohn, P. W., Page, C. S. et Bates, P. A. (2005a). Incorporation of flexibility into rigid-body docking : Applications in rounds 3–5 of CAPRI. *Proteins : Struct. Func. Bioinf.*, 60:263–268.

Smith, G. R. et Sternberg, M. J. E. (2002). Prediction of protein-protein interactions by docking methods. *Curr. Op. Struct. Biol.*, 12:28–35.

Smith, G. R., Sternberg, M. J. E. et Bates, P. A. (2005b). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.*, 347:1077–1101.

Soetens, J.-C., Millot, C., Chipot, C., Jansen, G., Angyan, J. et Maigret, B. (1997). Effect of polarizability on the potential of mean force of two cations. the guanidinium-guanidinium ion pair in water. *Journal of Physical Chemistry B*, 101(50):10910–10917.

Spolar, R. S. et Record, M. T. (1994). Coupling of local folding to site-specific binding of proteins to DNA. *Science*, 263(5148):777–784.

Stone, J. E., Phillips, J. C., Freddolino, P. L., Hardy, D. J., Trabuco, L. G. et Schulten, K. (2007). Accelerating molecular modeling applications with graphics processors. *J Comput Chem*, 28(16):2618–2640.

Strynadka, N. C., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B. K., Kuntz, I. D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M. et James, M. N. (1996). Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. *Nat Struct Biol*, 3(3):233–239.

Sumikoshi, K., Tereda, T., Nakamura, S. et Shimizu, K. (2005). A fast protein-protein docking algorithm using series expansions in terms of spherical basis functions. *Genome Inform.*, 16:161–173.

Sundberg, E. J., Li, H., Llera, A. S., McCormick, J. K., Tormo, J., Schlievert, P. M., Karjalainen, K. et Mariuzza, R. A. (2002). Structures of two streptococcal superantigens bound to TCR beta chains reveal diversity in the architecture of T cell signaling complexes. *Structure*, 10(5):687–699.

Swope, W. C., Andersen, H. C., Berens, P. H. et Wilson, K. R. (1981). *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules : Application to small water clusters*. Unknown.

Takagi, J., Yang, Y., Liu, J.-H., Wang, J.-H. et Springer, T. A. (2003). Complex between nidogen and laminin fragments reveals a paradigmatic beta-propeller interface. *Nature*, 424(6951):969–974.

Tama, F. et Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng*, 14(1):1–6.

Tang, C., Iwahara, J. et Clore, G. M. (2006). Visualization of transient encounter complexes in protein-protein association. *Nature*, 444(7117):383–386.

Tarini, M., Cignoni, P. et Montani, C. (2006). Ambient occlusion and edge cueing for enhancing real time molecular visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1237–1244.

Tasumi, M., Takeuchi, H., Ataka, S., Dwivedi, A. M. et Krimm, S. (1982). Normal vibrations of proteins : glucagon. *Biopolymers*, 21(3):711–714.

ten Dijke, P. et Hill, C. S. (2004). New insights into tgf-beta-smad signalling. *Trends Biochem Sci*, 29(5):265–73. 0968-0004 (Print)Journal ArticleReview.

ten Dijke, P., Miyazono, K. et Heldin, C. H. (2000). Signaling inputs converge on nuclear effectors in TGF-beta signaling. *Trends Biochem Sci*, 25(2):64–70.

Terashi, G., Takeda-Shitaka, M., Kanou, K., Iwadate, M., Takaya, D. et Umeyama, H. (2007). The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation. *Proteins*, 69(4):866–872.

Terrak, M., Kerff, F., Langsetmo, K., Tao, T. et Dominguez, R. (2004). Structural basis of protein phosphatase 1 regulation. *Nature*, 429(6993):780–4. 1476-4687 (Electronic)Journal Article.

Thouvenin, E., Schoehn, G., Rey, F., Petitpas, I., Mathieu, M., Vaney, M. C., Cohen, J., Kohli, E., Pothier, P. et Hewat, E. (2001). Antibody inhibition of the transcriptase activity of the rotavirus DLP : a structural view. *J Mol Biol*, 307(1):161–172.

Toledo, R. et Levy, B. (2004). Extending the graphic pipeline with new gpu-accelerated primitives. *Tech report*.

Tolstorukov, M. Y., Jernigan, R. L. et Zhurkin, V. B. (2004). Protein-DNA hydrophobic recognition in the minor groove is facilitated by sugar switching. *J Mol Biol*, 337(1):65–76.

Tovchigrechko, A. et Vakser, I. A. (2006). GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res*, 34(Web Server issue):W310–W314.

Tovchigrechko, A., Wells, C. A. et Vakser, I. A. (2002). Docking of protein models. *Protein Sci.*, 11:1888–1896.

Treger, M. et Westhof, E. (2001). Statistical analysis of atomic contacts at RNA-protein interfaces. *J Mol Recognit*, 14(4):199–214.

Tsai, C. J., Lin, S. L., Wolfson, H. J. et Nussinov, R. (1997). Studies of protein-protein interfaces : a statistical analysis of the hydrophobic effect. *Protein Sci*, 6(1):53–64.

Tsukazaki, T., Chiang, T. A., Davison, A. F., Attisano, L. et Wrana, J. L. (1998). SARA, a FYVE domain protein that recruits Smad2 to the TGFbeta receptor. *Cell*, 95(6):779–791.

Tuckerman, M., Berne, B. et Martyna, G. (1992). Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(3):1990–2001.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. et Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627.

Vajda, S. et Camacho, C. J. (2004). Protein-protein docking : is the glass half-full or half-empty ? *Trends in Biotechnology*, 22:110–116.

Vajda, S., Vakser, I. A., Sternberg, M. J. E. et Janin, J. (2002). Modeling of protein interactions in genomes. *Proteins : Struct. Func. Genet.*, 47(4):444–446.

Vakser, I. A. (1995). Protein docking for low-resolution structures. *Protein Eng.*, 8(4):371–377.

Vakser, I. A., Matar, O. G. et Lam, C. F. (1999). A systematic study of low resolution recognition in protein-protein complexes. *Proc. Natl. Acad. Sci.*, 96:8477–8482.

Valdar, W. S. et Thornton, J. M. (2001). Protein-protein interfaces : analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–124.

Valencia, A. et Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr. Op. Struct. Biol.*, 12:368–373.

van Dijk, A. D. J., Boelens, R. et Bonvin, A. M. J. J. (2005a). Data-driven docking for the study of biomolecular complexes. *FEBS J*, 272(2):293–312.

van Dijk, A. D. J. et Bonvin, A. M. J. J. (2006). Solvated docking : introducing water into the modelling of biomolecular complexes. *Bioinformatics*, 22(19):2340–2347.

van Dijk, A. D. J., de Vries, S. J., Dominguez, C., Chen, H., Zhou, H.-X. et Bonvin, A. M. J. J. (2005b). Data-driven docking : HADDOCK's adventures in CAPRI. *Proteins*, 60(2):232–238.

van Dijk, M. et Bonvin, A. M. J. J. (2008). A protein-DNA docking benchmark. *Nucleic Acids Res*, 36(14):e88.

van Gunsteren, W. F. et Berendsen, H. J. C. (1977). Algorithms for macromolecular dynamics and constraint dynamics. *Molecular Physics*, 34:1311–1327.

van Gunsteren, W. F. et Berendsen, H. J. C. (1990). Computer Simulation of Molecular Dynamics : Methodology, Applications, and Perspectives in Chemistry. *Angewandte Chemie International Edition in English*, 29(9):992–1023.

Varshney, A., Brooks, F. P. J. et Wright, W. V. (1994). Linearly scalable computation of smooth molecular, invited submission,. *IEEE Computer Graphics and Applications*.

Venter, J. C. et Celera Genomics (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.

Verlet, L. (1967). Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159(1):98–103.

Voet, D., Voet, J., Gaudemer, Y. et Dulieu, H. (1998). *Biochimie ($2^{ème}$ édition)*. De Boeck Université.

Vorobjev, Y. N. et Hermans, J. (1997). Sims : computation of a smooth invariant molecular surface. *Biophys J*, 73(2):722–32. 0006-3495 (Print)Journal ArticleResearch Support, U.S. Gov't, Non-P.H.S.Research Support, U.S. Gov't, P.H.S.

Wand, A. J. (2001). Dynamic activation of protein function : a view emerging from NMR spectroscopy. *Nat Struct Biol*, 8(11):926–931.

Wang, C., Schueler-Furman, O. et Baker, D. (2005). Improved side chain modeling for protein-protein docking. *Prot. Sci.*, 14:1328–1339.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. et Case, D. A. (2004). Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–1174.

Warner, D. R., Pisano, M. M., Roberts, E. A. et Greene, R. M. (2003). Identification of three novel smad binding proteins involved in cell polarity. *FEBS Lett*, 539(1-3):167–173.

Weiner, P. K. et Kollman, P. A. (1981). Amber : Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *Journal of Computational Chemistry*, 2(3):287–303.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. et Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784.

Westhof, E. et Fritsch, V. (2000). RNA folding : beyond Watson-Crick pairs. *Structure*, 8(3):R55–R65.

Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R. et Oschkinat, H. (2004). Quantification of pdz domain specificity, prediction of ligand affinity and rational design of super-binding peptides. *J Mol Biol*, 343(3):703–18. 0022-2836 (Print)Journal Article.

Wiehe, K., Pierce, B., Mintseris, J., Tong, W. W., Anderson, R., Chen, R. et Weng, Z. (2005). ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. *Proteins : Struct. Func. Bioinf.*, 60:207–213.

Wiehe, K., Pierce, B., Tong, W. W., Hwang, H., Mintseris, J. et Weng, Z. (2007). The performance of ZDOCK and ZRANK in rounds 6-11 of CAPRI. *Proteins*, 69(4):719–725.

Williamson, J. R. (2000). Induced fit in RNA-protein recognition. *Nat Struct Biol*, 7(10):834–837.

Winter, C., Henschel, A., Kim, W. K. et Schroeder, M. (2006). SCOPPI : a structural classification of protein-protein interfaces. *Nucleic Acids research*, 34:D310–D314.

Wodak, S. J. et Janin, J. (1978). Computer analysis of protein-protein interaction. *J. Mol. Biol.*, 124:323–342.

Wriggers, W. et Chacón, P. (2001). Using *situs* for the registration of protein structures with low-resolution bead models from X-ray solution scattering. *J. Appl. Cryst.*, 34:773–776.

Wu, G., Chen, Y. G., Ozdamar, B., Gyuricza, C. A., Chong, P. A., Wrana, J. L., Massagué, J. et Shi, Y. (2000). Structural basis of smad2 recognition by the smad anchor for receptor activation. *Science*, 287(5450):92–97.

Young, M. A., Gonfloni, S., Superti-Furga, G., Roux, B. et Kuriyan, J. (2001). Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-terminal tyrosine phosphorylation. *Cell*, 105(1):115–126.

Zacharias, M. (2003). Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Prot. Sci.*, 12:1271–1282.

Zauhar, R. J. (1995). SMART : a solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J Comput Aided Mol Des*, 9(2):149–159.

Zeev-Ben-Mordehai, T., Silman, I. et Sussman, J. L. (2003). Acetylcholinesterase in motion : visualizing conformational changes in crystal structures by a morphing procedure. *Biopolymers*, 68(3):395–406.

Zhang, C., Vasmatzis, G., Cornette, J. et DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.*, 267(3):707–726.

Zhang, Y. et Derynck, R. (1999). Regulation of Smad signalling by protein associations and signalling crosstalk. *Trends Cell Biol*, 9(7):274–279.

Zhou, H.-X. et Qin, S. (2007). Interaction-site prediction for protein complexes : a critical assessment. *Bioinformatics*, 23(17):2203–2209.

Résumé

Le *docking* protéine-protéine devient un outil incontournable pour répondre aux problématiques biologiques actuelles : approche basée sur l'assemblage des structures tridimensionnelles des protéines partenaires, elle peut permettre de guider la recherche expérimentale et de formuler des hypothèses quant à la nature et la fonction des complexes formés. Il reste cependant deux difficultés inhérentes aux méthodes de docking actuelles : 1) la majorité de ces méthodes ne considère pas les possibles déformations internes des protéines durant leur association. 2) Il n'est pas toujours simple de traduire les informations issues de la littérature ou d'expérimentations en contraintes intégrables aux programmes de docking.

Partant de ces conclusions, nous avons tenté de développer une approche permettant d'améliorer les programmes de docking existants. Pour cela nous nous sommes inspirés des méthodologies mises en place sur des cas concrets traités durant cette thèse. D'abord, à travers la création du complexe ERBIN PDZ/Smad3 MH2, nous avons pu tester l'utilité de la Dynamique Moléculaire en Solvant Explicite (DMSE) pour mettre en évidence des résidus importants pour l'interaction. Puis, nous avons étendu cette recherche en utilisant divers serveurs de docking puis la DMSE pour cibler un résultat consensus. Enfin, nous avons essayé le raffinage par DMSE sur une cible du challenge CAPRI et comparé les résultats avec des simulations courtes de Monte-Carlo.

Une autre partie de cette thèse portait sur le développement d'un nouvel outil de visualisation de la surface moléculaire. Ce programme, nommé MetaMol, permet de visualiser un nouveau type de surface moléculaire : la Skin Surface Moléculaire. La distribution des calculs à la fois sur le processeur de l'ordinateur (CPU) et sur ceux de la carte graphique (GPU) entraine une diminution des temps de calcul autorisant la visualisation, en temps réel, des déformations de la surface moléculaire.

Ainsi, le protocole serait divisé en 3 parties : d'abord l'utilisation de plusieurs programmes ou serveurs de docking pour mettre en évidence des solutions consensus. Puis, l'utilisation la DMSE pour raffiner ces résultats et mettre en évidence les résidus clés de l'interaction. Enfin, sur les modèles les plus prometteurs, laisser l'expert améliorer le modèle "à la main" en utilisant un programme de déformation de structure moléculaire comme, par exemple, le programme *SAMSON*. Ce type programme couplé à MetaMol permettrait de visualiser, en temps réel, les adaptations des surfaces moléculaires à l'interface des complexes.

Mots-clés: docking ; interactions protéiques ; interactions électrostatiques ; domaine PDZ d'ERBIN ; domaine MH2 de Smad3 ; Dynamique Moléculaire en Solvant explicite ; GPU ; Skin Surface Moléculaire ; déformations de surface ; MetaMol

Abstract

Protein-protein docking has become an extremely important challenge in biology : this approach, based on the three-dimensional structures of protein assemblies, can be used to guide experimental research and to formulate hypotheses about the nature and function of the resulting complexes. However, there remain two inherent difficulties in current docking methodologies : 1) most docking methods do not consider possible internal deformations of the proteins during their association ; 2) it is not always easy to translate information from the literature or from experiments into constraints suitable for use in protein docking algorithms.

Following these conclusions, we have developed an approach to improve existing docking programs. Firstly, through modelling the ERBIN PDZ / Smad3 MH2 complex, we have tested the utility of Molecular Dynamics with Explicit Solvent (MDSE) for elucidating the key residues in an interaction. We then extended this research by using several docking servers and the DMSE simulations to obtain a consensus result. Finally, we have explored the use of DMSE refinement on one of the targets from the CAPRI experiment and we have compared those results with those from short Monte-Carlo simulations.

Another aspect of this thesis concerns the development of a novel molecular surface visualisation tool. This program, named MetaMol, allows the visualisation of a new type of molecular surface : the Molecular Skin Surface.Distributing the surface calculation between a computer's central processing unit (CPU) and its graphics card (GPU) allows deformations of the molecular surface to be calculated and visualised in real time.

Thus the protocol is divided into three parts : first, several docking programs and servers are used to elucidate consensus solutions. Next, DMSE is used to refine these results and to obtain further evidence for key interaction residues. Finally, for the most promising models, an expert may improve the models "by hand" using a molecular structure deformation program such as, for example, Adaptive Molecular Dynamics. This program, coupled with MetaMol, would allow changes in the molecular surfaces at the interface of a complex to be visualised in real time.

Keywords: protein docking ; electrostatic interactions ; Erbin PDZ domain ; Smad3 MH2 domain ; Explicit Solvent Molecular Dynamic ; GPU ; Molecular Skin Surface ; MetaMol

---

Équipe ORPAILLEUR

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)

UMR 7503 - Campus Scientifique - BP 239 - 54506 Vandœuvre-les-Nancy Cedex