Prédiction structurale et ingénierie des assemblages macromoléculaires par bioinformatique

Hocine Madaoui

UNIVERSITÉ PARIS . DIDEROT
PARIS 7
ÉCOLE DOCTORALE INTER /// BIO

PRÉDICTION STRUCTURALE ET INGÉNIERIE DES ASSEMBLAGES MACROMOLÉCULAIRES PAR BIOINFORMATIQUE

THÈSE DE DOCTORAT

présentée et soutenue publiquement le 23 Novembre 2007

pour l'obtention du

DOCTORAT DE L'UNIVERSITÉ PARIS DIDEROT
ANALYSE DE GÉNOMES ET MODÉLISATION MOLÉCULAIRE

Par

Hocine MADAOUI

Composition du jury

Rapporteurs :	Dr. Anne POUPON Dr. Michael NILGES
Examinateurs :	Pr. Catherine ETCHEBEST , présidente Pr. Philippe MINARD Dr. Bernard JACQ
Directeurs de thèse :	Dr. Alain DESBOIS Dr. Raphaël GUEROIS

CEA/iBiTec-S/SB2SM/LBSR, URA 2096 CNRS CEA Saclay F91191 Gif-Sur-Yvette Cedex

Remerciements

Dans un premier temps, je remercie la Délégation Générale pour l'Armement, sans qui le projet de thèse que j'ai mené à bien n'aurait pas pu voir le jour. Je les remercie sincèrement de la confiance qu'ils m'ont accordée et pour avoir financé ce projet. Il se concrétise aujourd'hui par le développement d'une nouvelle méthodologie permettant de prédire rapidement la structure des complexes protéiques et ainsi d'exploiter les cartes d'interactions protéiques. En couplant notre approche aux stratégies développées par la société Hybrigenics, il sera désormais possible d'identifier les sites essentiels pour la fonction de nombreuses protéines cibles (agents pathogènes, protéines impliquées dans la réponse à divers stress) qui intéressent la Défense.

Je voudrais remercier Raphaël Guérois, qui m'a encadré tout au long de cette thèse. J'ai beaucoup appris en sa compagnie, et travailler avec lui a été un réel plaisir, car il m'a fait confiance, m'a donné l'opportunité de m'impliquer dans des projets qui me tenaient à cœur, et pour tout cela je l'en remercie. Je remercie également Alain Desbois, mon directeur de thèse, qui a toujours été disponible malgré ses fonctions de directeur de l'URA, ainsi que Pierre Legrain, pour m'avoir accueilli au sein de l'Institut, et pour les discussions scientifiques intéressantes que nous avons pu avoir.

Je tiens également à remercier tous mes rapporteurs et examinateurs, Michael Nilges, Anne Poupon, Philippe Minard, Bernard Jacq, et Catherine Etchebest, pour avoir accepté de juger mon travail, et pour leurs commentaires très instructifs. J'aimerais tout particulièrement remercier Catherine Etchebest, qui m'a suivi tout au long de ma formation universitaire, et qui m'a permis de suivre une formation de qualité, grâce à laquelle j'en suis arrivé là.

Merci à Grégoire Martin, Loïc Martin, Anne Descours, avec lesquels j'ai pu collaborer sur un projet ambitieux et motivant, le développement d'un vaccin contre le virus du sida. J'espère de tout cœur que ce projet, qui a été l'une des missions les plus motivantes au cours de ma thèse, aboutira.

Je tiens à remercier également tous les membres ou anciens membres du laboratoire de biologie structurale et de radiobiologie avec qui j'ai partagé tout ou partie de ma thèse. Tout d'abord Bernard Gilquin (l'un des « chefs » les plus sympathiques que je connaisse), Françoise Ochsenbein, Sophie Zinn, Jean-Baptiste Charbonnier, mais aussi Guillaume Letellier, Emmanuelle Becker, Béatrice Amigues, Morgane Agez, Karsten Seeger pour leur bonne humeur, et la bonne ambiance qu'ils ont entretenu au sein du laboratoire. Un grand merci également à tous les membres du laboratoire de Biochimie et de Génétique Moléculaire, et plus particulièrement à Marie-Claude Marsolier-Kergoat.

Je n'oublie pas tous mes amis rencontrés à l'université, en particulier au DEA AGM2, à savoir Anne, Mathieu, Thien-An, Julien, Thomas et Gaëlle, avec qui j'espère j'aurai l'occasion de travailler à nouveau.

Merci à tous mes potes, Julien, Eric, Pierrot ou encore Victor, pour avoir toujours cru en moi, et pour toute la publicité que vous pouvez me faire quand vous en avez l'occasion ;)

Un grand merci à mes potes de galère, Vincent et Thien-an, pour tous les bons moments passés au CEA, en particulier les grands moments de fous rire, les 'qualifs', ou encore les parties de NDS ;)

Concernant ma famille proche, je tiens à remercier tout particulièrement mes parents pour leurs conseils avisés, leur disponibilité, et pour m'avoir supporté pendant toutes ces années, ainsi que mes sœurs, et mon beau frère Mokrane, qui m'a pas mal aidé pour l'organisation du pot de thèse.

Enfin, je finirais par rendre hommage à ma petite nièce Sarah, disparue à l'âge de 1 an alors que je débutais ma thèse, et qui me manque énormément.

Liste des abréviations

AIR	=	Ambigous Interaction Restraint
ADN	=	Acide Désoxyribo Nucléique
ARN	=	Acide Ribo Nucléique
ASA	=	Accessible Surface Area
ATP	=	Adenosine triphosphate
Asf1	=	Anti-Silencing Function 1
CAPRI	=	Critical Assessment of PRedicted Interactions
CD4	=	Cluster of Differentiation 4
CSS	=	Conservation Shifting Sites
DSB	=	Double Strand Break
EEEF	=	Empirical Effective Energy Function
FTDock	=	Fourier Transform Dock
GO	=	Gene Ontology
HADDOCK	=	High Ambiguity Driven biomolecular DOCKing
Hsp90	=	Heat schock protein 90
ITC	=	Isothermal Titration Calorimetry
M	=	Molaire
MSA	=	Multiple Sequence Alignment
NBS	=	Nijmegen Breakage Syndrome
OMIM	=	Online Mendelian Inheritance in Man
PEEF	=	Physical Effective Energy Function
Pfam	=	Protein family database
RMN	=	Résonance Magnétique Nucléaire
RMSD	=	Root Mean Square Deviation (fr. écart quadratique moyen)
PDB	=	Protein Data Bank
PRM	=	Peptide Recognition Module
PSIMAP	=	Protein Structural Interactome Map
RFD	=	Random Forest Decision

ROC	=	Receiver Operating Characteristics
RPScore	=	Residue level Pair potential Score
RSS	=	Rate Shifting Sites
SCOP	=	Structural Classification of Proteins
SCOTCH	=	Surface Complementarity Trace in Complex History
SDS-PAGE	=	Sodium Dodécyl Sulfate-PolyAcrylamide Gel Electrophoresis
SEEF	=	Statistical Effective Energy Function
SPR	=	Surface Plasmon Resonance (fr. résonance plasmonique de surface)
UA	=	Unité Arbitraire
WT	=	Wild Type
gp120	=	glycoprotein 120
i2h	=	in silico two-hybrid
nM	=	nano-Molaire
p-Tyr	=	phospho-Tyrosine

Sommaire

Liste des abréviations **1**

Sommaire **3**

Chapitre 1 : Introduction Générale **9**

1.1. Les interactions protéine-protéine 15

1.1.1. Introduction 15

1.1.2. Nature et dynamique d'assemblage des complexes protéiques 15

1.1.3. Mise en évidence des interactions protéine-protéine 17

1.1.4. Les bases de données d'interactions protéine-protéine 21

1.2. Méthodes de prédiction des interactions protéines-protéines 26

1.2.1. Introduction 26

1.2.2. Les méthodes d'inférences génomiques 26

1.2.3. Les méthodes de co-évolution 29

1.2.4. Les méthodes de classification 32

1.3. Dynamique et logique moléculaire au sein des réseaux d'interaction protéine-protéine 32

1.3.1. Dynamique des réseaux contrôlés par l'expression des partenaires. 32

1.3.2. Dynamique des réseaux de signalisation et reconnaissance des motifs linéaires. 34

1.3.3. Intégration des signaux intra-cellulaires par des interactions de type domaine-domaine 37

1.3.4. Peut-on reproduire artificiellement la logique moléculaire ? 38

1.4. Caractérisation structurale et propriétés des complexes protéiques 40

1.4.1. Introduction 40

1.4.2. Propriétés géométriques des interfaces de complexes 40

1.4.3. Propriétés physico-chimiques des interfaces de complexes protéiques 41

1.4.4. Propriétés évolutives des interfaces de complexes protéiques 42

1.5. Méthodes bioinformatiques pour la prédiction des structures de complexes protéines-protéines 45

1.5.1. Introduction 45

1.5.2. Modélisation comparative des structures de complexes protéiques 45

1.5.3. Les techniques d'amarrage moléculaire ou docking 46

1.6. Design d'interactions protéines-protéines par des approches rationnelles 56

1.6.1. Introduction 56

1.6.2. Les méthodes automatiques et semi-automatiques de design 56
1.6.3. Design d'interfaces protéines-protéines 59

Chapitre 2 : Prédictions de sites fonctionnels par conservation différentielle 65

2.1. Introduction 67
2.2. Mise en évidence de régions fonctionnelles par analyse de conservation : la protéine Nbs1 68
2.2.1. La protéine NBS1 68
2.2.2. Mise en évidence de régions fonctionnelles au sein de la protéine Nbs1 par une analyse de conservation manuelle 69
2.2.3. Mise en évidence de régions fonctionnelles au sein de la protéine Nbs1 par une analyse de conservation automatique : la méthode rate4site 72
2.2.4. Conclusion 75
2.3. Prédiction des sites d'interactions protéiques des protéines kinases par analyse différentielle de conservation 77
2.3.1. Introduction 77
2.3.2. Principe général de la conservation différentielle 79
2.3.3. Prédiction des sites d'interactions intra-moléculaires par conservation différentielle : le cas des domaines régulateurs de kinases 80
2.4. L'analyse différentielle ou comment prédire des changements de spécificité fonctionnelle au sein des protéines : la protéine MsbA 83
2.4.1. La protéine P-gp et son homologue MsbA 83
2.4.2. Analyse par conservation différentielle 83
2.5. Discussion, Conclusions et Perspectives 86

Chapitre 3 : Analyse des propriétés évolutives des interfaces protéiques intra-moléculaires 89

3.1. Introduction 91
3.2. Constitution d'une base de données de complexes intra-moléculaires 92
3.3. Conservation des interfaces de complexes intra-moléculaires 92
3.4. Co-évolution des interfaces de complexes intra-moléculaires 94
3.4.1. Mutations corrélées et mutations compensatoires 94
3.4.2. Prédominance des événements de mutations compensatoires en périphérie des interfaces 96
3.5. Maintien de la complémentarité au sein des interfaces de complexes 98
3.5.1. Maintien de la complémentarité des paires de résidus en contact au sein des interfaces au cours de l'évolution 98
3.5.2. La prise en compte du contexte structural révèle un maintien de la complémentarité des paires de résidus en contact au cours de l'évolution 101

3.6. Conclusion 103

Chapitre 4 : Prédiction des assemblages macromoléculaires 105

4.1. Introduction 107
4.2. Constitution des bases de données de complexes protéiques 108
4.3. Evaluation des capacités discriminatives des méthodes évolutives : le cas des interactions intra-moléculaires 108
4.3.1. Evaluation des différentes approches évolutives 108
4.3.2. La complémentarité évolutive des interfaces protéiques est-elle un bon critère pour identifier les assemblages protéiques natifs ? 110
4.4. SCOTCH : une nouvelle méthode pour prédire la structure des assemblages protéiques 111
4.4.1. Apprentissage supervisé des caractéristiques évolutives des complexes : l'approche SCOTCH 112
4.4.2. Validation de l'approche SCOTCH pour le cas d'interactions intra-moléculaires 113
4.4.3. Validation de l'approche SCOTCH pour le cas d'interactions inter-moléculaires 116
4.4.4. Optimisation de l'approche SCOTCH par un score statistique 117
4.4.5. Capacités de filtrage de la méthode SCOTCH 120
4.4.6. Le programme RosettaDock 121
4.4.7. SCOTCH : une fonction de score efficace pour le docking protéine-protéine 123
4.5. Obtention de modèles d'assemblages à haute résolution 126
4.5.1. Introduction 126
4.5.2. Perturbations locales des structures de complexes 127
4.5.3. Profil énergétique des structures natives de complexes 132
4.5.4. Le profil énergétique permet-il d'identifier les structures natives de complexes parmi les structures générées à basse résolution ? 133
4.6. Applications de la méthode SCOTCH à des complexes étudiés au sein de l'équipe. 136
4.6.1. Application 1 : Modélisation structurale du complexe entre les domaines CS et ATPase de Hsp90 136
4.6.2. Application 2 : modélisation structurale de l'interaction Asf1 – Histone H3 141
4.7. Discussion 144

Chapitre 5 : De la structure à l'inhibition des complexes protéiques 147

5.1. Introduction 149
5.2. Inhibition de l'interaction Asf1-Histone H3-H4 149
5.2.1. Introduction 149
5.2.2. Optimisation d'un peptide inhibiteur de l'interaction Asf1-Histone H3-H4 151
5.2.3. Affinité du peptide H3-GAGG-H4 avec la protéine Asf1 156

5.3. Développements rationnels de vaccins 157
5.3.1. Introduction 157
5.3.2. Formation d'un complexe covalent entre la protéine gp120 et un peptide mimétique du récepteur CD4 159
5.3.3. Design d'une protéine chimère combinant gp120 et CD4M33 169
5.3.4. Conclusion 177

Chapitre 6 : Discussion – Perspectives 179

6.1. Mécanismes évolutifs et interfaces des complexes protéiques 181
6.2. De la co-évolution à la prédiction de structure des complexes protéiques. 183
6.3. Un double-hybride *in silico* basé sur le score SCOTCH ? 185
6.4. De la prédiction des interfaces protéiques à leur ingénierie 186

Chapitre 7 : Méthodes 189

7.1. Prédiction des sites d'interactions protéiques des protéines kinases par analyse différentielle de conservation 191
7.1.1. Création des alignements de séquences 191
7.1.2. Analyse de conservation 191
7.1.3. Analyse de conservation différentielle 192
7.2. Prédiction des assemblages macromoléculaires 192
7.2.1. Utilisation des fonctions de score du programme RosettaDock 192
7.2.2. Le programme SCOTCHer 193
7.2.3. Application 1 : Modélisation structurale des domaines CS et ATPase de Hsp90 195
7.2.4. Application 2 : modélisation structurale de l'interaction Asf1 – Histone H3 197
7.3. Inhibition de l'interaction Asf1-Histone H3-H4 198
7.4. Développements rationnels de vaccins 198
7.4.1. Formation d'un complexe covalent entre la protéine gp120 et le peptide **CD4M64** 198
7.4.2. Design d'une protéine chimère combinant gp120 et CD4M33 199

Chapitre 8 : Annexes 203

8.1. Paramètres des dynamiques moléculaires 205
8.2. Prédiction des assemblages macromoléculaires 205
8.2.1. Les quaternions 205
8.2.2. Le programme SCOTCHer 207
8.3. Inhibition de l'interaction Asf1-Histone H3-H4 208

Publications 209

Bibliographie..... 223

Chapitre 1 : Introduction Générale

Préambule

Depuis le début de la décennie, la caractérisation à grande échelle des cartes d'interactions protéiques a accompagné une avalanche de données biologiques allant du séquençage complet d'un nombre croissant d'organismes à la caractérisation de l'expression des gènes dans de multiples tissus. Cette masse d'informations conduit de façon irréversible à l'intrusion de l'informatique dans le domaine de la biologie. Les stratégies de comparaison, de classification, d'intégration et de prédiction à partir d'ensembles de données complexes et hétérogènes font désormais partie intégrante de la biologie post-génomique. Un des objectifs commun vers lequel tendent toutes ces activités est de caractériser, pour chaque protéine, les fonctions auxquelles elle est associée. Les interactions entre protéines constituent les nœuds essentiels au travers desquels l'information est transmise, l'architecture des machineries cellulaires est maintenue et l'adaptation des cellules à de multiples stimuli externes est orchestrée. Elles sont donc particulièrement fascinantes à étudier à l'échelle globale de l'organisation cellulaire.

En comparaison, les données sur la structure atomique des interactions protéine-protéine n'ont pas connu le même essor quantitatif. Pourtant, ces données sont des compléments essentiels à la connaissance des fonctions cellulaires des protéines car elles conditionnent souvent la compréhension des spécificités d'interactions et des mécanismes de régulation des machineries cellulaires. Un des objectifs de la bioinformatique structurale consiste à exploiter au mieux les masses de données issues des approches à grande échelle pour mieux prédire, par l'intermédiaire de leurs structures, les fonctions détaillées des protéines. L'objectif de ma thèse s'inscrit directement dans cette problématique et vise au développement de nouvelles approches permettant de prédire les caractéristiques fonctionnelles des protéines par la prédiction et la modélisation de leurs modes d'interactions.

Les masses de données collectées sur les réseaux d'interactions protéiques et l'analyse globale des propriétés évolutives de ces réseaux ont laissé plusieurs questions essentielles en suspens. A quel degré ces interactions exercent-elles des pressions de sélection négatives sur les séquences? Comment, pour une protéine donnée, les mutations au niveau d'un site d'interaction sont-elles tolérées par ses partenaires? Tous les types d'interactions, transitoires ou permanentes, sont-ils soumis aux mêmes types de contraintes évolutives ? Pour tenter de

répondre à ces questions, j'ai choisi de me focaliser sur les données structurales de complexes déjà disponibles. Elles constituent des systèmes très bien adaptés pour comprendre les mécanismes gouvernant les processus évolutifs au sein des interfaces. Le **chapitre 2** est consacré à une première stratégie d'analyse de ces pressions de sélection que j'ai développée au cours de mon stage de DEA et appliquée au cours de ma thèse.

D'un point de vue pratique, l'identification des mécanismes évolutifs assurant le maintien des assemblages protéiques doit fournir des contraintes structurales susceptibles de guider la prédiction des structures de complexes. Cette question est d'une actualité brûlante avec l'essor, dès la fin des années 1990s, des programmes de génomique structurale dans de nombreux pays (Chandonia and Brenner, 2006). En 2005, la moitié des structures déposées dans la PDB étaient issues des différents centres de génomique structurale. L'objectif de ces approches à grande échelle, outre de diminuer le coût d'obtention d'une structure, est également de couvrir le plus efficacement possible l'espace des repliements et d'attribuer une structure tridimensionnelle à chaque famille protéique. L'obtention d'une telle structure pour une seule protéine au sein d'une famille suffit à inférer des modèles structuraux détaillés pour l'ensemble des membres de cette famille. Comme mentionné précédemment, l'obtention de structures de complexes protéiques à haut débit reste encore du domaine de la prospective. Dans les prochaines années, les outils de prédiction de structures des assemblages protéiques sont donc appelés à jouer un rôle essentiel pour transformer la richesse des informations contenues dans les bases de données structurales en informations utiles pour analyser les mécanismes d'assemblages et étendre notre compréhension des réseaux d'interactions. Le **chapitre 3** et le **chapitre 4** sont dédiés à cette question et soulignent l'intérêt encore peu exploré des contraintes évolutives pour améliorer la prédiction systématique des structures de complexes. En particulier, je présenterai les potentialités d'une nouvelle approche, nommée SCOTCH (« *Surface Complementarity Trace in Complex History* »), pour guider la prédiction des structures de complexes protéiques.

Le domaine de la prédiction de structure tridimensionnelle, avant de devenir complètement mature, s'entraîne généralement sur des systèmes modèles, souvent idéalisés, qui permettent de mesurer les progrès effectués par la communauté scientifique. Les systèmes naturels font souvent preuve d'une complexité et d'une ingéniosité promptes à décourager la confrontation entre les approches prédictives de modélisation et la réalité expérimentale.

Une des voies qui permet néanmoins de susciter cette confrontation consiste à imiter les systèmes biologiques en utilisant les briques élémentaires que la nature nous fournit pour créer de nouvelles fonctions. Dans le cadre de la structure des protéines, ce champ de recherche est généralement appelé *design* rationnel de protéines. L'intérêt de ces approches dans le domaine des interactions protéine-protéine est double. D'une part, il constitue un test exigeant pour évaluer notre compréhension des processus de reconnaissance inter-moléculaire. D'autre part, il permet d'envisager de nouvelles stratégies pour le développement d'agents thérapeutiques qui visent non plus les sites catalytiques d'enzymes ou les sites actifs de transporteurs membranaires mais les interfaces de complexes protéiques. Le champ d'applications de l'inhibition des interfaces protéine-protéine est vaste et je développerai, au cours du **chapitre 5**, le cas de deux exemples pour lesquels j'ai cherché à concevoir des protéines ou peptides artificiels qui inhibent les interactions protéiques.

1.1. Les interactions protéine-protéine

1.1.1. Introduction

Récemment, les projets de séquençage menés sur différents génomes (Adams et al., 2000; Blattner et al., 1997; Lander et al., 2001), ont permis d'obtenir une meilleure caractérisation des différents gènes constitutifs d'un organisme. Désormais, l'objectif majeur est de mieux comprendre les fonctions de ces gènes, leurs régulations et leurs interactions, en s'intéressant plus particulièrement au produit de leur transcription, les ARNs, ou de leur traduction, les protéines.

L'ensemble des processus cellulaires est contrôlé par l'association et la dissociation des protéines. Ainsi, les interactions protéines-protéines sont au cœur de différentes fonctions essentielles, comme la réplication de l'ADN, l'organisation du cytosquelette, le contrôle du cycle cellulaire, ou encore la réponse à certains signaux extra-cellulaires. Leur caractérisation à l'échelle d'un organisme constitue une étape majeure pour une meilleure compréhension du rôle des protéines, et plus généralement du fonctionnement des cellules.

1.1.2. Nature et dynamique d'assemblage des complexes protéiques

En fonction de la durée de vie des complexes protéiques, les interactions protéines-protéines peuvent être regroupées en deux catégories : les interactions permanentes, et les interactions transitoires. À l'inverse des interactions permanentes, pour lesquelles les protéines n'existent que sous leur forme complexée, les complexes transitoires s'associent et se dissocient *in vivo*. Parmi ces interactions transitoires, nous pouvons distinguer les interactions transitoires faibles, où le complexe s'assemble et se dissocie en permanence, c'est le cas par exemple du complexe RhoA-RhoGAP impliqué dans les voies de la signalisation cellulaire, et les interaction transitoires fortes nécessitant l'intervention d'une autre molécule, comme les protéines G hétérotrimériques qui sont dissociées en sous unités Gα et Gβγ sous l'action du GTP et associés en présence de GDP (Nooren and Thornton, 2003).

Une des questions importante concerne le mécanisme d'assemblage des complexes permanents. En étudiant certains complexes macromoléculaires impliquant des interactions

permanentes entre protéines, tel que le ribosome, il apparaît que son assemblage est un processus dynamique. En effet, la biosynthèse des ribosomes est un processus ordonné qui implique la synthèse, la modification puis l'assemblage de différentes particules dans le noyau, avant leur exportation dans le cytoplasme (Tschochner and Hurt, 2003).

Un autre exemple de dynamique d'assemblage des protéines concerne les réponses cellulaires activées suite à un stress. À titre d'exemple, suite à des lésions dans l'ADN provoquées par des radiations ionisantes, telles que des cassures double brin, une cascade de signalisation se met en place faisant intervenir de multiples assemblages protéiques. Après la détection d'une cassure double brin, la transmission et l'amplification du signal vont être à l'origine d'une réponse adaptée. En fonction de la nature et de l'intensité des dommages causés à l'ADN, la cellule pourra déclencher un arrêt du cycle cellulaire, une réparation de l'ADN ou une mort cellulaire programmée par apoptose lorsque les dégâts sont trop importants (**Figure 1**).

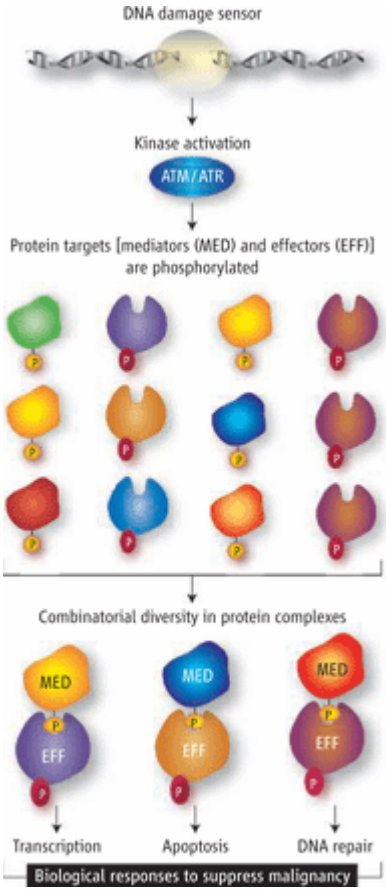


Figure 1. Exemple de réponses cellulaires activées suite à une cassure double brin (DSB). Figure extraite (Petrini, 2007).

Parmi les protéines intervenant dans la réponse cellulaire suite à une cassure double brin, les protéines kinases sont régulées de façon très stricte. Cette régulation est assurée par d'autres protéines par le biais d'interactions inter-moléculaires transitoires ou encore par des interactions intra-moléculaires pouvant impliquer différents domaines régulateurs (Rouse and Jackson, 2002). Ces systèmes de régulation peuvent donner lieu à des réponses extrêmement complexes qui mettent en œuvre de nombreux modules d'interactions protéine-protéine. Une étude récente, à grande échelle, a permis d'analyser les substrats phosphorylés par les deux kinases ATM et ATR, situées en amont de la détection des dommages de l'ADN (Matsuoka et al., 2007). L'analyse des résultats révèle que plus de 700 protéines humaines et murines se trouvent impliquées dans ces réponses cellulaires aux radiations ionisantes. Ce chiffre considérable est à mettre en rapport avec le nombre de substrats précédemment identifiés pour ATM et ATR, de l'ordre de 25. Ce changement d'échelle illustre bien la révolution conceptuelle que ces approches à grande échelle sont en train de créer dans notre appréhension des réseaux d'interaction protéine-protéine.

1.1.3. Mise en évidence des interactions protéine-protéine

Différentes méthodes peuvent être utilisées pour étudier ponctuellement certaines interactions protéines-protéines parmi lesquelles les techniques de chromatographie d'affinité, de co-immunoprécipitation ou encore de *phage-display* (Phizicky and Fields, 1995). De plus, des techniques d'analyses ont été proposées pour étendre l'identification des interactions protéiques à grande échelle, parmi lesquelles le système double-hybride de la levure (Fields and Song, 1989) (**Figure 2**)., ou encore l'analyse par spectrométrie de masse de complexes purifiés (Mann et al., 2001) (**Figure 3**).

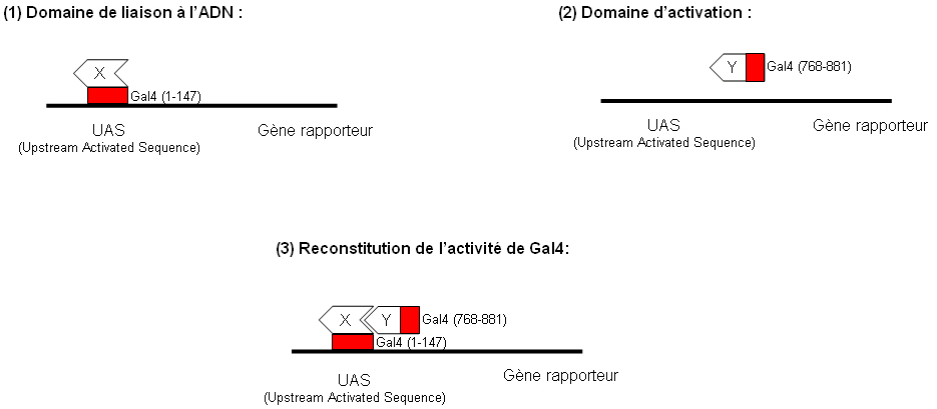


Figure 2. Principe du double hybride chez la levure. Une protéine « appât » X est fusionnée au domaine de liaison à l'ADN du facteur de transcription Gal4. Par ailleurs, une protéine « proie » Y est fusionnée au domaine activateur de ce même facteur de transcription. L'interaction entre les protéines X et Y va reconstituer le facteur de transcription Gal4, et stimuler l'expression d'un gène rapporteur sous contrôle du promoteur Gal4.

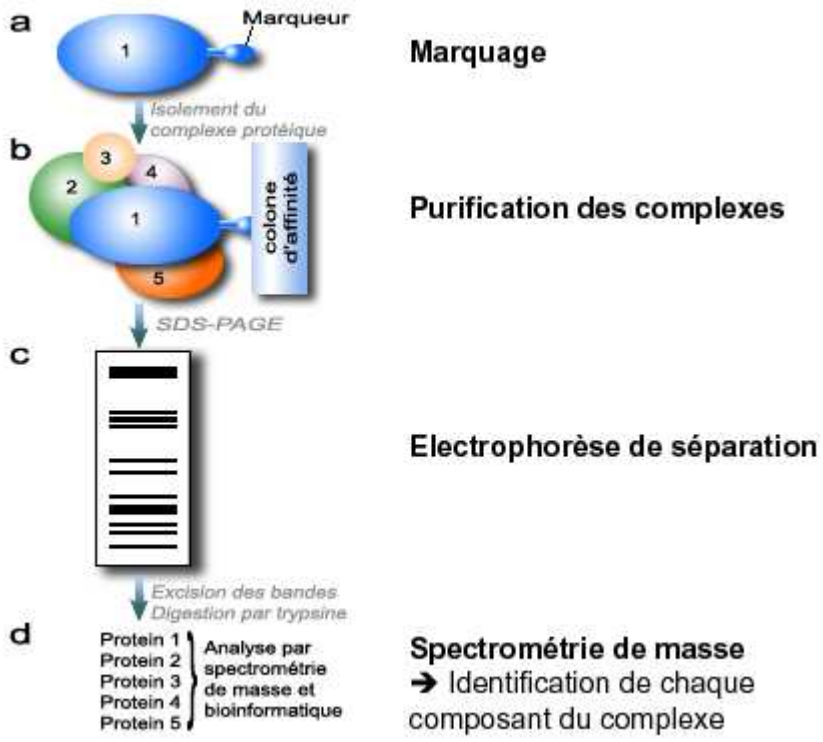


Figure 3. Analyse des complexes protéiques par spectrométrie de masse. (a) Marquage de la protéine d'intérêt. (b) La protéine cible est retenue avec ses partenaires protéiques sur une colonne d'affinitée. (c) Dissociation du complexe par électophorèse SDS-PAGE. (d) Identification des protéines impliquée dans le complexe par spectrométrie de masse. Figure adaptée de (Kumar and Snyder, 2002).

Identification des interactions protéines-protéines à grande échelle. Depuis quelques années, différents projets ont été menés pour caractériser les interactions protéines-protéines à grande échelle, et pour différents organismes. Les premières études, menées chez *S. cerevisiae* par des techniques de double hybride, ont ainsi permis l'identification de 957 interactions potentielles impliquant 1004 protéines (Uetz et al., 2000), puis 4549 interactions impliquant 3278 protéines (Ito et al., 2001). Ces études ont été suivies par d'autres expériences sur ce même organisme, réalisées cette fois-ci par des techniques de spectrométrie de masse (Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006).

D'autres cribles à grande échelle, basés sur la technique du double hybride, ont permis d'identifier un grand nombre d'interactions protéines-protéines chez d'autres organismes tels que *H. Pylori* (Rain et al., 2001), la mouche *D. melanogaster* (Giot et al., 2003), ou le vers *C. elegans* (Li et al., 2004). Plus récemment, les identifications d'interactions protéines-protéine ont pu être décrites chez l'homme, à partir de cribles à grande échelles utilisant la technique du double hybride (Lim et al., 2006; Rual et al., 2005; Stelzl et al., 2005).

À partir des données issues d'analyses à grande échelle, les premières cartes d'interactions protéines-protéines ou interactomes ont pu être proposées. Généralement, ces cartes d'interactions sont représentées sous la forme d'un graphe dont chaque nœud représente une protéine, et chaque arête reliant deux nœuds une interaction entre deux protéines. À la **Figure 4**, est présentée une partie de l'interactome humain impliquant 401 protéines et 911 interactions (Stelzl et al., 2005). En examinant cet interactome, nous pouvons constater que la majorité des protéines n'interagissent qu'avec un nombre limité de partenaires. Cependant, certaines protéines, appelées « *hubs* » sont au contraire impliquées dans plusieurs interactions protéines-protéines.

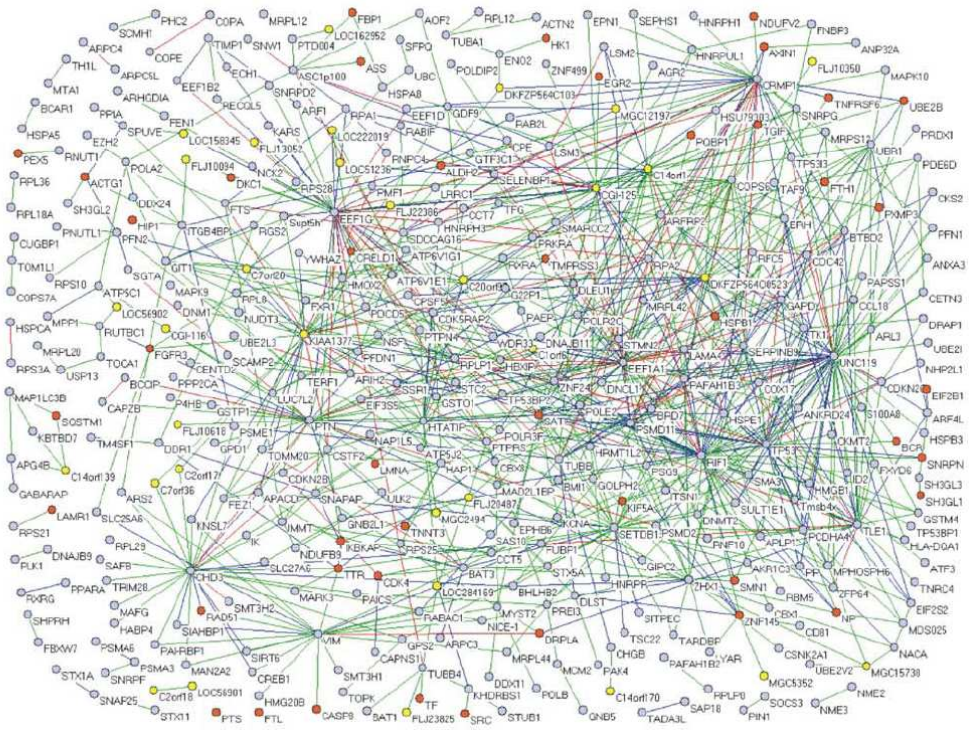


Figure 4. Carte d'interaction protéine-protéine impliquant 911 interactions caractérisées chez l'homme. Les cercles oranges représentent les protéines impliquées dans des maladies génétiques (d'après la base de données OMIM - http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM). Les cercles bleus représentent les protéines annotées par la banque « *Gene Ontology* »(GO) (Ashburner et al., 2000). Les protéines n'ayant pas été annotées selon les bases de données OMIM ou GO sont représentées par un cercle jaune. Les arêtes reliant les différents nœuds modélisent les interactions entre protéines. Ces arêtes sont colorées en fonction de leur score de confiance (score de confiance croissant du vert au violet en passant par le bleu et le rouge). Figure extraite de (Stelzl et al., 2005).

Dans l'exemple présenté, chaque protéine est colorée en fonction des annotations disponibles à partir de deux bases de données : (i) la base GO (« *Gene Ontology* ») structurée pour décrire les produits des gènes de différents organismes en termes de processus biologiques, fonctions moléculaires ou composants cellulaires (Ashburner et al., 2000) (ii) la banque OMIM (« *Online Mendelian Inheritance in Man* ») répertoriant les maladies génétiques humaines (http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM). Sur la base de ces annotations, 45 protéines impliquées dans 163 interactions protéiques sont associées à des désordres génétiques. Ces interactions protéines-protéines sont autant de cibles potentielles dont l'étude pourrait mener à une meilleure compréhension de ces maladies, ou au développement de certaines stratégies thérapeutiques.

1.1.4. Les bases de données d'interactions protéine-protéine

Pour collecter, classer et rendre accessible ces différentes données expérimentales issues de l'analyse des interactions protéines-protéines, différentes bases de données ont été développées. Le

Tableau 1 en mentionne 7 parmi les plus utilisées. Ces bases de données se distinguent par de nombreux critères : (i) sources des données collectées (extraites de la littérature ou des expériences à grande échelle), (ii) méthodes de nettoyage plus ou moins automatisées, (iii) enrichissement des informations dans certaines espèces etc... Par ailleurs, elles sont couplées à différents outils d'analyse de séquences, de domaines ou de préditions d'interaction extrêmement utiles pour accéder à un état des connaissances complet pour une protéine d'intérêt. Ces différences de traitements font que ces bases de données peuvent contenir des nombres très différents d'interactions. À titre d'exemple, la base de données BIND contient 188517 interactions, et la base de données DIP n'en possède que 56186.

Sigle	Base de donnée - description	Taille	URL
BioGRID	**Biological General Repository for Interaction Datasets**. Interactions entre biomolécules	198791	http://www.thebiogrid.org/
BIND	**Biomolecular Interaction Network Database** Interactions entre biomolécules	188517	http://www.bind.ca/
DIP	**Database of Interacting Proteins** Interactions entre protéines	56186	http://dip.doe-mbi.ucla.edu
IntAct	**Interaction Act** Interactions entre protéines	155533	http://www.ebi.ac.uk/intact/site/index.jsf
MINT	**Molecular INTeraction database** Interactions entre protéines	103808	http://mint.bio.uniroma2.it/mint/
HPRD	**The Human Protein Reference Database**. Interactions entre protéines	38167	http://www.hprd.org/
MIPS	**Munich Information center on Protein Sequences** Interactions entre protéines	15488	http://mips.gsf.de/genre/proj/mpact/

Tableau 1. Bases de données d'interactions protéine-protéine.

Les différentes expériences menées à grande échelle, en particulier chez *S. cerevisiae* ou chez *C. elegans*, ont largement contribué au développement de ces bases de données d'interactions protéiques. En 2003, la proportion de ce type de données dans la base de données BIND a été estimée à environ 80% (Salwinski and Eisenberg, 2003). Or ces données d'interactions proviennent pour la plupart d'expériences de double hybride, qui génèrent souvent un

grand nombre de faux positifs (Vidal and Legrain, 1999). Compte tenu du nombre de données à analyser, certaines méthodes automatiques ont donc été développées afin d'évaluer la validité biologique de ces interactions, en particulier, les méthodes EPR (Expression Profile Reliability) et PVM (Paralogous Verification Method) (Deane et al., 2002). À partir de ces méthodes d'évaluation, il a été estimé qu'environ 50% des 8000 interactions chez *S. cerevisiae* sont valides (Deane et al., 2002). Les interactions prédites comme valides ont été regroupées en un sous-ensemble représentant environ 30% des interactions de la base de données DIP, appelé CORE (Salwinski and Eisenberg, 2003). D'autres approches ont été développées afin d'enrichir le contenu des bases de données d'interactions protéiques. En particulier, certaines approches automatiques basées sur des techniques de fouilles de données (Donaldson et al., 2003; Marcotte et al., 2001), visent à extraire les articles scientifiques traitants des interactions protéiques caractérisées expérimentalement. Cependant, une analyse manuelle est nécessaire dans ce cas, avant d'intégrer ces données au sein des bases d'interactions protéiques.

Notons que cette question de la fiabilité des informations contenues dans les bases de données d'interactions est loin d'être tranchée et le pourcentage d'informations pertinentes contenu dans ces bases reste un sujet de débat. Pour le biologiste expérimental, les informations proposées, même à moitié fausses, contiennent souvent assez de nouveautés et de perles cachées pour susciter un intérêt pour sa recherche. La question se pose de façon plus critique lorsque ces données doivent servir à modéliser les comportements globaux des systèmes biologiques et à en extraire les lois générales du fonctionnement cellulaire. Un des paradoxes qui souligne la difficulté de cette question vient par exemple de l'analyse de la redondance effectuée sur les interactions entre protéines humaines disponibles dans certaines bases de données (**Tableau 2**). Non seulement le nombre d'interactions est très différent mais les taux de recouvrement sont également relativement faibles (excepté entre les bases MINT et IntAct) (Mathivanan et al., 2006). Une autre observation surprenante concerne l'analyse du groupe de Myers pour le développement de BioGRID. Ils ont effectué un travail de fouille de données considérable pour récupérer l'ensemble des données d'interactions protéine-protéine publiées dans la littérature pour *S. cerevisiae*, indépendamment des expériences à grande échelle. Pas moins de 33310 interactions ont été répertoriées par cette approche, ces valeurs approchant le nombre d'interactions obtenues avec les expériences à grande échelle. Néanmoins, entre ces deux sources de données, seules

14 % des interactions sont retrouvées en commun soulignant en partie les biais selon lesquels certaines protéines phares ont été sur analysées dans les laboratoires.

HPRD (36,617)								
BIND (6,621)	4,903							
DIP (1,067)	801	264						
MINT (11,367)	8,690	1463	379					
Reactome (5,960)	538	207	67	102				
IntAct (10,244)	8,031	1167	283	7,362	173			
MIPS (346)	307	294	28	65	14	43		
PDZ Base (101)	93	19	0	60	0	5	3	
	HPRD (36,617)	BIND (6,621)	DIP (1,067)	MINT (11,367)	Reactome (5,960)	IntAct (10,244)	MIPS (346)	PDZ Base (101)

Tableau 2. Analyse de la redondance des interactions entre protéines humaines entre les différentes bases de données d'interactions protéine-protéine.

Une des voies les plus prometteuses pour réduire le bruit inhérent aux bases de données d'interactions est probablement l'accumulation et la comparaison des cartes d'interactions issus de différents organismes mais également l'intégration avec d'autres réseaux générés à partir d'autres techniques à grande échelle (co-régulation, létaux synthétiques, CHIPonCHIP, etc …). Les problèmes conceptuels posés par l'intégration de ces données hétérogènes issues de multiples organismes font probablement partie des questions majeures auxquelles la bioinformatique aura à répondre dans les prochaines années (Sharan and Ideker, 2006).

Vers l'intégration des données structurales. Désormais la majorité des bases de données d'interactions protéine-protéine intègrent les données structurales concernant certains complexes protéiques connus, à partir de la PDB (Protein Databank). La Protein Data Bank est une base de données qui collecte les coordonnées atomiques de l'ensemble des structures tridimensionnelles expérimentales et des modèles théoriques des macromolécules biologiques (Bernstein et al., 1977) (**Tableau 3**). Dans sa version d'octobre 2007, la PDB recense 46557 structures protéiques avec en réalité environ 15000 chaînes non redondantes (pourcentage d'identité filtré à 70 %). Le décalage entre le nombre de séquences protéiques disponibles dans les bases de données de séquences et le nombre de structures disponibles dans la PDB s'explique par la difficulté de résoudre la structure des protéines et de leurs assemblages.

Sigle	Base de donnée - description	Taille	URL
PDB	**Protein Data Bank** Structure atomique des protéines et des complexes	46557	http://www.rcsb.org/pdb/
PSIMAP	**Protein Structural Interactome MAP** Structure des interactions entre domaines protéiques	1930	http://psibase.kobic.re.kr/
iPfam	**Protein Structural Interactome MAP** Structure des interactions entre domaines protéiques	3019	http://www.sanger.ac.uk/Software/Pfam/iPfam/
PQS	**Protein Quaternary Quaternary Structures** Structure quaternaire des protéines	48603	http://pqs.ebi.ac.uk/
ASEdb	**Alanine Scanning Energetics database** Interfaces protéiques (propriétés énergétiques)	3043	http://www.asedb.org

Tableau 3. Bases de données structurales de complexes protéine-protéine.

À partir des données structurales issues de la PDB, la base de données PSIMAP (« *Protein Structural Interactome MAP* ») a été développée dans le but de reconstruire une carte globale d'interactions qui décrive précisément les interactions entre domaines protéiques, aussi bien intra-moléculaires qu'inter-moléculaires. L'originalité de cette base de données consiste à considérer la définition SCOP (« *Structural Classification Of Proteins* ») des domaines protéiques, qui exploite des homologies structurales et fonctionnelles pour définir des familles ou superfamilles de domaines d'origine évolutive distincte. Ainsi, PSIMAP est la première base de données à proposer une vision globale des interactions entre domaines protéiques à l'échelle de la superfamille. À l'heure actuelle, la base de données PSIMAP répertorie 1930 interactions entre des superfamilles de la base de données de domaines structuraux SCOP. Une autre base de données particulièrement utilisée pour valider les prédictions d'interactions est la base iPfam qui rassemble l'ensemble des complexes entre domaines ayant été décrits dans la base Pfam (base de données de domaines construites sur la base des homologies entre séquences) (Bateman et al., 2002).

Les données déposées dans la PDB ne décrivent pas nécessairement de façon exhaustive le complexe étudié, et peuvent également inclure plusieurs copies d'une même molécule. Afin de reconnaître la présence de plusieurs copies pour une même molécule, et/ou la prise en compte de la symétrie des objets pour générer les coordonnées qui décrivent complètement

une macromolécule résolue par cristallographie aux rayons X, une méthode automatique a été développée et permet de répertorier les structures quaternaires des protéines au niveau d'un serveur : le serveur PQS (« *Protein Quaternary Server* »). Actuellement, le serveur PQS regroupe environ 48603 entrées, dont 17729 monomères, 16702 dimères et 14172 complexes d'ordre supérieur. D'autres bases de données sont également issues des données structurales d'interactions protéiques, et répertorient les propriétés énergétiques des interfaces.

1.2. Méthodes de prédiction des interactions protéines-protéines

1.2.1. Introduction

Face aux masses de données générées par les projets de séquençage de différents organismes, un des défis les plus importants de l'ère post-génomique consiste à exploiter ces informations, afin de prédire des interactions fonctionnelles entre protéines. Une approche globale consiste à utiliser des méthodes bioinformatiques pour comparer les protéomes de différents organismes et prédire des interactions fonctionnelles directes ou indirectes entre protéines.

Dans cette section, je présenterai différentes approches bioinformatiques destinées à prédire les interactions entre protéines : (i) les méthodes d'inférence génomique, (ii) les méthodes basées sur la co-évolution des protéines, (iii) les méthodes de classification.

1.2.2. Les méthodes d'inférences génomiques

Inférence d'interactions protéine-protéine : la notion d'interologues. En supposant que des protéines interagissant dans un organisme donné ont coévolué de façon à ce que leurs orthologues respectifs interagissent de la même façon au sein d'autres organismes, Walhout et collaborateurs ont introduit en 2000 la notion d'interaction conservée ou « interologues » (Walhout et al., 2000). La définition est la suivante : « en considérant deux protéines A et B connues pour interagir dans un organisme, si on trouve des orthologues A' et B' dans un autre organisme, A-B et A'-B' sont considérés comme des intérologues ». Chez *C. elegans*, Walhout et collaborateurs ont montré qu'environ 7% des interactions possédaient des interologues décrits dans la littérature.

Sur la base de cette considération, une étude réalisée par Matthews et collaborateurs, s'est focalisée sur une recherche d'interologues à grande échelle pour inférer des interactions protéine-protéine connues (Matthews et al., 2001). Les auteurs ont voulu évaluer à quel point une carte d'interactions protéiques décrite chez *S. cerevisiae* pouvait permettre la prédiction d'interactions par inférence chez *C. elegans*. Les auteurs ont estimé entre 16% et 31% le pourcentage minimum d'interactions vraies prédites par ce genre d'approche, ce qui a été confirmé par une étude plus récente (Yu et al., 2004).

Evènements de fusions de gènes : la méthode « *Rosetta Stone* ». Certaines études ont montré qu'on retrouvait des homologues entre certaines paires de protéines en interaction fusionnées en une seule protéine (Berger et al., 1996; Enright et al., 1999; Marcotte et al., 1999). Il a donc été proposé que deux protéines A et B dans un même organisme interagissent si leurs homologues se retrouvent exprimés au sein d'une même protéine dans une autre espèce (**Figure 5**).

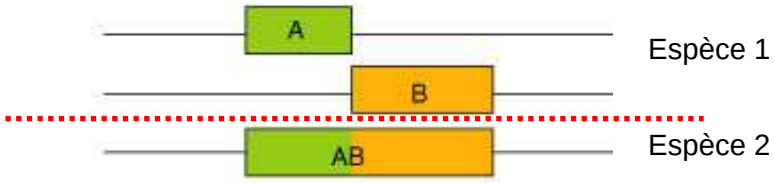


Figure 5. La méthode "*Rosetta Stone* ". Les protéines A et B sont exprimées séparément dans un organisme. Dans un autre organisme, les protéines A et B sont fusionnées en une seule protéine. L'existence de cette protéine de fusion permet de prédire l'interaction entre les protéines A et B. Figure adaptée de (Bowers et al., 2004).

Appliquée au génome d'*E. Coli*, cette méthode a permis de sélectionner 6809 paires de protéines potentiellement en interaction. Parmi les 3950 paires dont la fonction a été caractérisée, 68% des paires sont liées au niveau de leur fonction (Marcotte et al., 1999). De plus, les auteurs ont montré que 6.4% des interactions caractérisées sur le plan expérimental au niveau de la base de données DIP sont prédites par l'approche *Rosetta Stone*.

Conservation du voisinage physique des gènes. Après une étude de l'organisation des génomes de *H. influenza* et d'*E. Coli*, Tamames et collaborateurs ont conclut que les gènes liés fonctionnellement dans ces espèces étaient le plus souvent regroupés au sein de certaines régions proches particulièrement conservées, ce qui supposerait un lien direct entre une interaction physique et une conservation de l'ordre des gènes (Tamames et al., 1997). À titre d'exemple, dans les génomes procaryotes, certains gènes liés fonctionnellement, et codant des protéines qui interagissent potentiellement sont regroupées au sein de ce qu'on appelle des opérons.

En 1998, Dandekar et collaborateurs ont proposé que cette conservation de la proximité des gènes fonctionnellement liés, permettrait de prédire des interactions physiques entre

protéines (Dandekar et al., 1998). Le principe consiste à prédire une interaction entre les gènes physiquement proches dans les génomes de plusieurs organismes (**Figure 6**). Appliqué à l'étude de certains génomes bactériens et archaebactériens, les auteurs ont montré que 75% des paires de protéines prédites en interaction par cette approche interagissent physiquement. Des résultats similaires ont été également reportés dans d'autres génomes eucaryotes : *S. cerevisiae* et *C. elegans* (Teichmann and Babu, 2002).

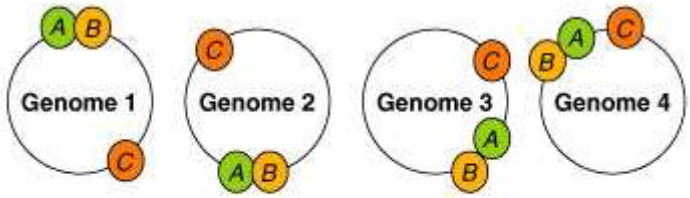


Figure 6. La méthode "*Gene Neighbor*". La méthode « *Gene Neighbor* » vise à identifier des paires de protéines proches dans les génomes de plusieurs organismes. Dans cet exemple, la paire A-B sera prédite en interaction par l'approche, contrairement à la paire A-C. Figure extraite de (Bowers et al., 2004).

Les profils phylogénétiques. En 1999, Pellegrini et collaborateurs ont introduit la notion de profils phylogénétiques pour identifier les protéines reliées à une même fonction (Pellegrini et al., 1999). Ce concept se base sur l'hypothèse qu'au cours de l'évolution, les protéines liées à une même fonction ont tendance à être préservées ou éliminées de façon corrélée au sein d'une espèce. Les auteurs décrivent cette propriété de co-évolution, en caractérisant chaque protéine par un « profil phylogénétique ». En considérant n espèces, le profil phylogénétique d'une protéine donnée est alors représenté sous forme d'un vecteur à n entrées, chaque entrée prenant comme valeur 1 ou 0 en fonction de la présence ou de l'absence d'un homologue de la protéine pour une espèce donnée (**Figure 7**).

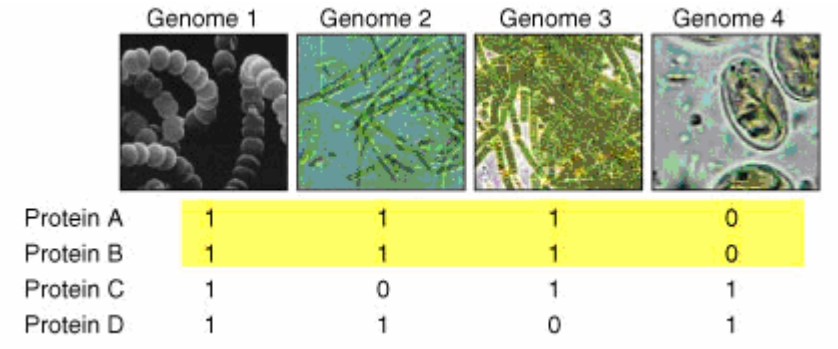


Figure 7. Méthode de profils phylogénétiques. En considérant 4 organismes différents, les protéines A, B, C et D sont recherchées au sein des ces espèces. Pour chaque protéine, un profil est construit, représenté sous la forme d'un vecteur dont chaque entrée prend comme valeur 1 ou 0 en fonction de la présence ou de l'absence d'un homologue dans un génome donné. Les protéines présentant un profil similaire (comme les protéines A et B dans cet exemple), sont prédites comme étant relié à la même fonction, et potentiellement en interaction. Figure extraite de (Bowers et al., 2004).

En comparant les profils phylogénétiques de toutes les paires d'un sous ensemble de 1231 protéines chez *E. coli* et 1131 protéines chez *S. cerevisiae* dont la fonction est connue, Date et Marcotte ont montré que la similarité des profils génétiques étaient hautement corrélée à une similarité fonctionnelle des protéines (Date and Marcotte, 2003).

1.2.3. Les méthodes de co-évolution

Similarité des arbres phylogénétiques. Différentes études ont montré que la similitude entre les arbres phylogénétiques de deux protéines en interaction était un marqueur de leur co-évolution (Fryxell, 1996). À titre d'exemple, un coefficient de corrélation de 0.79 a été reporté entre les arbres phylogénétiques des deux domaines protéiques N- et C-terminaux de la protéine kinase PGK (Goh et al., 2000). Des coefficients de corrélations très proches ont également été publiés par l'équipe de Valencia sur une base de données de 13 interactions domaine-domaine intra-moléculaires, ce qui suggère une pression évolutive et une co-adaptation très forte entre des domaines structuraux d'une même protéine (Pazos and Valencia, 2001). Les auteurs ont quantifié la similarité des arbres phylogénétiques des interactants en calculant le coefficient de corrélation linéaire entre les matrices de distances utilisées pour construire les arbres (**Figure 8**).

Par l'analyse de différentes bases de données d'interactions, les auteurs de cette étude ont observé le même type de signal de co-évolution entre des interactions de types intra-

moléculaires ou inter-moléculaires, et proposent un coefficient de corrélation seuil de 0.8 comme un bon critère pour la prédiction d'interactions protéine-protéine. Le potentiel de cette approche pour la prédiction d'interactions protéine-protéine a été confirmé par une étude portant sur six familles d'interactions ligand-récepteur (Goh and Cohen, 2002).

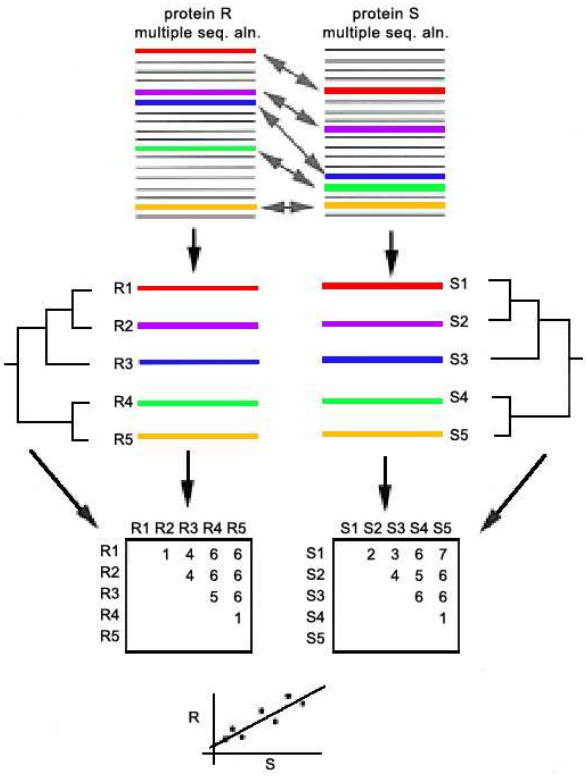


Figure 8. La méthode « *Mirror Tree* ». Les alignements multiples de séquences homologues à deux protéines R et S dont ont veut prédire l'interaction sont réduits à un ensemble commun aux deux organismes. Ces alignements réduits sont ensuite utilisés pour construire des matrices de distances entre chacune des séquences. Puis un coefficient de corrélation est calculé entre les matrices de distances correpondant aux protéines R et S. Figure adaptée de (Pazos and Valencia, 2001).

En 2005, une mise à jour de l'approche, consistant à intégrer explicitement l'histoire évolutive des espèces pour prendre en compte les similarités entre l'arbre phylogénétique des protéines et des espèces (Pazos et al., 2005), et plus récemment l'intégration de certains algorithmes basés sur des machines à vecteur de support (SVM) (Craig and Liao, 2007) ont considérablement amélioré les qualités prédictives de la méthode.

La méthode double hybride *in-silico* (i2h). La présence de mutations compensatoires au sein de complexes protéiques est un phénomène d'adaptation révélateur d'une trace de leur

co-évolution. Mises en évidence en dans les années 90 au sein de protéines globulaires (Gobel et al., 1994), elles permettraient de conserver la structure des protéines malgré leur taux d'évolution important, ce qui expliquerait que les structures des protéines puissent être plus conservées que leur séquences (Chothia and Lesk, 1986).

Le principe fondamental repose sur l'hypothèse qu'une mutation survenant à une position de la séquence d'une protéine, doit être compensée par une seconde mutation concernant une position voisine au niveau structural. Sur la base de ce principe, l'équipe de A. Valencia a développé en 2002 une nouvelle méthode basée sur la détection de mutations corrélées pour la prédiction *in silico* des interactions protéine-protéine : la méthode « *i2h* » (double hybride *in silico*) (Pazos and Valencia, 2002). L'objectif de cette méthode est non seulement de prédire les interactions physiques entre protéines, mais aussi d'identifier les régions impliquées dans ces interactions. Le principe de la méthode est illustré **Figure 9**. Les auteurs de cette étude ont définit un seuil empirique de 2.0 comme étant un bon indicateur d'une interaction physique entre deux protéines.

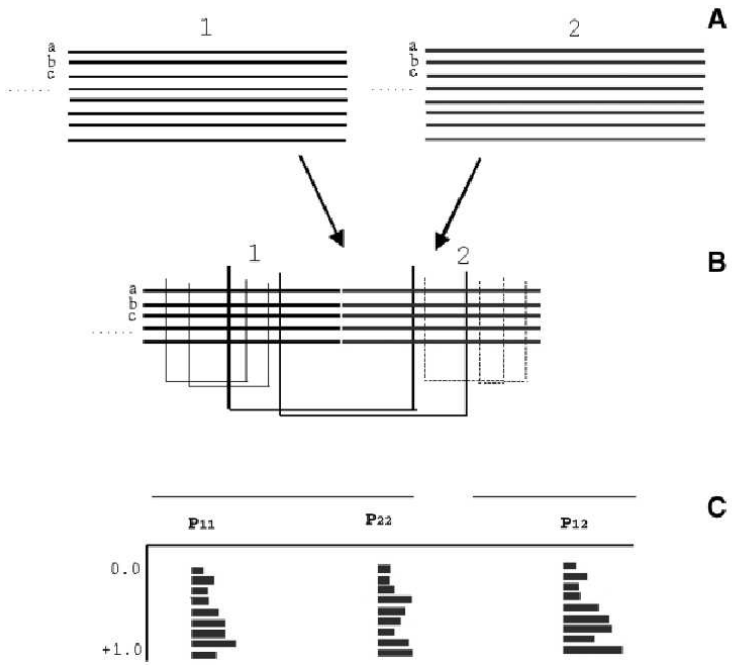


Figure 9. La méthode double hybride in-silico (*i2h*). (A) Alignements de séquences de deux protéines 1 et 2. (B) Concaténation des deux alignements de séquences, et calcul des coefficients de corrélation pour chaque paire de résidus au sein de la protéine 1 (P_{11}), de la protéine 2 (P_{22}), ou entre les protéines 1 et 2 (P_{12}). (C) À partir des distributions des valeurs de corrélation P_{11}, P_{22} et P_{12}, un indice d'interaction est calculé. Figure extraite de (Pazos and Valencia, 2002).

1.2.4. Les méthodes de classification

Les méthodes de classification se basent sur différentes sources de données pour réaliser un apprentissage des différentes caractéristiques qui distinguent des paires de protéines en interaction ou non. Dans le cas de la prédiction des interactions protéine-protéine, la comparaison de différentes méthodes de classification a montré que l'approche la plus efficace est la méthode « *Random Forest Decision* » (RFD) (Qi et al., 2006).

Le principe des méthodes RFD est la construction d'arbres de décision basés sur la composition en domaines de protéines interagissant ou n'interagissant pas entre elles (Chen and Liu, 2005). À partir d'un ensemble de données regroupant des paires de protéines connues pour interagir, la méthode construit un arbre de décision qui traduit la meilleure discrimination entre des protéines qui interagissent ou n'interagissent pas entre elles. En visitant l'arbre de décision, il est alors possible de prédire si deux protéines P_1 et P_2 interagissent ou non.

1.3. Dynamique et logique moléculaire au sein des réseaux d'interaction protéine-protéine

1.3.1. Dynamique des réseaux contrôlés par l'expression des partenaires.

Les cartes d'interactions protéine-protéine offrent une perception relativement statique des processus cellulaires. Pourtant, comme nous l'avons mentionné précédemment, les procesus d'assemblages s'avèrent souvent très dynamiques aussi bien dans la formation des machineries cellulaires que pour l'activation des complexes de signalisation. Un travail remarquable réalisé par les équipes de Bork et Brunak illustre particulièrement bien le caractère dynamique de ces réseaux d'interactions (de Lichtenberg et al., 2005).

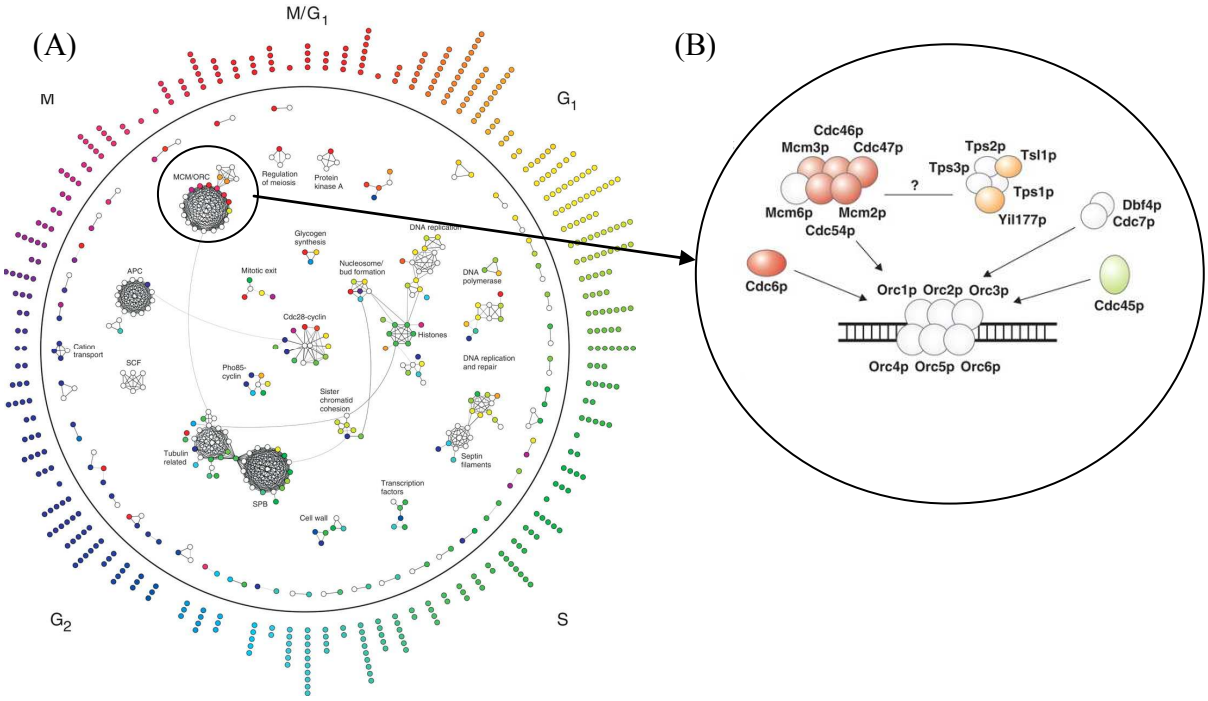


Figure 10. (A) Evolution temporelle du réseau d'interactions protéiques de *S. Cerevisiae* au cours du cycle cellulaire. Les protéines sont indiquées par des cercles. Les protéines du cycle impliquées dans des complexes sont placées à l'intérieur du cercle central, les autres à l'extérieur. Dans le cas des protéines « dynamiques », le moment d'expression maximal est indiqué par le code couleur correspondant à chaque phase du cycle. Les protéines « statiques » sont indiquées en blanc. (B) Zoom sur le schéma d'assemblage d'un des complexes, le complexe de pré-réplication. Ce complexe comprend 6 protéines statiques (Orc1p à Orc6p) liées aux origines de réplication. Un sous complexe de 6 protéines MCM est recruté sur les ORC en phase G1 par un mécanisme dépendant de Cdc6. Le recrutement final de la machinerie de réplication est dépendent de Cdc45 exprimée au tout début de la phase S. Figure extraite de (de Lichtenberg et al., 2005).

Les auteurs de cette étude ont combiné les cartes d'interactions protéiques et les données d'expression des protéines de *S. cerevisiae* impliquées dans le cycle cellulaire (**Figure 10-A**). Ils ont ainsi pu mettre en évidence deux classes de protéines, une « statique » pour laquelle l'expression est stable au cours du cycle et l'autre « dynamique » qui présente des pics d'expression. De façon remarquable, il apparaît fréquemment que le pic d'expression d'un des composants d'une machinerie cellulaire coïncide avec le moment précis où le complexe doit s'activer. L'exemple **Figure 10-B** illustre le cas la machinerie de pré-réplication pour lequel la protéine Cdc45 joue le rôle de déclencheur en phase S d'un complexe déjà préassemblé dès la phase G1. Ce contrôle temporel de l'assemblage des complexes par le contrôle de l'expression d'un de leur composant apparaît comme une propriété générale dans l'organisation des cellules eucaryotes. Néanmoins, de façon surprenante, le facteur limitant n'est pas toujours le même d'une espèce à l'autre. La **Figure 11** illustre cette observation pour le complexe de pré-réplication dans une étude à grande échelle comparant

les propriétés temporelles d'assemblage des complexes du cycle cellulaire dans trois organismes, *H. sapiens*, *S. cerevisiae* et *S. pombe* (Jensen et al., 2006). Pour être vraiment effectif ce contrôle transcriptionnel doit s'accompagner d'un contrôle au niveau de la dégradation des protéines. Ce dernier exemple indique qu'il doit exister des mécanismes de co-évolution entre les processus de régulation transcriptionnelle et de dégradation protéique afin de maintenir un contrôle de l'assemblage des complexes.

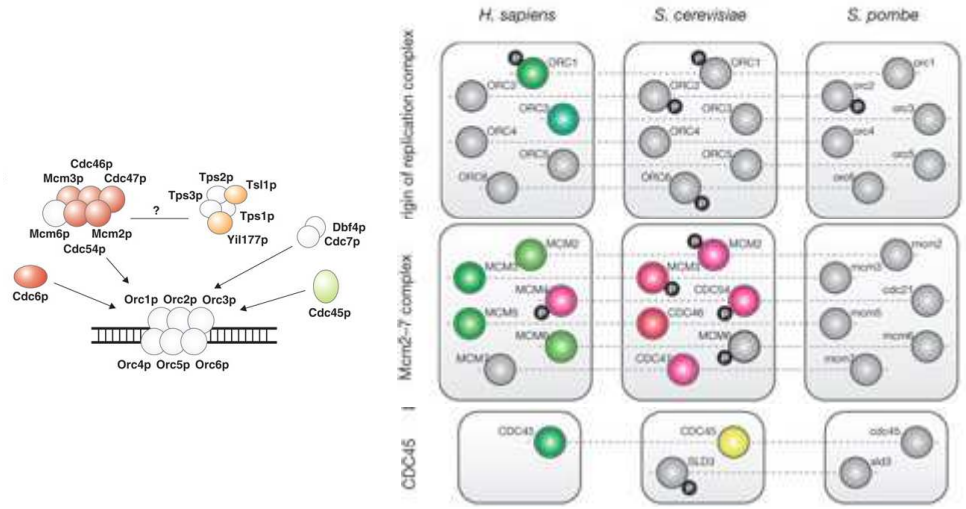


Figure 11. L'assemblage des complexes protéiques est contrôlé par différentes sous-unités dans chaque organisme. Exemple du complexe MCM/ORC présenté **Figure 10**, et comparaison des pics d'expressions pour les différentes sous-unités au cours du cycle par le même code couleur que dans la **Figure 10-A**.

1.3.2. Dynamique des réseaux de signalisation et reconnaissance des motifs linéaires.

Une des clés pour appréhender la plasticité des réseaux d'interactions protéiques au cours de l'évolution se situe probablement dans l'analyse des mécanismes de communication entre machineries cellulaires au niveau des mécanismes de transduction des signaux. Au sein de la cellule, la transduction des signaux est souvent médiée par des interactions entre des protéines modulaires. Il apparaît que certaines de ces interactions sont régulées par des domaines spécialisés, capables d'interagir avec des protéines, des phospholipides, des petites molécules ou des acides nucléiques. Il est maintenant bien établi que de tels domaines médiateurs d'interactions, appelés également domaines régulateurs, jouent un rôle

fondamental dans l'organisation et la coordination des réponses cellulaires suite à différents stimuli (Pawson and Nash, 2003).

Une des propriétés remarquable d'une majorité de ces domaines médiateurs d'interactions est de reconnaître de courts fragments protéiques (ces domaines sont appelés des« *Peptide Recognition Modules* » (PRMs)), ou encore de former des assemblages moléculaires de type domaine-domaine. Parmi les PRMs impliqués dans les voies de transduction du signal, plusieurs d'entre eux sont capables de reconnaître certains motifs peptidiques bien précis (**Figure 12**).

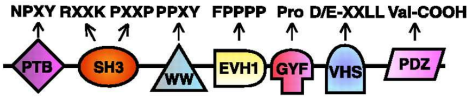


Figure 12. PRMs capables de reconnaître des motifs peptidiques. Un sous-ensemble de ces domaines régulateur est représenté. Figure extraite de (Pawson and Nash, 2003).

D'autres PRMs ayant également la faculté de reconnaître de courts motifs ancrés au sein d'une protéine cible, n'interagissent avec leur peptide cible que lorsque celui-ci a acquis une modification post-traductionnelle (**Figure 13**). Ce type de domaine, tel que le domaine SH2, possède souvent un site de liaison conservé, ainsi qu'une région plus variable responsable de la sélectivité entre différents peptides possédant une même modification post-traductionnelle.

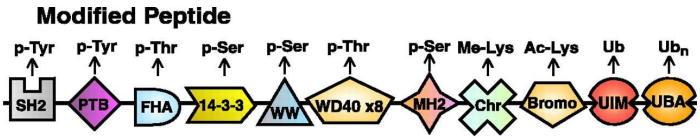


Figure 13. PRMs capables de reconnaître des motifs peptidiques modifiés. Un sous-ensemble de ces domaines régulateur est représenté. Figure extraite de (Pawson and Nash, 2003).

Plusieurs modes d'interactions dépendants de modifications post-traductionnelles ont été caractérisés. Ainsi, les interactions impliquant les PRMs peuvent être inductibles, coopératives, séquentielles, antagonistes, ou encore impliquer des mécanismes de régulation de type intra-moléculaires (Seet et al., 2006). Dans ce dernier cas, les protéines sont auto-

régulées par certains domaines régulateurs. En effet, de nombreuses protéines modulaires, qui utilisent leurs domaines régulateurs pour recruter leurs partenaires, régulent de surcroît leur activité par des interactions intra-moléculaires avec ces domaines, on parle alors d'allostérie modulaire (Dueber et al., 2004).

Dans le cas d'une allostérie modulaire, les protéines sont composées d'un domaine catalytique ainsi que d'autres domaines régulateurs dont le rôle sera d'inhiber l'activité de la protéine soit en bloquant l'accès au site actif du domaine catalytique (allostérie stérique), soit en contraignant la structure du domaine catalytique (allostérie conformationnelle). Le passage d'une forme inactive à active de la protéine se fait alors par un jeu de compétition, faisant par exemple intervenir d'autres ligands. Ces propriétés sont désormais bien documentées dans le cas des protéines kinases et phosphatases (Huse and Kuriyan, 2002) mais d'autres protéines telles que les nucléases, acétyl ou méthyl transférases sont susceptibles de fonctionner de même.

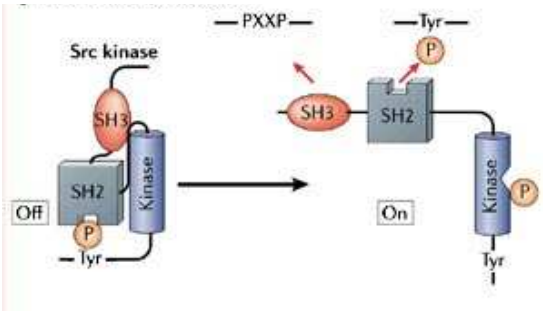


Figure 14. Régulation de la protéine kinase c-Src. La forme inactivée de la protéine est stabilisée par une interaction entre le domaine SH2 de la protéine, et une tyrosine phosphorylée (p-Tyr) présente au niveau de la queue C-terminale de la protéine. Cette conformation est également stabilisée par des interactions intra-moléculaires entre le domaine SH2 et le domaine kinase, ainsi qu'une interaction entre la boucle séparant les domaines SH2-kinase et le domaine SH3. La protéine bascule vers une forme activée par une déphosphorylation de la queue C-terminale de la protéine, ou par la présence de ligands activateurs (tels que des peptides riches en prolines, ou contenant des p-Tyr). Figure adaptées de (Seet et al., 2006).

La **Figure 14** décrit le cas de la protéine kinase c-Src illustrant parfaitement un cas d'allostérie conformationnelle. Dans cet exemple, la reconnaissance intra-moléculaire d'un motif phosphorylé par le domaine régulateur SH2 va contribuer à l'inactivation de la protéine. La dissociation de cette interaction par une déphosphorylation de ce motif, ou par la présentation d'autres partenaires capables de se lier à ce domaine régulateur, va activer la fonction catalytique de la protéine. C'est ce type d'interactions intra-moléculaires qui permet de coupler l'activité d'une protéine à la présence d'un ou plusieurs partenaires activateurs.

Elles constituent une des bases moléculaires qui régissent le fonctionnement des réseaux d'interactions.

1.3.3. Intégration des signaux intra-cellulaires par des interactions de type domaine-domaine

Certains cas d'interactions domaine-domaine impliquant des domaines régulateurs spécialisés, ont également été décrits dans la littérature, en particulier dans les voies de signalisation de l'apoptose ou de l'inflammation (**Figure 15**).

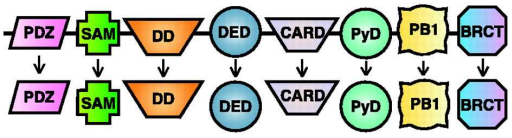


Figure 15. Domaines médiateurs d'interaction impliqués dans les voies de transduction du signal. Un sous-ensemble de ces domaines régulateur est représenté. Figure extraite de (Pawson and Nash, 2003).

C'est le cas des domaines CARD dont l'interaction est requise pour la dimérization et l'activation des caspases (**Figure 16**) (Hofmann, 1999; Martinon et al., 2002).

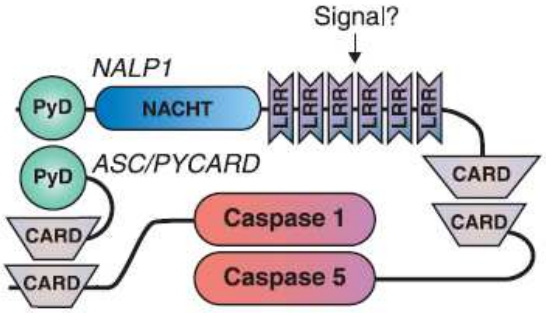


Figure 16. Activation des caspases par des interactions de type domaine-domaine. L'activation de NALP1 va provoquer la formation d'un complexe par des interactions de type domaine-domaine impliquant les domaines CARD, qui induit la dimérization et l'activation des caspases. Figure extraite de (Pawson and Nash, 2003).

Un autre exemple particulièrement intéressant concerne les domaines SAM de la protéine PcG chez la drosophile. Ces domaines ont la propriété de s'associer pour former une structure polymérique en hélice, qui apparaît essentielle à la régulation de la chromatine au

cours du développement de la drosophile. Dans ce cas, les interactions entre ces modules ne vont non pas impliquer de courts motifs peptidiques, comme c'est le cas des PRMs, mais des motifs structurés dont l'interaction est stabilisée par un réseau de liaisons hydrogènes et de ponts salins (Kim et al., 2002).

Dans certains cas, la classification de certains domaines comme liant des motifs peptidiques ou interagissant avec d'autres domaines structurés n'est pas claire. C'est le cas des domaines PDZ reconnaissant de courts motifs peptidiques d'environ 4 résidus à l'extrémité C-terminale de leurs partenaires, mais également capables de reconnaître d'autres motifs PDZ au sein de protéines partenaires (Hillier et al., 1999; Songyang et al., 1997).

1.3.4. Peut-on reproduire artificiellement la logique moléculaire ?

Les réponses cellulaires orchestrées par les protéines modulaires apparaissent souvent très sophistiquées et on peut s'interroger sur les propriétés évolutives de ces systèmes d'intégration et de transmission des signaux. Dans une étude réalisée en 2003, l'équipe de W. Lim est parvenue, en jouant avec des mécanismes évolutifs simples, à mimer de façon spectaculaire des processus de reprogrammation de la régulation d'une protéine modulaire, la protéine N-WASP (Dueber et al., 2003). La protéine N-WASP est normalement régulée de façon allostérique par deux domaines : un domaine GDB et un domaine basique (B). Dans cette étude, les auteurs ont remplacé les domaines régulateurs de la protéine N-WASP par d'autres domaines régulateurs, les domaines SH3 et PDZ. De cette façon, ils ont généré différents variants de la protéine N-WASP, dont l'activité pouvait être contrôlée par les motifs de reconnaissance des domaines SH3 et PDZ (**Figure 17**). En modulant l'affinité des domaines pour leurs ligands, l'organisation des interactions intra-moléculaires et la longueur des connecteurs inter-domaines, ils sont parvenus à générer des protéines modulaires présentant une grande versatilité fonctionnelle. Plus précisément, ces protéines se comportent comme des intégrateurs de signaux similaires à ceux utilisés en électronique et permettant d'effectuer des opérations logiques élémentaires telles que des portes AND, OR, ou plus sophistiquées telles que des NAND-NOT (**Figure 17**).

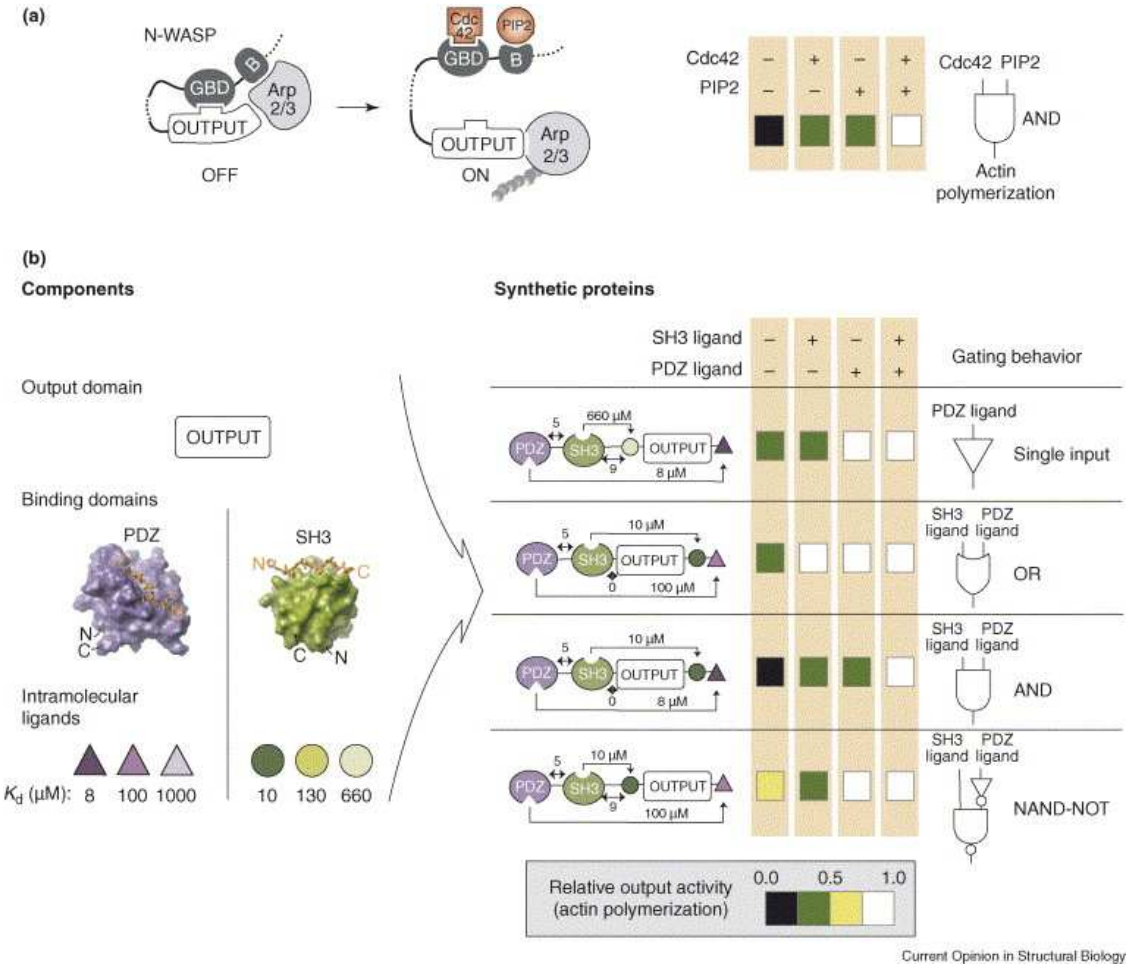


Figure 17. Modification du mode de régulation de la protéine N-WASP. (a) Régulation naturelle de la protéine N-WASP par auto-inhibition, impliquant les domaines GBD et basique (B). (b) Régulation artificielle de l'activité de la protéine N-WASP par les domaines régulateurs SH3 et PDZ. Différentes portes logiques ont pu être définies en fonction de la combinaison des différents domaines, ou de l'affinité des domaines régulateurs pour leur ligand. Figure extraite de (Dueber et al., 2004).

Cet exemple illustre que de nombreuses combinaisons logiques peuvent être échantillonnées au cours de l'évolution par la modification de propriétés d'affinité ou d'organisation structurale des protéines composant les complexes. La généralisation de ces mécanismes évolutifs à l'ensemble des protéines et pas seulement aux protéines de la signalisation reste encore à explorer. Il n'en reste pas moins que pour comprendre l'origine des spécificités de reconnaissances et des variations d'affinités cruciales dans les processus d'assemblage, l'analyse structurale et énergétique des interfaces de complexes constitue une étape majeure.

1.4. Caractérisation structurale et propriétés des complexes protéiques

1.4.1. Introduction

La présence d'un nombre croissant de structures de complexes au sein de la PDB a permis à différentes équipes d'étudier les interfaces de complexes protéiques, afin d'en extraire les propriétés géométriques, physico-chimiques mais aussi évolutives.

1.4.2. Propriétés géométriques des interfaces de complexes

La taille des interfaces de complexes protéiques est généralement mesurée par la différence entre l'aire de la surface accessible (ASA) du complexe et celle des partenaires libres. Si on considère deux protéines A et B s'associant sous forme d'un complexe, la surface de contact est définie par :

$$S = ASA_{protA} + ASA_{protB} - ASA_{complexeAB}$$

En 1999, l'équipe de J. Janin a analysé les caractéristiques structurales de 75 complexes protéiques, et plus particulièrement la taille de leurs interfaces. Les complexes analysés regroupent entre autres des complexes de type protéase-inhibiteur, anticorps-antigène, enzyme-inhibiteur ou encore d'autres complexes impliqués dans la transduction du signal (Lo Conte et al., 1999). Par cette étude, les auteurs ont montré qu'une grande proportion de ces assemblages, principalement des complexes de type protéase-inhibiteur ou antigène-anticorps, ont une interface dont la taille est de 1600 (+/-400) Å^2. D'autres complexes ont la particularité de présenter des interfaces de liaison particulièrement importantes, allant de 2000 à 4660 Å^2. Les protéines constitutives de ces complexes ont de plus la caractéristique de subir des changements conformationnels importants au cours de leur assemblage, à l'inverse des complexes protéiques de taille standard (Lo Conte et al., 1999).

En outre, il semblerait que la taille des interfaces de complexes soit reliée au caractère transitoire ou permanent des assemblages protéiques. En effet, il apparaît que les tailles des interfaces de complexes permanents sont plus importantes que celles des complexes transitoires (Jones and Thornton, 1996; Lo Conte et al., 1999).

1.4.3. Propriétés physico-chimiques des interfaces de complexes protéiques

Les propriétés physico-chimiques des interfaces de complexes protéiques ont été étudiées afin de déterminer les contributions majeures à la stabilité des interfaces.

Glaser et collaborateurs ont démontré, en analysant une base de données de 621 interfaces de complexes protéiques, regroupant entre autres des oligomères ou des complexes enzyme-inhibiteur, que la plupart des contacts entre paires de protéines aux interfaces de complexes impliquaient de larges résidus hydrophobes (Glaser et al., 2001). En outre, les auteurs ont remarqué une différence significative de la composition des interfaces entre de petites et de grosses interfaces. En effet, il apparaît que les résidus hydrophobes sont souvent retrouvés au sein de grosses interfaces, alors que les résidus polaires prévalent dans les petites interfaces. Compte tenu de ces résultats, il apparaît que l'effet hydrophobe, jouant un rôle primordial dans le repliement des protéines, a également un rôle majeur dans la stabilisation des complexes protéiques (Glaser et al., 2001). Les interactions électrostatiques et hydrophobes seraient les propriétés physico-chimiques qui contribuent le plus à la stabilité des complexes protéiques (McCoy et al., 1997; Sheinerman et al., 2000).

Un des aspects non appréhendé par l'étude de Glaser et collaborateurs concerne la distinction entre les complexes permanents et les complexes transitoires. Or en fonction du type de complexes, les forces qui régissent leur interaction sont différentes. En effet, il a été montré que les interfaces de complexes permanents étaient généralement plus hydrophobes que les interfaces de complexes transitoires (Carugo and Argos, 1997; Larsen et al., 1998).

Récemment, De et collaborateurs sont allés plus loin dans la comparaison des complexes permanents et transitoires, en analysant les différences de propriétés au cœur et en périphérie de ces interfaces (De et al., 2005). Ils ont en particulier montré que la périphérie des interfaces de complexes permanents et transitoires est de nature plus polaire que le cœur de ces interfaces. En outre, ils ont confirmé que les interfaces de complexes transitoires sont plus polaires que les interfaces de complexes permanents, aussi bien au cœur qu'en périphérie. Cette propriété peut s'expliquer par le fait que les résidus à l'interface des complexes transitoires interagissent avec le solvant lorsque les protéines constitutives de ce type de complexe existent sous leur forme libre.

Afin d'analyser la contribution de chaque type de résidus à la stabilisation d'un complexe protéique, Bogan et Thorn ont étudié plus précisément leur influence sur l'énergie de liaison (Bogan and Thorn, 1998). Pour cela, les auteurs ont analysé au sein d'une base de données de complexes protéiques d'hétérodimères, la contribution énergétique de chaque résidu dans la stabilisation de ces complexes. Par cette étude, ils ont montré que l'énergie de liaison n'est pas distribuée de façon uniforme sur une interface, mais qu'il existe certains résidus, appelés « *hot spots* », dont la mutation conduit à une chute significative de l'affinité de liaison. Il apparaît que ces « *hot spots* », principalement des tryptophanes, tyrosines et arginines, sont localisés préférentiellement au cœur des interfaces de complexes, et entourés de résidus dont le rôle serait de les exclure du solvant. Cette exclusion du solvant permettrait ainsi une diminution de la constante diélectrique, et aurait donc pour effet d'augmenter la force des interactions électrostatiques et des liaisons hydrogènes (Bogan and Thorn, 1998). Des analyses complémentaires concernant ces positions particulièrement importantes pour la spécificité de reconnaissance des assemblages protéiques, ont par la suite montré que ces résidus étaient particulièrement bien conservés (Halperin et al., 2004; Hu et al., 2000).

1.4.4. Propriétés évolutives des interfaces de complexes protéiques

Le nombre croissant de données génomiques et structurales disponibles dans les banques de données a permis récemment d'entreprendre les premières études systématiques dans le but d'explorer les relations entre similarité de séquences et similarité de structures des complexes protéiques. L'objectif de ces études était alors de mieux appréhender les mécanismes d'évolution des interfaces de complexes protéiques.

En 2003, l'équipe de Russell a entrepris l'étude à grande échelle de structures de complexes protéiques domaine-domaine intra-moléculaires et inter-moléculaires caractérisés sur le plan expérimental (Aloy et al., 2003). En comparant les structures de complexes homologues selon la classification SCOP (Murzin et al., 1995), cette étude a révélé que les homologues proches (à plus de 30% en identité de séquence) interagissent de la même façon sur le plan structural. Les résultats de cette étude, portant sur plus de 30000 interactions, suggèrent ainsi qu'en dessous de 30% d'identité, les interactions entre protéines homologues peuvent différer complètement (**Figure 18**). Une seconde étude plus récente, portant sur plus de 4000 interactions domaine-domaine non redondantes, a confirmé ce résultat, et conclut que les

positions relatives des résidus situés aux interfaces de complexes protéiques sont généralement conservés même chez des homologues lointains (Kim and Ison, 2005).

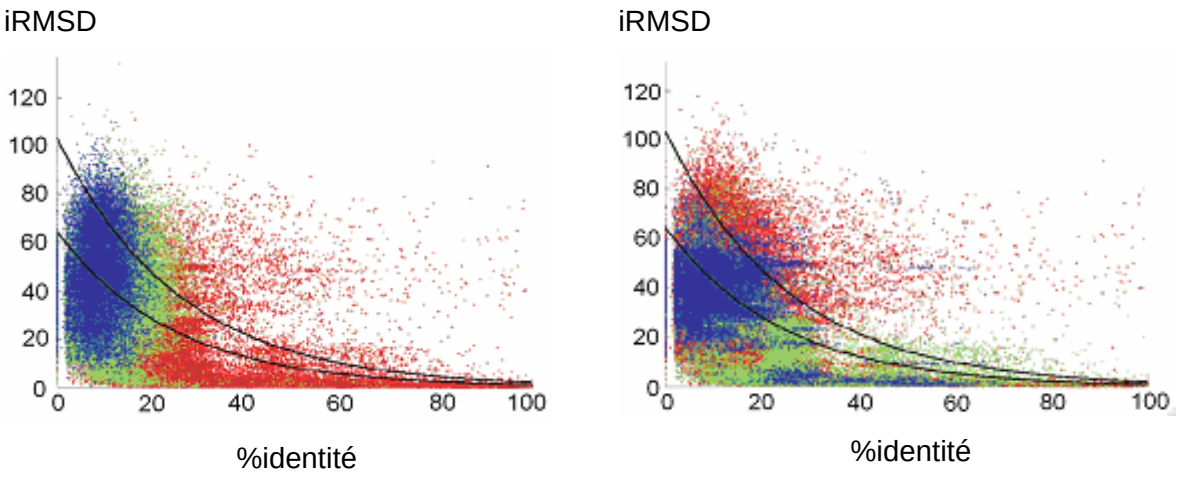


Figure 18. Relation entre similarité sequentielle et structurale des interactions protéiques. (A) Chaque interaction est coloré en fonction de leur classification SCOP (famille en rouge, superfamille en vert et type de repliement en bleu). (B) Chaque interaction est colorée en fonction de sa nature : inter-moléculaire en rouge, intra-moléculaire en vert et fusion de gène en bleu. Figure extraite de (Aloy et al., 2003). L'axe des abscisses représente le pourcentage d'identité de séquences des structures de complexes comparées. L'axe des ordonnées représente un critère géométrique quantifiant la différence entre les orientations des domaines en interaction (Aloy et al., 2003).

En se focalisant non plus sur la conservation structurale des interfaces de complexes protéiques, mais sur la conservation séquentielle des résidus situés à leur interface, différents groupes ont confirmé l'existence d'une pression évolutive (Caffrey et al., 2004; Jimenez, 2005; Mintseris and Weng, 2005). Ces études ont souligné certains faits importants, en particulier, les régions les plus divergentes sont généralement retrouvées de façon prédominante à l'extérieur des interfaces de complexes, tandis que les sites d'interaction sembleraient plus conservés à l'exception des complexes du type antigènes-anticorps (**Figure 19**) (Jimenez, 2005). Pour cette dernière catégorie de complexes, le site de liaison de l'anticorps est très variable, ce qui résulte du besoin que présente un organisme d'avoir un répertoire assez large pour cibler différents antigènes.

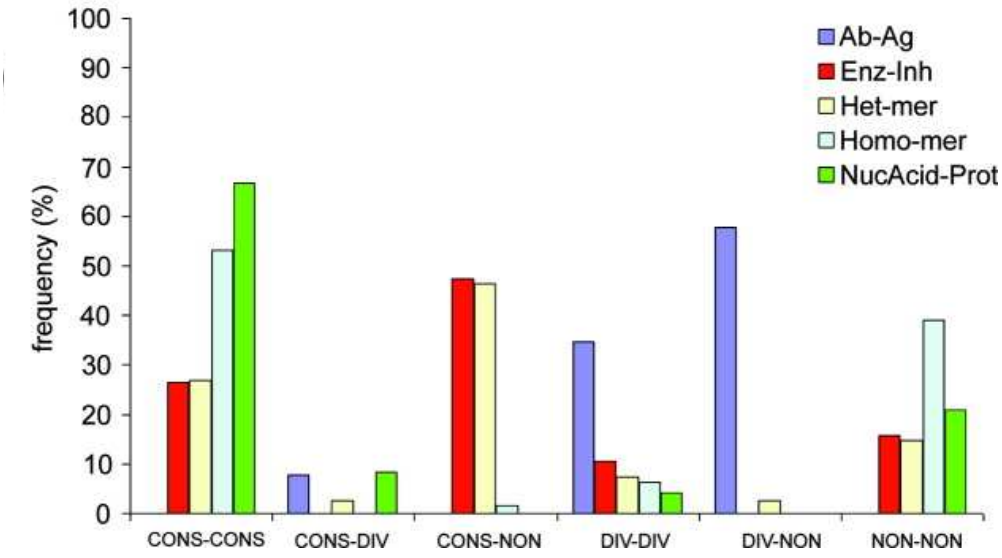


Figure 19. Conservation en fonction du type d'interface. Pour chaque type de complexes, les interfaces des interactants ont été classées comme conservées (CONS), divergentes (DIV). Une troisième classe (NON) définit les cas ni conservés ni divergents. Les sous-ensembles représentés correspondent à des complexes acides-nucléiques-protéines (vert), homo-oligomères (cyan), antigène-anticorps (violet), enzyme-inhibiteur (rouge), ou autres hétéro-oligomères (jaune). Figure extraite de (Jimenez, 2005).

Une des hypothèses à la base des méthodes de prédictions des interactions protéine-protéine, concerne la co-évolution des partenaires en interactions. Un des concepts issu de cette théorie, l'existence de mutations corrélées, repose sur l'hypothèse qu'une mutation d'un résidu à l'interface d'un complexe protéique doit être compensée par une seconde mutation d'une position voisine au niveau structural (voir section 1.2.3).

Une des premières études visant à mesurer le degré de co-évolution des complexes protéiques par l'existence de mutations corrélées a été publiée en 1997 par le groupe de A. Valencia (Pazos et al., 1997). Les auteurs de cette étude ont en effet montré que les paires corrélées ont tendance à se regrouper aux interfaces de complexes protéiques intra-moléculaires. En étendant la recherche de mutations corrélées à un jeu de données regroupant 212 complexes transitoires, et 115 complexes permanents de complexes protéiques inter-moléculaires, Mintseris et Weng ont prouvé que cet effet concernait uniquement les complexes permanents (Mintseris and Weng, 2005). Les auteurs de cette étude ont ainsi conclut que les résidus situés aux interfaces de complexes permanents co-évoluent, alors que la plasticité inhérente aux complexes transitoires ne permet pas de

détecter la présence de mutations corrélées à leurs interfaces. Ces résultats ont été confirmés par l'étude d'Halperin et collaborateurs (Halperin et al., 2006).

1.5. Méthodes bioinformatiques pour la prédiction des structures de complexes protéines-protéines

1.5.1. Introduction

Actuellement, avec le développement de méthodes de plus en plus efficaces pour la prédiction des interactions entre protéines, l'écart entre les données moléculaires et structurales devient de plus en plus important. De plus, il est particulièrement difficile de déterminer la structure tri-dimentionnelle de complexes protéiques. C'est pourquoi, des approches basées sur des techniques bioinformatiques ont vu le jour, et ont pour objectif de prédire la structure tridimensionnelle des complexes protéines-protéines.

1.5.2. Modélisation comparative des structures de complexes protéiques

Au niveau structural, il est clairement établi que des protéines homologues se replient souvent de la même façon, ce qui est la base des études de modélisation comparative. De façon analogue, les interactions protéine-protéine peuvent également être modélisées par des techniques de modélisation par homologie, par le biais de structures de complexes protéiques dont les composantes sont homologues aux protéines cibles. Comme nous l'avons mentionné précédemment, les homologues proches de complexes protéiques, c'est-à-dire à plus de 30% d'identité de séquence, interagissent de la même façon sur le plan structural (Aloy et al., 2003; Kim and Ison, 2005).

À titre d'exemple, d'excellents résultats ont été obtenus pour la modélisation du complexe superantigen toxin – chaîne beta du récepteur T cell par une technique de modélisation comparative (Mendez et al., 2003).

1.5.3. Les techniques d'amarrage moléculaire ou docking

Les techniques de docking ont été développées afin de prédire la structure tridimensionnelle de complexes protéiques, à partir des structures isolées de chaque monomère. Ces méthodes reposent sur l'hypothèse que les partenaires en interaction présentent un certain degré de complémentarité de forme, notion introduite par E. Fischer en 1895, pour expliquer les mécanismes de liaison enzyme-substrat.

Ces approches reposent dans un premier temps sur une recherche des différents modes de liaison des partenaires, puis sélectionnent le modèle le plus probable suivant certains critères spécifiques. Dans certains cas, il est possible d'orienter la recherche par des données expérimentales, évolutives ou statistiques. Cependant, bien souvent, aucune donnée autre que la structure des interactants ne permet de restreindre la zone de recherche autour des partenaires.

Principe des méthodes de docking. Différentes méthodes de docking ont été publiées dans la littérature ces 20 dernières années. Bien que ces méthodes reposent le plus souvent sur des algorithmes spécifiques, elles se basent toutes sur certaines quatre étapes majeures : (i) la description de la surface ou du volume des protéines, (ii) un algorithme de recherche destiné à représenter les différents modes d'association des partenaires, (iii) une étape de filtration destinée à réduire le nombre de solutions potentielles, (iv) une étape de raffinement ne concernant que les modèles les plus probables (**Figure 20**). Souvent, la dernière étape est suivie d'un processus de « *clustering* » des solutions retenues.

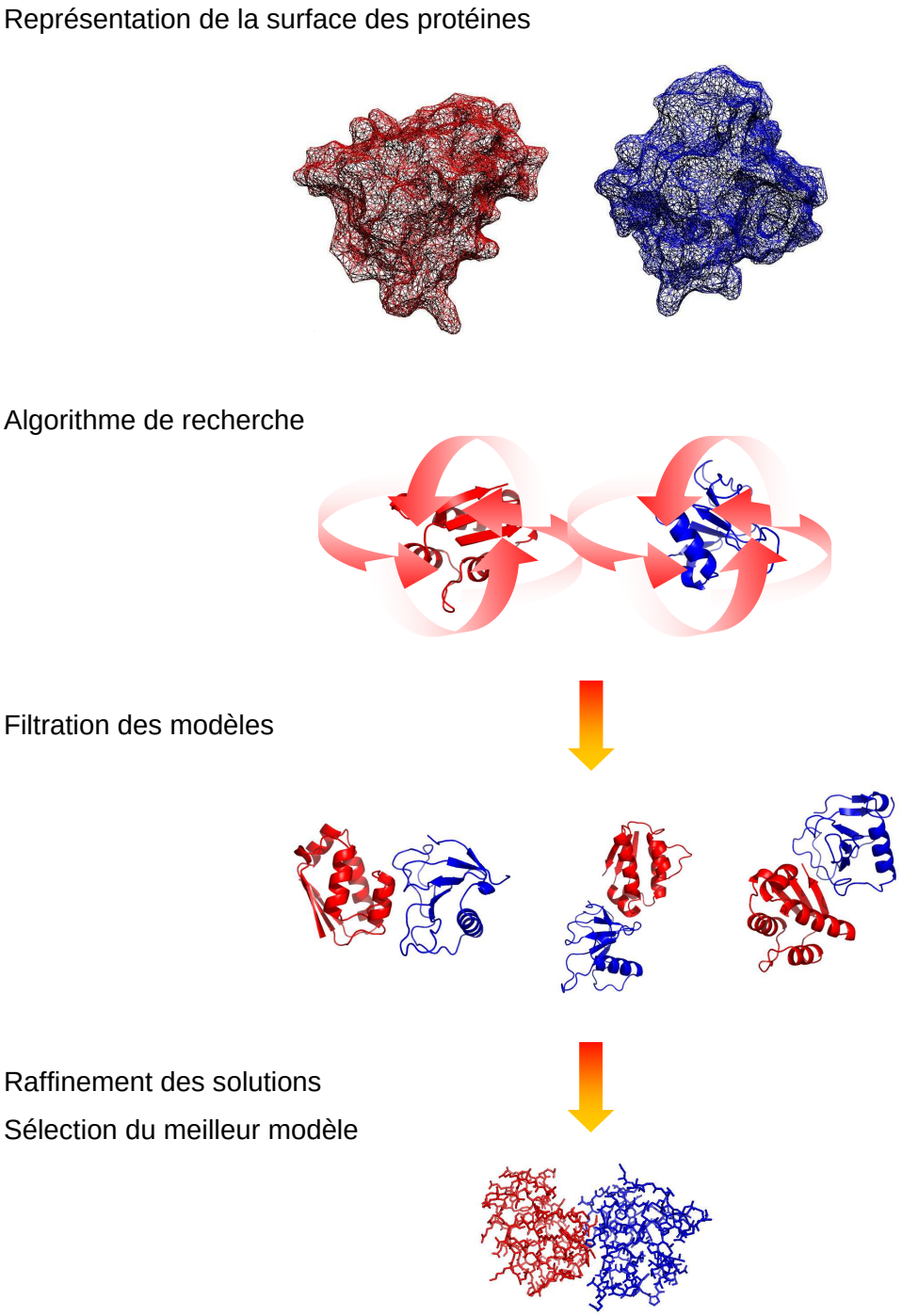


Figure 20. Principales étapes des méthodes de docking. A l'issue des deux premières étapes, un grand nombre de solution sont générés, puis évaluées par une fonction de score propre à chaque programme. Les meilleures solutions sur la base de ce score sont retenues pour une dernière étape d'optimisation, suivie à nouveau d'une évaluation des solutions et de la sélection d'un ou plusieurs modèles d'assemblage.

Représentation de la surface des protéines. Un des problèmes des programmes de docking réside dans la représentation de la surface des protéines. En effet, peu de méthodes de docking utilisent une représentation explicite des chaînes latérales lors des premières étapes de recherche (Fernandez-Recio et al., 2002), en raison du coût important que cela engendre en terme de temps de calcul. Généralement, le choix de représentation des chaînes latérales est également lié à l'algorithme de recherche conformationnelle, comme c'est le cas pour les programmes de recherche sur grilles (Katchalski-Katzir et al., 1992), ces derniers représentant la structure des protéines en les projetant sur une grille tridimentionnelle. Certains programmes modélisent la surface des protéines par des harmoniques sphériques ou encore par une représentation simplifiée basée sur les surfaces de Conolly. Ces représentations de la surface des protéines peuvent être " adoucies " pour permettre une certaine prise en compte de la flexibilité des chaînes latérales de façon implicite, en autorisant par exemple un certain degré de recouvrement des surfaces. Enfin, d'autres approches utilisent des approximations, telle qu'une représentation simplifiée des chaînes latérales par des centroïdes (Gray et al., 2003) durant les premières étapes de docking à basse résolution, pour éviter la prise en compte explicite de la flexibilité des chaînes latérales.

Algorithmes de recherche. À l'exception de peu de programmes, les approches de docking se basent sur une recherche de type " corps rigide " lors des premières étapes. Cependant, malgré cette approximation, l'espace conformationnel à explorer est conséquent. Il existe en effet 6 degrés de libertés (3 pour la translation, 3 pour la rotation). Afin d'entreprendre cette recherche, différents algorithmes ont été développés, généralement associés à un mode de représentation de la surface des protéines donné. En particulier, lorsque les protéines sont projetées sur des grilles tri-dimentionnelles, la stratégie de recherche la plus courante est basée sur les transformées de Fourier. Cette technique introduite par les travaux de Katchalski-Katzir (Katchalski-Katzir et al., 1992) se base sur un calcul de corrélation entre différents points de la grille, réduisant considérablement le temps de calcul nécessaire à une exploration quasi-exhaustive de l'espace des conformations. Cette technique a été introduite dans différents programmes de docking tels que les programmes FTDock (Gabb et al., 1997), GRAMM (Vakser, 1995), DOT (Mandell et al., 2001), ZDOCK (Chen et al., 2003), ou encore le programme HEX dont l'originalité est l'utilisation d'harmoniques sphériques (Ritchie and Kemp, 2000).

Certains programmes de docking font appels, lors de leur processus de recherche, à des algorithmes d'optimisation non-déterministes. Les méthodes non-déterministes, en faisant appel à des tirages de nombres aléatoires, permettent d'explorer l'espace des conformations plus efficacement. C'est le cas des méthodes de Monte-Carlo, implémentées dans les programmes RosettaDock (Gray et al., 2003) et ICM (Fernandez-Recio et al., 2002), ou un grand nombre de modèles de complexes sont générés aléatoirement. D'autres programmes, tes que les programmes GAPDOCK (Gardiner et al., 2001) ou DARWIN (Taylor and Burnett, 2000) implémentent une autre méthode d'optimisation non déterministe : l'algorithme génétique. Le principe est alors de simuler l'évolution d'une population d'individus divers (ici une population de modèles de complexes) auxquels on applique différents opérateurs génétiques et que l'on soumet à chaque génération à une sélection. Les opérateurs génétiques vont consister à réaliser certaines modifications structurales des modèles, puis le processus de sélection va consister à évaluer l'adéquation des modèles avec une fonction d'évaluation.

Filtration et Raffinement. À l'issue des premières étapes de docking, un certain nombre de complexes sont sélectionnés par des fonctions d'évaluation dérivées de fonctions statistiques, énergétiques ou par le biais de contraintes expérimentales. Par la suite, lors des dernières étapes de docking, certains procédés de raffinement peuvent être introduits. La plupart des méthodes prennent dès lors en compte la flexibilité des chaînes latérales à l'interface des complexes modélisés. Certains programmes, tels que les programmes HADDOCK (Dominguez et al., 2003) ou RosettaDock (Gray et al., 2003) introduisent également de façon simultanée un certain degré de flexibilité au niveau du squelette peptidique. Il apparaît que cette technique soit efficace lorsque les changements conformationnels dus à l'interaction entre les partenaires sont de faible amplitude. En outre, certains procédés visent à optimiser (Fernandez-Recio et al., 2002) ou minimiser les chaînes latérales (Jackson et al., 1998), ou encore à introduire de façon explicite les effets de la solvatation (Dominguez et al., 2003). Ces techniques de raffinement et d'optimisation sont très importantes, en particulier lorsque les fonctions de score utilisées se basent sur des critères énergétiques, très sensibles à la qualité des modèles.

Les fonctions de score. Tout au long du processus de docking, les différents modèles générés sont évalués par différentes fonctions de score, destinées à sélectionner les modèles les plus

probables. Ces fonctions de score se basent généralement sur des critères (i) géométriques, (ii) statistiques, (iii) énergétiques.

i. Certains programmes se basent sur une complémentarité géométrique des partenaires en interaction. En particulier, le système de score des amarrages sur grilles vise à optimiser la complémentarité de surfaces entre les deux partenaires, une fois représentés sur une grille tri-dimensionnelle. La complémentarité de surface reflète ainsi l'extension des interactions de van der Waals à l'interface des assemblages protéiques. Cependant, la complémentarité géométrique à elle seule est incapable de discriminer des modèles proches de la structure native de faux modèles d'assemblage. En particulier, les complexes pour lesquels l'énergie d'interaction électrostatique joue un rôle dominant sont plus difficiles à prédire par ces fonctions de score. En général, les méthodes basées sur une recherche sur grille considèrent en plus la contribution éléctostatique (Gabb et al., 1997; Mandell et al., 2001). Récemment, une fonction d'évaluation basée sur une représentation de la surface par des Voronoï pour décrire les propriétés de complémentarité géométriques et physico-chimiques des interfaces a été publiée par le groupe de A. Poupon (Bernauer et al., 2007) (**Figure 21**).

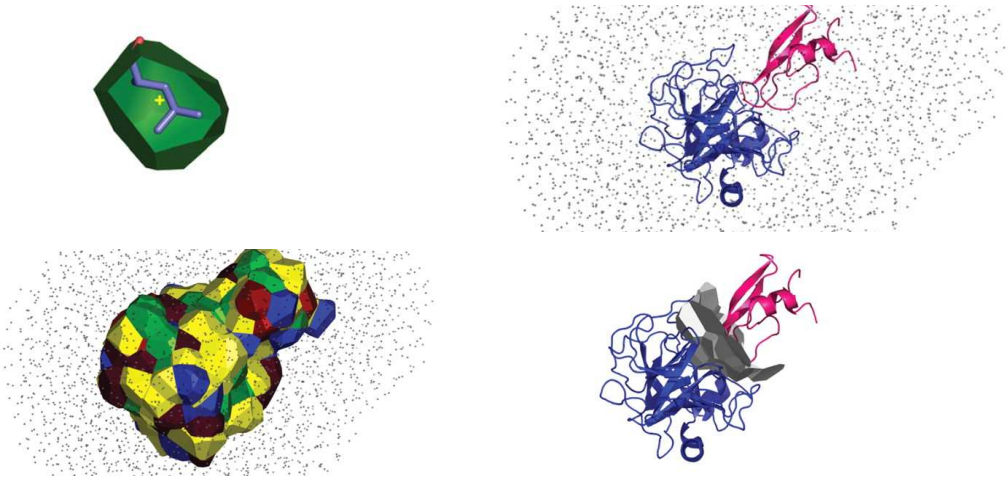


Figure 21. Représentation de Voronoï d'un complexe protéique. (A) Cellule de Voronoï d'un résidu Leucine. La croix jaune indique le centroïde. (B) Complexe protéique représenté sous forme de rubans (code PDB 1p2k). Les deux protéines sont représentées en bleu et rose. (C) Représentation de Voronoï du complexe (code pdb 1p2k). (D) Représentation de l'interface du complexe (gris). Figure extraite de (Bernauer et al., 2007).

ii. Certaines approches implémentent dans leur fonction de score des potentiels empiriques issus de l'analyse de structures de complexes caractérisés

expérimentalement. C'est le cas du package 3D-Dock (Gabb et al., 1997), du programmes BIGGER (Palma et al., 2000), ou encore du programme RosettaDock (Gray et al., 2003). Concernant le package 3D-dock, la fonction statistique utilisée, RPScore (« *Residue level Pair potential Score* »), est basée sur les statistiques de contact entre paires de résidus de structures issues de la PDB, qui exprime la probabilité d'interaction entre deux résidus donnés (Moont et al., 1999). Ce potentiel empirique issu de l'analyse de 103 interfaces de complexes non homologues, est utilisé au sein du package 3D-dock afin de filtrer certains complexes générés lors de la phase d'exploration par le programme FTDock (Gabb et al., 1997), et ainsi de réduire le nombre de modèles à analyser par des techniques plus coûteuses en temps de calcul. De façon similaire, la fonction de score du programme RosettaDock lors de la première phase de recherche à basse résolution, est basée sur un potentiel statistique d'interaction résidu-résidu. Ainsi, RosettaDock implémente deux fonctions de score en fonction de la précision de la phase de recherche exploratoire : une fonction basée sur des critères statistiques lors d'une recherche à basse résolution où les chaînes latérales sont représentées grossièrement sous forme de centroïde, une fonction plus précise basée sur des critères énergétiques lors de la seconde phase, où les chaînes latérales sont modélisées de façon explicite.

iii. Le plus souvent, les fonctions d'énergies sont prises en compte par des potentiels sur grille, mais elles peuvent aussi être définies plus finement. C'est le cas du programme RosettaDock, pour lequel la fonction de score à haute résolution est basée entre autres sur certains termes électrostatiques, de van der Waals, de solvatation, d'énergie de liaison hydrogène ainsi qu'un terme de probabilité rotamérique (Gray et al., 2003).

Il a été montré que la combinaison de différents termes pouvait permettre d'augmenter la qualité des prédictions par les approches de docking (Murphy et al., 2003). C'est pourquoi de plus en plus de programme de docking ne se basent plus sur un seul type de fonction de score, mais plutôt sur la combinaison de ces différentes fonctions.

Introduction de données biologiques ou évolutives lors du processus de docking. L'existence de données expérimentales, telles que des données de mutagenèse ou des données RMN, conditionne souvent le succès des programmes de docking. Lorsque ces

données sont disponibles, elles peuvent être intégrées de deux façons dans les programmes de docking. La méthode la plus conventionnelle consiste à filtrer les solutions qui ne satisfont pas ces données expérimentales, c'est le cas dans la suite 3D-Dock (Gabb et al., 1997). Une méthode plus optimale, HADDOCK (Dominguez et al., 2003), développée au sein du groupe de A. Bonvin, consiste à guider le processus de recherche en fonction de ces contraintes expérimentales, ce qui a pour intérêt d'enrichir le nombre de conformations proches de la structure native.

La philosophie du programme de docking HADDOCK (« *High Ambiguity Driven biomolecular DOCKing* ») consiste à intégrer différentes sources d'informations expérimentales sur les interfaces de complexes protéiques pour pouvoir les modéliser en 3D. Ces informations sont encodées sous forme de contraintes ambigües (« *Ambigous Interaction Restraints* » ou AIRs) par analogie aux contraintes ambigües communément utilisées pour la détermination de structure par RMN (Nilges and O'Donoghue, 1998). Les AIRs sont ajoutées comme un terme additionnel à la fonction d'énergie que le programme HADDOCK essayera de minimiser au cours de l'exploration de l'espace des conformations.

De façon intéressante, même en l'absence de données expérimentales précises, le programme HADDOCK peut se révéler particulièrement efficace dans la détermination de la structure des assemblages de complexes protéiques. À titre d'illustration, des données de protection contre la digestion enzymatique couplées à des données évolutives de conservation ont suffit au programme HADDOCK pour prédire avec succès la structure trimérique de la glycoprotéine E du virus TBE (van Dijk et al., 2005).

Avec le déséquilibre croissant entre les données génomiques et expérimentales, les données issues des banques de séquences génomiques sont bien souvent les seules informations disponibles pour guider les programmes de docking. C'est pourquoi différents groupes se sont intéressés aux propriétés évolutives des interfaces de complexes protéiques, telles que les propriétés de conservation des résidus enfouis à l'interface des complexes (Caffrey et al., 2004; Jimenez, 2005; Mintseris and Weng, 2005), ou encore leurs propriétés de co-évolution (Mintseris and Weng, 2005; Pazos et al., 1997). Sur la base de ce type de données évolutives, différentes méthodes dont l'objectif est la prédiction des sites d'interactions protéiques ont ainsi vu le jour (Bradford and Westhead, 2005; Fariselli et al., 2002; Neuvirth et al., 2004;

Ofran and Rost, 2003; Wang et al., 2006). De plus, différentes études ont prouvé l'intérêt de ce type d'information pour guider les programmes de docking moléculaire (Chelliah et al., 2006; de Vries et al., 2006; Duan et al., 2005; Tress et al., 2005; van Dijk et al., 2005).

Evaluation des méthodes de docking : le concours CAPRI. L'idée de confronter les différentes approches de prédiction des structures de complexes protéiques a donné naissance au concours CAPRI (« *Critical Assessment of PRedicted Interactions* »). Le concours CAPRI débute dès qu'une cible est proposée. Les cibles correspondent à des structures de complexes protéiques résolues par des méthodes expérimentales, et dont la structure n'a pas encore été déposée dans la PDB. L'objectif des participants est alors de prédire la structure de ces complexes à partir des coordonnées des composants libres. Les prédictions sont ensuite comparées aux structures natives des complexes. Les structures de départ des partenaires proviennent soit de la PDB, donc résolues dans leur configuration non liée, ou de modèles reconstruits par homologie. On parle alors d'assemblages « non liés ». À défaut, la structure extraite du complexe peut être proposée pour l'un des partenaires, ce qui correspond à des assemblages « lié-non lié ».

Les modèles proposés par les participants sont évalués selon différents critères. Ces critères correspondent à la proportion de contacts natifs correctement prédits, ou encore au déplacement du ligand par rapport à son site de liaison. Suivant ces critères, la qualité de ces modèles est définie comme haute, bonne, acceptable ou incorrecte (**Tableau 4**).

Classement	f_{nat} [a]	L_rms [b]	I_rms [c]
Haut	>= 0.5	<= 1.0	ou <= 1.0
Bon	>= 0.3	1.0 < x <= 5.0	ou 1.0 < x <= 2.0
Acceptable	>= 0.1	5.0 < x <= 10.0	ou 2.0 < x <= 4.0
Incorrect	< 0.1		

[a] F_{nat} : Proportion de contacts natifs du modèle sur le nombre de contacts dans le complexe natif
[b] L_rms : RMSD calculé sur la chaîne principale du ligand, après superposition des récepteurs du complexe natif et du modèle
[c] I_rms : RMSD entre le complexe natif et le modèle calculé sur la chaîne principale des résidus situés à l'interface

Tableau 4. Critères d'évaluation du concours CAPRI. Extrait de (Mendez et al., 2003).

À ce jour (octobre 2007), 13 cycles ont été effectués, et 29 cibles soumises depuis 2001. Lors des différents cycles du concours CAPRI, certains enseignements ont pu êtres apportés,

pouvant conduire à une amélioration des techniques de docking. En particulier, la cible 12 correspondant à un complexe dockerin-cohesin, a démontré tout l'intérêt de prendre en compte de façon explicite la flexibilité des chaînes latérales au cours du processus de docking (Wang et al., 2005). De plus, les données biologiques permettent souvent de guider efficacement la recherche, en restreignant l'espace conformationnel à rechercher (van Dijk et al., 2005). Dans certains cas, les informations évolutives issues d'alignements de séquences des protéines du complexe à étudier sont utilisées pour guider les approches de docking. Ainsi, en se basant sur ce genre d'informations, certaines cibles ont pu être prédites avec un niveau de qualité acceptable lors des cycles 3, 4 et 5 (Tress et al., 2005).

Lors des derniers cycles du concours CAPRI, il est apparu que les obstacles majeurs auxquels se confrontent actuellement les programmes de docking sont les changements conformationnels importants que subissent les protéines si l'on compare leur forme liée et non liée. Ces défauts ont été particulièrement mis en évidence lors des cycles 6 à 12, dont les cibles comprennent un nombre plus réduit d'assemblages rigides de type enzyme-substrat ou antigène-anticorps, et beaucoup plus de cibles présentant des variations importantes de leur conformation au cours de leur amarrage, comme des protéines de la signalisation (Chaudhury et al., 2007). Un exemple qui démontre la différence importante de qualité prédictive des algorithmes de docking lorsque les protéines subissent certains changements conformationnels concerne le complexe dockerin-cohesin (Carvalho et al., 2003). Deux cibles concernaient ce complexe : la cible T11 correspondant à un cas « non liés » et la cible T12 à un cas « lié-non lié ». La majorité des groupes ont soumis des modèles de haute ou de bonne qualité pour la cible T12 contrairement à la cible T11.

Prise en compte des changements conformationnels du squelette peptidique. L'une des solutions pour mieux appréhender la flexibilité des protéines lors du processus de docking consiste à représenter les protéines par un ensemble de structures, qui seront ensuite utilisées par un algorithme de docking de type corps-rigide. Si on considère un complexe formé de deux protéines A et B, cette procédure consiste à amarrer chaque conformation de la protéine A à chaque conformation de la protéine B. Pour générer les différentes conformations des protéines A et B, il est alors courant d'utiliser des techniques de dynamique moléculaire. Cette procédure a en particulier été utilisée lors de la modélisation du complexe UbcH5B-CNOT4 (Dominguez et al., 2003). Récemment, les équipes de M. Nilges et de P. Bates ont

analysé sur un grand nombre de complexes protéiques, l'efficacité d'une telle approche pour retrouver des conformations proches des structures natives (Grunberg et al., 2004; Smith et al., 2005). De par ces études, il apparaît que cette procédure permet d'augmenter la probabilité de trouver des conformations proches des structures natives. Toutefois, cette technique produit plusieurs paires de conformations complémentaires (Grunberg et al., 2004). De plus, elle augmente considérablement le nombre de solutions à analyser, et donc la difficulté à distinguer ces différentes orientations possibles par les fonctions de score.

D'autres approches se basent sur l'hypothèse que les changements conformationnels ont lieu consécutivement à l'amarrage des molécules. Lors du cycle 9 du concours CAPRI, le groupe de D. Baker a utilisé une approche consistant à optimiser simultanément la conformation du squelette et l'orientation de la protéine ARHGAP10 par rapport à Arf1, correspondant à la cible 24. Cependant, bien que cette approche permette de modéliser des changements conformationnels importants, elle n'a pu permettre de produire un modèle d'assemblage correct du complexe (Chaudhury et al., 2007).

1.6. Design d'interactions protéines-protéines par des approches rationnelles

1.6.1. Introduction

Une des façons d'explorer les bases de la reconnaissance moléculaire et des spécificités d'interactions des assemblages protéiques consiste à analyser les effets de certaines mutations au niveau de leurs interfaces. Dans ce sens, les structures de complexes protéiques constituent une base incontournable pour l'application de telles stratégies.

Le *design* informatique de protéines est un procédé visant à modifier la séquence de la protéine dans le but de changer certaines de ses propriétés comme sa stabilité, sa fonction, ou sa spécificité d'interaction. La prédiction des mutations à effectuer découle d'une analyse bioinformatique de la structure de la protéine. En se basant sur les méthodes de *design*, les différents succès reportés dans la littérature illustrent le champ d'applications de ces approches. En effet, plusieurs études ont montré que la stabilisation de protéines basée sur leurs structures (Dahiyat and Mayo, 1997; Filikov et al., 2002; Korkegian et al., 2005) ou encore l'ingénierie de nouveaux repliements protéiques (Kuhlman et al., 2003) pouvaient être envisagées efficacement en utilisant des approches automatiques de *design*. Ces différents succès prouvent que les techniques de *design* de protéines sont désormais assez efficaces pour être appliquées mais aussi étendues à l'étude des interactions protéines-protéines.

1.6.2. Les méthodes automatiques et semi-automatiques de design

Les progrès réalisés dans ce domaine montrent que le succès des méthodes de *design* repose sur deux facteurs : des algorithmes efficaces permettant de gérer l'exploration de l'espace des séquences et sa combinatoire exponentielle, et des fonctions d'énergie (ou d'évaluation) dont le rôle est de trier les séquences relativement à leur adéquation avec le repliement de la protéine. Du fait de la précision requise par ce type d'approche, leur efficacité dépend directement des fonctions d'énergie sur lesquelles elles s'appuient.

Le développement de fonctions de score précises est donc l'un des enjeux majeurs de la bioinformatique structurale, particulièrement en ce qui concerne le *design* de protéines. À l'heure actuelle, différents programmes de *design* des protéines ont été rendus accessibles à la

communauté scientifique (Madaoui et al., 2006) (**Tableau 5**). Les fonctions d'énergies utilisées par ces programmes se divisent en trois sous-catégories (Lazaridis and Karplus, 2000) : les méthodes statistiques (SEEF pour « *Statistical Effective Energy Function* »), les méthodes basées sur un champ de force physique (PEEF pour « *Physical Effective Energy Function* ») et enfin une troisième classe de méthode basée sur l'utilisation de données expérimentales (EEEF pour « *Empirical Effective Energy Function* »).

	Méthode	Description	Site web
a	PoPMuSiC	Prédiction de l'effet stabilisant de mutations ponctuelles Potentiel statistique issu de l'analyse de bases de données	http://babylone.ulb.ac.be/popmusic
	I-Mutant	Prédiction de l'effet stabilisant de mutations ponctuelles Version 1.0 basée sur un réseau de neurones Version 2.0 basée sur un SVM Apprentissage à partir de données thermodynamiques	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi
b	EGAD	Ingénierie des protéines Prédiction de l'effet stabilisant de mutations ponctuelles Basée sur le champ de force OPLS-AA	http://egad.berkeley.edu/software.php
c	Dfire-Dmutant	Prédiction de l'effet stabilisant de mutations ponctuelles Potentiel distance dépendant, résidu-spécifique	http://sparks.informatics.iupui.edu/
	Foldx	Prédiction de l'effet stabilisant de mutations ponctuelles Evaluation de l'énergie libre ΔG Paramétrisation à partir de mutants expérimentaux	http://foldx.embl.de
	Rosetta[c]	Ingénierie des protéines Prédiction de l'effet stabilisant de mutations ponctuelles Evaluation de l'énergie de la protéine cible	http://rosettadesign.med.unc.edu

Tableau 5. Différentes méthodes de *design* de protéines. Quelques méthodes de design automatiques basées sur des fonctions d'énergie SEEF (a), PEEF (b) ou EEEF (c) sont présentées. Ces méthodes sont analysées plus en détail dans (Madaoui et al., 2006).

Méthodes SEEF. Les potentiels statistiques sont dérivés des données de la PDB. Le premier programme dans cette catégorie est Prosa (Sippl, 1993). Son principe est d'identifier au sein de la PDB quelles distances peuvent séparer chaque paire de résidus, et de comparer ces données statistiques à la structure à évaluer. Il s'agit donc d'un potentiel statistique entre

paires de résidus. De nombreux autres programmes ont été développés depuis la version initiale de Prosa. La caractéristique commune à toutes ces méthodes SEEF est avant tout de se servir exclusivement de statistiques collectées au sein de la PDB pour évaluer les structures proposées en fonction de leur adéquation avec ces données statistiques. De ce fait, les méthodes SEEFs ont l'avantage d'être très rapides et peu sensibles aux petites erreurs de positionnement des atomes ; c'est la raison pour laquelle elles sont fréquemment utilisées dans le domaine de la modélisation par homologie (Lazaridis and Karplus, 2000).

Méthodes PEEF. Les potentiels physiques combinent des fonctions d'énergies issues de la mécanique moléculaire, ainsi que des modèles prenant en compte les effets de la solvatation sur l'énergie libre du système. La principale différence par rapport aux méthodes statistiques concerne la paramétrisation de ces méthodes qui n'est pas dérivée de la structure de protéines, mais de la mesure de paramètres physiques. Ces méthodes sont largement utilisées dans le domaine de la dynamique moléculaire, pour simuler le comportement des protéines.

Méthodes EEEF. Ces méthodes combinent une description physique des interactions et une connaissance basée sur des données expérimentales et statistiques. À titre d'exemples, l'algorithme AGADIR (Munoz and Serrano, 1995), le programme FOLDX (Guerois et al., 2002; Schymkowitz et al., 2005), ainsi que le programme RosettaDesign (Kuhlman et al., 2003) sont trois méthodes empiriques couramment utilisées pour le *design* de mutation permettant d'augmenter la stabilité des protéines. En particulier, différents succès de *design* de protéines ont été reportés aussi bien par le programme FOLDX (van der Sloot et al., 2004) que par le programme RosettaDesign (Dantas et al., 2003; Kuhlman et al., 2003; Kuhlman et al., 2001; Kuhlman et al., 2002; Nauli et al., 2001).

La méthode RosettaDesign. L'objectif principal de ce programme est l'identification de séquences protéiques permettant la stabilisation de structures protéiques cibles. À partir de la structure d'une protéine cible, et de la position des résidus à optimiser, RosettaDesign détermine spécifiquement les mutations permettant d'augmenter la stabilité de la protéine. Pour cela, RosettaDesign couple une fonction d'énergie d'évaluation, et une procédure d'optimisation destinée à l'exploration de l'espace des séquences.

Concernant la phase d'optimisation, RosettaDesign utilise un algorithme de recuit simulé. En partant d'une séquence protéique aléatoire, des mutations ponctuelles des résidus à modifier, couplées à des modifications de leur rotamères associés sont acceptées selon un critère de Métropolis. Les chaînes latérales des acides aminés à modifier vont ainsi adopter un ensemble discret de rotamères, tirés de la librairie de rotamères de Dunbrack (Dunbrack and Cohen, 1997). Certaines variations de ces conformations préférentielles sont autorisées pour les résidus enfouis au sein de la protéine, ceci par de légères modifications des valeurs de chi1 et chi2.

L'originalité du programme RosettaDesign consiste en l'incorporation de la flexibilité du squelette peptidique dans le processus d'optimisation de séquence décrit précédemment. L'introduction de la flexibilité permet d'explorer un champ plus large de possibilités au niveau de l'espace des séquences, et représente donc une étape essentielle pour le design de protéines non naturelles de topologie inconnue. C'est probablement cette originalité qui a permis la synthèse d'une nouvelle protéine de 93 résidus baptisée Top7, dont la topologie était jusqu'alors inconnue (Kuhlman et al., 2003) (**Figure 22**).

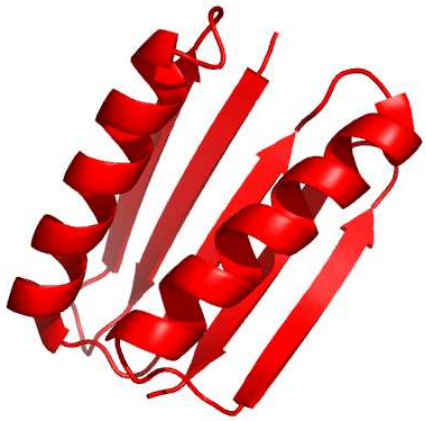


Figure 22. Structure cristallographique de la protéine Top7 (code PDB 1QYS).

1.6.3. Design d'interfaces protéines-protéines

Jusqu'à récemment, la compréhension des mécanismes régissant les interactions protéines-protéines était principalement issue d'analyses de mutagenèse dirigée, ou encore de criblages à haut débit, destinées à sélectionner des ligands de haute affinité pour certains récepteurs donnés. Cependant, ces techniques sont lourdes, et ne permettent pas

d'appréhender de façon détaillée les principes physiques régissant les interactions protéines-protéines.

Les différents succès rencontrés au cours de ces dernières années par les approches de *design* des protéines dans des domaines aussi variés que sont la stabilisation de protéines ou encore la synthèse de protéines artificielles, ont ainsi motivé l'expansion de leur champ d'application à l'étude des interactions protéine-protéine. Dans cette section, sont présentées différentes applications des méthodes de *design*, aussi bien pour altérer que pour mimer des assemblages protéiques.

Modification des spécificités d'interactions de complexes protéiques. Les interactions protéine-protéine sont impliquées dans des réseaux très complexes et souvent, une même protéine interagit avec plusieurs partenaires. Afin de modifier ou de moduler ces réseaux complexes au niveau cellulaire, les approches bioinformatiques de *design* doivent nécessairement intégrer les bases moléculaires qui font la spécificité de ces assemblages protéiques. Dans ce sens, plusieurs études récentes ont confronté ces techniques de design pour modifier les spécificités d'interactions des complexes protéiques.

En 2002, Shifman et Mayo ont montré que par ce type d'approche, il était possible de modifier la spécificité d'interaction de complexes protéines-ligands (Shifman and Mayo, 2002). Les auteurs ont étudié un système particulièrement intéressant impliquant la calmoduline, capable de reconnaître une grande variété de ligands, dont le peptide smMLCK. L'objectif de l'étude était alors de modifier la calmoduline pour qu'elle ne reconnaisse que le peptide smMLCK. Se basant sur la structure du complexe calmoduline-smMLCK (code PDB 1CDM), les auteurs, en modifiant son interface par des techniques de *design*, ont réussi à produire un variant capable de lier spécifiquement le peptide smMLCK (l'affinité pour ce peptide était alors 86 fois supérieure aux autres peptides connus pour interagir avec la calmoduline sauvage).

Reina et collaborateurs sont allés encore plus loin, en proposant le *design* de nouvelles paires de complexes protéines-peptides (Reina et al., 2002). Leur modèle d'étude, les domaines PDZ, sont de petits domaines globulaires capables de reconnaître les 4-7 derniers résidus de l'extrémité C-terminale de protéines cibles. Par des techniques de *design*, les auteurs de cette

étude ont réussi à créer un domaine PDZ capable de reconnaître spécifiquement un peptide, le peptide Eg5 correspondant à l'extrémité C-terminale d'une protéine « *kinésine-like* ». En confirmant l'interaction *in vitro* et *in vivo*, les auteurs de cette étude démontrent ainsi qu'il est possible de modifier les domaines PDZ pour qu'ils reconnaissent spécifiquement certaines protéines cibles. L'extrémité C-terminale étant le plus souvent déstructurée et accessible au solvant pour la plupart des protéines (Berezovsky et al., 1999), cette approche constitue donc une stratégie générale pour cible ou purifier les protéines (Reina et al., 2002).

Synthèse de protéines chimères. Une autre application intéressante des méthodes de *design* de protéines consiste à synthétiser des protéines chimères combinant plusieurs modules au sein d'une même protéine. Nous avons vu précédemment qu'en combinant certains domaines au sein d'une protéine, il était possible de reprogrammer le comportement d'une protéine régulée (section 1.3.4). En outre, une étude récente, menée par Chevalier et collaborateurs, a exploré la possibilité de créer, par des techniques de *design* de protéines, des protéines chimères en combinant certains domaines connus de liaison à l'ADN (Chevalier et al., 2002). Dans cette étude, deux domaines, capables chacun de reconnaître des motifs d'ADN spécifiques, ont été fusionnés, puis 14 résidus à l'interface entre ces domaines ont été optimisés par des techniques de Monte Carlo (**Figure 23**). La protéine chimère résultante, baptisée E-Drel, reconnaît spécifiquement une séquence d'ADN chimère, contenant les deux sites de liaisons des domaines séparés. Non seulement un complexe protéique stable a pu être synthétisé par cette procédure (sa surface d'interaction étant de 1400 $Å^2$) mais en plus la protéine chimère conserve sa fonction catalytique de clivage de l'ADN.

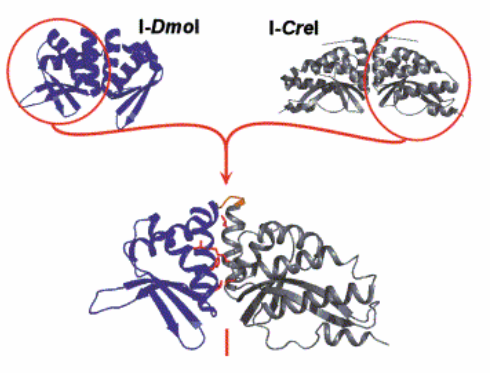


Figure 23. Création de la protéine chimère E-Drel. La combinaison des domaines I-Dmol et I-Crel en une seule protéine chimère, et l'optimisation de l'interface a été réalisée par des techniques de *design* de protéines. Figure extraite de (Chevalier et al., 2002).

Synthèse de peptides mimétiques d'une interaction. Récemment, la structure du complexe entre la protéine gp120 du virus du SIDA, et la protéine CD4 humaine a été cristallisée (Kwong et al., 1998). Sur la base de cette structure cristallographique, une étude récente, menée par Claudio Vita, a consisté à exploiter la connaissance de la zone d'interaction entre les deux protéines pour synthétiser de façon rationnelle un peptide mimétique de la protéine CD4, qui pourrait inhiber l'interaction entre le récepteur CD4 et la protéine virale gp120 (Vita et al., 1999). Sur la base de cette étude, différents peptides ont été proposés, parmi lequel le peptide CD4M33, dont la structure cristallographique en complexe avec la protéine gp120 a été résolue en 2005 par l'équipe de P. Kwong en 2005 (code pdb : 1YYL). En examinant la surface d'interaction de ce complexe, il apparaît qu'elle implique la même région que pour le complexe gp120-CD4 (code PDB 1GC1, voir **Figure 24**).

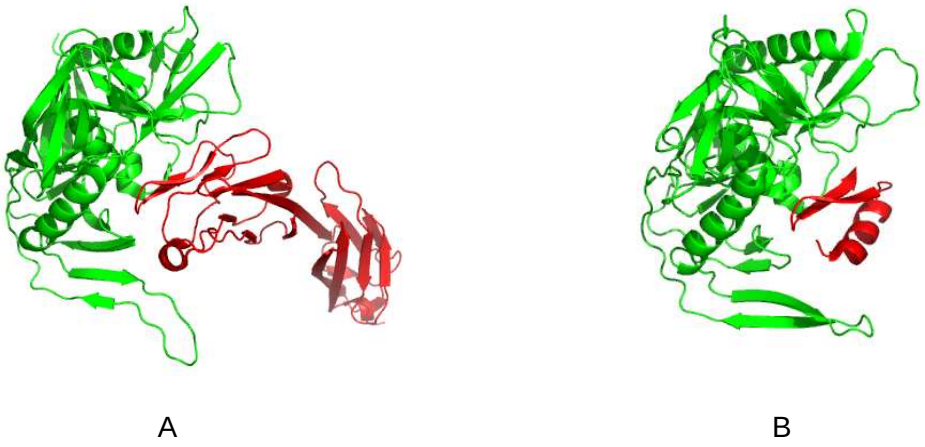


Figure 24. Structure cristallographique des complexe gp120-CD4 et gp120-CD4M33. (A) Structure du complexe gp120-CD4 (code PDB 1GC1). En vert, la protéine H1V-1 YU2 gp120 . En rouge, la protéine CD4. (B) Structure du complexe gp120-CD4M33 (code PDB 1YYL). En vert, la protéine H1V-1 YU2 gp120 . En rouge, le peptide CD4M33.

De façon intéressante, le peptide CD4M33 est capable de se fixer à la protéine gp120 provenant de différentes souches virales, telles que SF2 (X4), IIIB (X4), JRFL (R5) ou Bal (R5) avec des valeurs de constantes de dissociation (Kd) de l'ordre de 1 à 20 nM, affinités comparables avec celles reportées pour la protéine native CD4 (Stricher et al., 2006).

Inhibition des interactions protéines-protéine. La plupart des inhibiteurs d'interaction conçus ces dernières années se basent sur de petites molécules capables de s'adapter dans la poche de liaison d'une protéine dont la structure est connue. Les enzymes présentant de profondes cavités ont été largement ciblées dans ce but. Récemment, il est apparu que

l'inhibition des interfaces de complexes protéiques pouvait également constituer des cibles thérapeutiques d'intérêt dans plusieurs pathologies. Un des exemples particulièrement intéressant d'inhibition des interactions protéines-protéines par de petites molécules concerne les nutlins, capables d'inhiber spécifiquement l'interaction entre MDM2 et p53 (Vassilev et al., 2004). Cette molécule, antagoniste de la protéine MDM2, se lie à MDM2 au niveau de la poche de liaison de p53 (**Figure 25**), et active ainsi p53 qui va entraîner l'arrêt du cycle cellulaire, et l'apoptose des cellules cancéreuses. Cependant, l'absence de poche de liaison aux interfaces de la plupart des complexes protéiques se révèle être un obstacle à la généralisation de ce type d'approche.

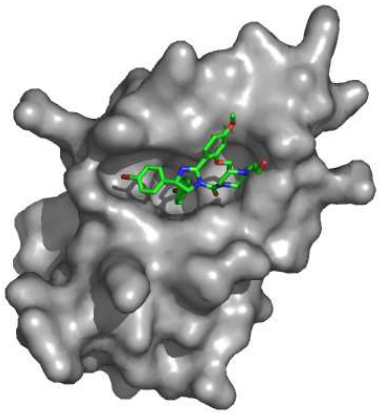


Figure 25. Inhibition de MDM2. La protéine MDM2 est en surface grise L'inhibiteur est représenté sous forme de bâton. (Code PDB : 1RV1).

Une autre stratégie d'inhibition du complexe MDM2-p53, faisant intervenir cette fois-ci non plus une petite molécule mais un peptide, a récemment fait l'objet d'une étude (Zhong and Carlson, 2005). Dans cette étude, les auteurs ont utilisé une procédure combinant des approches de dynamique moléculaire et de *design* de protéines, afin de produire un peptide mimétique de la région de p53 impliquée dans l'interaction. De façon intéressante, ce peptide possède une similarité structurale surprenante avec les nutlins découvertes par criblage de différentes banques de petites molécules. La comparaison de ces deux types d'inhibiteurs permet désormais d'envisager différentes optimisations par des approches de *design* de protéines, qui pourront conduire à l'amélioration de leurs capacités d'inhibition.

Chapitre 2 : Prédictions de sites fonctionnels par conservation différentielle

2.1. Introduction

Afin de mieux comprendre les mécanismes d'assemblage entre protéines, il est important d'identifier précisément les régions d'interaction et d'évaluer la contribution individuelle de chaque résidu dans les processus de reconnaissance. Les méthodes structurales telles que la cristallographie des rayons X ou la RMN constituent des techniques très puissantes pour obtenir ces informations. Cependant, elles sont parfois délicates à mettre en œuvre ou inadaptées pour certains complexes protéiques. Avec l'avènement des différents projets de séquençage à grande échelle, de nouvelles sources de données peuvent désormais orienter la sélection de résidus fonctionnellement importants, par l'analyse de séquences homologues à la protéine d'intérêt. En effet, les régions importantes pour la fonction d'une protéine, telles que les sites catalytiques ou les sites d'interactions avec les partenaires cellulaires, sont susceptibles d'être soumises à une pression de sélection particulière qui limite le taux de mutations dans ces sites.

Dans de chapitre, je m'appuierai sur une des approches évolutives utilisée pour quantifier la pression de sélection s'exerçant à la surface des protéines : l'analyse de conservation. Dans un premier temps, j'illustrerai le principe général à travers l'analyse d'une protéine que j'ai étudiée au cours de ma thèse : la protéine Nbs1. Je présenterai comment, en analysant de façon manuelle ou automatique un alignement de séquences de protéines homologues, on peut détecter certains sites fonctionnels au sein d'une protéine. L'analyse de cette protéine révèle l'existence de changements de fonctions pouvant être exploités pour identifier plus précisément certaines sous-familles d'homologues. Comme nous le verrons dans une seconde partie, ces variations fonctionnelles au sein d'une sous-famille d'homologues, peuvent permettre l'identification spécifique de sites d'interaction entre protéines. Deux exemples seront présentés, consacrés à l'étude des modes de régulation des protéines kinases, ou de certains résidus clés impliqués dans la reconnaissance spécifique de certains composés par la protéine MsbA.

2.2. Mise en évidence de régions fonctionnelles par analyse de conservation : la protéine Nbs1

2.2.1. La protéine NBS1

Le complexe MRN chez l'homme, comprenant les protéines Mre11, Rad50 et Nbs1 (Petrini and Stracker, 2003; van den Bosch et al., 2003; Zhang et al., 2006), joue un rôle essentiel dans différents processus cellulaires. En particulier, ce complexe intervient dans la détection des cassures doubles brins suite à un stress génotoxique en permettant le recrutement du complexe ATM (« *Ataxia Telangiectasia Mutated* ») (Kobayashi et al., 2004; Lee and Paull, 2005). Une des protéines du complexe MRN, la protéine Nbs1, est composée de 754 résidus et est constituée de différents domaines protéiques. Les outils de bioinformatique standard tels que SMART (Schultz et al., 1998) ou Pfam (Bateman et al., 2002), détectent dans la région N-terminale un domaine FHA suivi d'un domaine BRCT. Dans la région C-terminale, des expériences de biochimie ont identifiées deux sites de liaison à Mre11 (Desai-Mehta et al., 2001) et à la protéine ATM (Falck et al., 2005) (**Figure 26**). Une mutation de la protéine Nbs1 est à l'origine du syndrome de la cassure de Nijmegen ou « *Nijmegen Breakage Syndrome* » (NBS), une maladie rare et récessive associée à des caractéristiques cliniques telles que des microcéphalies, un faciès spécifique, un retard de croissance, un retard mental et une prédisposition aux cancers lymphoïdes (van der Burgt et al., 1996).

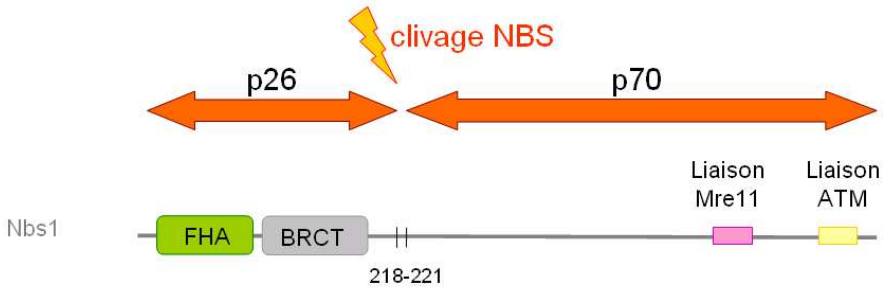


Figure 26. Composition en domaines de la protéine Nbs1 humaine. Quatre domaines protéiques peuvent être identifiés : un domaine FHA suivi d'un domaine BRCT du côté N-terminal, un domaine de liaison à Mre11 suivi d'un domaine de liaison à ATM du côté C-terminal. Chez les patients atteints de NBS, une délétion de 5 paires de bases entre les positions 218 et 221 clive la protéine en deux fragments : p26 et p70.

Récemment, en utilisant de nouvelles techniques permettant la comparaison de séquences à basse identité, Vincent Meyer a pu construire un alignement multiple de 25 séquences d'homologues lointains de la protéine humaine Nbs1. Cet alignement a permis la détection

d'un second domaine BRCT formant un tandem avec le premier (Becker et al., 2006) (**Figure 27**). L'identité de séquences dans cette région ne dépasse pourtant pas 10% entre les séquences de Xrs2 de levure et de Nbs1 humaine.

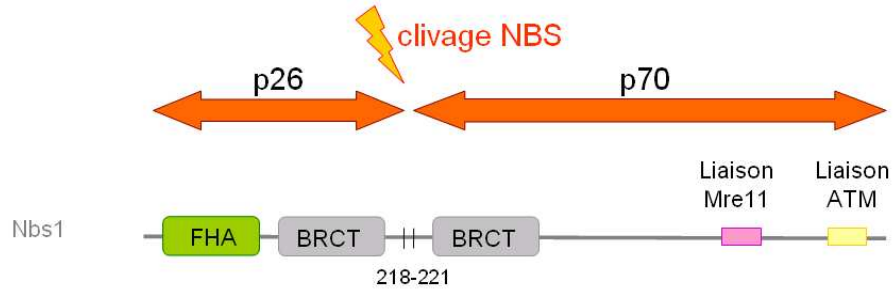


Figure 27. Composition en domaines de la protéine Nbs1 d'après (Becker et al., 2006). Cinq domaines protéiques peuvent être identifiés : un domaine FHA suivi d'un tandem de domaines BRCT du côté N-terminal, un domaine de liaison à Mre11 suivi d'un domaine de liaison à ATM du côté C-terminal. Chez les patients atteints de NBS, une délétion de 5 paires de bases entre les positions 218 et 221 clive la protéine en deux fragments : p26 et p70, précisément entre les deux domaines BRCT.

De façon intéressante, il apparaît que la mutation caractérisée chez les patients atteints du syndrome NBS, clive la protéine Nbs1 précisément entre les deux domaines BRCT, mettant en exergue le rôle fonctionnel de ce tandem. Ces données suggèrent que le clivage du tandem BRCT pourrait altérer la spécificité de reconnaissance des cibles de la protéine Nbs1, affectant ainsi la cascade de signalisation nécessaire à une réponse efficace aux dommages de l'ADN.

2.2.2. Mise en évidence de régions fonctionnelles au sein de la protéine Nbs1 par une analyse de conservation manuelle

En examinant plus précisément l'alignement des 25 séquences homologues de la protéine Nbs1, j'ai étudié si une analyse de la conservation des résidus pouvait permettre d'expliquer le rôle fonctionnel du tandem BRCT mis en évidence par Vincent Meyer.

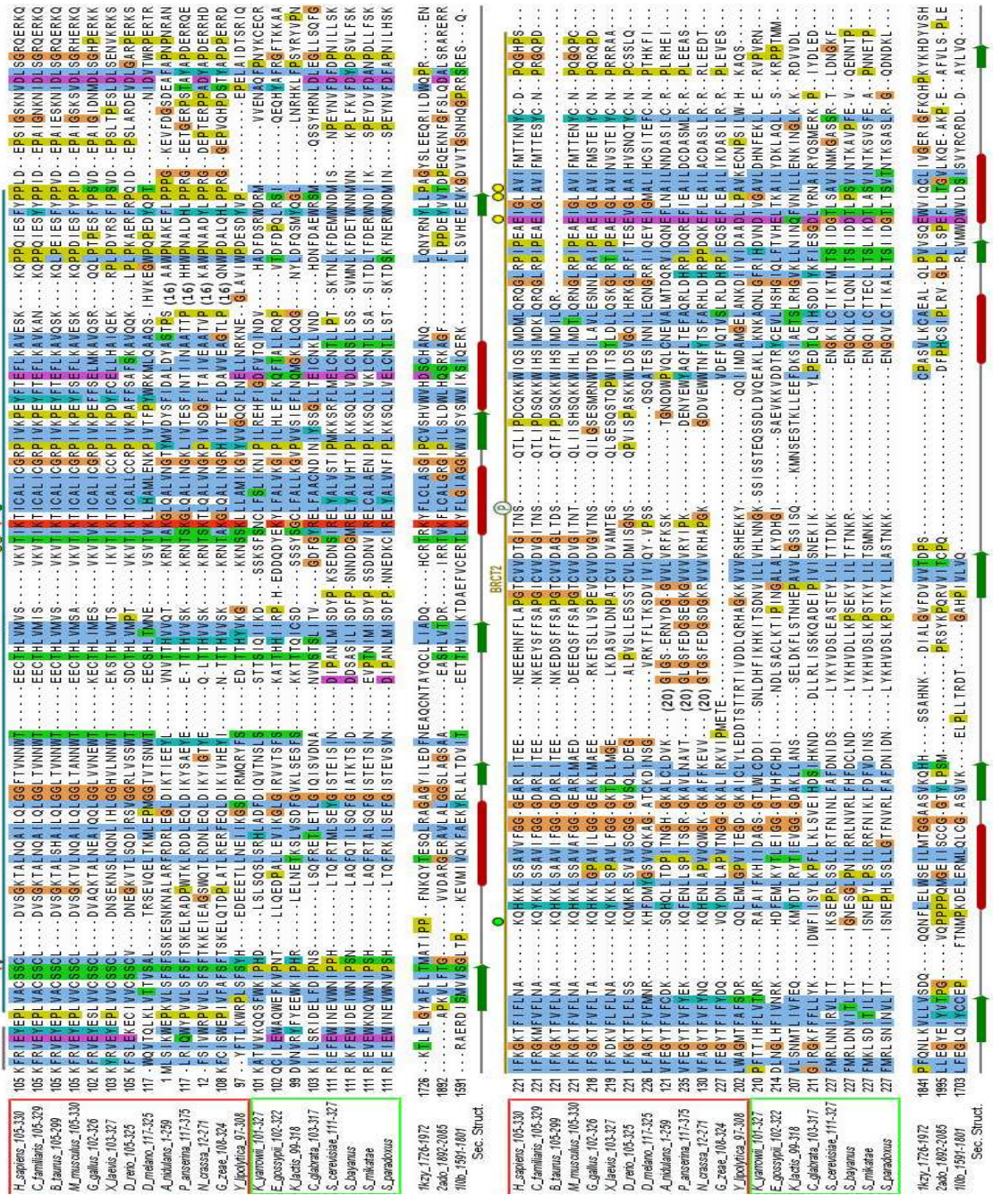


Figure 28. Alignement multiple des séquences des domaines BRCT de la protéine Nbs1. Deux sous-familles de séquences sont représentées par un rectangle rouge (séquences proches de la protéine humaine Nbs1), et un rectangle vert (séquences proches de la protéine de levure Xrs2, homologues de Nbs1 chez *S. cerevisiae*). Sur la ligne du bas sont représentées les structures secondaires (hélices en rouge, brins en vert) d'après les structures connues des tandems de domaines BRCT (1kzy, 2ado, 1l0b). Les triangles figurent les résidus en contact avec une sérine phosphorylée dans les structures connues (triangles rouges lorsque les contacts impliquent la chaîne latérale, triangles blancs lorsque les contacts se font par le squelette). Figure extraite de (Becker et al., 2006).

Différentes études ont montré que les tandems BRCT sont généralement impliqués dans la liaison des résidus phospho-sérines *in vitro* (Manke et al., 2003; Yu et al., 2003). Au niveau de la séquence, cette capacité à lier des phospho-sérines se caractérise par une signature spécifique, à savoir un motif [S/T-G] dans la première boucle du premier domaine BRCT, ainsi qu'un second motif [S/T-x-K] dans la seconde hélice du même domaine BRCT (Glover et al., 2004) (position des triangles rouges **Figure 28**). Concernant le premier motif, la glycine n'a probablement qu'un rôle mineur, car certains tandems de domaines BRCT ont la capacité de lier une phospho-sérine malgré la présence d'un autre résidu à cette position, comme c'est le cas du tandem BRCT de la protéine 53Bp1.

En analysant l'alignement de séquences (**Figure 28**), nous retrouvons un motif [S-C/F] dans la boucle β1/α1 ainsi que le motif [S/T-x-K] dans l'hélice α2 du premier domaine BRCT, strictement conservés, de l'homme à *Y. lipolytica*, ce qui suggère que ces protéines ont la capacité de lier des phospho-sérines. Pour des organismes plus distants, de *K. yarrowii* à *S. paradoxus* (**Figure 28**, **Figure 29**), certaines de ces positions, en particulier les positions pour lesquelles le contact avec la sérine phosphorylée se fait par la chaîne latérale (représentées par des triangles rouges sur la **Figure 28**) sont très conservées. Cependant, ces protéines ne respectent pas les motifs caractéristiques d'un site de liaison aux sérines phosphorylées. En particulier, pour la protéine Xrs2, homologue lointain de Nbs1 chez *S. cerevisiae*, le motif de la boucle β1/α1 est remplacé par un motif [P-P], celui de l'hélice α2 par un motif [S-x-R]. Cependant, certains indices laissent penser que le tandem BRCT de la protéine Xrs2 pourrait quand même lier des sérines phosphorylées. Premièrement, le domaine BRCT de la ligase III, capable de lier des sérines phosphorylées *in vitro*, contient une proline au lieu d'une sérine dans la boucle β1/α1 (Yu et al., 2003) comme c'est le cas pour Xrs2. Deuxièmement, une arginine au lieu d'une lysine est retrouvée dans le motif de l'hélice α2 chez la ligase IV, capable elle aussi de lier des sérines phosphorylées *in vitro* (Yu et al., 2003).

En analysant ainsi de façon manuelle l'alignement de séquences de protéines homologues à Nbs1, il apparaît que certains sites essentiels à la reconnaissance des sérines phosphorylées par le tandem BRCT, subissent certaines modifications dans leur composition, tout en gardant leurs fonctions. Ces sites correspondent à des positions conservées au sein de sous-familles de protéines, mais dont la nature des résidus varie entre ces mêmes sous-familles.

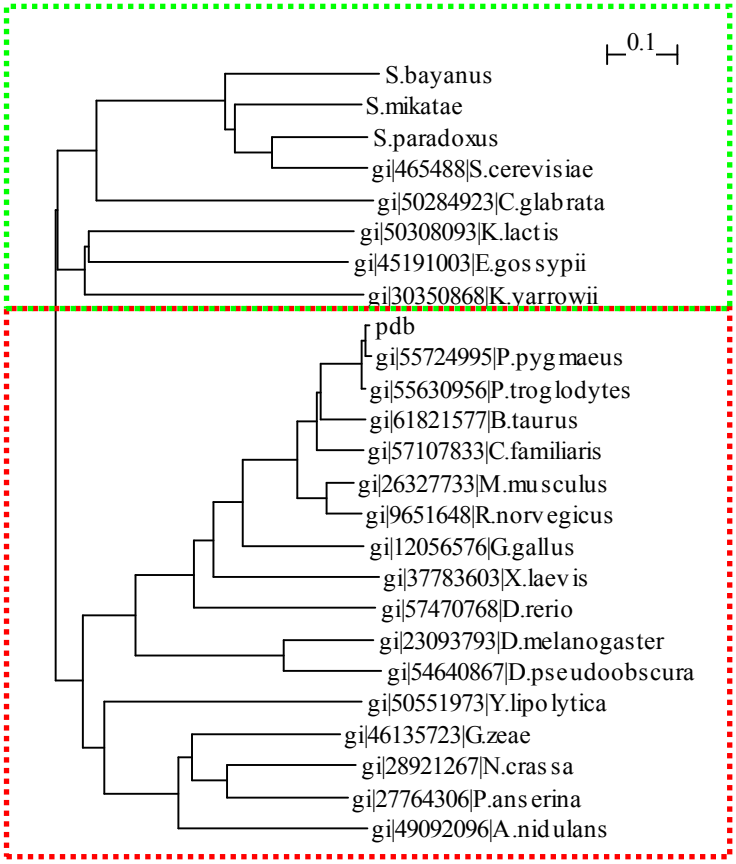


Figure 29. Arbre phylogénétique de la protéine Nbs1. Deux familles de séquences peuvent être identifiées. En vert, les séquences proches de la protéine de levure Xrs2 (*S. cerevisiae*), en rouge les séquences proches de la protéine humaine Nbs1. L'arbre phylogénétique a été construit avec le programme Clustalw (Thompson et al., 1994).

2.2.3. Mise en évidence de régions fonctionnelles au sein de la protéine Nbs1 par une analyse de conservation automatique : la méthode rate4site

Dans la section précédente, nous avons vu qu'en recoupant les informations issues des alignements de séquences, de la phylogénie des séquences, ainsi que de certaines données bibliographiques, il était possible de détecter des régions fonctionnelles importantes au sein de la protéine Nbs1. Cependant, une analyse manuelle telle qu'elle a été réalisée ne peut évidemment pas être envisagée à très grande échelle.

Certaines méthodes développées récemment, ont pour objectif de prédire ce type de région fonctionnelle de façon automatique (Landgraf et al., 2001; Lichtarge et al., 1996; Pupko et al., 2002). Parmi ces approches figure le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002), dont l'originalité consiste à intégrer l'histoire évolutive des séquences afin d'en inférer

les taux d'évolution par site. Pour cela, rate4site estime le taux d'évolution des positions d'un alignement multiple de séquences en considérant la topologie et la longueur des branches de l'arbre phylogénétique associé à ces séquences. Si on considère un alignement de séquences et deux arbres phylogénétiques possibles (**Figure 30**), la prise en compte de l'arbre phylogénétique des séquences peut se révéler être d'un intérêt majeur. Dans l'exemple de la **Figure 30**, la différence entre les arbres phylogénétiques 1 et 2 correspond à la longueur des branches associées aux deux dernières séquences. Sachant que la probabilité qu'une mutation survienne est plus forte sur des branches longues, le site analysé (pointé par une flèche) est donc plus conservé dans le scénario évolutif du second arbre que dans celui du premier.

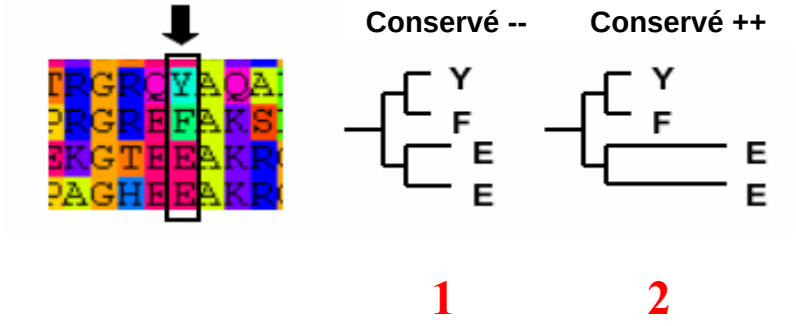


Figure 30. Influence de la phylogénie des séquences dans le calcul du score de conservation par le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002).

Comme nous l'avons illustré au travers de l'exemple de la protéine Nbs1, la prise en compte de l'histoire évolutive est essentielle, car elle permet la mise en évidence de sites fonctionnels apparemment variables si l'on considère l'alignement de séquence global, mais conservés si l'on analyse plus précisément les sous-familles de protéines représentées au sein de cet alignement de séquences. C'est pourquoi, j'ai souhaité tester la capacité du programme rate4site, à détecter ces régions fonctionnelles dans cette protéine, en particulier le site de liaison aux peptides phosphorylés mis en évidence par une analyse manuelle des séquences.

Après avoir calculé un score de conservation basé sur les taux d'évolution du programme rate4site pour chacune des positions de la protéine Nbs1, j'ai projeté ces scores de conservation à la surface d'un fragment de cette protéine, comprenant le tandem BRCT ainsi que le domaine FHA modélisé par E. becker (Becker et al., 2006) (**Figure 31**).

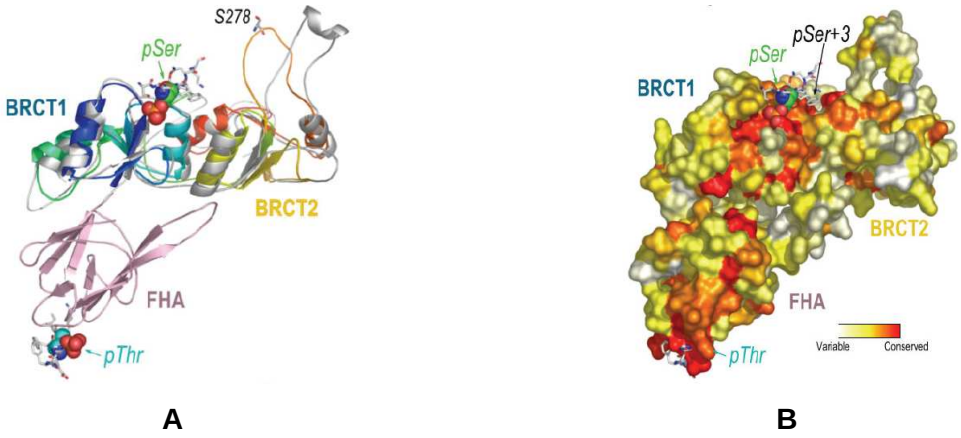


Figure 31. Modèle d'assemblage du complexe FHA – tandem BRCT de la protéine Nbs1. (A) Représentation en rubans du complexe. Les peptides phosphorylés, caractérisés dans les structures connues de domaines FHA et de tandems BRCT avec leur ligand, sont représentés sous forme de sphère (une phospho-sérine est reconnue par le premier domaine BRCT, et une phospho-thréonine est reconnue par le domaine FHA). (B) Conservation à la surface du complexe FHA – tandem BRCT. Les taux d'évolution ont été calculés par site avec le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002), et projetés sur la surface de la protéine. La couleur de chaque résidu va du blanc au rouge pour les positions des moins conservées aux plus conservées. Figure extraite de (Becker et al., 2006).

Nous pouvons constater de façon intéressante que les sites de liaison aux peptides phosphorylés apparaissent soumis à une pression évolutive très forte (peptides pThr et pSer représentés en sphères en contact de la surface rouge dans la **Figure 31**). Malgré les variations importantes en termes de motif de reconnaissance, le programme rate4site a réussi à détecter cette région comme fonctionnellement importante grâce à la prise en compte de l'arbre phylogénétique des séquences. En analysant par le programme rate4site le premier domaine FHA de la protéine Nbs1, il apparaît également que le motif de liaison aux thréonines phosphorylés caractéristique des domaines FHA est également très conservé (**Figure 31-B**). De façon assez surprenante, l'analyse de la structure du modèle d'assemblage entre le domaine FHA et le tandem de domaines BRCT, révèle la présence d'une région très conservée chevauchant le domaine FHA et le premier domaine BRCT (**Figure 31-B**). Cette région particulièrement étendue, pourrait correspondre à un site de liaison spécifique à la protéine Nbs1, car certains des résidus impliqués dans cette zone ne sont par ailleurs pas particulièrement bien conservés dans d'autres sous-familles de domaines FHA (**Figure 32-B**).

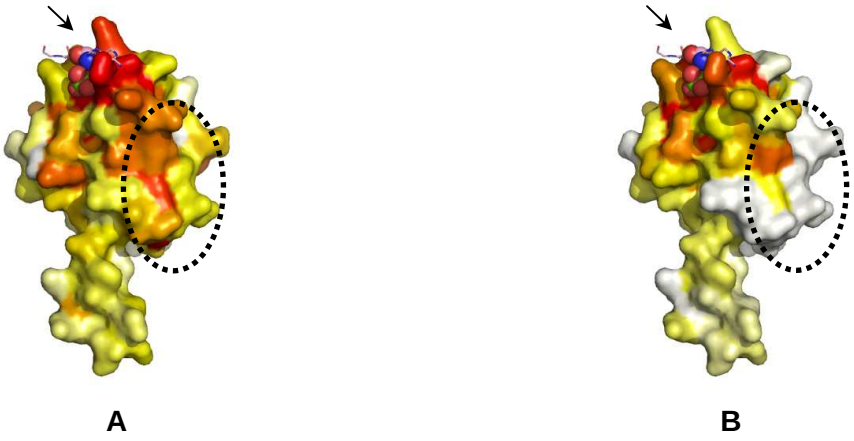


Figure 32. Conservation des domaines FHA. (A) Conservation du domaine FHA associé à un tandem BRCT. L'alignement de séquence contient 25 séquences homologues à la protéine humaine Nbs1. (B) Conservation des domaines FHA d'après l'alignement de séquences des domaines FHA proposé par la banque de domaines Pfam (Bateman et al., 2002). Les analyses de conservation ont été réalisées par le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002). La zone entourée par des pointillés comprend certains résidus conservés spécifiquement dans la famille de séquences homologues à la protéine Nbs1. Nous observons que le site de liaison aux thréonines phosphorylées caractéristique des domaines FHA, figuré par une flèche, est conservé dans les 2 cas.

2.2.4. Conclusion

En analysant la conservation de la protéine Nbs1, et en disséquant les différents modules qui la composent, il apparaît que certains motifs canoniques sont conservés au sein de cette protéine. Une des caractéristiques de certains de ces modules, en particulier le domaine FHA, est de reconnaître spécifiquement les peptides phosphorylés. Les sites de reconnaissance correspondant ont donc été particulièrement bien conservés au cours de l'évolution, résultat d'une pression évolutive forte concernant ces sites.

Cependant, dans le cas du domaine FHA et du tandem BRCT, certains sites apparaissent conservés, bien que ne correspondant pas à des motifs canoniques. Une des hypothèses est que ces sites pourraient correspondre à des sites d'interactions protéiques intra-moléculaires ou inter-moléculaires spécifiques aux domaines FHA suivis d'un tandem de domaines BRCT. Il y aurait alors des profils de conservation différentiels suivant les différentes combinaisons de modules, qui pourraient permettre la détection de certains sites de liaison, en particulier dans le cas des interactions intra-moléculaires. L'étude de certaines protéines particulièrement bien caractérisées sur le plan structural, les protéines kinases, peut

permettre de vérifier cette hypothèse selon laquelle des profils de conservation différentiels pourraient correspondre à des sites d'interactions protéiques.

2.3. Prédiction des sites d'interactions protéiques des protéines kinases par analyse différentielle de conservation

2.3.1. Introduction

Parmi les protéines intervenant dans la réponse cellulaire suite à une cassure double brin, les protéines kinases sont régulées de façon très stricte. Cette régulation peut être assurée par d'autres protéines partenaires dans le cadre de régulations inter-moléculaires, ou par des interactions intra-moléculaires pouvant impliquer différents types de domaines régulateurs. Ces mécanismes de régulations intra-moléculaires regroupent deux types d'interactions : (i) les interactions de type domaine – « *linker* » (régions non globulaires situées aux extrémités ou entre les domaines), (ii) les interactions de type domaine - domaine.

Par exemple, dans la kinase C-Src, des interactions intra-moléculaires entre le domaine SH3 et la boucle séparant les domaines SH2 et le domaine catalytique, ainsi qu'une interaction de type domaine-domaine entre les domaines SH2 et le domaine catalytique stabilisent la forme inactive de la protéine (**Figure 33**). Ce mécanisme d'autorégulation, particulièrement bien caractérisé sur le plan structural, permet d'intégrer des signaux cellulaires au niveau des sites de liaison des domaines SH2 et SH3 et d'activer le domaine kinase en fonction de la simultanéité de ces signaux. De nombreuses autres protéines kinases apparaissent aujourd'hui régulées par des processus similaires. D'autres protéines modulaires associant un domaine catalytique et des domaines de liaison ont également été décrits comme fonctionnant avec ces stratégies d'autorégulation (Bhattacharyya et al., 2006; Pawson, 2007).

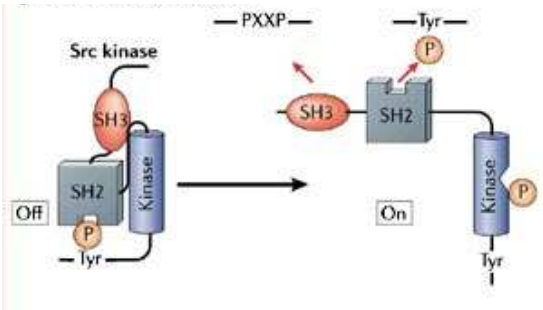


Figure 33. Régulation de la protéine kinase c-Src. La forme inactivée de la protéine est stabilisée par une interaction entre le domaine SH2 de la protéine, et une tyrosine phosphorylée (p-Tyr) présente au niveau de la queue C-terminale de la protéine. Cette conformation est également stabilisée par des interactions intra-moléculaires entre le domaine SH2 et le domaine kinase, ainsi qu'une interaction entre la boucle séparant les domaines SH2-kinase et le domaine SH3. La protéine bascule vers une forme activée par une déphosphorylation de la queue C-terminale de la protéine, ou par la présence de ligands activateurs (tels que des peptides riches en prolines, ou contenant des p-Tyr). Figure adaptées de (Seet et al., 2006).

Dans le cas de la protéine C-Src, un domaine de type SH2 intervient dans la stratégie d'auto-inhibition de cette protéine. Si l'on émet l'hypothèse qu'une pression de sélection forte s'exerce à l'interface entre le domaine SH2 et le domaine catalytique qui interagissent (le domaine catalytique est ici appelé domaine kinase), on peut s'attendre à ce que cette interaction laisse une trace évolutive à la surface du domaine SH2, potentiellement détectable par des analyses de conservation. Pour vérifier cette hypothèse, j'ai appliqué une analyse de conservation sur le domaine SH2, après avoir récupéré des séquences homologues de la protéine C-Src dans une base de données de séquences (section 7.1).

La **Figure 34** montre que différentes régions sont conservées au sein du domaine SH2 de la protéine C-Src. En particulier, le site de liaison des peptides phosphorylés représente la région la plus conservée du domaine (**Figure 34-A**), la capacité de lier des peptides phosphorylées étant une caractéristique commune des domaines de type SH2.

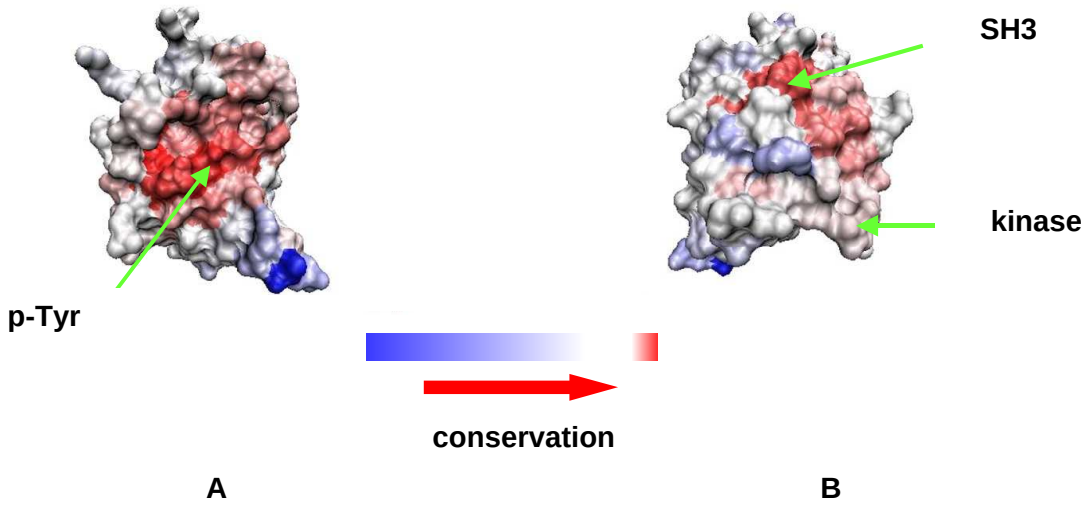


Figure 34. Identification des régions conservées du domaine SH2 de la protéine C-Src par le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002). Les régions mises en évidence par conservation sont : (A) le site de reconnaissance des tyrosines phosphorylées (p-Tyr), (B) les sites d'interaction aux domaines SH3 et au domaine kinase.

En outre, le site de liaison au domaine SH3 apparaît lui aussi bien conservé au cours de l'évolution (**Figure 34-B**). Cependant, le site d'interaction avec le domaine kinase est relativement peu conservé en comparaison des sites de liaison au domaine SH3 ou aux peptides phosphorylés. Dans ce contexte, la question posée est la suivante : comment détecter de façon spécifique un site de liaison particulier par l'analyse des taux d'évolution de chaque position de ce domaine ?

Le domaine SH2 est assez répandu dans les protéines eucaryotes. Compte tenu de l'aspect modulaire des protéines, ce domaine peut être retrouvé associé à d'autres domaines protéiques que les domaines kinase. Notre hypothèse est qu'en analysant de façon différentielle les taux d'évolution au sein de domaines SH2 associés ou non à des domaines kinases, il est possible que le site d'interaction impliquant les domaines SH2 et kinase dans la protéine C-Src apparaisse de façon plus spécifique.

En tenant compte de l'ensemble de ces considérations, j'ai développé un protocole d'analyse de conservation différentielle, applicable à la prédiction des sites d'interactions intra-moléculaires impliquant des domaines protéiques régulateurs. Après avoir introduit brièvement le principe de la conservation différentielle, je présenterai les résultats de cette approche sur la classe des protéines kinases pour lesquelles les stratégies d'auto-inhibition ont été le mieux caractérisées expérimentalement.

2.3.2. Principe général de la conservation différentielle

Le principe de la conservation différentielle, décrit à la **Figure 35**, vise à déterminer de façon spécifique le site d'interaction entre deux domaines protéiques. Dans le cas d'un mécanisme d'auto-inhibition d'un domaine catalytique de type kinase par un domaine régulateur, le principe est de séparer en deux sous-ensembles les séquences protéiques comprenant les deux domaines et les séquences ne comprenant que le domaine régulateur associé à d'autres domaines. Après un alignement des séquences, les scores de conservation sont déterminés pour les deux sous-ensembles de séquences par le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002). Lors d'une dernière étape, un taux de conservation différentiel est calculé pour chaque site par différence des taux d'évolution au niveau des deux sous-ensembles de séquences (voir section 7.1.3).

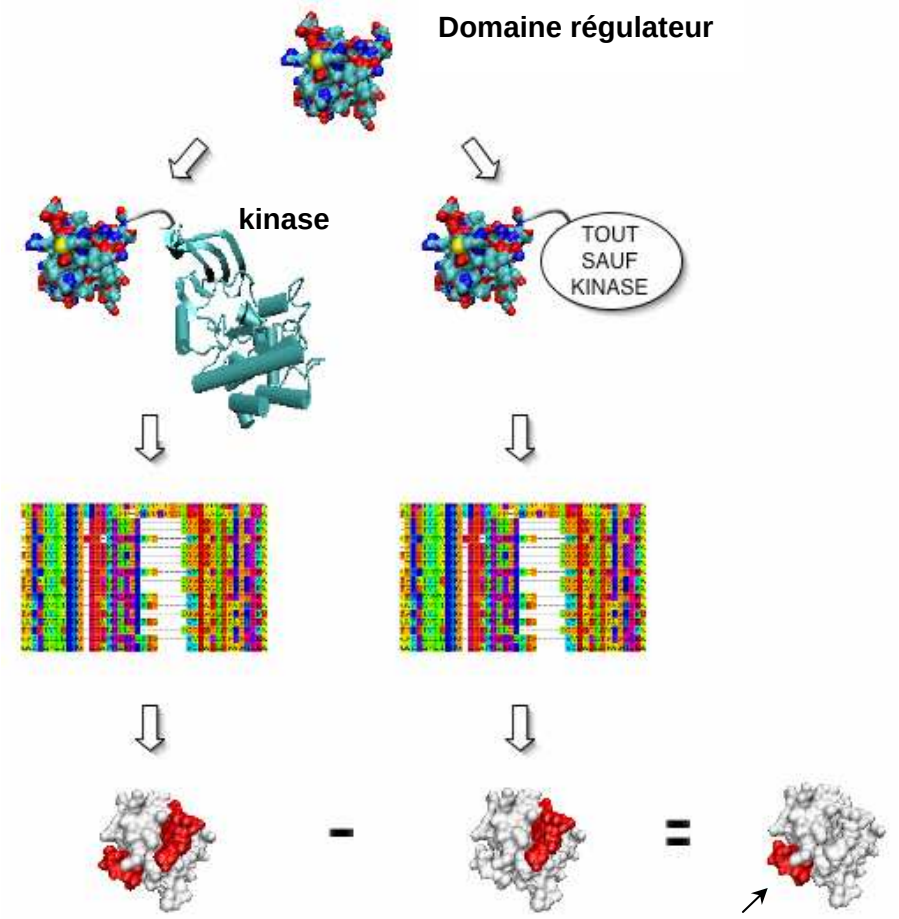


Figure 35. Application du principe de conservation différentielle au cas des domaines régulateurs de protéines kinase. Dans un premier temps, les séquences homologues au domaine régulateur d'intérêt sont recherchées dans une base de données de séquences. Deux pools de séquences sont constitués en fonction de la présence unique du domaine régulateur, ou de la présence simultanée de ce domaine et d'un domaine kinase. Après alignement des deux pools de séquences correspondant, la conservation est calculée par le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002) pour chacun des sites en considérant les deux alignements de séquences (les résidus conservés sont représentés en rouge), puis la différence de conservation est quantifiée par conservation différentielle (les résidus différentiellement conservés sont pointés d'une flèche).

2.3.3. Prédiction des sites d'interactions intra-moléculaires par conservation différentielle : le cas des domaines régulateurs de kinases

Afin de détecter la région impliquée dans l'interaction entre un domaine régulateur de kinase et son domaine catalytique, je me suis intéressé à cinq structures de protéines kinases auto-régulées par des interactions de type domaine-domaine et dont la structure a été résolue (**Tableau 6**). Parmi ces 5 structures, un des cas ne permet pas d'envisager une étude par conservation différentielle : le récepteur TGF-Beta. En effet, son domaine régulateur, le domaine GS, n'est retrouvé qu'en association à des protéines kinases.

Protéine	PDB	Domaine(s) régulateur(s)
C-Src	1fmk	SH3 (82-140), SH2 (148-246)
Récepteur Beta-adrenergique	1omw	RGS (30-185)
PAK1	1f3m	IS (87-136)
Twitchin	1koa	IG (6263-6356)
Récepteur TGF-Beta	1b6c	GS (176-204)

Tableau 6. Protéines kinases régulées par des interactions intra-moléculaires de type domaine-domaine. L'identifiant PDB ainsi que la position des domaines régulateurs au sein de la protéine sont indiqués respectivement dans les 2ème et 3ème colonnes.

En appliquant le principe de conservation différentielle au cas du domaine SH2 de la protéine kinase C-Src, deux résidus sont mis en évidence comme spécifiquement conservés dans les protéines kinases : K_{152}, et R_{160}. De façon intéressante, Le résidu R_{160} est situé à l'interface entre le domaine régulateur, et le domaine kinase (**Figure 36**).

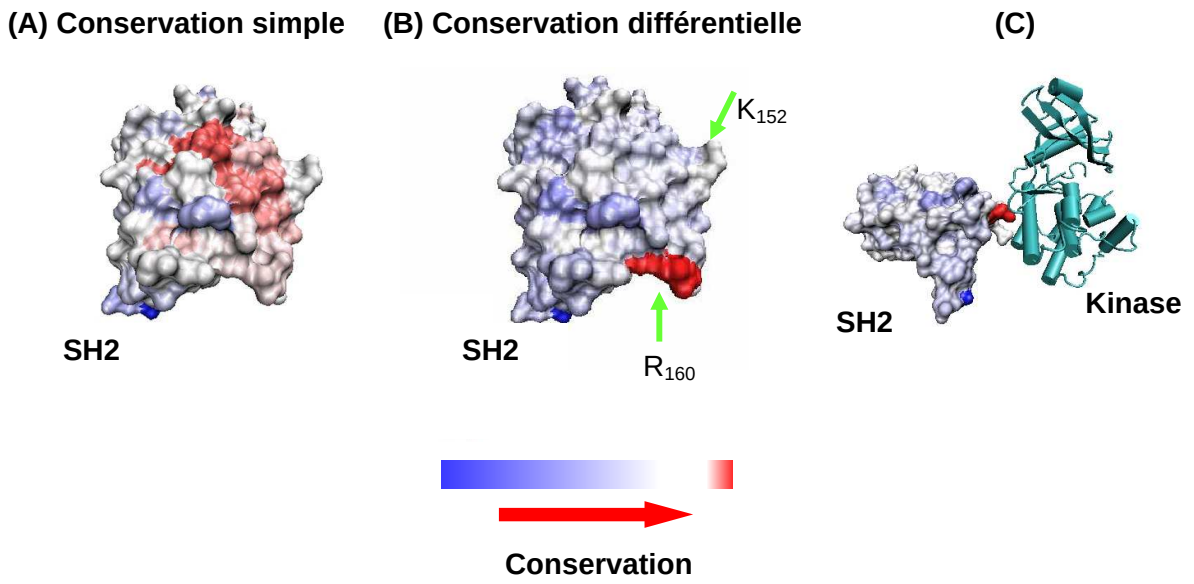


Figure 36. Application de la conservation différentielle au domaine SH2 de la proteine kinase C-Src. (A) conservation des domaines SH2 associés à un domaine kinase. (B) Conservation différentielle des domaines SH2 associés à un domaine kinase. Les résidus mis en évidence par ce procédé, R_{160} et K_{152} sont figurés par une flèche. (C) Interaction entre le domaine SH2 et le domaine catalytique de la protéine kinase C-src (code PDB : 1fmk). Le domaine SH2 est représenté sous forme de surface, le domaine catalytique sous forme de rubans.

Pour les quatre complexes intra-moléculaires, j'ai analysé les résidus de l'interface (un résidu est à l'interface entre 2 protéines si sa surface accessible est modifiée par l'association de ces deux partenaires). La **Figure 37** présente les résultats de cette analyse de conservation différentielle appliquée aux domaines régulateurs de kinases. Pour ces résidus, la sensibilité de la prédiction est en moyenne de 17 % ce qui indique qu'un nombre restreint de résidus

présents à l'interface est détectable par cette approche. En revanche, 67% en moyenne des résidus prédits se situent effectivement à l'interface. Parmi les résidus prédits, ceux qui sont effectivement situés aux interfaces de complexes ont tendance à se grouper dans des régions particulières de la surface. En couplant l'information de proximité à la surface avec la prédiction par conservation différentielle, il devient possible d'améliorer encore la précision des prédictions. Dans la suite de ce chapitre, nous avons souhaité tester cette stratégie de conservation différentielle à d'autres types de fonctions biologiques.

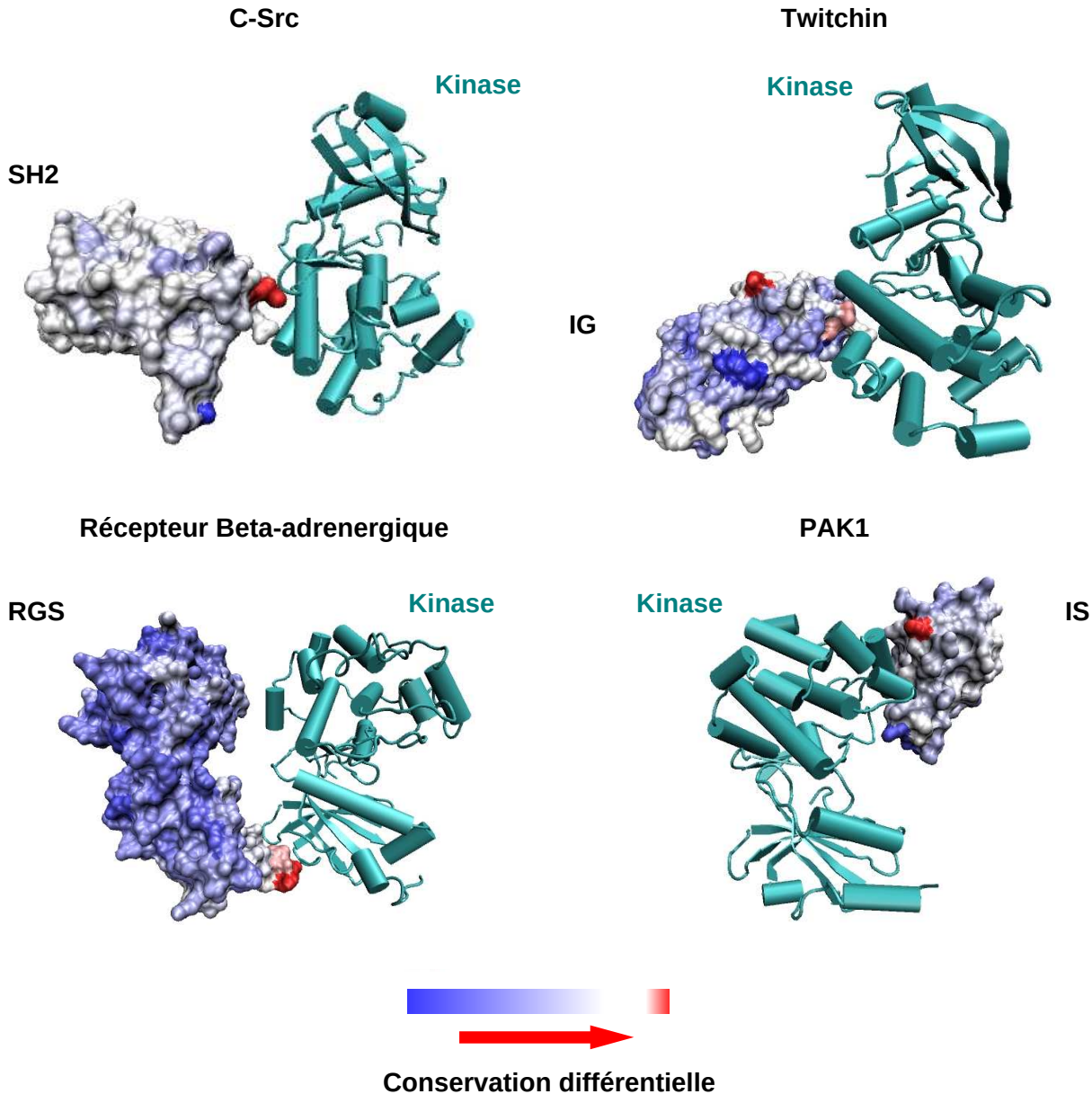


Figure 37. Application de la conservation différentielle aux domaines régulateurs de kinase. Les domaines régulateurs sont représentés sous forme de surface, les domaines catalytiques des protéines kinases sont représentés sous forme de rubans.

2.4. L'analyse différentielle ou comment prédire des changements de spécificité fonctionnelle au sein des protéines : la protéine MsbA

2.4.1. La protéine P-gp et son homologue MsbA

Une collaboration avec l'équipe de S. Orlowski au CEA de Saclay nous a fourni un exemple d'étude tout à fait intéressant, celui des protéines de la famille des ABC transporteurs, responsables chez l'homme de l'efflux des drogues anti-cancéreuses et de la plupart des résistances aux chimiothérapies. Une des protéines étudiées dans cette équipe est la P-glycoprotéine humaine (MDR1 ou P-gp), protéine membranaire, impliquée dans la réponse des cellules à différents stress chimiques. C'est elle qui est le principal acteur dans le phénomène de résistance multi-drogues (MDR) observé dans les cellules cancéreuses. La P-gp est responsable du transport actif d'agents anti-tumoraux vers l'extérieur de la cellule, ce qui empêche leur accumulation.

En 2001, la structure cristallographique d'un homologue bactérien (identité de séquence avec MDR1 de l'ordre de 35 %), la protéine MsbA d'*E. Coli*, a été résolue par cristallographie des rayons X (Chang and Roth, 2001). La structure de MsbA se présente comme un homodimère, chaque sous-unité étant composée d'une partie transmembranaire composée de six hélices, et d'un domaine nucléotidique (NBD), site de fixation de l'ATP. MsbA intervient dans le transport vers l'extérieur des membranes procaryotes de certains lipides, comme le lipide A (Doerrler et al., 2001) ou de drogues (Reuter et al., 2003; Woebking et al., 2005). Cette homologie fonctionnelle fait de la protéine MsbA un modèle bactérien de choix pour l'étude de la P-gp (Higgins and Gottesman, 1992) (Romsicki and Sharom, 2001).

2.4.2. Analyse par conservation différentielle

La protéine MsbA représente un cas intéressant pour étudier l'évolution fonctionnelle des protéines de type ABC transporteur, en particulier par des analyses différentielles de conservation. Afin de mieux comprendre la façon dont a évolué la fonction de MsbA, j'ai donc procédé à une analyse évolutive de cette protéine (Haubertin et al., 2006).

Après récupération d'homologues proches de la protéine MsbA par une procédure Psi-Blast (Altschul et al., 1997) appliquée à la banque de données de séquences « nr », j'ai constitué

deux groupes de séquences : (i) A_1, restreint aux homologues fonctionnels de MsbA, comprenant la protéine MsbA d'*E. Coli* ainsi que des homologues proches de cette protéine (42% d'identité moyenne avec *E.coli* MsbA), (ii) A_2 comprenant des homologues lointains de la protéine MsbA d'*E. Coli* correspondant à des protéines ABC transporteurs récupérées par la procédure Psi-Blast (20% d'identité moyenne avec *E.coli* MsbA).

J'ai ensuite procédé à une analyse ce conservation du groupe de séquences A_1 et A_2 (Haubertin et al., 2006). La **Figure 38** présente les résultats de cette analyse de conservation simple appliquée aux homologues proches de la protéine MsbA. Il apparaît que les régions de la protéine les plus conservées sont localisées à l'intérieur de la cavité du dimère et à la charnière entre chaque monomère.

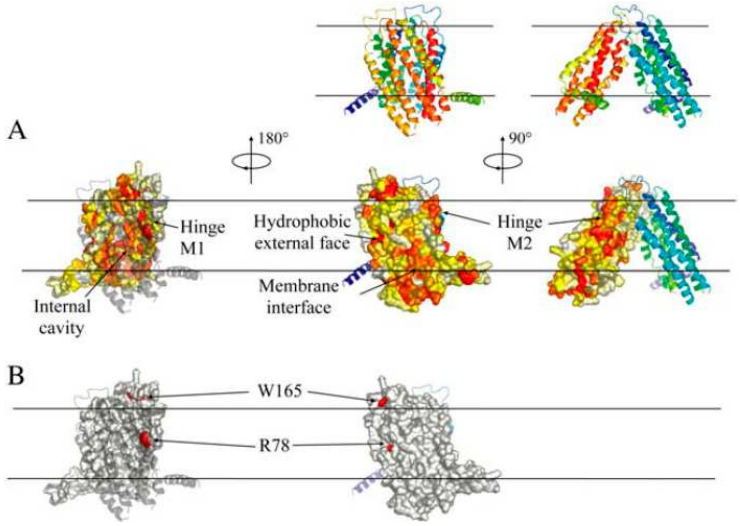


Figure 38. Conservation de la protéine E. Coli MsbA. (A) Conservation de la protéine MsbA. La protéine MsbA est représentée sous forme de surface et colorée en fonction de sa conservation (l'alignement considéré pour cette analyse est l'alignement de séquences A_1, homologues proches de E. Coli MsbA). La couleur de chaque résidu va du blanc au rouge pour les positions des moins conservées aux plus conservées. (B) Conservation différentielle de la protéine MsbA. Deux résidus sont mis en évidence par conservation différentielle, et sont représentés en rouge : R_{78} et W_{165}. Figure extraite de (Haubertin et al., 2006).

Dans un second temps, j'ai procédé à une analyse de conservation différentielle de la protéine. J'ai pour cela analysé de façon indépendante chacune des positions des deux alignements correspondant aux sous-groupes A_1 et A_2. En comparant les intervalles de confiance à 95% des taux d'évolution de chaque position pour les deux sous-groupes, j'ai extrait les positions dont le taux d'évolution est significativement plus lent dans le groupe 1 que dans le groupe 2 (Haubertin et al., 2006). En suivant cette procédure, deux résidus sont

sélectionnés comme potentiellement importants pour rendre compte de la spécificité de fonction du premier sous-ensemble d'ABC transporteurs : R_{78} et W_{165}.

En parallèle de cette étude, David Haubertin a procédé à une analyse de la protéine MsbA par des techniques de dynamique moléculaire. Il a pour cela réalisé une simulation de 50ns de la partie transmembranaire de MsbA dans une bicouche lipidique, et en solvant explicite. Son analyse a révélé une conformation particulière de MsbA, stabilisée par deux phospholipides qui interagissent avec la région charnière de la protéine (Haubertin et al., 2006). L'analyse de la dynamique moléculaire a mis en évidence que deux résidus clés participent à la stabilisation de cette conformation en interagissant avec ces phospholipides : R_{78} et T_{81} (Haubertin et al., 2006).

Rappelons que l'analyse par conservation différentielle avait identifié le résidu R_{78} comme soumis à une pression évolutive particulière dans les protéines homologues proches de MsbA. La comparaison entre ces deux analyses, phylogénétique et structurale, suggère que la région charnière pourrait être importante pour la reconnaissance spécifique du lipide A pour les protéines de la famille MsbA. Ce même type d'analyses est actuellement en cours pour comparer entre eux d'autres transporteurs de la famille ABC. En particulier, certaines familles de transporteurs, homologues à la protéine MRP1, semblent posséder un site de liaison au glutathion important pour leur fonction qui pourrait être identifié par ce type d'approches prédictives (Cole and Deeley, 2006). Des collaborations entre les équipes de S. Orlowski et M. Vivaudou devraient également permettre d'analyser les spécificités fonctionnelles de différents membres des transporteurs ABC tels que les protéines SUR (Moreau et al., 2000) et de comparer les prédictions effectuées aux nombreux mutants fonctionnels décrits dans la littérature (Shilling et al., 2006).

2.5. Discussion, Conclusions et Perspectives

L'analyse des taux d'évolution spécifiques pour chaque résidu d'une protéine s'avère une approche performante pour identifier les pressions de sélection exercées à la surface des protéines. Par les différents exemples présentés, nous avons pu montrer l'importance de la prise en compte de la phylogénie des séquences pour la mise en évidence de ces régions fonctionnelles au sein d'une protéine. D'autres méthodes alternatives au programme rate4site, intégrant cette notion, ont été proposées dans la littérature en particulier l'analyse de la trace évolutive (appelée ET pour « *Evolutionary Trace* ») développée par le groupe de O. Lichtarge (Madabushi et al., 2002; Yao et al., 2003). Cette approche semble donner des résultats significatifs quand les positions au sein d'un alignement sont relativement bien conservées. Cependant, en comparaison à l'algorithme implémenté dans rate4Site, cette méthode n'est pas assez sensible pour mettre en évidence des régions soumises à des pressions de sélection plus faibles, comme c'est le cas du site d'interaction intra-moléculaire du domaine SH2 pour le domaine kinase de la protéine C-Src.

Dans ce chapitre, j'ai montré que les taux d'évolutions calculés par le programme rate4site pouvaient être combinés par une analyse différentielle afin de mettre en évidence des propriétés fonctionnelles spécifiques au sein d'un sous-ensemble d'homologues. Le potentiel d'applications de cette stratégie est vaste puisqu'elle peut s'appliquer pour la prédiction des sites d'interactions entre domaines protéiques mais également pour la prédiction des sites d'interaction entre les protéines et d'autres molécules telles que les lipides, les acides nucléiques ou des petits métabolites. D'autres approches publiées plus récemment ont également pour objectif de détecter ces variations de spécificités fonctionnelles au travers des alignements de séquences. Pour cela, les changements de fonctions au niveau de sous-familles de protéines peuvent être prédits en identifiant deux types de positions au sein d'un alignement multiple de séquences : les CSS (« *Conservation-Shifting Sites* ») ou RSS (« *Rate-Shifting Sites* ») (Abhiman and Sonnhammer, 2005). Les CSS correspondent à des positions conservées au sein de sous-familles de protéines, mais dont la nature des résidus varie entre ces mêmes sous-familles. Les RSS correspondent quand à eux à des positions dont les taux d'évolution entre deux sous-groupes d'une famille protéique diffèrent de façon significative. Les calculs de CSS et RSS ont été utilisés à grande échelle pour détecter les résidus fonctionnellement importants et pour définir des sous-familles au sein de familles de

domaines protéiques issus de la banque de domaines Pfam (Bateman et al., 2002). Ces données sont répertoriées au sein de la base de données FunShift (Abhiman and Sonnhammer, 2005). L'utilité de la base FunShift pour l'analyse de domaines tels que les domaines SH2 se révèle néanmoins décevante car la granulosité de la classification est trop importante. À titre d'exemple, pas moins de 141 sous-familles de domaines SH2 sont répertoriées dans la base FunShift et de nombreuses sous-familles ne regroupent que 4 séquences très similaires, ce qui ne permet donc pas d'envisager des stratégies d'analyses évolutives sur la base de ces données.

De façon générale, le découpage en sous-ensembles de séquences homologues présentant des variations de spécificités fonctionnelles, n'est pas toujours possible ou facile à effectuer. En effet, comme nous l'avons vu au cours de ce chapitre, une des protéines kinases ne peut être analysée par conservation différentielle car son domaine régulateur (le domaine GS du récepteur TGF-Beta) n'est jamais associé à des domaines autres que le domaine kinase. De plus, si la variation fonctionnelle implique la même région dans chacun des deux sous-ensembles, il ne sera pas possible de la mettre en évidence par des analyses différentielles. Dans la suite de ma thèse, j'ai donc cherché à développer des approches qui ne souffrent pas de ces limitations. Ces approches seront présentées dans les deux chapitres suivants.

Chapitre 3 : Analyse des propriétés évolutives des interfaces protéiques intra-moléculaires

3.1. Introduction

D'un point de vue évolutif, différentes études ont montré que la structure des interfaces de complexes protéiques est bien conservé entre des homologues lointains, dont le pourcentage d'identité peut descendre jusqu'à 30% (Aloy et al., 2003; Kim and Ison, 2005). Un taux de conservation en séquence aussi bas suggère que les interfaces de complexes ont la capacité d'évoluer, tout en maintenant une certaine spécificité entre les partenaires protéiques. Pour rendre compte de cette conservation structurale, on peut imaginer qu'au cours de leur évolution, les interfaces de complexes protéiques ont subi certaines pressions de sélection conduisant à des empreintes spécifiques.

Dans le chapitre précédent, nous avons vu que les propriétés évolutives des protéines conduisent souvent à des signatures caractéristiques à leurs interfaces, et qu'on pouvait détecter ces signatures par des analyses de conservation des résidus. Parmi une autre hypothèse évolutive pouvant expliquer la conservation structurale des assemblages protéiques, la co-évolution des partenaires en interaction est l'une des théories qui a été la plus soutenue ces 10 dernières années. En se basant sur cette théorie, certaines études ont montré qu'il était possible de prédire les sites d'interaction protéiques, en détectant certaines mutations concertées impliquant des résidus proches dans l'espace (Pazos et al., 1997). Dans ces approches, ce n'est donc plus la recherche de sites conservés mais de sites variables qui permet la détection des sites d'interaction protéiques.

Dans ce chapitre, je m'intéresserai tout d'abord aux capacités prédictives de ces deux types d'approches conservation et mutations concertées, pour identifier les sites d'interaction protéiques. Afin d'évaluer les propriétés évolutives des partenaires en interaction, j'ai choisi d'analyser une base de données d'interactions intra-moléculaires, pour lesquelles les domaines en interaction ont évolué au sein d'une même séquence. Dans le cas où les domaines protéiques en interaction auraient évolué au sein de différentes protéines, il est en effet plus délicat de s'assurer de leur relation d'orthologie[1]. Dans un second temps, je présenterai une approche alternative pour rendre compte des phénomènes de co-évolution à l'interface des complexes protéiques. Cette nouvelle stratégie d'analyse possède des

[1] Des orthologues sont des groupes de gènes ou de protéines de différents organismes qui ont la même fonction.

implications générales pour la prédiction des sites d'interactions protéiques que je discuterai par la suite.

3.2. Constitution d'une base de données de complexes intra-moléculaires

À partir de la banque de données PSIMAP, regroupant les interactions entre domaines protéiques caractérisés sur le plan structural (Kim et al., 2004), j'ai construit une base de données de complexes intra-moléculaire afin d'en étudier les propriétés évolutives. Dans sa version initiale, la base de données PSIMAP contenait au total 12758 interactions de type intra-moléculaire. Dans un premier temps, j'ai extrait un sous-ensemble de 319 interactions non redondantes. J'ai par la suite développé un outil automatique permettant pour chacun des deux partenaires de chaque complexe, de récupérer des séquences protéiques homologues pour différentes espèces. Ces séquences ont été alignées à l'aide de programmes d'alignements de séquences.

En ne considérant que les assemblages protéiques pour lesquels le nombre de protéines homologues disponibles dans les bases de données de séquences est suffisant pour entreprendre l'étude de leurs propriétés évolutives, 124 cas ont été retenus. J'ai ensuite éliminé les complexes dont la surface de contact était inférieure à 700 $Å^2$, susceptibles de correspondre à des contacts cristallins non spécifiques. Au final, un ensemble de 86 complexes intra-moléculaires a été sélectionné.

3.3. Conservation des interfaces de complexes intra-moléculaires

L'analyse de conservation est classiquement utilisée afin de détecter d'éventuelles pressions de sélection à la surface des protéines. En utilisant le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002) décrit dans le chapitre précédent, j'ai analysé la base de données de 86 complexes intra-moléculaires extraits de la banque de données PSIMAP (Kim et al., 2004). Afin de préciser l'état de conservation de chaque résidu, leurs taux d'évolution ont été normalisés pour définir 3 classes de résidus : (i) résidus conservés, (ii) résidus moyennement conservés, (iii) résidus divergents (Madaoui and Guerois, 2007).

Sur la base de la structure des 86 complexes de domaines protéiques, j'ai déterminé la proportion de résidus conservés par rapport à l'ensemble des résidus situés à l'interface. Cette information mesure la sensibilité de l'analyse de conservation pour la détection des sites d'interaction. En représentant la distribution de cette sensibilité pour l'ensemble des complexes analysés, nous pouvons constater qu'une grande partie des résidus situés à l'interface de complexes intra-moléculaires sont conservés (**Figure 39-A**). En effet, plus d'un tiers de l'interface apparaît conservée pour la très grande majorité des cas traités.

Cependant, cette analyse ne renseigne pas sur la spécificité de l'analyse de conservation. C'est pourquoi j'ai également évalué cette spécificité, en rapportant le nombre de résidus conservés d'une interface donnée à l'ensemble des résidus conservés à la surface de chacun des partenaires (**Figure 39-B**). De façon intéressante, nous pouvons observer que la distribution est relativement étalée. La spécificité de conservation peut varier de 0 à 80 %, avec un nombre majoritaire de cas pour lesquels elle varie entre 15 et 50%. Ce résultat traduit la faible spécificité de l'analyse de conservation pour la détection de sites d'interactions intra-moléculaires. Elle peut s'expliquer par la présence d'autres régions conservées, qui ne sont pas situées à l'interface des deux partenaires en interactions. Ces régions peuvent correspondre aussi bien à des sites actifs, certains motifs structuraux, ou encore à d'autres sites d'interaction alternatifs avec d'autres partenaires protéiques (interactions qui peuvent être aussi bien intra-moléculaire qu'inter-moléculaire).

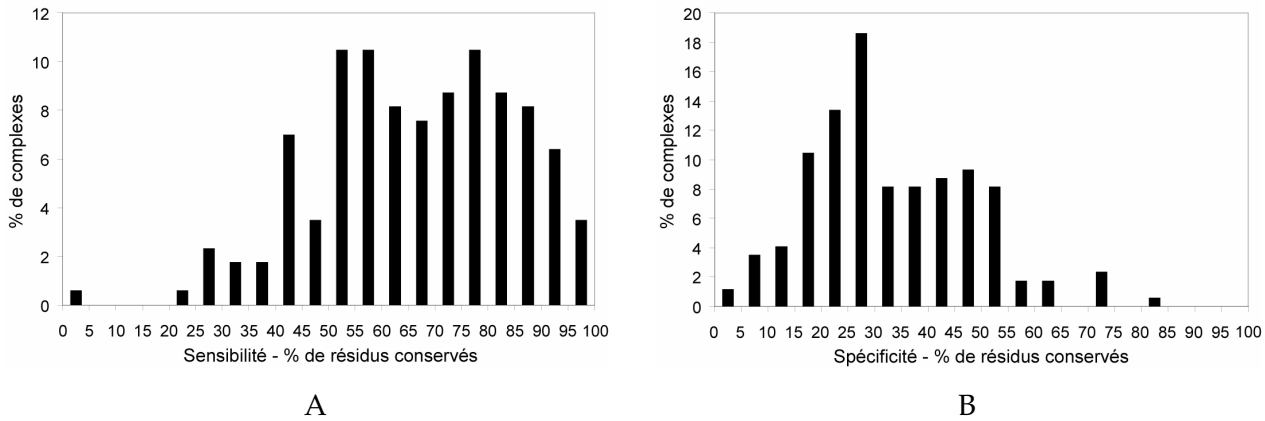
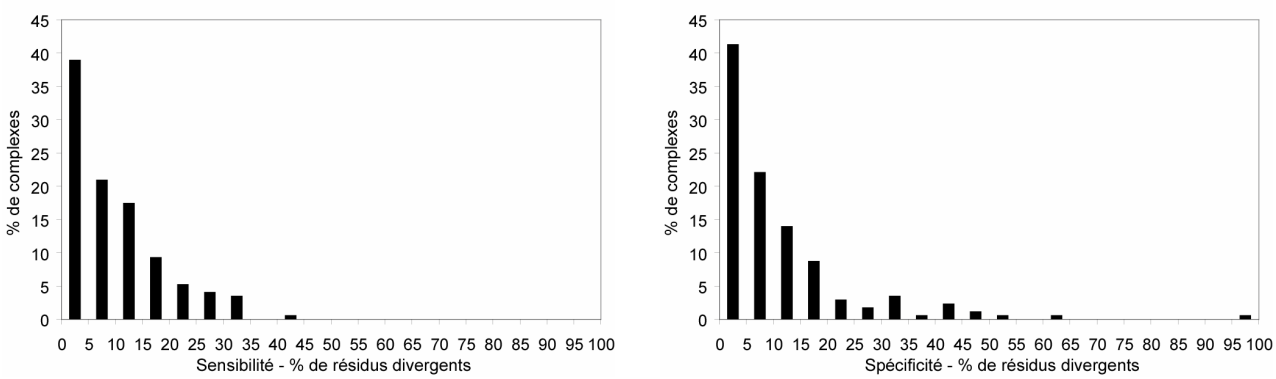


Figure 39. Conservation à l'interface des complexes intra-moléculaires. (A) Sensibilité : pourcentage de résidus conservés en fonction du nombre de résidus à l'interface des complexes. (B) Spécificité : pourcentage de résidus conservés à l'interface des complexes en fonction du nombre de résidus conservés sur l'ensemble de la surface. Figure adaptée de (Madaoui and Guerois, 2007).

Dans un second temps, j'ai analysé la répartition des résidus les plus variables à la surface des domaines en interaction. L'objectif était d'évaluer si les résidus divergents sont exclus des zones d'interaction protéiques. La **Figure 40** montre que les résidus divergents sont minoritaires à l'interface des complexes protéiques intra-moléculaires (**Figure 40-A**) et qu'ils sont situés de façon prédominante à l'extérieur des interfaces de complexes (**Figure 40-B**).



A B

Figure 40. Divergence à l'interface des complexes intra-moléculaires. (A) Sensibilité : pourcentage de résidus divergents en fonction du nombre de résidus à l'interface des complexes. (B) Spécificité : pourcentage de résidus divergents à l'interface des complexes en fonction du nombre de résidus divergents sur l'ensemble de la surface. Figure adaptée de (Madaoui and Guerois, 2007).

Cette analyse montre que les propriétés de conservation des résidus situés aux interfaces de complexes intra-moléculaires sont similaires à celles observées pour les interactions inter-moléculaires (Caffrey et al., 2004; Jimenez, 2005). Il apparaît ainsi que les régions d'interaction sont majoritairement conservées. Cependant, un point majeur qui ressort de cette analyse concerne la faible spécificité d'une analyse de conservation quant à la détection de résidus situés aux interfaces de complexes intra-moléculaires. Ce résultat rejoint les conclusions tirées lors d'une précédente analyse, qui s'était focalisée sur une base de données de complexes inter-moléculaires (Jimenez, 2005).

3.4. Co-évolution des interfaces de complexes intra-moléculaires

3.4.1. Mutations corrélées et mutations compensatoires

La recherche de mutations concertées représente un second type d'analyses évolutives, pouvant permettre l'identification de sites d'interactions protéiques. Elle se distingue d'une

analyse de conservation dans la mesure où elle vise à identifier des relations d'évolution entre paires de résidus appartenant à deux domaines distincts (au contraire, une analyse de conservation analyse indépendamment les partenaires en interaction). Différentes études ont montré qu'au sein de protéines globulaires, deux positions proches dans l'espace sont susceptibles de muter de façon corrélée, le changement de propriété physico-chimique d'une position étant compensé par la mutation de la seconde position. La compensation peut consister en une inversion de charge électrostatique, de caractère hydrophobe/hydrophile ou encore de taille de résidu (Neher, 1994).

Afin de quantifier la co-évolution de domaines protéiques par ce type d'approche, j'ai implémenté deux méthodes d'analyse d'alignements multiples de séquences, l'une visant à détecter des événements de mutations corrélées au cours de l'évolution, la méthode de Göbel (Gobel et al., 1994), l'autre à identifier des événements de mutations traduisant une compensation d'une propriété physico-chimique, la méthode d'Afonnikov (Afonnikov et al., 2001) (voir (Madaoui and Guerois, 2007)). La recherche de mutations corrélées se base sur l'utilisation de matrices de substitution (Gobel et al., 1994), alors que la recherche de mutations compensatoires exploite des matrices de propriétés physico-chimiques des acides aminés (Afonnikov et al., 2001; Neher, 1994).

Pour analyser les performances des analyses de co-variation entre paires de résidus pour la détection de positions proches dans l'espace, j'ai analysé la distribution des distances entre toutes les paires de résidus qui co-évoluent au sein de la base de données de 86 complexes intra-moléculaires en utilisant les deux méthodes citées précédemment. Je les ai ensuite comparé à la distribution des distances entre toutes les paires de résidus. Les distributions présentées à la **Figure 41** comparent le pourcentage de paires variant de façon corrélée identifiées par la méthode de Göbel (**Figure 41-A**, barres noires) ou présentant des processus de compensation identifiées par la méthode d'Afonnikov (**Figure 41-B**, barres noires) en fonction de leur distance dans le complexe. Pour les deux approches implémentées, les paires de résidus proches dans l'espace correspondent plutôt à des paires co-évoluant. Cependant, ces méthodes sont très bruitées car de nombreux résidus éloignés dans l'espace co-évoluent également. En effet, selon la méthode utilisée, seuls 2 et 4% des paires de résidus sélectionnés correspondent effectivement à des acides aminés en contact (un contact étant

définit entre deux résidus si la plus courte distance entre leurs atomes respectifs est inférieure à 4.5Å).

Malgré le nombre important de faux positifs générés par une recherche de mutations compensatoires par la méthode d'Afonnikov, cette méthode se révèle plus spécifique qu'une recherche de mutations corrélées se basant sur des matrices de substitution entre acides aminés.

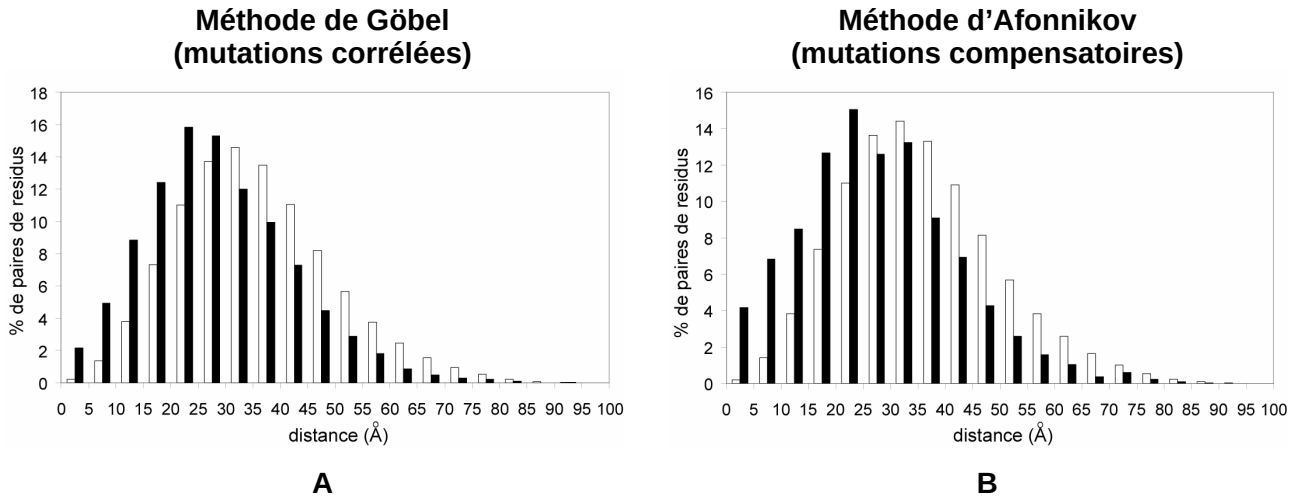


Figure 41. Mutations concertées aux interfaces de complexes intra-moléculaires. Les distributions représentent les distances entre toutes les paires de résidus possibles (blanc) et entre les paires de résidus subissant des mutations concertées en considérant (A) la méthode de Göbel (recherche de mutations corrélées (Gobel et al., 1994), et (B) la méthode d'Afonnikov (recherche de mutations compensatoires (Afonnikov et al., 2001). Quatre complexes n'ont pas été traités par la méthode d'Afonnikov (aucune paire ne présentant d'évènements de compensation). Figure adaptée de (Madaoui and Guerois, 2007). Les distances mesurées entre deux résidus correspondent à la plus courte distance entre leurs atomes respectifs.

3.4.2. Prédominance des événements de mutations compensatoires en périphérie des interfaces

Un des points étonnants des méthodes d'analyses des mutations compensatoires concerne le fait nombre de paires de résidus en contact soumises à ce type de phénomène. Afin de mieux comprendre l'origine de cette faible sensibilité, j'ai analysé les caractéristiques géométriques des paires de résidus en contact détectées comme co-évoluant par cette approche.

Dans un premier temps, j'ai analysé si les paires de résidus présentant des phénomènes de mutations compensatoires se situaient plutôt à la périphérie ou dans le cœur des interfaces. Pour cela, j'ai défini deux classes de paires de résidus au sein de l'interface: (i) les paires de

résidus situées en périphérie (moins de 50% des résidus avoisinants présentent une modification de leur accessibilité au solvant par la formation du complexe), (ii) les paires de résidus enfouies dans l'interface (plus de 50% des résidus avoisinants présentent une modification de leur accessibilité au solvant par la formation du complexe). Au total, 254 paires de résidus sont assimilées à la classe des périphériques, et 11 paires de résidus sur ces 254 paires sont prédites en contact par l'analyse de mutations compensatoires. Pour les 3631 paires de résidus affiliées à la seconde classe des paires enfouies, 68 paires sont prédites en contact par l'analyse de mutations compensatoires.

Nous pouvons déduire de cette analyse que les paires de résidus situées en périphérie des interfaces de complexes intra-moléculaires, sont significativement mieux prédites en contact par une analyse de mutations compensatoires. En effet, un test du Chi2 se révèle significatif pour un seuil de 1%.

Une des raisons pouvant expliquer cette différence pourrait être la densité de contact autour des paires de résidus analysés. En effet, les résidus enfouis présentent une densité de contact plus importante que les résidus situés en périphérie, et participent donc à plus d'interactions, ce qui accroît la difficulté de déterminer avec quels résidus l'événement de mutation compensatoire s'est produit au cours de l'évolution. Il a été effectivement noté que la réponse à une mutation ponctuelle pouvait être distribuée sur un groupe de résidus plutôt que sur un seul au cours de l'évolution des séquences protéiques (Gobel et al., 1994), ce qui peut limiter considérablement l'efficacité de telles approches, basées sur l'analyse de paires de résidus plutôt que de groupes de résidus.

L'origine de cette différence peut s'expliquer autrement, en considérant les propriétés de conservation des résidus situés en périphérie ou au cœur des interfaces. De façon générale, il apparaît en effet que les résidus situés au centre des interfaces de complexes sont plus conservés que les résidus situés en périphérie dans le cas des complexes permanents (Caffrey et al., 2004). Or la recherche de mutations concertées se base avant tout sur l'existence d'événements de mutations, et ne prend donc pas en compte les positions strictement conservées au cours de l'évolution.

3.5. Maintien de la complémentarité au sein des interfaces de complexes

Nous avons précédemment mis en évidence certaines faiblesses des analyses de mutations concertées pour la détection de paires de résidus en contact au sein des interfaces de complexes intra-moléculaires. En effet, ces méthodes se révèlent peu spécifiques, et inadaptées à la détection de résidus en contacts enfouis au cœur des interfaces.

Un des problèmes majeurs des méthodes basées sur la recherche de mutations concertées concerne leur incapacité à analyser des positions complètement conservées au cours de l'évolution. De plus, ces méthodes reposent sur l'hypothèse qu'au cours de l'évolution, les mutations concertées s'établissent entre paires de résidus en contact. Cette hypothèse est soutenue par l'analyse des bases de données de structures de complexes protéiques qui montre que la plupart des résidus en contact impliquent des résidus dont les propriétés physico-chimiques sont complémentaires (Glaser et al., 2001). Néanmoins l'étude de Glaser et collaborateurs ne renseigne pas sur la façon dont la complémentarité a été maintenue au cours de l'évolution des séquences. Cette analyse fait l'objet de la section suivante.

3.5.1. Maintien de la complémentarité des paires de résidus en contact au sein des interfaces au cours de l'évolution

Pour analyser la complémentarité des interfaces, nous avons considéré un modèle très simple impliquant trois types de complémentarité physico-chimique :

- classe 1 : complémentarité entre résidus hydrophobes (résidus G, A, V, L, I, M, C, F, P, W, Y),
- classe 2 : complémentarité entre résidus polaires (résidus S, T, N, Q),
- classe 3 : complémentarité entre résidus de charges opposées (résidus chargés positivement K, R, H ou négativement D, E),

Afin de simplifier l'analyse de la complémentarité, chaque acide aminé est associé à une classe de façon exclusive. Par exemple, bien que la tyrosine ou le tryptophane possède une fonction polaire, ils sont exclusivement considérés dans la classe des hydrophobes qui correspond à leurs propriétés physico-chimiques majeure. Notons que notre définition de la complémentarité en trois classes correspond à un schéma *a priori* assez restrictif car elle exclut des interactions entre résidus polaires et chargés qui s'établissent souvent *via* des

réseaux de liaisons hydrogènes. Pour évaluer la pertinence de cette définition simplifiée, j'ai analysé la complémentarité physico-chimique pour 86 interfaces de complexes intra-moléculaires extraits de la base de données PSIMAP (voir section 3.2). Pour chacun de ces 86 complexes intra-moléculaires impliquant deux domaines A et B, j'ai déterminé l'ensemble des paires de résidus (i,j) en contact entre ces deux domaines (i et j étant respectivement des résidus des domaines A et B). L'analyse de la complémentarité de ces paires de résidus montre qu'en moyenne 62% des résidus d'une interface sont en contact avec au moins un autre résidu qui lui est complémentaire (**Figure 42**).

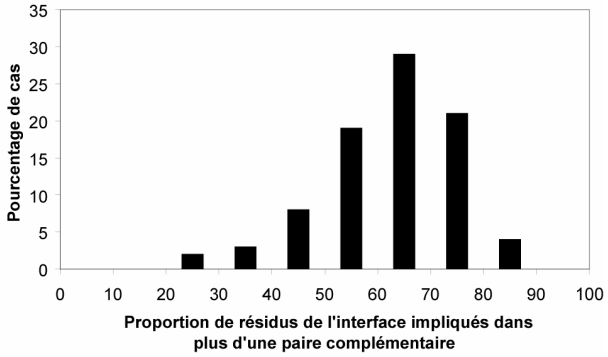


Figure 42. Proportion des résidus situés à l'interface de complexes protéiques intra-moléculaires impliqués dans au moins une paire complémentaire. La complémentarité entre les paires de résidus en contact est évaluée en considérant trois classes de complémentarité physico-chimiques. Les complexes analysés correspondent aux structures natives de 86 complexes intra-moléculaires extraits de la base de données PSIMAP (Kim et al., 2004). Figure adaptée de (Madaoui and Guerois, 2007).

Précédemment, nous avons signalé que deux études statistiques des structures de complexes d'homologues lointains ont montré que la position relative des résidus situés aux interfaces est conservée (Aloy et al., 2003; Kim and Ison, 2005). Il m'a donc paru intéressant d'étendre l'analyse de complémentarité, et d'étudier dans quelle mesure la complémentarité des paires de résidus en contact a pu être maintenue au cours de l'évolution. Ainsi, pour chaque paire de résidus en contact (i,j), un paramètre nommé *ratio* de complémentarité, qui quantifie le degré de conservation de la complémentarité au cours de l'évolution, a été défini. Ce descripteur mesure la proportion de séquences d'un alignement multiple, pour lesquelles la complémentarité entre deux résidus i et j en contact a été maintenue (**Figure 43**) (Madaoui and Guerois, 2007). Si cette proportion est supérieure à 95%, les résidus i et j sont considérés comme « significativement complémentaires » au cours de leur évolution.

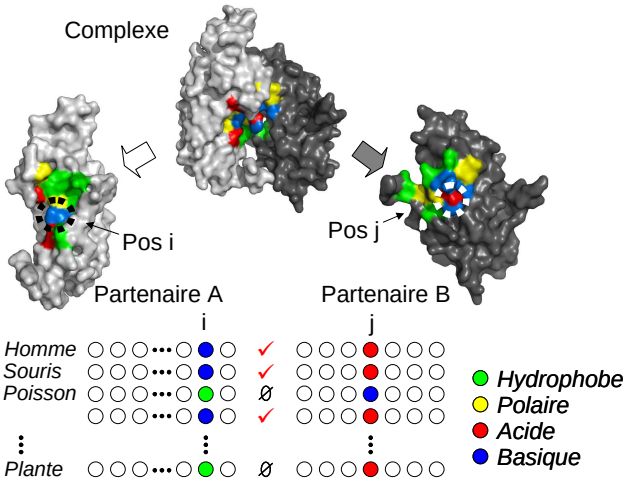


Figure 43. Analyse de la complémentarité évolutive à l'interface d'un complexe protéique. Les surfaces d'interactions de deux protéines A et B connus pour interagir et une représentation schématique des alignements de séquences de leurs protéines homologues sont représentés. Les positions de l'alignement, représentées sous forme de cercle, sont colorées en fonction de la propriété physico-chimique des résidus. L'existence ou l'absence d'une complémentarité dans un organisme donné est schématisé par un trait rouge ou un zéro barré. Pour les séquences de poisson et de plante, aucune complémentarité n'existe. Pour les cinq séquences représentées, le ratio de complémentarité est de 3/5 soit 60%. Figure adaptée de (Madaoui and Guerois, 2007).

En considérant la base de données intra-moléculaire de 86 complexes (voir section 3.2), nous pouvons constater de façon surprenante que la complémentarité physico-chimique n'a été maintenue que pour une faible proportion de paires de résidus en contact. En effet, seules 11% des paires de résidus sont évaluées comme « significativement complémentaires » au cours de leur évolution (**Figure 44**).

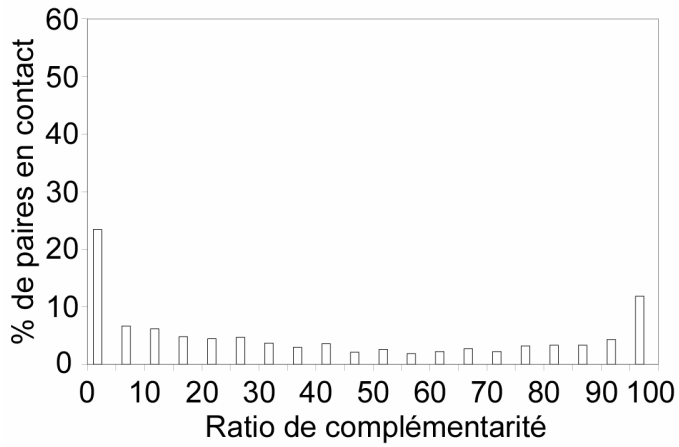


Figure 44. *Ratio* de complémentarité à l'interface de complexes intra-moléculaires. La distribution représente la proportion de paires de résidus en contact à l'interface des complexes protéiques intra-moléculaires, présentant un certain *ratio* de complémentarité. Le *ratio* de complémentarité est la proportion de séquences dans un alignement multiple de séquences pour lesquelles la complémentarité physico-chimique entre deux positions est maintenue (voir **Figure 43**). Figure adaptée de (Madaoui and Guerois, 2007).

3.5.2. La prise en compte du contexte structural révèle un maintien de la complémentarité des paires de résidus en contact au cours de l'évolution

Une des limites de l'analyse précédente est qu'elle ne prend pas en compte l'environnement structural local des paires de résidus en contact. En effet, on peut raisonnablement émettre l'hypothèse qu'en considérant une paire de résidu complémentaire (i,j), la mutation du résidu i à l'origine d'une rupture de complémentarité avec le résidu j, puisse être compensée par une seconde mutation d'un résidu i_1 proche de i au niveau structural (**Figure 45**).

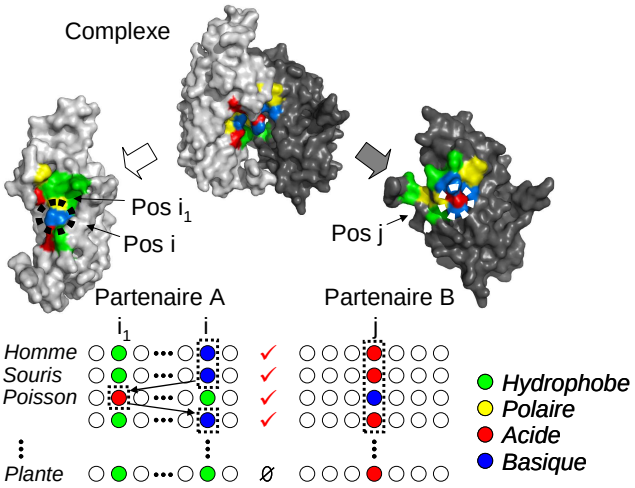


Figure 45. Analyse de la complémentarité évolutive à l'interface d'un complexe protéique, prenant en compte le voisinage structural. Les surfaces d'interactions de deux protéines A et B connus pour interagir et une représentation schématique des alignements de séquences de leurs protéines homologues sont représentés. Les positions de l'alignement, représentées sous forme de cercle, sont colorées en fonction de la propriété physico-chimique des résidus. L'existence ou l'absence d'une complémentarité dans un organisme donné est schématisé par un trait rouge ou un zéro barré. La séquence de poisson illustre un cas pour lequel un voisin structural i_1 de la position analysée i est requis pour le maintien de la complémentarité. Pour les cinq séquences représentées, le ratio de complémentarité est de 4/5 soit 80%. Figure adaptée de (Madaoui and Guerois, 2007).

Afin de prendre en compte cette éventualité, j'ai analysé de nouveau la complémentarité évolutive des paires de résidus en contact, en considérant cette fois leurs voisins structuraux respectifs. Pour chaque paire de résidu en contact (i,j), les deux voisins structuraux respectifs de i et de j ont donc été pris en compte lors du calcul du *ratio* de complémentarité (Madaoui and Guerois, 2007). En analysant de nouveaux les paires de résidus en contact aux interfaces de complexes intra-moléculaires, il apparaît désormais que près de 50% de ces paires sont

définies comme « significativement complémentaires » au cours de leur évolution, contre seulement 11% si l'on ne considère pas le voisinage structural de ces paires (**Figure 46**).

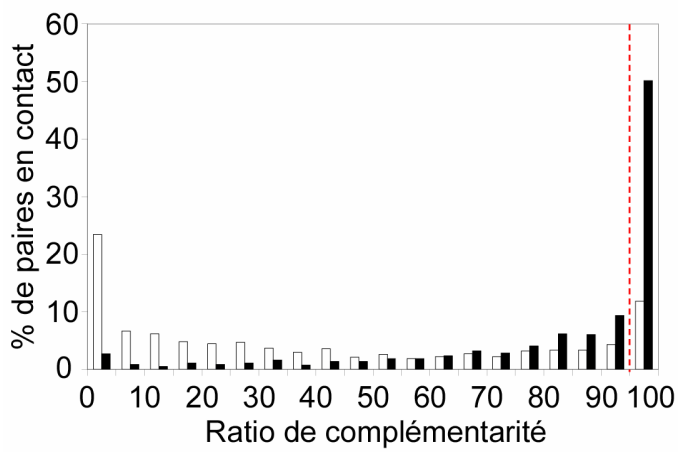


Figure 46. Influence du voisinage structural sur le *ratio* de complémentarité à l'interface de complexes intra-moléculaires. La distribution représente la proportion de paires de résidus en contact à l'interface des complexes protéiques intra-moléculaires, présentant un certain *ratio* de complémentarité. Le *ratio* de complémentarité est la proportion de séquences dans un alignement multiple de séquences pour lesquelles la complémentarité physico-chimique entre deux positions est maintenue. Le *ratio* de complémentarité est évalué en considérant le voisinage structural des paires de résidus en interaction (barres noires, voir **Figure 45**) ou non (barres blanches, voir **Figure 43**). En considérant le voisinage structural, les 2 plus proches voisins respectifs des paires de résidus en interaction sont pris en compte (k = 2). Figure adaptée de (Madaoui and Guerois, 2007).

Afin d'évaluer l'influence du nombre de voisins structuraux, le nombre de voisins a été modifié de 0 à 3, puis les analyses ont été réitérées en considérant toujours les paires de résidus en contact. À titre de comparaison, j'ai également appliqué ce protocole à toutes les paires de résidus localisées en dehors des sites d'interactions protéiques (Madaoui and Guerois, 2007). Cette dernière analyse qui concerne des paires de résidus éloignées rend compte de la proportion de paires « significativement complémentaires » détectées de façon fortuite. La **Figure 47** montre que la proportion de paires de résidus en contact détectées comme « significativement complémentaires » augmente en fonction du nombre de voisins, jusqu'à atteindre un plateau au-delà de deux voisins (histogramme rouge). En revanche, pour les paires éloignées en dehors de l'interface, la proportion augmente de façon monotone, cohérente avec la notion d'un signal fortuit (ligne noire). En fixant le nombre de voisins structuraux à deux, il apparaît qu'en moyenne 45% des résidus d'une interface interviennent dans au moins une paire de résidus « significativement complémentaires » au cours de l'évolution. Parmi ces résidus, environ 60% sont détectés seulement lorsqu'on prend en compte leurs voisins structuraux, ce qui souligne l'importance de la prise en compte des

voisins structuraux dans la caractérisation des processus évolutifs des interfaces de complexes protéiques.

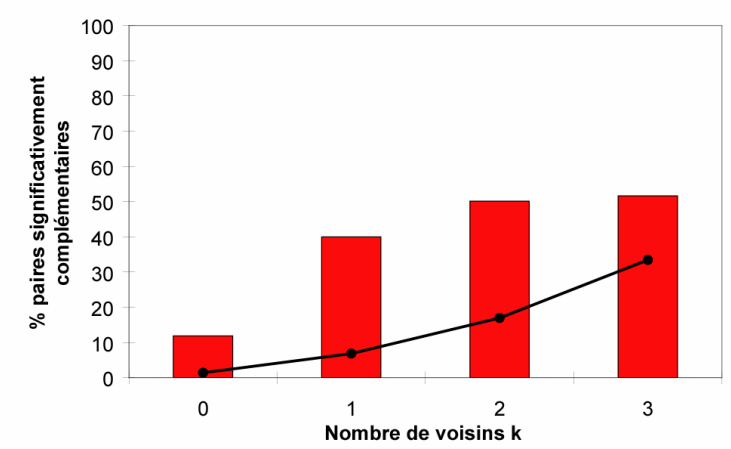


Figure 47. Pourcentage de paires de résidus détectées comme étant significativement complémentaires en fonction du nombre de voisins structuraux *k*. Les paires considérées correspondent à toute les paires de résidus en contact entre deux domaines protéiques en interaction (histogramme rouge), et toutes les paires de résidus possibles, prises aléatoirement en dehors des interfaces et n'interagissant donc pas (courbes noire). La base de données analysé est la base de 86 complexes intra-moléculaires extraits de la banque PSIMAP (Kim et al., 2004). Figure adaptée de (Madaoui and Guerois, 2007).

3.6. Conclusion

Afin de caractériser l'existence d'une pression de sélection aux interfaces de complexes protéine-protéine, différentes études ont été proposées dans le but d'analyser les interfaces de ces assemblages, en se basant principalement sur deux hypothèses évolutives : (i) la conservation des sites d'interaction, (ii) la co-évolution des résidus localisés au niveau de ces régions d'interaction. Ces analyses ont ainsi mis en évidence certaines propriétés des interfaces de complexes protéiques, comme une conservation de leurs interfaces (Caffrey et al., 2004) ou l'existence de paires de résidus dont les mutations sont concertées au cours de leur évolution (Pazos et al., 1997). Cependant, ces mêmes études soulignent que les régions d'interaction ne sont pas les seules régions conservées à la surface des protéines, et que deux résidus distants peuvent également montrer une tendance à co-évoluer.

Dans ce chapitre, l'étude d'une base de données de 86 complexes protéiques intra-moléculaires, a permis de confirmer le manque de spécificité de ce genre d'approche. Plus encore, cette analyse a permis de révéler un manque évident de sensibilité, particulièrement

lorsque les sites protéiques analysés sont enfouis au cœur des interfaces protéiques. Or il est bien connu que les résidus essentiels à la spécificité de reconnaissance des assemblages protéiques, les *hots spots*, sont souvent enfouis au cœur des interfaces protéiques (Halperin et al., 2004). De plus, ces résidus sont bien conservés (Halperin et al., 2004), ce qui rend leur analyse par ces approches particulièrement difficile étant donnée que ces méthodes se basent avant tout sur la présence de mutations.

Une autre hypothèse qui permettrait d'expliquer le manque de sensibilité des méthodes basées sur la détection de mutations concertées au cœur des interfaces de complexes serait l'existence de processus de compensation n'impliquant non pas des paires de résidus, mais des groupes de résidus. Les paires de résidus en contact au sein des interfaces de complexes sont en effet entourées d'autres résidus pouvant participer eux-mêmes au processus de co-évolution. Cela pourrait alors expliquer pourquoi ces méthodes sont particulièrement spécifiques dans le cas de la prédiction d'appariements de type Watson-Crick des ARNrs, où les processus d'appariement impliquent des paires de bases.

Afin de vérifier cette dernière hypothèse, j'ai introduit dans ce chapitre un nouveau mode d'analyse de la co-évolution, basé sur les propriétés de complémentarité des paires de résidus en interaction. D'après les résultats obtenus, il apparaît qu'au cours de l'évolution, un maintien strict de la complémentarité peut être observé pour des paires de résidus en contact. En revanche, ce maintien de la complémentarité ne peut s'expliquer qu'en prenant en compte non pas des paires de résidus, mais des groupes de résidus proches au niveau de la structure. De plus, ce phénomène, de maintien de la complémentarité apparaîtrait comme particulièrement spécifique, si l'on compare sa prédominance aux interfaces des complexes protéiques, en comparaison à des zones de la protéine n'interagissant pas. Il serait donc intéressant de vérifier si l'introduction d'un tel descripteur évolutif peut permettre de prédire efficacement les assemblages de complexes protéiques, ce que je propose d'analyser dans le chapitre suivant.

Chapitre 4 : Prédiction des assemblages macromoléculaires

4.1. Introduction

Dans le chapitre précédent, nous avons présenté différentes méthodes s'appuyant sur des propriétés évolutives pour identifier des sites d'interaction. Pour juger si ces stratégies sont utiles pour la prédiction structurale des assemblages, elles peuvent être couplées aux méthodes d'amarrage moléculaire ou docking. Ces méthodes comportent toutes deux phases principales : (i) une première phase exploratoire, le plus souvent « corps rigide », qui vise à explorer différents modes d'assemblage possibles entre les deux molécules en interaction, (ii) une seconde phase de raffinement, intégrant plus de flexibilité, afin de raffiner les modèles présélectionnés (voir section 1.5.3).

Dans la littérature, les méthodes évolutives ont été citées comme efficaces pour filtrer un grand nombre de solutions, et limiter la recherche exploratoire dans la phase de raffinement (Duan et al., 2005; Pazos et al., 1997). Cependant, la comparaison systématique des capacités de discrimination des différentes approches évolutives sur une même base de données n'a pas, à notre connaissance, été réalisée.

Dans ce chapitre, les approches évolutives seront évaluées pour leur capacité à discriminer, les modèles d'assemblages proches des complexes natifs au sein d'un vaste ensemble de faux complexes. Je présenterai dans un second temps une nouvelle approche pour la prédiction des sites d'interaction protéiques, l'approche SCOTCH (« *Surface COmplementarity Trace in Complex History* »), basée sur la conservation de la complémentarité physico-chimique des interactions entre paires de résidus. L'efficacité de cette approche pour la prédiction structurale des assemblages protéiques intra-moléculaires et inter-moléculaires sera analysée, puis comparée aux autres méthodes publiées dans la littérature. Enfin, je présenterai son application à deux cas d'études qui intéressent particulièrement notre équipe : les complexes SGT1-Hsp90 et Asf1-Histone H3-H4.

4.2. Constitution des bases de données de complexes protéiques

Afin de comparer différentes approches évolutives pour la prédiction structurale des assemblages protéiques, j'ai compilé deux bases de données protéiques contenant des interactions de type intra-moléculaire et inter-moléculaire. La première banque de données contient 86 complexes correspondant à des interactions intra-moléculaires de type domaine-domaine, issus de la banque de données PSIMAP (Kim et al., 2004). La constitution de cette banque de données a été précédemment décrite dans le chapitre 3 (voir section 3.2).

La seconde banque de données contient des complexes inter-moléculaires caractérisés dans leur état « liés » ou « non liés » (lorsque les structures sont extraites du complexe, on parle d'assemblage « liés », alors que lorsque les structures des deux partenaires protéiques correspondent à leur forme libre, on parle d'assemblage « non liés »). Ces complexes structuraux proviennent de la banque de données de WENG (Chen et al., 2003), fréquemment utilisée pour évaluer les capacités prédictives des programmes de docking. La banque de données de WENG contient initialement 84 complexes protéiques. En éliminant les complexes impliquant plus d'une chaîne par protéine, et les cas pour lesquels le nombre de séquences protéiques homologues est insuffisant pour entreprendre des analyses évolutives, 11 complexes protéiques ont été retenus (Madaoui and Guerois, 2007).

4.3. Evaluation des capacités discriminatives des méthodes évolutives : le cas des interactions intra-moléculaires

4.3.1. Evaluation des différentes approches évolutives

Pour évaluer la capacité discriminative des différentes méthodes évolutives, j'ai généré pour chacun des 86 cas de la banque de données intra-moléculaire, 10000 orientations possibles pour les deux partenaires en interaction, en utilisant le programme FTDock (Gabb et al., 1997). Une solution est considérée comme proche de la structure native si son RMSD par rapport au complexe natif est inférieur à 3 Å (le RMSD étant calculé sur l'ensemble du complexe en considérant son squelette peptidique). Pour chacun des cas d'étude, il existe un nombre très restreint de modèles corrects proches de la structure native du complexe. En moyenne, 7.4 solutions proches de la structure native sont générés sur 10000 modèles.

Pour chacun des 860000 modèles générés, différents scores ont été calculés à partir des 4 méthodes d'analyse évolutive introduites dans le chapitre 3 : (i) proportion de résidus conservés aux interfaces, (ii) proportion de résidus divergents aux interfaces, (iii) proportion de paires de résidus présentant des événements de mutations compensatoires (détectés par la méthode d'Afonnikov (Afonnikov et al., 2001)), (iv) proportion de paires de résidus « significativement complémentaires » au cours de l'évolution des interfaces (voir chapitre 3), que l'on appellera « score de complémentarité évolutive » (avec et sans prise en compte du voisinage structural). Pour les deux premiers descripteurs, les pourcentages de résidus conservés et divergents ont été déterminés d'après les taux d'évolution par site calculés par le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002), de la même façon que dans le chapitre précédent. Pour chacun des 86 cas analysés, les 10000 modèles ont été classés en fonction des différents scores évolutifs. Pour chaque score évolutif, le modèle le plus probable est celui qui maximise le score (à l'exception du descripteur (ii) analysant la divergence des interfaces, pour lequel le modèle le plus probable est celui qui minimise ce score).

Pour chacune des méthodes évolutives, j'ai déterminé ensuite le nombre de cas dans la base de données pour lesquels au moins une solution correcte est retenue parmi les solutions les mieux classées. Pour quatre complexes de la base de données, le programme FTDock ne génère aucune solution correcte parmi les 10000 modèles générés. Les analyses portent donc sur les 82 complexes restant. La question posée est la suivante : parmi les x% meilleures solutions classées suivant un score évolutif y, peut-on sélectionner une solution proche de la structure native du complexe d'étude ? Le **Tableau 7** répond à cette question, en considérant différentes valeurs pour x : 1, 5 et 10 %.

En examinant les résultats présentés au **Tableau 7**, nous constatons que les méthodes reposant sur la complémentarité évolutive obtiennent les meilleurs résultats. En particulier, la méthode d'analyse de complémentarité évolutive avec prise en compte du voisinage structural se révèle être la méthode la plus performante, particulièrement lorsqu'on sélectionne les 100 meilleures solutions produites par FTDock (top 1%).

	[a]	[b]	[c]	[d]	[d]
Top 1%	14 / 82	21 / 82	18 / 82	34 / 82	41 / 82
Top 5%	41 / 82	44 / 82	47 / 82	60 / 82	62 / 82
Top 10%	54 / 82	55 / 82	59 / 82	67 / 82	68 / 82

[a] Divergence (pourcentage de résidus divergents aux interfaces)
[b] Conservation (pourcentage de résidus conservés aux interfaces)
[c] Mutations Compensatoires (méthode d'Afonnikov (Afonnikov et al., 2001))
[d] Complémentarité évolutive sans prise en compte du voisinage structural (k = 0)
[e] Complémentarité évolutive avec prise en compte du voisinage structural (k = 2)

Tableau 7. Comparaison des méthodes d'analyses évolutives pour la sélection de bons modèles d'assemblage intra-moléculaires. Pour chacune des méthodes analysées, le nombre de cas pour lesquels un modèle correct a été classé dans les 100, 500 ou 1000 meilleurs modèles a été déterminé (un modèle est définit comme correct si son RMSD par rapport à la structure native du complexe est inférieur à 3 Å). Pour quatre complexes, le programme FTDock n'a généré aucun modèle correct parmi les 10000 solutions générées, les résultats portent donc sur 82 complexes issus de la base de données PSIMAP. La variable *k* design le nombre de voisins structuraux considérés dans l'analyse de la complémentarité évolutive.

4.3.2. La complémentarité évolutive des interfaces protéiques est-elle un bon critère pour identifier les assemblages protéiques natifs ?

Afin d'examiner plus en détail les performances de la méthode mesurant la complémentarité évolutive, j'ai séparé les 860000 solutions générées par FTDock (10000 modèles pour chacun des 86 complexes de la base de données) en deux groupes : (i) modèles corrects, (ii) modèles erronés. Un modèle est considéré comme correct s'il possède un RMSD inférieur à 3 Å par rapport à la structure native, conformément à la précédente analyse.

En traçant la distribution des scores des deux groupes de structures, la discrimination entre les solutions vraies et fausses apparaît très clairement (**Figure 48**). Il existe un décalage net de la distribution des scores de complémentarité évolutive vers des valeurs importantes pour les solutions vraies. Ce résultat particulièrement intéressant, est assez informatif sur les capacités de filtrage de cette approche, lors des premières étapes de docking moléculaire. En effet, en fixant un seuil égal à 20, environ 80% des fausses solutions et seulement 15% des vraies solutions sont éliminées.

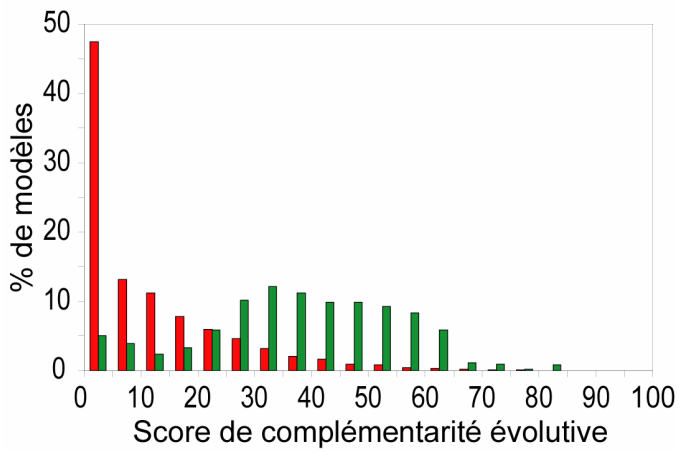


Figure 48. Distribution des scores de complémentarité évolutive pour les fausses solutions (rouge), et les vraies solutions (vert) générées par le programme FTDock (Gabb et al., 1997). Les vraies solutions sont à moins de 3 Å des complexes natifs, les fausses solutions sont à plus de 3 Å des complexes natifs (le RMSD étant mesuré sur l'ensemble du complexe en considérant le squelette peptidique). Figure adaptée de (Madaoui and Guerois, 2007).

4.4. SCOTCH : une nouvelle méthode pour prédire la structure des assemblages protéiques

Afin d'augmenter la discrimination entre vraies et fausses solutions, certaines améliorations peuvent être apportées à cette analyse de complémentarité. Premièrement, l'intensité du score de complémentarité évolutive calculé pour des régions fortement variables peut être biaisée par la qualité locale de l'alignement de séquences. Dans les régions très variables, un événement de complémentarité a une probabilité non négligeable d'être observé de façon fortuite. D'autre part, deux interfaces présentant des scores de complémentarité évolutive identiques et respectivement 10 ou 100 paires significativement complémentaires ne seront pas différenciées alors que la probabilité que la complémentarité soit fortuite sur 10 paires est plus forte que sur 100 paires de résidus. Pour rendre compte de ces deux propriétés, le score de complémentarité évolutive précédemment introduit a été combiné à (i) un score caractérisant la variabilité de certaines positions dans l'alignement (proportion de résidus divergents à l'interface du complexe) et (ii) au nombre de paires de résidus en contact analysées au sein de l'interface (les paires de résidus en contact sont toutes les paires entre les domaines protéiques en interaction à moins de 4.5 Å). Pour cela, j'ai recherché la combinaison optimale de ces trois facteurs par une procédure d'apprentissage.

4.4.1. Apprentissage supervisé des caractéristiques évolutives des complexes : l'approche SCOTCH

La combinaison des différents paramètres que sont le score de complémentarité évolutive, la variabilité des positions de l'alignement et le nombre de paires en contact dans les interfaces peut être réalisée de différentes façons. Une des méthodes de combinaison consiste à pondérer ces différents paramètres à partir d'une procédure d'apprentissage.

Afin d'obtenir une règle de décision pour la classification des solutions en deux groupes (vraies et fausses solutions), j'ai donc divisé la banque de données de 86 complexes intra-moléculaires en deux sous ensembles : (i) un sous ensemble de 58 complexes représentant 2/3 de la base de données initiale pour l'échantillon d'apprentissage, (ii) un sous-ensemble de 28 complexes soit 1/3 de la base de données initiale pour l'échantillon de validation. Au sein de l'échantillon d'apprentissage, 361 solutions produites par FTDock sont proches des complexes natifs respectifs. J'ai sélectionné ces 361 solutions, ainsi qu'un nombre égal de faux complexes pris aléatoirement parmi les solutions restantes. Au total, 722 complexes ont été sélectionnés. Ensuite, un apprentissage par régression logistique a été réalisé, ce qui m'a permis de déterminer la combinaison linéaire optimale des 3 variables décrites précédemment (Madaoui and Guerois, 2007). Cette combinaison définit une nouvelle fonction de score : SCOTCH (« *Surface COmplementarity Trace in Complex History* »). La fonction SCOTCH (**Équation 1**) estime ainsi la probabilité qu'un modèle donné soit vrai, en fonction des trois paramètres décrits précédemment, x étant le score de complémentarité évolutive, y le score de variabilité par position et z le score associé au nombre de paires significativement complémentaires.

$$F(x,y,z) = -2.1 + 0.036\,x - 0.16\,y + 0.082\,z$$

Équation 1. Fonction de score SCOTCH. x correspond à la proportion de paires significativement compémentaires, y au pourcentage de paires de résidus variables, z au nombre de paires en interaction au sein de l'interface. Les poids de chaque variable ont été optimisés par une régression logistique d'après les caractéristiques de 722 complexes vrais ou faux générés par une approche de docking.

4.4.2. Validation de l'approche SCOTCH pour le cas d'interactions intra-moléculaires

Afin de tester le pouvoir discriminant de la méthode SCOTCH sur la base de données de validation de 28 complexes intra-moléculaires, j'ai généré 10000 modèles d'assemblages pour chacun des 28 cas à l'aide du programme FTDock (Gabb et al., 1997). J'ai ensuite calculé le score SCOTCH de chacune des solutions, puis classé ces solutions en fonction de ce score. De la même façon, j'ai classé ces différentes solutions en fonction des autres méthodes évolutives introduites précédemment, à savoir, les méthodes basées sur des analyses de conservation ou sur la détection de mutations concertées par les approches de Göbel et d'Afonnikov.

Concernant l'évaluation des modèles par une analyse de conservation, le score de conservation est toujours basé sur le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002). Pour l'évaluation des solutions par les méthodes de détection de mutations concertées, j'ai implémenté une autre fonction de score précédemment introduite par Pazos lors d'une étude portant sur 21 complexes intra-moléculaires (Pazos et al., 1997) (pour plus de détails, voir en annexe (Madaoui and Guerois, 2007)). Dans cette étude, les auteurs identifient dans un premier temps par la méthode de Göbel (Gobel et al., 1994) l'ensemble des paires de résidus dont les mutations sont corrélées au cours de l'évolution, et ceci pour chaque paire de domaine protéique interagissant. Ensuite, après avoir généré un nombre important de modèles d'assemblage de chaque paire de domaines protéiques en interaction, les auteurs évaluent ces différents modèles à partir d'une fonction harmonique, appelée *Xd*, destinée à quantifier pour un modèle donné la prédominance de paires de résidus corrélées à son interface (Pazos et al., 1997). Ainsi, plus le paramètre *Xd* d'un modèle de complexe est fort, plus le nombre de paires corrélées à son interface est important. Les auteurs ont montré lors de cette étude que les modèles pour lesquels le paramètre *Xd* est le plus élevé, correspondent le plus souvent à des solutions proches des structures natives de complexes. J'ai donc calculé le paramètre *Xd* de chacune des solutions générées par le programme FTDock (Gabb et al., 1997), en recherchant les mutations concertées entre domaine protéiques interagissant soit en considérant la méthode de Göbel (*Xd_Göbel*), soit en considérant la méthode d'Afonnikov (*Xd_Afonnikov*) (Madaoui and Guerois, 2007).

Les qualités prédictives de chaque méthode d'évaluation peuvent être quantifiées en dénombrant le nombre de cas dans la base de validation pour lesquels un complexe vrai est sélectionné parmi : (i) les 10 meilleures solutions (Top 10), (ii) les 100 meilleures solutions (Top 100), (iii) les 1000 meilleures solutions (Top 1000), (iv) les 10000 solutions (Top 10000). La capacité prédictive de chaque approche est alors considérée comme excellente, bonne, acceptable ou mauvaise en fonction des 4 classes respectivement définies.

Suivant ce critère d'évaluation des différentes méthodes, nous constatons que le score SCOTCH donne les meilleurs résultats pour les cas d'interaction intra-moléculaires, avec une grande majorité de cas pour lesquels la discrimination est excellente ou bonne (**Figure 49**). L'amélioration est particulièrement remarquable si l'on considère uniquement les 10 meilleures solutions sur les 10000 produites par le programme FTDock, car l'approche SCOTCH sélectionne au moins une solution correcte pour près de la moitié de la base de données test (13 cas sur 28), contrairement aux autres approches.

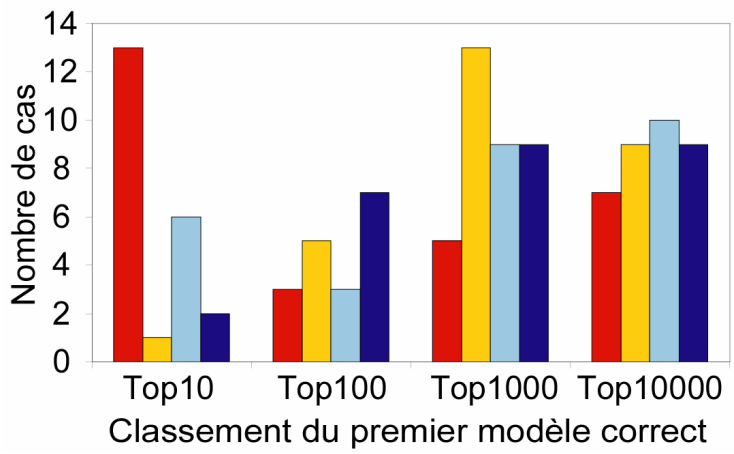


Figure 49. Distribution du nombre de cas pour lesquels une vraie solution est sélectionnée parmi les 10, 100, 1000, 10000 meilleures solutions générées par docking suivant les scores évolutifs (i) SCOTCH (rouge), (ii) Conservation (jaune), (iii) *Xd*_Göbel (bleu clair), (iv) *Xd*_Afonnikov (bleu foncé). Les scores *Xd*_Göbel et *Xd*_Afonnikov sont basés respectivement sur les méthodes de Göbel (Gobel et al., 1994) et d'Afonnikov (Afonnikov et al., 2001). La base de données considérée est la base de données de validation intra-moléculaires de 28 complexes issus de la banque PSIMAP (Kim et al., 2004). Les modèles d'assemblages évalués ont été générés par le programme FTDock (Gabb et al., 1997), et une vraie solution correspond à un modèle dont le RMSD par rapport à la structure native du complexe est inférieur à 3Å. Figure adaptée de (Madaoui and Guerois, 2007).

J'ai également évalué les différentes méthodes évolutives par des courbes ROC (« *Receiver Operating Characteristics* »). Les courbes ROC, en représentant la fraction de vrais et de faux positifs pour différentes valeurs seuils, rendent compte de façon très visuelle de la sensibilité

et de la spécificité d'une méthode. La capacité prédictive d'une méthode est directement corrélée à l'aire sous la courbe (égale à 1 lorsque la discrimination fausses solutions / bonnes solutions est parfaite). Ce mode de représentation va ainsi permettre d'évaluer le pourcentage de mauvaises solutions éliminées par les différentes approches, et leur capacité à retenir les vraies solutions.

En examinant les courbes ROCs relatives aux différentes méthodes évolutives (**Figure 50**), nous confirmons que la méthode SCOTCH donne de meilleurs résultats en terme de pouvoir discriminant. Il est intéressant de noter que les méthodes basées sur des analyses de mutations concertées (méthodes de Göbel ou d'Affonikov) entre des partenaires protéiques donnent de meilleurs résultats qu'une analyse de conservation. La présence dans la base de données d'étude d'un certain nombre de domaines possédant de multiples sites d'interaction avec d'autres domaines (comme les complexes 1a9x, 1ay0, 1bgl ou 1ea0, (Madaoui and Guerois, 2007), pourrait être une des explications de ce résultat. En effet, les différents sites d'interaction conduisent à l'existence de multiples sites conservés à la surface d'un domaine et la méthode de conservation ne permet pas de sélectionner spécifiquement un site d'interaction parmi d'autres.

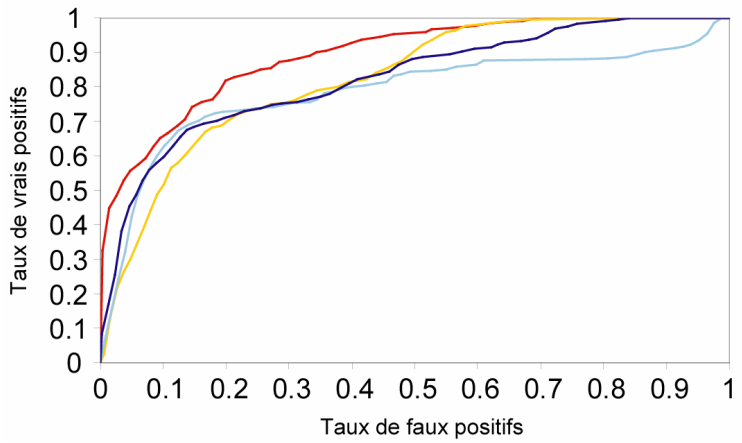


Figure 50. Comparaison des courbes ROCs pour les scores évolutifs SCOTCH (rouge), (ii) Conservation (jaune), (iii) *Xd*_Göbel (bleu clair), (iv) *Xd*_Afonnikov (bleu foncé). Les scores *Xd*_Göbel et *Xd*_Afonnikov sont basés respectivement sur la méthode de Göbel (Gobel et al., 1994) et d'Afonnikov (Afonnikov et al., 2001). La base de données considérée est la base de données de validation intra-moléculaires de 28 complexes issus de la banque PSIMAP (Kim et al., 2004). Les modèles d'assemblages évalués ont été générés par le programme FTDock (Gabb et al., 1997), et une vraie solution correspond à un modèle dont le RMSD par rapport à la structure native du complexe est inférieur à 3Å. Figure adaptée de (Madaoui and Guerois, 2007).

4.4.3. Validation de l'approche SCOTCH pour le cas d'interactions inter-moléculaires

La même procédure d'évaluation a été appliquée en considérant la seconde base de données de 11 complexes inter-moléculaires non redondants dans leur état « liés » (voir section 4.2), issue de la banque de données de WENG (Chen et al., 2003). La **Figure 51** montre que les prédictions réalisées par l'approche SCOTCH sont très bonnes, meilleures que dans le cas des interactions intra-moléculaires. En effet, pour 10 cas sur 11, le pouvoir prédictive est considéré comme bon ou très bon par notre système d'évaluation.

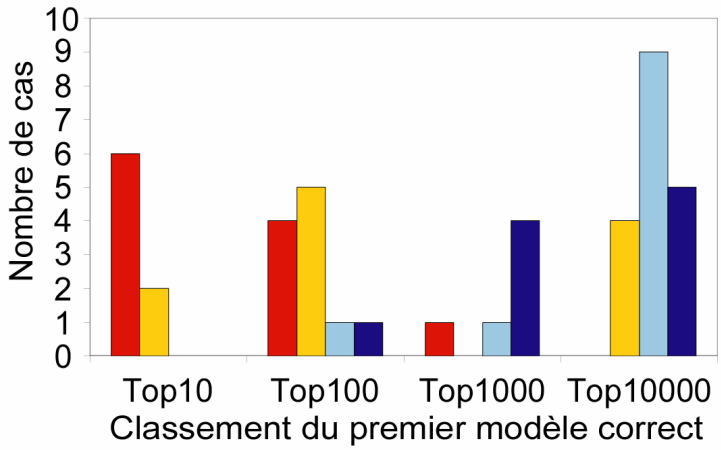


Figure 51. Distribution du nombre de cas pour lesquels une vraie solution est sélectionnée parmi les 10, 100, 1000, 10000 meilleures solutions générées par docking suivant les scores évolutifs (i) SCOTCH (rouge), (ii) Conservation (jaune), (iii) *Xd*_Göbel (bleu clair), (iv) *Xd*_Afonnikov (bleu foncé). Les scores *Xd*_Göbel et *Xd*_Afonnikov sont basés respectivement sur la méthode de Göbel (Gobel et al., 1994) et d'Afonnikov (Afonnikov et al., 2001). La base de données considérée est la base de données de validation inter-moléculaires de 11 complexes issus de la banque WENG dans leur état « liés » (Chen et al., 2003). Les modèles d'assemblages évalués ont été générés par le programme FTDock (Gabb et al., 1997), et une vraie solution correspond à un modèle dont le RMSD par rapport à la structure native du complexe est inférieur à 3Å. Figure adaptée de (Madaoui and Guerois, 2007).

Deux résultats importants sont à souligner. D'une part, nous constatons une nette amélioration de la capacité de détection de la méthode basée sur la conservation des sites d'interfaces protéiques, en comparaison aux résultats intra-moléculaires. Dans le cas d'interactions intra-moléculaires, plusieurs cas d'études ont la particularité de posséder plusieurs sites d'interactions intra- ou inter-moléculaires, ce qui n'est probablement pas le cas pour la majeure partie des complexes inter-moléculaires considérés ici. Un autre résultat intéressant concerne les méthodes basées sur des analyses de mutations concertées. En effet, nous constatons que la capacité discriminative de ces approches se révèle particulièrement

mauvaise, nettement moins bonne que dans le cas des interactions intra-moléculaires. Seul 1 cas sur 11 est considéré comme étant bien prédit par notre système d'évaluation par ces méthodes (**Figure 51**).

Une analyse par courbe ROC confirme le manque de spécificité des méthodes basées sur la recherche de mutations concertées (**Figure 52**), soulignant une différence de capacité discriminative très importante entre les cas intra-moléculaires et inter-moléculaires (**Figure 50** et **Figure 52**). A l'opposé, la capacité discriminative de la méthode SCOTCH reste satisfaisante, aussi bien dans le cas de complexes intra-moléculaires qu'inter-moléculaires, ce qui laisse entrevoir son application pour la prédiction de tout type d'assemblage moléculaire.

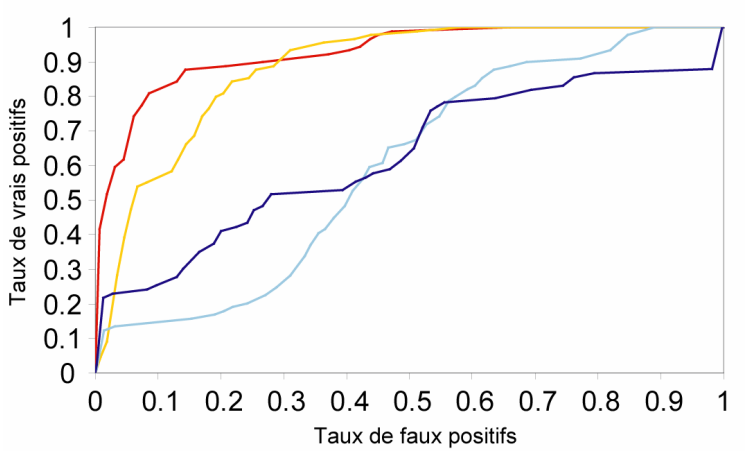


Figure 52. Comparaison des courbes ROCs pour les scores évolutifs SCOTCH (rouge), (ii) Conservation (jaune), (iii) *Xd*_Göbel (bleu clair), (iv) *Xd*_Afonnikov (bleu foncé). Les scores *Xd*_Göbel et *Xd*_Afonnikov sont basés respectivement sur la méthode de Göbel (Gobel et al., 1994) et d'Afonnikov (Afonnikov et al., 2001). La base de données considérée est la base de données de validation inter-moléculaires de 11 complexes issues de la banque WENG dans leur état « liés » (Chen et al., 2003). Les modèles d'assemblages évalués ont été générés par le programme FTDock (Gabb et al., 1997), et une vraie solution correspond à un modèle dont le RMSD par rapport à la structure native du complexe est inférieur à 3Å. Figure adaptée de (Madaoui and Guerois, 2007).

4.4.4. Optimisation de l'approche SCOTCH par un score statistique

D'autres méthodes se basent sur une description statistique des interactions entre résidus afin d'évaluer des modèles de complexes protéiques. Parmi elles, la fonction RPScore (« *Residue level Pair potential Score* ») (Moont et al., 1999) est dérivée d'une analyse statistique de 103 interactions inter-moléculaires non redondantes. Le RPScore est une mesure

empirique de la probabilité d'observer certains types de contact aux interfaces de complexes protéiques.

En comparant cette approche statistique avec la méthode SCOTCH, nous pouvons noter certains points intéressants. Tout d'abord, l'approche SCOTCH possède une meilleure capacité à discriminer entre des vrais et des faux complexes, aussi bien pour des modes d'interactions intra-moléculaires ou inter-moléculaires (**Figure 53** et **Figure 54**). La différence de capacité prédictive est particulièrement importante dans le cas des interactions intra-moléculaires. Dans ce type de complexe, le pouvoir prédictif de l'approche RPScore est considéré comme très bon pour seulement 4 complexes sur 28 (**Figure 53-A**). Le fait que la fonction RPScore soit exclusivement basée sur des statistiques effectuées sur une base de données de complexes inter-moléculaires pourrait expliquer ce manque de discrimination pour d'autres modes d'assemblages.

Après avoir comparé ces deux approches qui évaluent toutes deux la complémentarité physico-chimique des interfaces de complexes protéiques, j'ai intégré le potentiel RPScore dans la fonction d'évaluation SCOTCH, afin de mesurer l'apport d'une méthode statistique sur une procédure d'analyse évolutive. Pour cela, j'ai intégré le potentiel RPScore comme un paramètre supplémentaire de la régression logistique (voir section 4.4.1), et recalculé les poids optimaux des différents paramètres de cette fonction par la même procédure (Madaoui and Guerois, 2007). En examinant les performances de cette nouvelle méthode hybride, nous pouvons constater une amélioration sensible de la capacité discriminative, aussi bien pour des complexes intra-moléculaires qu'inter-moléculaires (**Figure 53**, barres noires). En comparaison de l'approche SCOTCH, le nombre de cas considérés comme bien prédits par la méthode hybride augmente de 16 à 19 pour les interactions intra-moléculaires, et le nombre de cas considérés comme très bien prédits augmente de 6 à 9 pour les interactions inter-moléculaires (**Figure 53**, barres noires). Ce résultat démontre ainsi que le signal détecté par l'approche évolutive SCOTCH peur être renforcé par une représentation statistique de la complémentarité.

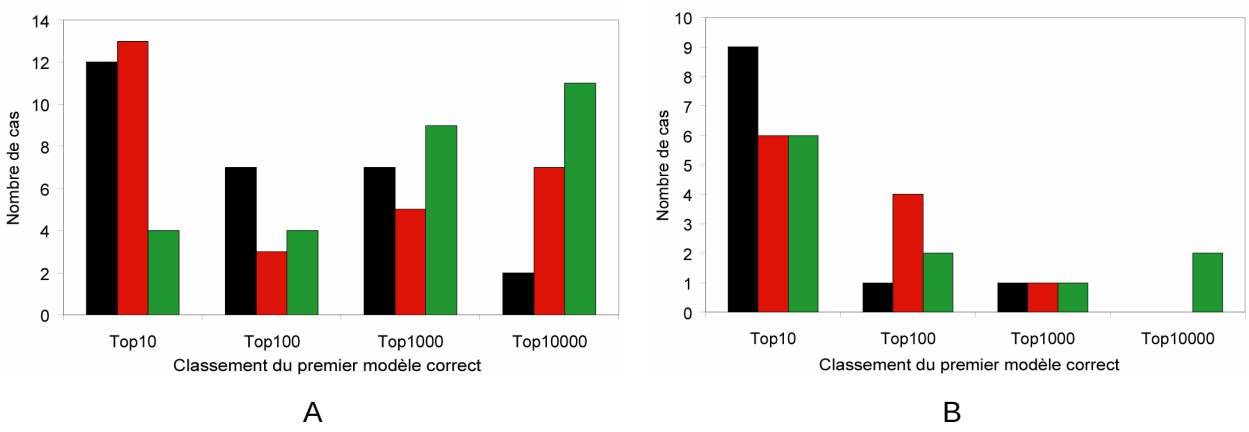


Figure 53. Distribution du nombre de cas pour lesquels une vraie solution est sélectionnée parmi les 10, 100, 1000, 10000 meilleures solutions générées par docking suivant les scores (i) SCOTCH (rouge), (ii) RPScore (vert), (iii) SCOTCH optimisé par RPScore (noir). (A) La base de données considérée est la base de données de validation intra-moléculaires de 28 complexes issus de la banque PSIMAP (Kim et al., 2004). (B) La base de données considérée est la base de données de validation inter-moléculaires de 11 complexes issus de la banque WENG dans leur état « liés » (Chen et al., 2003). Les modèles d'assemblages évalués ont été générés par le programme FTDock (Gabb et al., 1997), et une vraie solution correspond à un modèle dont le RMSD par rapport à la structure native du complexe est inférieur à 3Å. Figure adaptée de (Madaoui and Guerois, 2007).

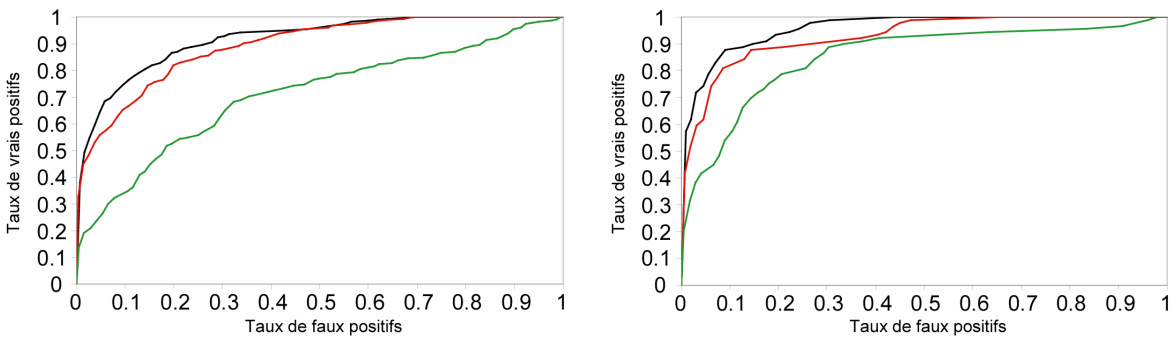


Figure 54. Comparaison des courbes ROC pour les scores SCOTCH (rouge), (ii) RPScore (vert), (iii) SCOTCH optimisé par RPScore (noir). (A) La base de données considérée est la base de données de validation intra-moléculaires de 28 complexes issus de la banque PSIMAP (Kim et al., 2004). (B) La base de données considérée est la base de données de validation inter-moléculaires de 11 complexes issus de la banque WENG dans leur état « liés » (Chen et al., 2003). Les modèles d'assemblages évalués ont été générés par le programme FTDock (Gabb et al., 1997), et une vraie solution correspond à un modèle dont le RMSD par rapport à la structure native du complexe est inférieur à 3Å. Figure adaptée de (Madaoui and Guerois, 2007).

4.4.5. Capacités de filtrage de la méthode SCOTCH

L'ensemble des analyses effectuées jusqu'à maintenant montre que l'approche SCOTCH est la plus efficace pour discriminer entre des complexes vrais et des complexes faux générés par docking. Pour évaluer cette capacité discriminative, j'ai utilisé le score absolu de la fonction SCOTCH, qui correspond à la probabilité qu'un modèle soit correct (pour être précis, la fonction SCOTCH, qui est une équation de régression logistique, prédit les « *log odds* » de cette probabilité). Il est également possible de classer les modèles en deux groupes en fonction de cette probabilité. Si cette probabilité est supérieure à 0.5, le modèle est dans ce cas prédit comme correct, sinon, il est prédit comme faux.

Base de données	Méthode	% Faux Positifs filtrés		% Vrais Positifs récupérés	
		moyenne	écart type	moyenne	écart type
INTRA[a]	SCOTCH	86.7	7.6	67.6	43.2
	SCOTCH optimisée*	89.4	5.5	75.2	36.9
INTER[b]	SCOTCH	92.9	4.3	86.5	18.6
	SCOTCH optimisée*	94.1	4.1	81.8	19.6

[a] base de données de validation intra-moleculaire
[b] base de données de validation inter-moleculaire ("état liés")

Tableau 8. Capacité de discrimination de l'approche SCOTCH. Deux variantes de l'approche ont été testées : l'approche SCOTCH standard, et l'approche SCOTCH optimisée par le score RPScore (Moont et al., 1999) (signifiée par une étoile dans le tableau, voir Madaoui et Guerois 2007). Les deux bases de données considérées sont (i) la base de données de validation intra-moléculaires de 28 complexes issus de la banque PSIMAP (Kim et al., 2004) et (ii) la base de données de validation inter-moléculaires de 11 complexes issus de la banque WENG dans leur état « lié » (Chen et al., 2003). Les modèles d'assemblages ont été générés par le programme FTDock (Gabb et al., 1997), puis classés comme bons ou faux par les approches SCOTCH et SCOTCH optimisée en fonction de leur scores. Le tableau transcrit les capacités de filtrage des fausses solutions, ainsi que les capacités de reconnaissance des vraies solutions. Tableau adapté de (Madaoui and Guerois, 2007).

En classant les différents modèles générés par le programme FTDock comme vrais ou faux par la fonction SCOTCH, j'ai voulu évalué les capacités de filtre de cette approche, lors des premières étapes de docking moléculaire. Comme nous pouvons le voir sur le **Tableau 8**, les capacités de filtrage des solutions incorrectes par la méthode SCOTCH sont très satisfaisantes. En effet, 87% et 93% des fausses solutions sont filtrées si l'on considère respectivement les bases de données intra- et inter-moléculaires. De plus, une très large majorité des modèles proches des structures natives de complexes sont reconnues comme correctes (68% dans le cas des interactions intra-moléculaires, et 87% dans le cas des interactions inter-moléculaires). En considérant la méthode hybride (méthode SCOTCH

optimisée par RPScore), une amélioration de la capacité de filtrage est observée (d'environ 3% pour les complexes intra-moléculaires, et 1% pour les complexes inter-moléculaires), ainsi qu'une amélioration substantielle de la sensibilité dans le cas des complexes intra-moléculaires (avec une augmentation du nombre de vrais solutions reconnues de près de 8%).

4.4.6. Le programme RosettaDock

Parmi les méthodes les plus efficaces de docking à l'heure actuelle, figure le programme RosettaDock (Gray et al., 2003). Ce programme, dispose de deux fonctions de score : (i) une fonction de score à basse résolution, issue d'un couplage entre un potentiel statistique, et une fonction d'énergie à basse résolution qui représente les chaînes latérales d'un résidu sous forme d'un centroïde, (ii) une fonction de score à haute résolution, reposant sur une fonction d'énergie prenant en compte explicitement chaque atome de chaque résidu. Au cours des premières étapes de la procédure RosettaDock basées sur une recherche de type corps-rigide, la fonction à basse résolution, peu coûteuse en temps de calcul, est utilisée. Après cette recherche à basse résolution, une optimisation locale de la structure avec une précision atomique est réalisée. Lors de cette étape, plus coûteuse en temps de calcul, la fonction d'énergie à haute résolution est utilisée (Gray et al., 2003).

Pour comparer l'efficacité de ces deux fonctions de score de RosettaDock (score « centroïde » pour la basse résolution et score « tout atome » pour la haute résolution), j'ai tout d'abord évalué leur capacité à reconnaître la structure native d'un complexe, parmi un ensemble de conformations alternatives. Pour cela, j'ai considéré à nouveau la base de données inter-moléculaire de 11 complexes, ainsi que les 10000 modèles générés par FTDock (Gabb et al., 1997) pour chacun de ces complexes. J'ai ensuite calculé pour chaque structure native, et chacun des modèles, le score centroïde ainsi que le score tout atome par le programme RosettaDock (Gray et al., 2003) (voir section 7.2.1). Le **Tableau 9** présente, pour chaque cas, les pourcentages de modèles mieux classés que la structure native. Comme nous pouvons le constater, la structure native est toujours mieux classée par la fonction à haute résolution du programme RosettaDock, ce qui traduit le fait que cette fonction est très efficace pour reconnaître la structure native des complexes (**Tableau 9**). La fonction à basse résolution, bien que moins efficace, parvient cependant à classer dans la plupart des cas la structure native parmi les meilleures. En effet, très peu de modèles sont mieux classés que la structure

native (moins de 5% des modèles sont mieux classés que la structure native pour 9 cas sur 11, voir **Tableau 9**).

Méthodes	Identifiant PDB										
	1AY7	1B6C	1BUH	1E6E	1EAW	1EWY	1FQJ	1GRN	1I2M	1IBR	2PCC
[a]	0.13	0.06	6.99	0.05	1.41	0.01	3.17	0	0.54	0.05	21.21
[b]	0	0	0	0	0	0	0	0	0	0	0

[a] Fonction de score à basse résolution de RosettaDock (Centroïde)
[b] Fonction de score à haute résolution de RosettaDock (Tout atome)

Tableau 9. Pourcentage de modèles mieux classés que la structure native par les fonctions de score du programme RosettaDesign (Gray et al., 2003). La base de données considérée est la base de validation inter-moléculaires de 11 complexes issus de la banque WENG dans leur état « liés » (Chen et al., 2003). Pour chacun de ces cas, 10000 modèles d'assemblage ont été générés par le programme FTDock (Gabb et al., 1997), puis évalués par les fonctions de score à basse résolution et à haute résolution du programme RosettaDesign (Gray et al., 2003). Pour chacun des cas, la structure native a ensuite été évaluée de la même façon, et le pourcentage de solutions mieux classées qu'elle a été déterminé.

Dans un second temps, j'ai évalué la capacité de ces deux fonctions de score, à reconnaître des structures proches des complexes natifs, générées par des approches de docking. Lors des premières étapes de docking moléculaire, un ensemble de solutions sont générées, mais leur précision n'est pas toujours très bonne. Il est donc essentiel de sélectionner par des fonctions de score discriminantes, les solutions les plus proches des structures natives de complexes. Pour chacun des cas de la base de données test, j'ai donc évalué les 10000 modèles produits par le programme FTDock avec les deux fonctions de score du programme RosettaDock, puis déterminé le classement du modèle le mieux classé, à moins de 3 Å de la structure native. Nous pouvons constater que cette fois, la fonction à basse résolution du programme RosettaDock est beaucoup plus efficace que la fonction à haute résolution. En effet, pour 9 cas sur 11, un modèle proche de la structure native est classé parmi les 100 meilleurs modèles par la fonction centroïde de RosettaDock, alors que seuls 2 cas sur 11 remplissent ce critère si l'on considère la fonction à haute résolution du programme (**Tableau 10**).

Methode	Identifiant PDB										
	1AY7	1B6C	1BUH	1E6E	1EAW	1EWY	1FQJ	1GRN	1I2M	1IBR	2PCC
[a]	5	34	43	48	80	42	28	8	107	65	189
[b]	162	408	905	2732	98	446	157	320	396	62	3759

[a] Fonction de score à basse résolution de RosettaDock (Centroïde)
[b] Fonction de score à haute résolution de RosettaDock (Tout atome)

Tableau 10. Classement du premier modèle proche de la structure native du complexe par les fonctions de score du programme RosettaDesign (Gray et al., 2003). La base de données considérée est la base de validation inter-moléculaires de 11 complexes issus de la banque WENG dans leur état « liés » (Chen et al., 2003). Pour chacun de ces cas, 10000 modèles d'assemblage ont été générés par le programme FTDock (Gabb et al., 1997), puis évalués par les fonctions de score à basse résolution et à haute résolution du programme RosettaDesign (Gray et al., 2003). Le classement du modèle le mieux classé, à moins de 3 Å de la structure native est indiqué dans le tableau pour chacun des cas testé.

Les fonctions d'énergie à haute résolution, très efficaces pour reconnaître les structures natives de complexes protéiques, vont pénaliser de façon importante les éventuelles interpénétrations atomiques impliquant des chaînes latérales ou des atomes du squelette peptidique. Ces approches ne sont donc généralement pas adaptées lors des premières étapes de docking, au cours desquelles les modèles produits ne sont pas de très bonne qualité.

4.4.7. SCOTCH : une fonction de score efficace pour le docking protéine-protéine

Un des intérêts des méthodes à basse résolution pour prédire la structure d'assemblages protéiques, est que ces méthodes sont moins sensibles aux détails structuraux que d'autres approches basées sur des fonctions d'énergies précises. Parmi les approches à basse résolution, la fonction centroïde du programme RosettaDock, ou la méthode SCOTCH que j'ai développée au cours de ma thèse, apparaissent comme des solutions potentiellement intéressantes lors des premières étapes de docking moléculaire, où il est nécessaire de pouvoir évaluer rapidement des modèles dont la qualité n'est pas toujours très bonne.

Afin de mesurer le potentiel de la méthode hybride SCOTCH optimisée par RPScore, en comparaison avec la fonction centroïde à basse résolution du programme RosettaDock, j'ai utilisé deux bases de données de structures : (i) la base de données inter-moléculaire de 11 complexes résolus dans leur état « liés », (ii) la même base de données de 11 complexes mais résolus dans leur état « non liés » (cette seconde base de données correspond à des structures de départ plus réalistes pour le docking de protéines dont on ne connaît pas la structure, car

les structures « liées » des partenaires sont souvent différentes des structures « non liées »). Pour chacun des cas, j'ai généré 10000 modèles d'assemblages par le programme FTDock (Gabb et al., 1997), puis évalué ces différents modèles par les fonctions de score à basse résolution du programme RosettaDock, puis par la méthode hybride SCOTCH optimisée par RPScore. J'ai ensuite déterminé pour chacun des cas d'étude, le classement de la première solution à moins de 3 ou 5 Å de la structure native du complexe pour chacune des fonctions de score (le RMSD étant mesuré sur l'ensemble du complexe en considérant le squelette). Les résultats sont synthétisés au niveau du **Tableau 11**.

	RMSD[c]	Méthode	Identifiant PDB										
			1AY7	1B6C	1BUH	1E6E	1EAW	1EWY	1FQJ*	1GRN*	1I2M	1IBR*	2PCC
Liés[a]	3 Å	ROSETTA[d]	5	34	43	48	80	42	28	8	107	65	189
		SCOTCH[e]	2	1	2	3	9	48	267	1	1	1	6
	5 Å	ROSETTA[d]	5	34	43	48	5	11	28	8	51	65	57
		SCOTCH[e]	2	1	2	1	9	5	267	1	1	1	6
Non-liés[b]	3 Å	ROSETTA[d]	1123	1	76	90	4	23	-	-	914	-	382
		SCOTCH[e]	19	1	11	1	458	12	-	-	43	-	123
	5 Å	ROSETTA[d]	68	1	76	2	4	23	201	678	87	-	47
		SCOTCH[e]	6	1	11	1	97	2	305	775	43	-	15

[a] base de données de validation inter-moleculaire (état « liés »)
[b] base de données de validation inter-moleculaire (état « non-liés »)
[c] RMSD seuil définissant une structure proche du complexe natif (3 ou 5 Å)
[d] Fonction de score à basse resolution du programme RosettaDock
[e] Fonction de score SCOTCH optimisée par RPScore

Tableau 11. Classement du premier modèle proche de la structure native du complexe par les fonctions de score du programme RosettaDesign (Gray et al., 2003) et de la fonction SCOTCH optimisée par RPScore (Moont et al., 1999). La base de données considérée est la base de validation inter-moléculaires de 11 complexes issus de la banque WENG dans leur état « liés » ou « non liés » (Chen et al., 2003). Pour chacun de ces cas, 10000 modèles d'assemblage ont été générés par le programme FTDock (Gabb et al., 1997), puis évalués par les 2 fonctions de score. Un trait signifie les cas pour lesquels le programme FTDDock n'a pas réussi à générer un modèle proche de la structure native du complexe. Un complexe est considéré proche de la structure native s'il est à moins de *x* Å de la structure native du complex (avec *x* = 3 ou 5).

En considérant le premier jeu de données (complexes résolus dans leur état « liés ») et un seuil de 3 Å, la méthode SCOTCH optimisée par RPScore sélectionne un complexe proche de la structure native parmi les 10 meilleurs modèles pour 9 cas sur 11, contre 2 cas sur 11 pour la fonction du programme RosettaDesign. En considérant le second jeu de données

(complexes résolus dans leur état « non liés »), nous pouvons noter que pour 3 cas sur 11, le programme FTDock ne génère aucune solution à moins de 3 Å de la structure native. En se focalisant sur les 8 cas pour lesquels au moins une structure native est générée, le score SCOTCH sélectionne une structure proche de la structure native parmi les 20 meilleures structures pour 5 cas sur 8, contre 2 cas sur 8 si l'on considère la fonction à basse résolution du programme RosettaDesign.

Deux conclusions peuvent être tirées de cette étude. D'une part, il apparaît que la méthode SCOTCH est capable de reconnaître des modèles structuraux proches des structures natives de complexes, et ceci de façon plus efficace que la méthode à basse résolution du programme RosettaDock. De plus, la qualité des prédictions de SCOTCH apparaît relativement peu sensible à la précision des structures initiales. Cette insensibilité à la qualité de la structure initiale peut se mesurer par l'analyse de l'aire sous la courbe ROC. Pour les deux jeux de données (structures de complexes résolus dans leur état « liés » ou « non liés ») la précision mesurée par l'aire sous la courbe est identique (égale à 0.96) (**Figure 55**). Si l'on mesure les capacités de filtrage de la méthode SCOTCH, il apparaît que 94% des faux modèles sont filtrés, et 85% des modèles proches des structures natives de complexe sont reconnus si l'on considère les complexes dans leur état « non liés » (Madaoui and Guerois, 2007).

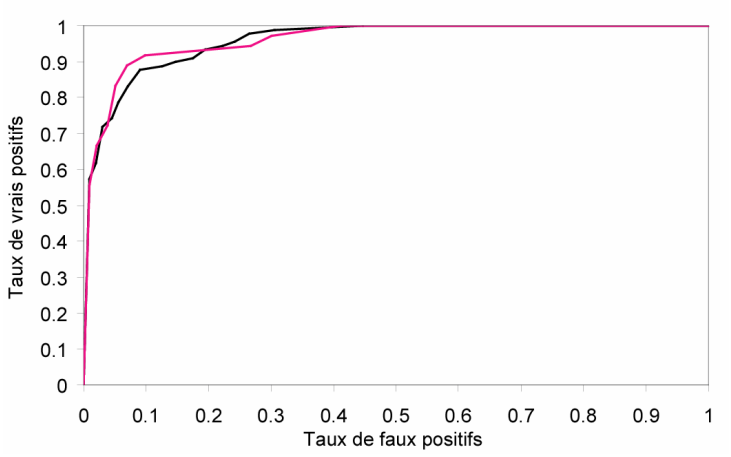


Figure 55. Capacité discriminative de la méthode SCOTCH optimisée par RPScore mesurée par des courbes ROCs. Les bases de données considérées sont la base de données de validation inter-moléculaires de 11 complexes issus de la banque WENG (Chen et al., 2003) dans leur état « liés » (noir) ou « non liés » (rose). Pour chacun de ces cas, 10000 modèles d'assemblage ont été générés par le programme FTDock (Gabb et al., 1997), puis évalués par la fonction de score SCOTCH optimisée par RPScore (Moont et al., 1999). Une solution est considérée comme proche du complexe natif si son RMSD par rapport à la structure native du complexe est inférieur à 3Å. Figure adaptée de (Madaoui and Guerois, 2007).

4.5. Obtention de modèles d'assemblages à haute résolution

4.5.1. Introduction

Dans les sections précédentes, j'ai présenté l'approche SCOTCH, qui se révèle très efficace lors des premières étapes de docking destinées à localiser les sites d'interaction protéiques ainsi que l'orientation relative des partenaires en interaction. Il apparaît qu'en sélectionnant un échantillon restreint de conformations, cette approche permet de récupérer des modèles très proches de la structure native des complexes.

Au cours de ma thèse, j'ai développé un programme autonome, SCOTCHer, destiné à remplacer la première étape de docking par FTDock. L'objectif de ce programme est d'intégrer les contraintes évolutives directement au cours de l'exploration corps rigide à basse résolution et pas seulement comme filtre *a posteriori* des solutions de docking. Sur la base de l'outil SCOTCHer, j'ai également développé un module permettant d'explorer plus finement les modèles de complexes les plus probables identifiés par l'approche SCOTCH. Le principe consiste à perturber chaque modèle retenu, en échantillonnant différentes conformations autour de son site de liaison, et ceci afin de mieux caractériser le paysage énergétique autour de cette région d'interaction. Cette technique est inenvisageable lors des premières étapes de docking car elle est trop coûteuse en temps de calcul. Néanmoins, elle a été décrite comme très pertinente pour évaluer la qualité des différents modèles et discriminer entre des modèles corrects ou erronés. Différents travaux ont montré que des perturbations autour des complexes natifs donnaient lieu à la présence d'un profil énergétique caractéristique en forme d'entonnoir, ou *funnel*, présentant une corrélation entre l'énergie calculée et le RMSD (le RMSD étant lié à la perturbation réalisée autour du complexe de départ, voir **Figure 56**) (Camacho et al., 2000).

L'analyse du profil énergétique d'une structure de complexe passe par plusieurs étapes essentielles : (i) génération de plusieurs modèles par perturbations locales autour de la structure cible, (ii) évaluation de chacun des modèles par une fonction de score énergétique, (iii) détermination de la corrélation entre l'intensité de la perturbation (quantifiée par un critère de RMSD par rapport à la structure à analyser) et l'énergie d'interaction.

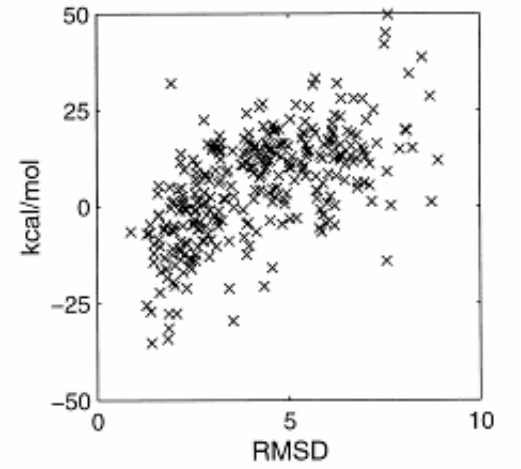


Figure 56. Illustration du principe de « *funnel* ». Nous pouvons observer une corrélation entre l'énergie (mesurée en Kcal/mol) et le RMSD. Figure extraite de (Camacho et al., 2000).

4.5.2. Perturbations locales des structures de complexes

Pour effectuer la première étape, destinée à perturber la structure du complexe à analyser, SCOTCHer procède à des opérations de translations et de rotations de la structure, suivies d'une optimisation de la conformation des chaînes latérales au niveau de la région d'interaction. Ainsi, après avoir perturbé légèrement la structure du plus petit des deux partenaires par une opération de translation (suivant l'axe passant par le centre de gravité des deux molécules), une rotation aléatoire est appliquée à cette molécule. En ce qui concerne les opérations de rotation, j'ai opté pour une méthode mathématique plus efficace que la méthode d'Euler : la méthode des quaternions. Au lieu de tourner un objet à partir d'une série de rotations successives comme c'est le cas pour les angles d'Euler, les quaternions permettent de le tourner autour d'un axe arbitraire et d'un angle quelconque. Un premier intérêt de la méthode des quaternions est qu'elle permet d'obtenir des transformations beaucoup moins redondantes qu'en utilisant la méthode d'Euler (Kuffner, 2004) (**Figure 57**). Par ailleurs, il est possible de calculer très rapidement les transformations spatiales à effectuer sur les coordonnées d'une protéine pour la faire tourner comme une rotule autour de la surface de son partenaire (méthode d'interpolation appelée « *slerp* » pour « *spherical linear interpolation* » illustrée **Figure 58**).

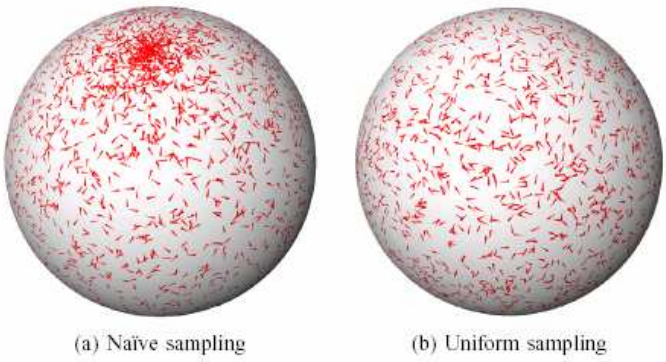


Figure 57. Comparaison entre (A) un échantillonnage naïf des rotations par variation systématique des angles d'Euler (B) échantillonnage des rotations obtenu par variation systématique des quaternions. Figure extraite de (Kuffner, 2004).

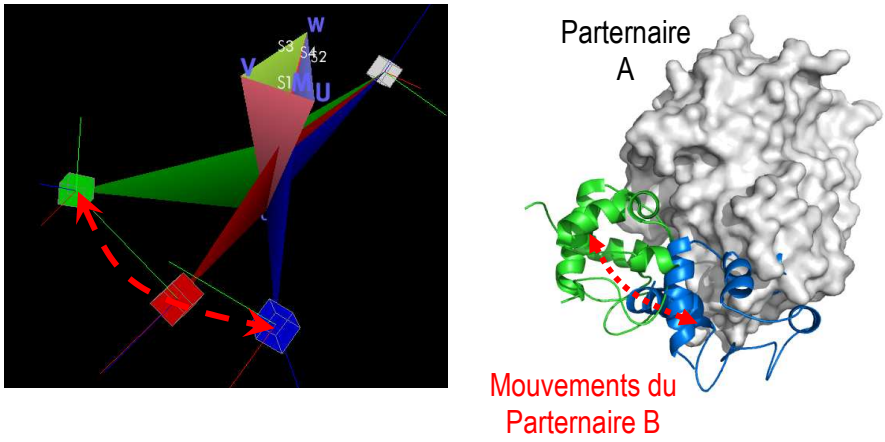


Figure 58. La méthode d'interpolation « *slerp* » appliquée aux quaternions permet de calculer très rapidement les transformations à effectuer sur les coordonnées d'une protéine pour effectuer des rotations le long d'une trajectoire sur une sphère entre deux états symbolisés ici par les cubes vert et bleu (la position des protéines correspondantes est indiquée sur le schéma de droite).

À chaque pas de perturbation sur la structure d'un complexe, le programme SCOTCHer procède à une optimisation de la conformation des chaînes latérales par des techniques issues de la théorie des graphes. Pour cela, les contacts entre les chaines latérales de l'interface de la structure du complexe sont représentés sous la forme d'un graphe décrivant les interactions entre chaque résidu (**Figure 59**). Cette représentation permet de hiérarchiser la recherche combinatoire de la meilleure orientation des chaînes latérales ou rotamères[2] et conduit à un gain sensible dans la vitesse d'exploration conformationnelle.

[2] Un rotamère est un élément d'un ensemble de conformères résultant d'une rotation restreinte autour d'une liaison simple

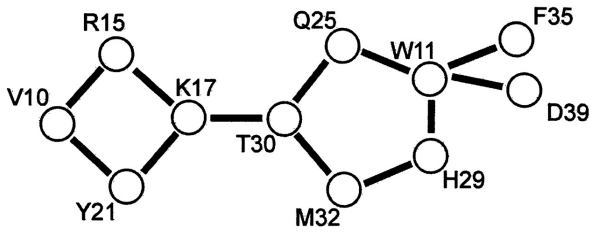


Figure 59. Représentation des contacts entre paires de résidus sous la forme d'un graphe. Figure extraite de (Canutescu et al., 2003).

Pour l'optimisation des chaînes latérales à l'interface des partenaires en interaction, deux étapes sont nécessaires. La première consiste à rechercher les composantes connexes du graphe qui sont des ensembles maximaux tel qu'il existe un chemin entre deux sommets quelconque leur appartenant. L'identification de composantes connexes permet de réduire le coût algorithmique en subdivisant le graphe initial en sous graphes qui peuvent être optimisés indépendamment.

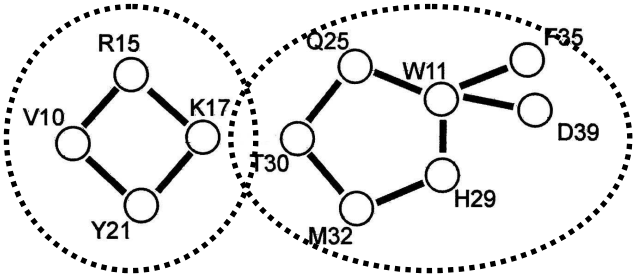


Figure 60. Représentation des contacts entre paires de résidus sous la forme d'un graphe. Les composantes connexes du graphe sont entourées par des pointillés. Figure adaptée de (Canutescu et al., 2003).

Dans l'exemple exposé à la **Figure 60**, deux composantes connexes sont présentes. Lorsque le nombre de nœud d'un graphe connexe est important, comme c'est le cas dans l'exemple de la **Figure 59** (11 nœuds), le temps de calcul nécessaire pour déterminer la combinaison optimale des rotamères devient trop lourd. Un nouvel algorithme, proposé dans la dernière version du programme SCWRL (Canutescu et al., 2003), a donc été implémenté lorsque le nombre de nœud est supérieur à 5. Il consiste à dissocier les graphes connexes en sous

graphes par la recherche des composantes biconnexes[3]. En corollaire, lorsque aucun cycle ne passe par une arête, celle-ci forme une composante biconnexe à elle seule). Les graphes biconnexes peuvent aussi bien être des cycles que des paires de résidus connectées par une arête. À titre d'exemple, le graphe de la **Figure 61** peut être subdivisé en 5 composantes biconnexes : A (résidus V10, R15, Y21, et K17), B (résidus K17 et T30), C (résidus T30, Q25, M32, H29, et W11), D (résidus W11 et F35) et E (résidus W11 et D39).

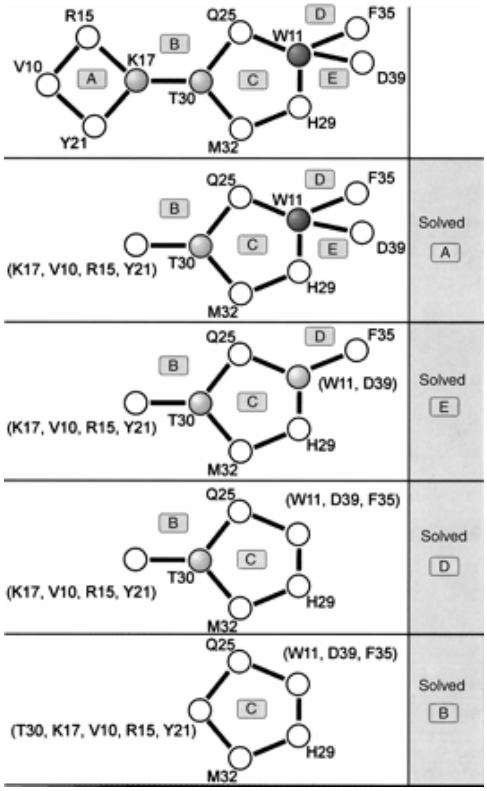


Figure 61. Résolution d'un problème d'optimisation par la recherche des composantes biconnexes d'un graphe. L'énergie minimum associée au graphe représenté à la **Figure 59** est déterminée par l'identification successive de l'énergie minimum associée à chacune des composantes biconnexes du graphe. Chacune des composantes biconnexes sont résolues, comme indiqué dans la colonne de droite, puis agrégées pour ne former plus qu'une seule composante. Figure extraite de (Canutescu et al., 2003).

[3] Une composante biconnexe est un ensemble maximal d'arêtes (non orientées) qui vérifient la propriété suivante : « Soient *a* et *b* deux arêtes distinctes, alors il existe un cycle simple (qui ne passe pas deux fois par le même sommet) qui passe par *a* et *b* ».

Les nœuds apparaissant dans plus d'une composante biconnexe sont appelés « points d'articulation[4] ». L'élimination des points d'articulation va donc avoir pour effet la subdivision du graphe initial en sous graphes. Au niveau du graphe représenté à la **Figure 59**, les résidus K17, T30 et W11 sont des points d'articulation. Afin de trouver les composantes biconnexes d'un graphe, ainsi que leur points d'articulation, j'ai implémenté dans SCOTCHer l'algorithme développé par Tarjan (Tarjan, 1972). Cet algorithme, de la même manière que pour la recherche des composantes connexes d'un graphe, utilise un algorithme de parcours en profondeur de l'arbre. Une fois les composantes biconnexes mises en évidence, le protocole va consister à déterminer la combinaison optimale de rotamères pour chacune des composantes indépendamment, puis à les prendre en compte simultanément. Les composantes biconnexes sont traitées en fonction du nombre de point d'articulation qu'elles contiennent (les composantes ayant le plus petit nombre de points d'articulation sont donc traitées en premier). Ainsi, en considérant la **Figure 61**, la première composante à traiter est la composante A qui ne contient qu'un seul point d'articulation. Pour cette composante, la combinaison de rotamères donnant la plus basse énergie est sélectionnée (et ceci pour chacun des rotamères du point d'articulation), puis le point d'articulation est mis à jour en lui agrégeant la combinaison de rotamères optimales ainsi déterminée (le point d'articulation peut être considéré comme un « super résidu », contenant des combinaisons de rotamères ou « super-rotamères »). Ainsi, en traitant successivement chacune des composantes biconnexes, nous arrivons à déterminer la combinaison optimale de rotamères pour l'ensemble du graphe.

Lors d'une dernière étape, l'énergie d'interaction de chacun des modèles issus de la perturbation de la structure de départ est évaluée avec la fonction d'énergie FOLDX (Guerois et al., 2002; Schymkowitz et al., 2005). Enfin, après avoir déterminé le RMSD de chacun des modèles perturbés par rapport à la structure de départ, un graphe analysant l'énergie d'interaction des modèles en fonction de leur RMSD par rapport à la structure cible (« *binding funnel plot* ») est construit, puis la corrélation entre ces deux variables est mesurée.

[4] Les points d'articulation sont les sommets du graphe qui sont partagés par plusieurs composantes biconnexes

4.5.3. Profil énergétique des structures natives de complexes

La première condition importante à tester pour évaluer l'intérêt de l'approche par perturbations consiste à vérifier que le programme SCOTCHer est à même de distinguer un *funnel* pour des structures de complexes natives. Pour cela j'ai analysé les « *binding funnel plot* » de chacune des structures natives de complexes de la banque de données inter-moléculaire introduite dans la section 4.2. Pour chaque cas, 2000 structures perturbées ont été générées puis évaluées par la procédure introduite dans la section précédente (voir section 7.2.2). La **Figure 62** montre que pour la majorité des complexes analysés, il existe une corrélation entre l'énergie FOLDX et le RMSD pour les structures située à moins de 3 Å de la structure native qui correspond à la définition du *funnel* (Wollacott et al., 2007).

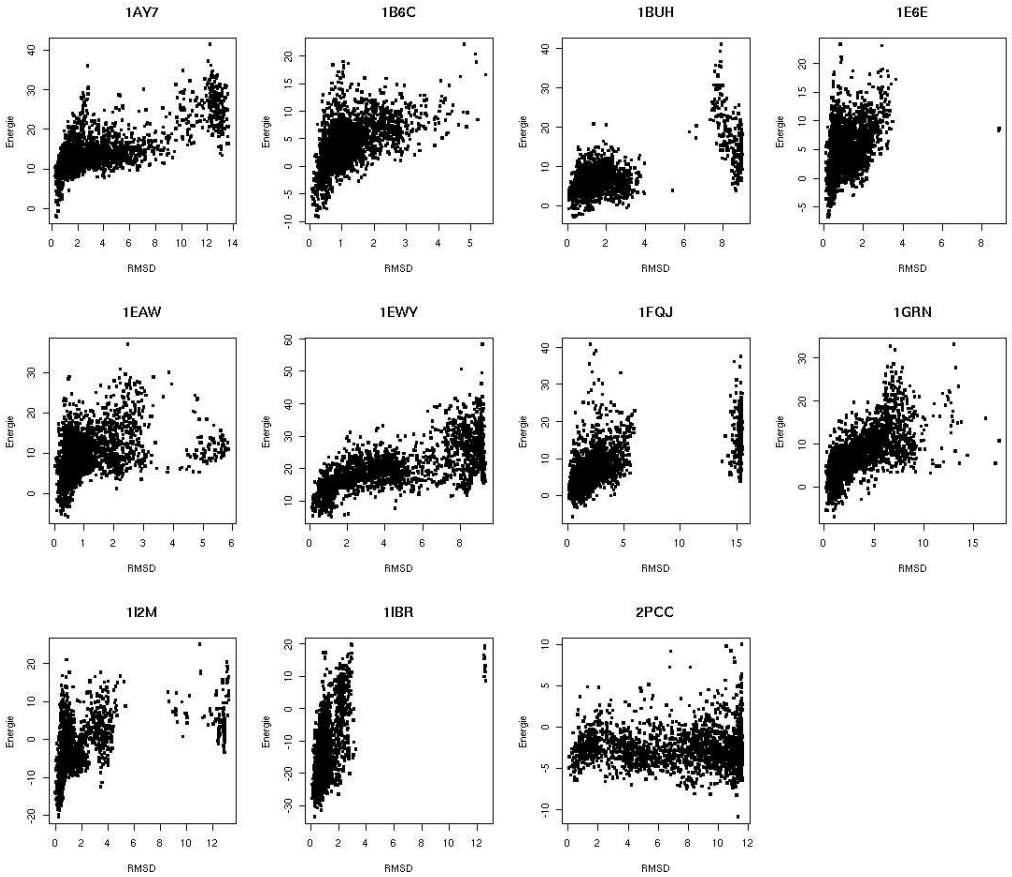


Figure 62. Profil énergétique des structures natives de complexes. La base de données considérée est la base de données de validation inter-moléculaires de 11 complexes issus de la banque WENG (Chen et al., 2003) dans leur état « liés ». Pour chaque complexe natif, 2000 structures proches ont été générées par perturbation, puis un graphe a été généré représentant l'énergie de ces structures en fonction de leur RMSD par rapport au complexe natif. Les structures ont été produites par le programme SCOTCHer, et l'énergie de ces structures a été évaluée par le programme FOLDX (Guerois et al., 2002; Schymkowitz et al., 2005).

Dans un seul cas, celui dont le code est 2PCC, aucune corrélation n'est observée entre RMSD et énergie. En examinant plus particulièrement ce cas, il apparaît que ce complexe présente certaines caractéristiques inhabituelles des complexes protéiques (Chen et al., 2003). En particulier, de larges cavités sont présentes au niveau du site de liaison de ce complexe, alors que la densité de contact des autres complexes analysés dans cette étude est beaucoup plus importante.

4.5.4. Le profil énergétique permet-il d'identifier les structures natives de complexes parmi les structures générées à basse résolution ?

Nous avons vu qu'une analyse du bassin énergétique de structures de complexes natifs se caractérisait par un profil caractéristique, appelé *funnel*, qui correspond à une dépendance entre l'énergie d'interaction d'un complexe et les légères perturbations appliquées à leurs structures. Il a été proposé que ce type d'analyse pouvait permettre de discriminer les vrais structures natives de fausses solutions alternatives (Wollacott et al., 2007). Si cette hypothèse était vérifiée, une telle approche pourrait alors permettre de filtrer les prédictions faites avec l'approche SCOTCH (qui sélectionne en effet un nombre très restreint de solutions possibles parmi lesquelles la structure native est souvent présente) et fournir un modèle unique d'assemblage pour le complexe étudié de bonne qualité.

Afin de vérifier cette hypothèse, j'ai procédé à l'étude du complexe barnase-barstar résolu dans son état lié (code PDB 1AY7). Si on génère 10000 modèles d'assemblage possible par le programme FTDock (Gabb et al., 1997), et qu'on sélectionne les 10 meilleurs modèles suivant le score SCOTCH, 3 modèles proches de la structure native, situées respectivement à 0.531, 0.684 et 1.028 Å de la structure native du complexe sont retenus et classés respectivement en 2ème, 4ème et 10ème position. Pour chacun de ces modèles, j'ai analysé leur bassin énergétique, en construisant leur « *binding funnel plot* » à l'aide du programme SCOTCHer (**Figure 63**).

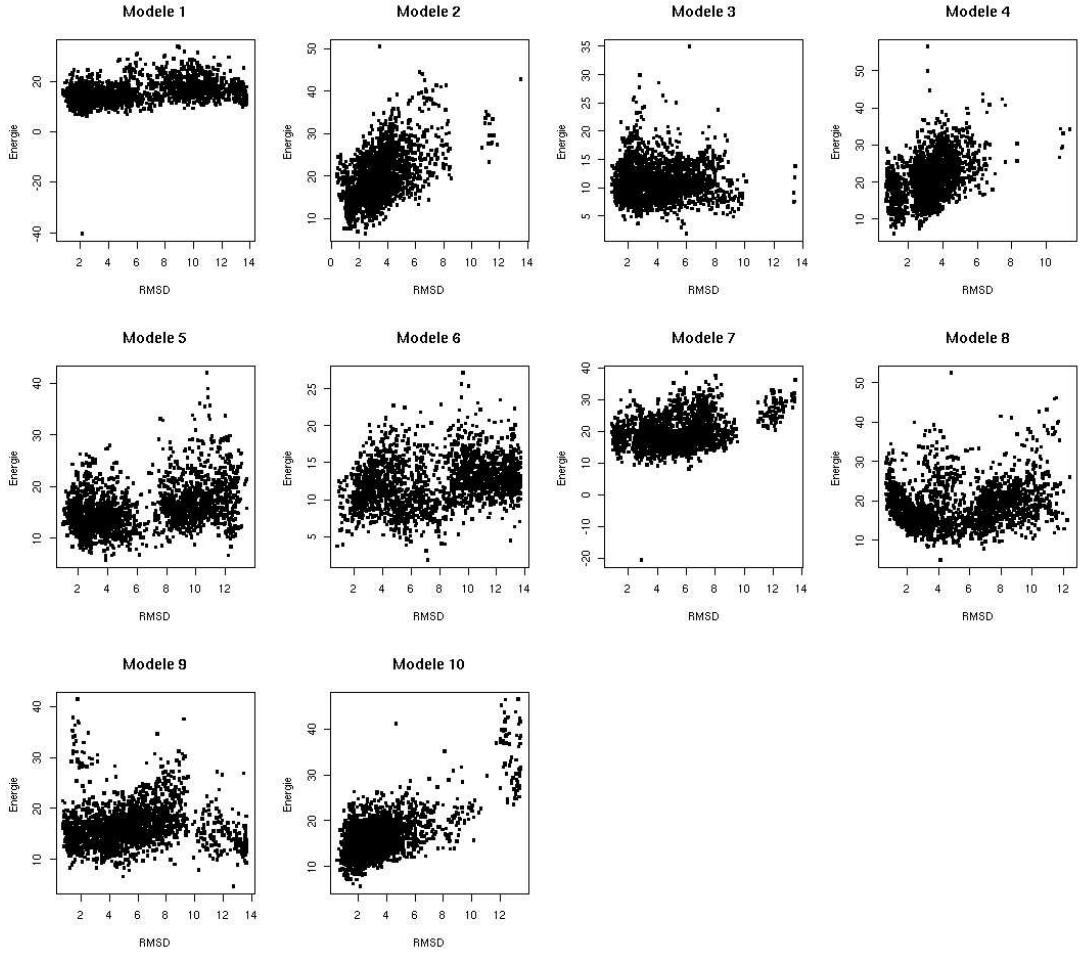


Figure 63. Profil énergétique des meilleurs modèles retenus sur la base du score SCOTCH optimisé par RPScore (Moont et al., 1999). Le complexe étudié est le complexe Barnase-Barstar (code PDB 1AY7) résolu dans son état « liés ». Pour ce complexe, 10000 modèles d'assemblage ont été générés par le programme FTDock (Gabb et al., 1997), puis évalués par la fonction de score SCOTCH optimisée par RPScore. Les dix meilleurs modèles sur la base de ce score ont été retenus, puis soumis à une analyse par perturbation par le programme SCOTCHer. Pour chacun de ces modèles, 2000 structures proches ont été générées par perturbation, puis un graphe a été généré représentant l'énergie de ces structures en fonction de leur RMSD par rapport aux modèles de départ. Les structures ont été produites par le programme SCOTCHer, et l'énergie de ces structures a été déterminé par le programme FOLDX (Guerois et al., 2002; Schymkowitz et al., 2005).

De façon intéressante, il apparaît que seules les solutions proches de la structure native présentent un profil énergétique caractéristique d'un *funnel*, ce qui est quantifié par le calcul des corrélations « énergie-RMSD » de chaque modèle (**Tableau 12**). De plus, si l'on représente maintenant l'intensité de cette corrélation en fonction du RMSD de chaque modèle par rapport à la structure native, nous constatons une anti-corrélation très forte de ces deux variables, égale à -0.82 (**Figure 64**). Notons que l'énergie absolue calculée avec FOLDX pour ces différents modèles n'aurait pas permis de discriminer entre bonnes et

mauvaises solutions. L'existence du « *funnel* » énergétique est donc une propriété tout à fait remarquable.

classement[a]	RMSD[b]	correlation[c]
1	6,152	0,30
2	0,531	0,53
3	11,667	0,033
4	0,684	0,46
5	4,699	0,38
6	7,664	0,29
7	3,042	0,38
8	4,24	0,18
9	5,009	0,083
10	1,028	0,69

[a] classement par la fonction de score SCOTCH
[b] RMSD par rapport à la structure native du complexe
[c] Corrélation entre perturbation et énergie d'interaction

Tableau 12. Corrélation entre RMSD et énergie des meilleurs modèles d'assemblage du complexe 1AY7. Les 10 meilleurs modèles d'assemblage des protéines Barnase et Barstar ont été sélectionnés par l'approche SCOTCH optimisée par RPScore (Moont et al., 1999).

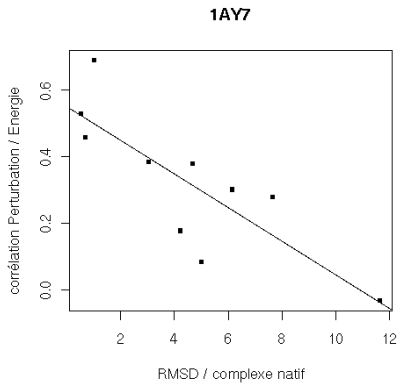


Figure 64. Intensité du *funnel* en fonction de la distance au complexe natif. Les 10 meilleurs modèles d'assemblage des protéines Barnase et Barstar ont été sélectionnés par l'approche SCOTCH optimisée par RPScore (Moont et al., 1999). Pour chacun de ces modèles, un « *binding funnel plot* » a été construit, et l'intensité du *funnel* quantifiée. Le graphique représente l'intensité de ce *funnel*, en fonction du RMSD de chacun des modèles par rapport à la structure native du complexe.

Ces résultats démontrent qu'une analyse à haute précision des modèles d'assemblages protéiques permet de discriminer la structure native d'autres solutions incorrectes. Dans l'idéal, cette analyse devrait être réalisée pour tous les modes d'assemblages possibles entre deux partenaires, lors des premières étapes de docking. Cependant, le temps de calcul nécessaire à sa réalisation (en moyenne deux jours de calcul sur un processeur Xeon 3Ghz par modèle), principal facteur limitant, ne permet son application qu'à un nombre limité de modèles.

4.6. Applications de la méthode SCOTCH à des complexes étudiés au sein de l'équipe.

Dans cette section, j'ai souhaité évaluer l'intérêt que pouvait présenter la méthode SCOTCH pour la caractérisation de complexes sur des cas pratiques étudiés ces dernières années au sein de l'équipe. De la mise en évidence des interactions à la résolution des structures de ces complexes, il s'est écoulé entre 1 et 2 ans de travail expérimental. L'approche SCOTCH pourrait constituer une alternative pour obtenir un premier modèle de structure, et guider rapidement la proposition de mutants pour les biologistes cellulaires et les généticiens qui souhaitent exploiter ces structures.

4.6.1. Application 1 : Modélisation structurale du complexe entre les domaines CS et ATPase de Hsp90

Les protéines chaperons telles que la protéine Hsp90, sont impliquées dans un grand nombre de voies de signalisation et constituent un des centres d'intérêt de notre équipe. Certaines protéines, décrites sous le terme de co-chaperons, participent à la régulation et à la spécificité de reconnaissance de la protéine Hsp90. Parmi ces protéines, SGT1 apparaît comme un acteur essentiel de certaines voies de signalisation telles que les voies d'ubiquitination dans la levure, de formation du kinétochore chez l'homme et la levure ou encore des voies de résistance aux agents pathogènes chez les plantes. En 2003, Takahashi et collaborateurs ont montré que la protéine SGT1 interagissait physiquement avec la protéine Hsp90 chez *A. thaliana*, et que deux domaines étaient impliquées dans cette interaction : le domaine central CS de SGT1 et le domaine ATPase de Hsp90 (Takahashi et al., 2003) (**Figure 65**).

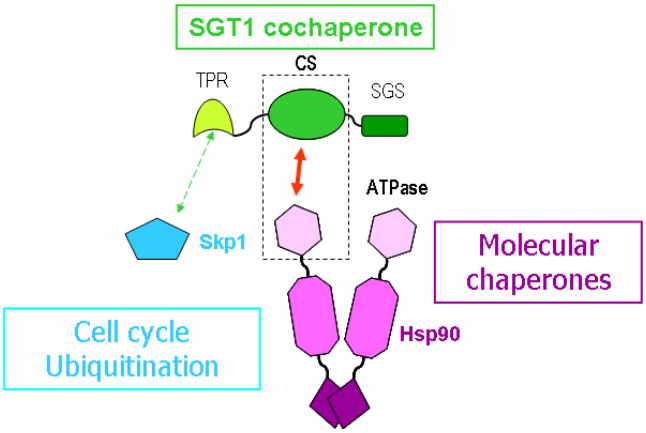


Figure 65. Interaction entre la protéine chaperon Hsp90 et son co-chaperon SGT1. Les domaines CS de SGT1 et ATPase de Hsp90 qui interagissent entre eux sont encadrés.

Les auteurs de cette étude ont alors initié une collaboration avec certains membres de notre équipe, afin de déterminer la structure du complexe impliquant ces deux domaines protéiques. Béatrice Amigues, au cours de son travail de thèse, a caractérisé expérimentalement la structure du complexe entre ces deux domaines, par des techniques de résonance magnétique nucléaire (RMN), et a identifié ainsi de façon précise le site d'interaction impliquant ces deux domaines protéiques (Boter et al., 2007). Cependant, l'orientation relative des partenaires et les contacts exacts entre acides aminés de l'interface n'étaient pas connus. À partir de ces données expérimentales, j'ai donc utilisé l'outil de docking dédié HADDOCK, décrit dans la section 1.5.3, pour construire des modèles du complexe.

Le complexe impliquant les domaines ATPase de Hsp90 et CS de SGT1 apparaît comme un modèle d'étude intéressant pour tester les capacités prédictives de l'approche SCOTCH. En effet, ces domaines protéiques sont conservés chez tous les eucaryotes, de la levure à l'homme (Lee et al., 2004), ce qui rend accessible son étude par des techniques évolutives. De plus, les données structurales et biochimiques obtenues par Béatrice Amigues permettent de modéliser la structure du complexe, et ainsi de valider ou réfuter les prédictions obtenues avec la méthode SCOTCH. D'un point de vue méthodologique, le cas Hsp90/SGT1 est intéressant car il fait intervenir deux protéines avec des caractéristiques évolutives très différentes. La séquence de Hsp90 est beaucoup plus conservée (70 % d'identité entre l'homme et *S. cerevisiae*) que celle du domaine CS de SGT1 (33 % d'identité entre homme et *S. cerevisiae*)

Détermination de la structure du complexe par l'intégration de données RMN dans le logiciel de docking HADDOCK. La structure à haute résolution du complexe entre les domaines CS de SGT1 et ATPase de Hsp90 n'est pas encore connue. On dispose seulement, pour valider les prédictions de la méthode SCOTCH, d'un ensemble de contraintes expérimentales obtenues par Béatrice Amigues par des analyses RMN couplées à des expériences de mutagénèse ciblées. Une première étape de la validation a donc consisté à générer l'ensemble des structures de complexes compatibles avec ces contraintes expérimentales. Par des techniques de RMN, Béatrice Amigues a au cours de sa thèse, identifié différents résidus impliqués dans l'interaction entre les domaines CS de SGT1 et

ATPase de Hsp90. Pour cela j'ai utilisé le programme dédié à ce type de problématique, nommé HADDOCK (Dominguez et al., 2003). J'ai introduit les données RMN et de mutagénèse sous forme de contraintes ambigües, afin de produire des modèles d'assemblages compatibles avec ces données. Pour cela, j'ai définit deux classes de résidus : les résidus actifs affectés par la présence du partenaire protéique en RMN, et les résidus passifs, voisins de ces résidus au niveau structural, et accessibles au solvant (voir section 7.2.3). Ces résidus sont présentés pour chacun des deux domaines à la **Figure 66**.

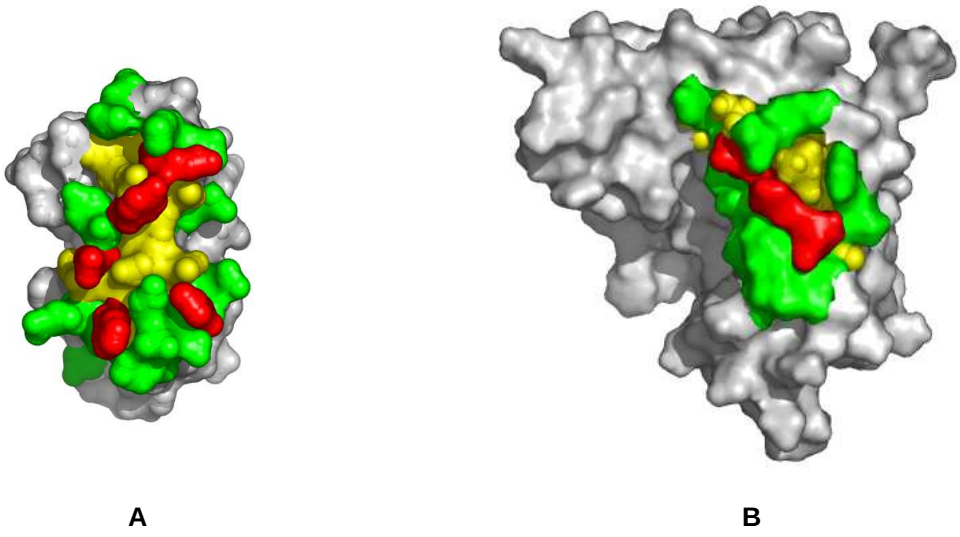


Figure 66. Définition des résidus actifs et passifs pour le docking moléculaire sous HADDOCK pour le domaine CS (A) et pour le domaine ATPase de Hsp90 (B). Les résidus actifs sont représentés en rouge, les résidus passifs en vert, et les résidus peu accessibles au solvant en jaune.

Après la première phase de docking rigide, 1000 modèles on été générés, puis classés suivant leur énergie d'interaction. Les 200 meilleurs modèles sont ensuite soumis à la seconde phase de recuit simulé semi-flexible, puis à la troisième phase de raffinement en solvant explicite. À l'issue de ces différentes étapes, les 200 meilleures structures ont été classées en sous ensembles à partir d'un critère de RMSD calculé par paires de complexes, au niveau des interfaces (le seuil de classification étant de 3Å).

En analysant les structures obtenues à l'issue de cette procédure, nous avons constaté que ces différents modèles se regroupaient en deux principaux clusters structuraux (**Figure 67**). Pour

chacun des deux clusters, le modèle de plus basse énergie a été retenu comme représentatif du cluster (**Figure 68**).

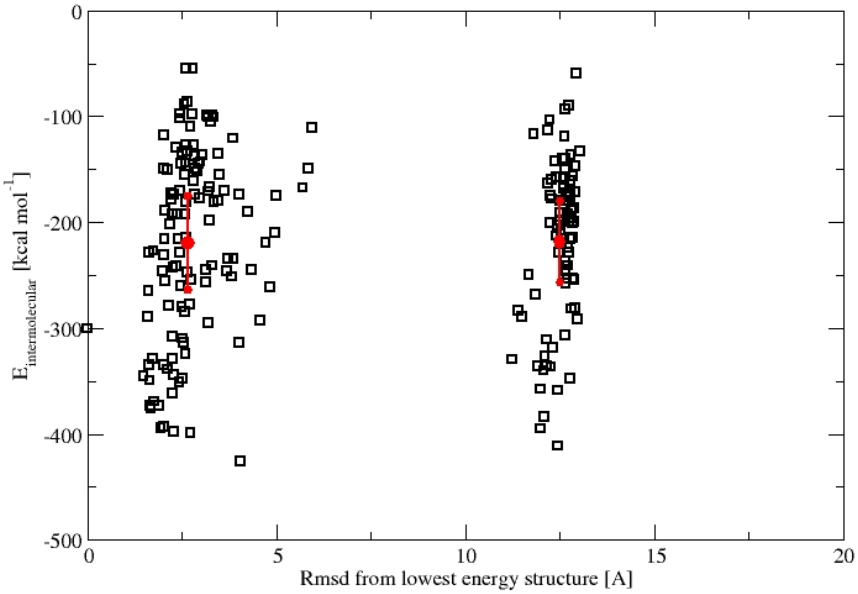


Figure 67. Energies d'interaction inter-moléculaires des modèles en fonction de leur RMSD par rapport au modèle de plus basse énergie. Les carrés correspondent aux modèles, et les cercles aux moyennes des deux principaux clusters avec leur écart type.

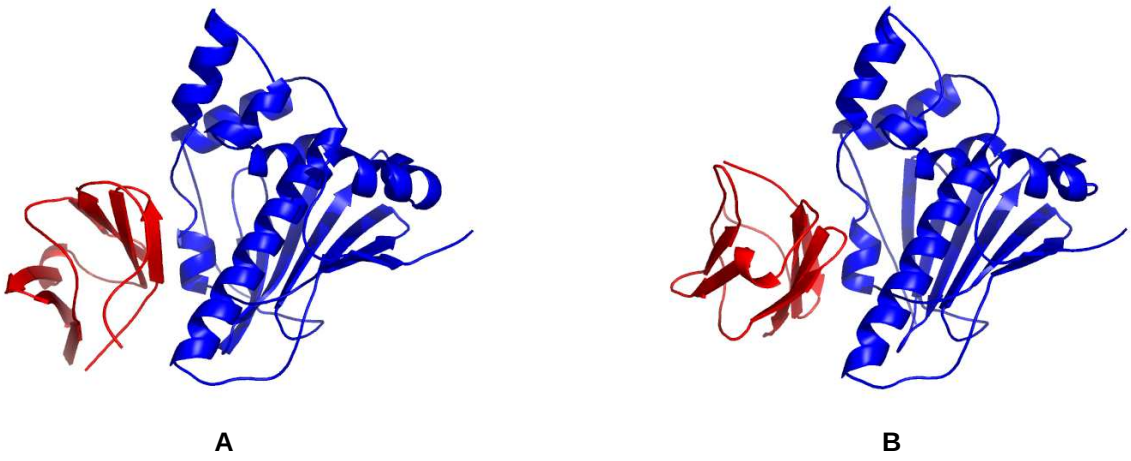


Figure 68. Modèles d'assemblages des domaines CS (en rouge) et du domaine ATPase de Hsp90 (en bleu). (A) modèle de plus basse énergie du cluster 1. (B) modèle de plus basse énergie du cluster 2. Ces modèles d'assemblages correpondent aux deux modes d'assemblage possibles identifiés par le programme HADDOCK (Dominguez et al., 2003).

Les contraintes expérimentales obtenues par la RMN et les premières données de mutagénèse ne permettent donc pas de discriminer entre deux modes d'assemblage très différents. Ce cas de figure particulier est probablement dû à une symétrie dans la répartition des résidus chargés et hydrophobes exposés au niveau de la région du domaine CS

impliquée dans l'interaction. Pour lever l'ambiguïté entre ces deux orientations, de nouvelles mutations ont été proposées à nos collaborateurs expérimentalistes. L'objectif de ces nouveaux mutants n'était plus d'abroger l'interaction entre SGT1 et Hsp90 mais de restaurer l'interaction perdue par un premier mutant. De tels mutants sont appelés compensatoires et impliquent généralement des acides aminés chargés dont on inverse la charge. Une paire de mutations sur les résidus E223 de SGT1 et K88 de Hsp90 permet d'observer ce phénomène de compensation et constitue donc une indication précise d'une interaction entre ces deux résidus (**Figure 69**). Cette interaction n'est possible que dans le cas du cluster 1, et indique que le modèle représentatif de ce cluster correspond à la structure la plus proche de la structure native.

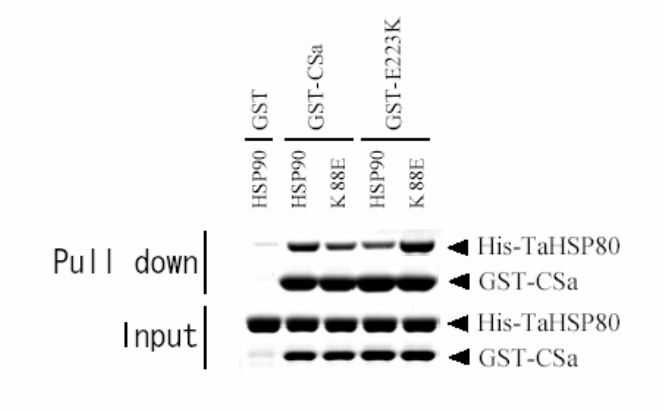


Figure 69. Expérience de GST-pulldown effectuée en liant soit le domaine CS de SGT1 sauvage (*A. thaliana*) soit un mutant ponctuel de ce domaine (E223K) à une colonne de GSH-agarose grâce à une fusion avec la GST. Une forme sauvage de Hsp90 (HSP80 chez *A. thaliana*) ou une forme mutée sur la position K88E est plus ou moins retenue en fonction de l'affinité entre les protéines. Dans ce cas, le mutant Hsp90-K88E restaure l'interaction avec GST-CS-E223K et correspond donc à une mutation compensatoire.

Comparaison avec les modèles d'assemblage les plus probables sélectionnés par l'approche SCOTCH. Par une recherche de séquences protéiques homologues aux domaines CS et ATPase de Hsp90 dans les bases de données de séquences protéiques, j'ai identifié 41 espèces présentant des orthologues potentiels de ces protéines. J'ai ensuite aligné ces séquences en vue d'une analyse par la méthode SCOTCH (voir section 7.2.3).

La méthode SCOTCH optimisée par RPScore a ensuite été appliquée afin de prédire l'orientation entre les domaines CS de SGT1 et ATPase de Hsp90. Pour cela, 10000 orientations ont été générées par le programme FTDock, puis les 20 meilleurs modèles ont été retenus sur la base de leur score évolutif.

Parmi les solutions retenues par l'approche SCOTCH, le modèle n°12 est située à 3.9 Å de la solution correspondante au cluster 1 du programme HADDOCK (**Figure 70**). Aucune solution proposée par SCOTCH ne s'apparente au mode d'assemblage sélectionné dans le cluster 2. Ainsi, les prédictions obtenues avec la méthode SCOTCH apparaissent en très bon accord avec les données expérimentales dont on dispose à l'heure actuelle. De plus, l'obtention d'un mutant compensatoire est extrêmement intéressante d'un point de vue physiologique pour comprendre le rôle de l'interaction spécifique entre SGT1 et Hsp90. Dans le cas de ce complexe, je n'ai malheureusement pas eu le temps d'explorer si la structure sélectionnée à basse résolution donnait lieu à un *funnel* énergétique quand on la perturbait de quelques angströms (voir section 4.5). C'est une des étapes très intéressantes de ce projet qu'il reste à effectuer.

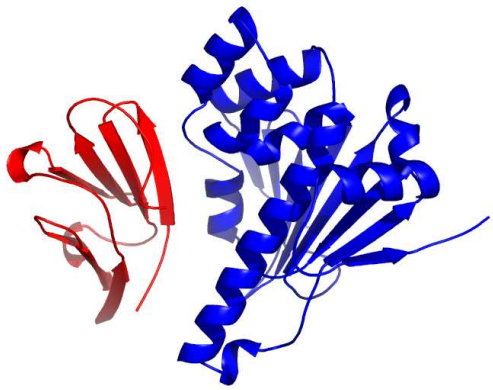


Figure 70. Modèle d'assemblages des domaines CS (A) et du domaine ATPase de Hsp90 prédit par la méthode SCOTCH optimisée par RPScore (Moont et al., 1999). 10000 modèles d'assemblages ont été produits par le programme FTDock (Gabb et al., 1997), puis les 20 meilleures solutions ont été retenues. Le modèle présenté a été classé en 12ème position par l'approche SCOTCH optimisée par RPScore. Le domaine CS est représenté en rouge, le domaine ATPase est représenté en bleu.

4.6.2. Application 2 : modélisation structurale de l'interaction Asf1 – Histone H3

Une autre protéine chaperon qui intéresse notre équipe est la protéine Asf1, dont l'une des fonctions essentielles est d'être un chaperon d'histones. En effet, il a été montré qu'Asf1 était impliquée dans l'assemblage du nucléosome de par ses interactions *in vivo* avec les histones H3/H4. Au sein de l'équipe, les travaux de thèse de Florence Mousson et d'Aurélie Lautrette

ont permis d'une part de résoudre la structure tridimensionnelle d'Asf1 et d'autre part d'identifier la région d'interaction d'Asf1 avec l'hélice C-terminale de l'histone H3 (Mousson et al., 2005). Plus récemment, deux structures du complexe impliquant la protéine Asf1 et les protéines H3-H4 ont été résolues par cristallographie aux rayons X (English et al., 2006; Natsume et al., 2007).

De même que pour le cas du complexe impliquant les domaines CS de SGT1 et ATPase de Hsp90, le nombre de séquences disponibles dans les bases de données génomiques peuvent permettre l'analyse de ce complexe protéique par la procédure SCOTCH. Jusqu'à très récemment, la connaissance de la région d'interaction au niveau de l'histone H3 était la seule information disponible par les biologistes (Mousson et al., 2005). En se basant sur cette information partielle, une des questions que l'on peut se poser est donc de savoir si elle est suffisante à la procédure SCOTCH pour permettre l'obtention d'un modèle unique d'interaction entre ces deux protéines.

Sélection des modèles d'assemblage les plus probables par l'approche SCOTCH. Par une procédure identique à celle décrite pour le complexe impliquant les domaines CS et ATPase de Hsp90, la méthode SCOTCH a été appliquée afin de prédire l'orientation de la protéine Asf1 et de l'histone H3. Pour cela, 10000 orientations ont été générées par le programme FTDock, puis les 20 meilleurs modèles ont été retenus sur la base de leur score SCOTCH (voir section 7.2.4).

Validation de l'approche SCOTCH. Après avoir sélectionné les 20 meilleurs modèles sur la base du score SCOTCH, j'ai recherché parmi ces différentes solutions s'il existait un modèle de complexe impliquant la région C-terminale de l'histone H3. Un seul modèle, classé en 5ème position par cette procédure, répond effectivement à ce critère de sélection. En comparant plus précisément ce modèle avec la structure du complexe Asf1 - Histone H3-H4 résolue par l'équipe de Tyler (English et al., 2006) (code PDB 2HUE), il apparaît que ce modèle est très proche de la structure native du complexe protéique (RMSD de 1.83 Å par rapport à la structure native si l'on considère le squelette peptidique, voir **Figure 71**).

A **B**

Figure 71. Structure prédite du complexe Asf1 – Histone H3-H4. (A) Structure cristallographique du complexe Asf1-Histone H3-H4 (code PDB 2HUE). (B) Modèle retenu par l'approche SCOTCH optimisée par RPScore (Moont et al., 1999).

4.7. Discussion

Ces dernières années, différentes approches basées sur l'analyse des propriétés évolutives des interfaces de complexes protéiques ont été proposées afin de guider la prédiction des assemblages de complexes macromoléculaires. Ainsi, deux types d'approches, basées sur l'hypothèse d'une conservation des interfaces de complexes protéiques, ou encore sur leur co-évolution, ont été largement utilisées pour prédire aussi bien des complexes de type intra-moléculaires, qu'inter-moléculaires (Chelliah et al., 2006; de Vries et al., 2006; Duan et al., 2005; Pazos et al., 1997; Tress et al., 2005). Cependant, d'autres études, basées quant à elles sur une analyse descriptive des interfaces de complexes protéiques, tendent à prouver qu'aussi bien les analyses de conservation (Jimenez, 2005), que les analyses de co-évolution (Halperin et al., 2006; Mintseris and Weng, 2005), ne sont pas assez spécifiques pour prédire les sites d'interaction protéiques.

Dans le chapitre précédent, j'ai introduit une nouvelle forme d'analyse de la co-évolution des protéines, par l'analyse du maintien de leur complémentarité physico-chimique au cours de leur évolution. C'est pourquoi, j'ai voulu confronter les capacités prédictives d'une telle méthode, avec des approches standard de conservation ou de co-évolution fréquemment utilisées pour guider la prédiction des assemblages protéiques. Pour cela, j'ai développé une nouvelle fonction de score basée sur ce principe de complémentarité évolutive des interfaces de complexes protéiques, l'approche SCOTCH.

En analysant les capacités prédictives de ces différentes approches évolutives, il apparaît que la méthode SCOTCH apparaît comme particulièrement efficace pour la prédiction structurale de complexes protéiques. En effet, couplée à une procédure de docking standard, la méthode SCOTCH sélectionne de façon spécifique les modèles proches des structures natives de complexes, même dans le cas où les structures de départ ne correspondent pas aux formes « liées » des protéines. En effet, en analysant une base de données de complexes transitoires inter-moléculaires, il apparaît que plus de 80% des modèles proches des structures natives sont reconnues, alors que près de 95% des faux modèles sont rejetés. De plus, l'efficacité de cette fonction de score est également très bonne dans le cas de complexes intra-moléculaires, suggérant l'applicabilité de cette approche à différentes formes d'assemblages protéiques. Au contraire, les principales défauts des méthodes évolutives

couramment utilisées, sont mises en évidence soit par l'analyse de complexes intra-moléculaires, soit par l'analyse de complexes inter-moléculaires. Ainsi, l'analyse d'une base de données intra-moléculaires souligne la difficulté qu'ont les analyses de conservation à distinguer différents sites de liaison, alors que la détection de mutations concertées ne permettrait quant à elle, pas de prédire efficacement les résidus en contact dans des complexes de types transitoires (Mintseris and Weng, 2005).

En analysant par la nouvelle méthode SCOTCH deux cas d'études qui intéressent particulièrement notre laboratoire, il apparaît que cette méthode pourrait être une alternative intéressante lorsque les données biologiques dont on dispose sont réduites. En effet, en se basant sur des données partielles, un seul modèle d'assemblage est sélectionné par l'approche SCOTCH dans le cas du complexe Asf1 - Histone H3-H4, résolu très récemment par cristallographie aux rayons X. De plus, lorsque les données biologiques elle-même ne permettent pas de trancher entre plusieurs modes d'assemblages possibles entre deux protéines, cette approche évolutive peut se révéler assez complémentaire, comme nous l'avons vu dans le cas du complexe impliquant les domaines CS de SGT1 et ATPase de Hsp90 non résolu à ce jour.

De par ces cas d'études, nous avons pu montrer le potentiel des méthodes évolutives pour guider vers l'obtention d'un modèle d'assemblage relativement précis entre deux protéines. Par des techniques de raffinement faisant appel à des fonctions d'énergies plus précises mais plus coûteuses en temps de calcul, il est possible comme nous avons pu le voir, de préciser les prédictions effectuées par une première approche évolutive, et ne proposer ainsi qu'un seul modèle d'assemblage. Ce modèle pourra alors servir de base de départ à l'application de certaines stratégies rationnelles pour inhiber ou imiter les assemblages de complexes protéiques : les méthodes de *design* de protéines.

Chapitre 5 : De la structure à l'inhibition des complexes protéiques

Chapitre 6 : Discussion – Perspectives

6.1. Mécanismes évolutifs et interfaces des complexes protéiques

La majeure partie de mon travail de thèse a été consacrée à explorer comment la pression de sélection s'est exercée aux interfaces de complexes protéiques au cours de l'évolution. De nombreuses études ces dernières années ont profité de l'expansion des bases de données de structures et de domaines pour rechercher les facteurs essentiels à la formation et au maintien de la stabilité des complexes protéiques. Ainsi, il a été proposé que la conservation des séquences et l'existence de mutations concertées au niveau des interfaces protéiques constituaient des descripteurs évolutifs importants permettant de distinguer les interfaces des autres régions de la surface des protéines.

Dans le cas des analyses de conservation, une des principales limites concerne les assemblages protéiques présentant de multiples sites de liaison (Jimenez, 2005). Par exemple, lorsque les cibles analysées sont des domaines protéiques exposant à la fois des sites d'interactions intra- et inter-moléculaires, il est particulièrement difficile d'identifier la région d'interaction spécifique à un partenaire protéique donné. À l'inverse, les méthodes basées sur la co-évolution des protéines doivent permettre de contourner cet obstacle, en identifiant plus précisément les relations de dépendance entre les séquences des partenaires protéiques.

Néanmoins, les méthodes analysant les mutations concertées pour rendre compte des processus de co-évolution aboutissent à des résultats mitigés en fonction de la nature des cibles analysées. En effet, elles s'avèrent assez performantes pour la prédiction des modes d'assemblages permanents (Mintseris and Weng, 2005; Pazos et al., 1997), alors que leur application pour la détection de paires de résidus en contact au sein d'assemblages protéiques transitoires souffre d'une très faible spécificité (Mintseris and Weng, 2005). Les auteurs de cette dernière étude ont ainsi émit l'hypothèse que des taux d'évolution relativement bas pour les résidus situés à l'interface des complexes permanents leurs permettaient de co-évoluer contrairement aux résidus situés aux interfaces de complexes transitoires beaucoup plus plastiques.

Dans mon étude j'ai analysé deux bases de données de complexes protéiques correspondant approximativement à chacune de ces catégories, permanente et transitoire. Les complexes de

type intra-moléculaires sont pour la majorité des assemblages de type permanent, tandis que notre base de données de complexes inter-moléculaires est exclusivement composée d'interactions transitoires. Par une approche alternative aux mutations concertées, nous avons cherché à caractériser les évènements de co-évolution assurant le maintien d'une stricte complémentarité physico-chimique aux interfaces de complexes protéiques. L'idée sous-jacente est que toute mutation ne respectant pas cette complémentarité serait susceptible d'abolir l'interaction. Ces évènements de co-évolution nous ont conduit à développer un nouveau descripteur, baptisé SCOTCH, possédant des capacités de reconnaissance des interfaces protéiques remarquables. En effet, le descripteur SCOTCH, se révèle performant pour la reconnaissance des sites d'interactions aussi bien pour les complexes intra- qu'inter-moléculaires. Contrairement à l'hypothèse de Mintseris et Weng, les phénomènes de co-évolution semblent donc exister aussi bien pour les complexes transitoires que pour les complexes permanents. L'absence de signal de co-évolution dans leur étude peut s'expliquer par deux hypothèses :

- les analyses de mutations concertées se basent sur la recherche de mutations, et donc de modifications au sein des alignements de séquences. Le problème d'une telle approche est qu'elle est particulièrement sensible à d'éventuelles différences dans la vitesse d'évolution aux interfaces de complexes protéiques. Le signal de co-évolution analysé sous cet angle est donc susceptible d'être particulièrement bruité dans le cas des complexes transitoires.

- les analyses de mutations concertées se restreignent à une analyse par paires de résidus. Or, en étudiant les propriétés évolutives des interfaces, nous avons pu constater qu'une conservation des propriétés physico-chimiques des interfaces pour différents homologues lointains ne pouvait s'expliquer que par la prise en compte de groupes de résidus plutôt que de paires de résidus. C'est probablement la raison pour laquelle les techniques d'analyse de mutations concertées sont très efficaces pour la prédiction des appariements de type Watson-Crick dans le cas des ARN ribosomiques (Dutheil et al., 2005). Dans le cas des interactions protéine-protéine, la densité de contact autour des paires de résidus en interaction est probablement un obstacle majeur pour ce type de recherche par paires, dans la mesure où plusieurs

positions peuvent être impliquées dans la compensation de certaines ruptures de complémentarité au cours de l'évolution (Gobel et al., 1994).

En tenant compte de ces deux aspects fondamentaux, l'approche SCOTCH parvient à extraire des contraintes spécifiques dans l'évolution des interfaces protéiques. Le modèle très simple de complémentarité formalisé dans l'approche SCOTCH montre que même chez des homologues lointains, les différentes mutations qui ont pu survenir n'ont quasiment jamais aboli les interactions fondamentales qui contribuent le plus à la stabilité des complexes, à savoir les interactions électrostatiques et les interactions hydrophobes (McCoy et al., 1997; Sheinerman et al., 2000).

Les résultats obtenus sont d'autant plus encourageants, que certaines notions évolutives importantes comme celles des arbres phylogénétiques des séquences n'ont pas encore été implémentées. Ainsi, on peut imaginer qu'en intégrant l'histoire évolutive des séquences dans la procédure SCOTCH on puisse gagner en spécificité. L'intégration de cette notion par différents systèmes de pondération susceptibles d'augmenter le pouvoir prédictif de la méthode est actuellement en cours. En effet, au cours du chapitre 2, nous avons souligné l'intérêt de prendre en compte l'arbre phylogénétique des séquences pour identifier des variations de spécificités fonctionnelles entre différents homologues.

6.2. De la co-évolution à la prédiction de structure des complexes protéiques.

La prédiction à haute résolution des structures de complexes protéiques implique la prise en compte explicite des interactions atomiques. Ce niveau de détail requiert cependant une exploration exhaustive de plusieurs degrés de liberté au niveau du squelette peptidique et des chaînes latérales des protéines en interaction. Cependant, cette exploration conformationnelle est très coûteuse en temps de calcul, et ne peut donc être envisagée lors d'une première étape sans une connaissance au préalable des sites d'interaction. En l'absence de données expérimentales permettant de cibler ces sites d'interaction, il est nécessaire de disposer de méthodes dites « à basse résolution » peu coûteuses en temps de calcul, et peu sensibles aux détails atomiques, mais efficaces pour filtrer différents modes d'association, et retenir des solutions proches de la structure native. Il a été établi que pour être efficaces, ces

méthodes « à basse résolution » devaient être capables de retenir des solutions à moins de 3 Å de la structure native. Cette valeur de 3 Å correspond approximativement au rayon de convergence des méthodes de prédiction à haute résolution (Wollacott et al., 2007).

L'intérêt des méthodes de prédiction exploitant les données de l'évolution est qu'elles sont peu sensibles aux détails des conformations structurales. L'évaluation des performances de la méthode SCOTCH comme méthode de prédiction « à basse résolution » des complexes est tout à fait encourageante. Sur quelques exemples précis, nous avons montré qu'un nombre très limité de données biologiques suffisaient à identifier un modèle unique d'assemblage de complexe très proche de la réalité. Actuellement, différentes voies d'amélioration de la procédure sont étudiées. En particulier, une des caractéristiques fondamentales des interfaces de complexes protéiques, est de présenter un profil énergétique caractéristique, appelé *funnel*. La mise en évidence de cette propriété pour un modèle structural de complexe, nécessite une exploration conformationnelle minutieuse et donc un temps de calcul inenvisageable si tous les modes d'association possibles doivent être considérés. C'est pourquoi, une sélection préalable de différents modèles d'assemblage par le programme SCOTCH, particulièrement rapide car elle ne nécessite qu'en moyenne 1 à 2 secondes par modèle, pourrait être intégrée en amont. Les résultats présentés au cours de cette étude ont montré que l'approche SCOTCH était capable de détecter des modèles d'assemblages suffisamment proches des structures natives de complexes, pour permettre à ces procédures d'analyse plus fines de détecter la présence d'un *funnel*.

A cet effet, nous développons actuellement, le programme SCOTCHer, une approche de docking autonome, indépendante du programme FTDock, dont l'objectif à long terme sera la prédiction à grande échelle de différentes cibles extraites des cartes d'interactions protéiques. Pour cela, trois conditions majeures devront être satisfaites : (i) maintien de la qualité des prédictions si on utilise non plus des structures expérimentales mais des modèles obtenus par modélisation comparative (ii) réduction du temps de calcul, (iii) prise en compte explicite des mouvements du squelette peptidique.

Une des solutions rapide à mettre en œuvre pour réduire le temps de calcul, sera de générer non pas 10000 modes d'assemblages pour les filtrer ultérieurement mais plutôt de guider dès le départ l'orientation des assemblages en fonction du score SCOTCH ou d'autres propriétés

évolutives. Cette procédure, existant déjà dans d'autres programmes tels que HADDOCK, permettra de limiter la procédure de recherche conformationnelle. Concernant la seconde phase du docking, au cours de laquelle on cherche, par de petits mouvements, à caractériser l'existence d'un *funnel*, le programme SCOTCHer permet d'obtenir des résultats satisfaisants en termes de vitesse et de qualité d'échantillonnage. Les approximations effectuées et les algorithmes implémentés pour l'optimisation des interactions entre chaînes latérales permettent de générer dans un temps raisonnable un ensemble de 2000 structures. L'intégration de la flexibilité du squelette peptidique dans la procédure de docking est susceptible d'augmenter sensiblement ce temps de calcul. Pourtant, comme le soulignent les derniers rapports du concours CAPRI, ce développement est une priorité majeure pour une prédiction réaliste de différents types de complexes protéiques (voir section 1.5.3). L'intégration des mouvements du squelette peptidique repose sur deux modèles sensiblement différents : (i) on suppose que la conformation liée correspond à un état de haute énergie en équilibre avec la forme libre et que cette conformation peut être atteinte par des recherches conformationnelles autour de la forme libre indépendemment du partenaire. Différents exemples ont été publiés sur la base de ces modèles et pourront être intégrés dans un premier temps (Grunberg et al., 2004). (ii) l'autre hypothèse est que la modélisation correcte de la structure liée et des mouvements à effectuer dépend du partenaire et qu'il est nécessaire de l'intégrer dans les calculs. Ce dernier modèle est à la base des derniers développements proposés par le groupe de D. Baker pour prendre en compte les mouvements de grande amplitude du squelette (Wang et al., 2007). En se basant sur ce principe, nous pourrons donc par exemple envisager d'étudier plus particulièrement la co-évolution de certaines portions de protéines plutôt qu'une interface entière, et orienter le mouvement de ces régions en fonction de leur co-évolution pour la prédiction de modèles d'assemblages protéiques à haute résolution.

6.3. Un double-hybride *in silico* basé sur le score SCOTCH ?

Une des questions majeures de ce travail concerne l'utilité du score SCOTCH pour des applications à grande échelle dans la prédiction des interactions entre protéines. Cette problématique originellement posée par le groupe de Valencia sur la base de l'analyse des mutations concertées pourrait trouver un regain d'intérêt au vu des analyses comparatives que nous avons effectuées entre les deux approches (Madaoui and Guerois, 2007). Ici, la

Chapitre 7 : Méthodes

7.1. Prédiction des sites d'interactions protéiques des protéines kinases par analyse différentielle de conservation

7.1.1. Création des alignements de séquences

Afin d'obtenir des séquences homologues à un domaine régulateur donné, une requête avec le programme PSI-BLAST (Altschul et al., 1997) a été effectuée en utilisant la banque de données SPTREMBL. À partir des séquences obtenues, deux ensembles ont été constitués : (i) séquences comprenant le domaine régulateur d'intérêt ainsi qu'un domaine kinase (*MSA 1*), (ii) séquences comprenant le domaine régulateur, et tout autre domaine différent du domaine kinase (*MSA 2*). La recherche de domaines a été effectuée à partir de la base de données Pfam (Bateman et al., 2002). Par la suite, les séquences redondantes (plus de 90% d'identité de séquences), ainsi que les séquences tronquées ont été retirées. La recherche de séquences homologues par le programme PSI-BLAST a été étendue à la banque de données « nr », lorsque l'un des deux groupes comprenait moins de 30 séquences. Le récepteur TGF-Beta a été exclu de cette étude, dans la mesure où le domaine GS est spécifique des protéines kinases (aucune séquence protéique ne contient ce domaine associé à un domaine autre que kinase). Enfin, dans chacun des cas d'étude, un alignement de séquence a été réalisé à partir du programme HMMER (Eddy, 1998).

7.1.2. Analyse de conservation

Les analyses de conservations ont été appliquées en considérant pour chacun des cas d'étude, les deux alignements de séquences générés dans la précédente étape *MSA 1* et *MSA 2*. Pour chacune des positions des alignements *MSA 1* et *MSA 2*, le taux d'évolution a été calculé par le programme rate4site (Mayrose et al., 2004; Pupko et al., 2002) en utilisant les paramètres par défaut du programme. Les taux d'évolution calculés par le programme rate4site sont des mesures relatives de la conservation à chaque position de l'alignement de séquences : les scores les plus bas correspondent aux positions les plus conservées. Un score de conservation a été calculé entre 0 et 99 pour chaque position *i* de l'alignement en utilisant la formule suivante :

$$Rscore = 99 - (99 * (S – Low\_score) / \Delta score)$$

où *Rscore* est le score de conservation à la position *i*, *S* le taux d'évolution calculé par rate4site, *Low_score* le plus bas taux d'évolution, et *Δscore* l'amplitude des taux d'évolution calculés par rate4site. Les scores de conservation ont ensuite été mappés sur la structure de la protéine en remplaçant la colonne des B-facteurs dans le fichier PDB. Les résidus les plus conservés correspondent aux plus aux *Rscore*.

7.1.3. Analyse de conservation différentielle

Conservation différentielle. Pour chacune des positions *i* de l'alignement *MSA 1*, un score de conservation différentiel est calculé de la façon suivante :

$$Diff\_score = 99 - (99 * (\Delta cons – Low\_\Delta cons) / \Delta\Delta cons)$$
$$\text{avec } \Delta cons = Rscore_{MSA\ 1} – Rscore_{MSA\ 2}$$

où *Diff_score* est le score de conservation différentiel pour la position *i*, *Δcons* la différence de conservation en considérant les alignements *MSA 1* et *MSA 2* pour la position *i*, *Low_Δcons* le plus bas *Δcons* si l'on considère toutes les positions de l'alignement, et *ΔΔcons* l'amplitude des *Δcons*. Les scores de conservation sont ensuite été mappés sur la structure de la protéine en remplaçant la colonne des B-facteurs dans le fichier PDB. Les positions présentant un taux de conservation différentiel supérieur à 90%, sont considérées comme spécifiquement conservées dans l'une des deux sous-familles. Les positions présentant un taux de gap de plus de 50% ne sont pas considérées lors de cette procédure de conservation différentielle.

7.2. Prédiction des assemblages macromoléculaires

7.2.1. Utilisation des fonctions de score du programme RosettaDock

Les fonctions de score à basse et à haute résolution du programme RosettaDock ont été utilisées pour évaluer différents modèles d'assemblages générés par des approches de docking rigide. Pour cela, le programme RosettaDock (Gray et al., 2003) a été utilisée. Par défaut, la fonction de score utilisée pour évaluer une structure par l'option « –score » est la

fonction à basse résolution du programme. En utilisant l'option « -dockFA », la seconde fonction à haute résolution est prise en compte. L'option « -dock_score_norepack » a été utilisée dans la procédure d'évaluation.

7.2.2. Le programme SCOTCHer

Le programme SCOTCHer, est un programme développé en python, un langage de script orienté objet. Le diagramme de classes représentant l'architecture et les différentes fonctions de ce programme sont présentées à la section 8.2.2.. SCOTCHer possède deux fonctions essentielles : générer des modèles d'assemblages possibles entre deux protéines connues pour interagir (module de docking corps-rigide), perturber des structures dont on veut évaluer le profil énergétique. A l'heure actuelle (octobre 2007), le premier module est encore en phase de développement. C'est pourquoi, je ne détaillerai que le second module.

Le second module vise, à partir d'une structure de départ, à produire plusieurs solutions très proches de cette structure de départ, puis d'analyser la corrélation entre le RMSD de ces solutions par rapport à la structure de départ, et l'énergie de ces solutions. Ainsi, deux aspects principaux sont à mentionner : (i) générations de solutions proches de la structure de départ, (ii) évaluation de l'énergie d'interaction de ces solutions.

Génération de solutions proches de la structure de départ. Pour générer plusieurs solutions autour de la structure de départ, le programme SCOTCHer procède à une perturbation de cette structure par des mouvements de translation et de rotation. Tout d'abord, une translation aléatoire de plus ou moins 2 Å est appliquée à la plus petite des deux protéines en interaction, l'axe de translation passant par les centres de gravité des deux protéines. Ensuite, deux rotations aléatoires sont combinées par un produit de quaternions : (i) une première rotation autour de l'axe de translation (l'angle de rotation étant choisi aléatoirement entre 0 et 360°), (ii) une seconde rotation qui serait non plus une précession autour de l'axe de translation, mais un « *tilt* » (basculement) par rapport à cet axe. Cette seconde rotation est effectuée à partir d'une interpolation sphérique, dont le principe est présenté à la section 4.5.2. L'angle de tilt maximum autorisé a été fixé à 15°.

Une fois la translation et la rotation appliquées, le modèle ainsi généré est évalué par une fonction d'énergie simplifiée. La contribution énergétique de tous les atomes de chaque paire

Publications

Article 1

Madaoui H. and Guerois R. (2007) Surface Complementarity Trace in Complex History (SCOTCH), a robust method to discriminate native protein complex interfaces. *PNAS*. in revision.

Surface Complementarity Trace in Complex History (SCOTCH), a robust method to discriminate native protein complex interfaces.

Hocine Madaoui* & Raphaël Guerois*

Author affiliation: * CEA, iBiTecS, Laboratoire de Biologie Structurale et Radiobiologie, Gif-sur-Yvette, F-91191, France.

Corresponding author:
Raphaël GUEROIS, Point courrier 22, CEA, iBiTecS, SB2SM, Laboratoire de Biologie Structurale et Radiobiologie, Gif-sur-Yvette, F-91191, France.
Tel : +33 1 69 08 67 17
Fax : + 33 1 69 08 47 12
raphael.guerois@cea.fr

Manuscript information:
22 pages, 4 figures, 1 table.

Word and character counts:
174 words in abstract
46931 characters

Abbreviations footnote:
RMSD : Root Mean Square Deviation
ROC : Receiver Operating Characteristics
SCOTCH : Surface Complementarity Trace in Complex History

Abstract

Protein surfaces are under significant selection pressure to maintain interactions with their partners throughout evolution. Capturing how selection pressure acts at the interfaces of protein-protein complexes is a fundamental issue with high interest for the structural prediction of macromolecular assemblies. We tackled this issue under the assumption that, throughout evolution, mutations should minimally disrupt the physico-chemical compatibility between specific clusters of interacting residues. This constraint drove the development of the so-called Surface Complementarity Trace in Complex History method (SCOTCH) which was found to discriminate with high efficiency the structure of biological complexes. SCOTCH performances were assessed not only with respect to other evolutionary-based approaches, such as conservation and coevolution analyses but also with respect to statistically-based scoring methods. Validated on a set of 97 complexes of known structure extracted from both intra- and inter-molecular interactions, SCOTCH appears as a novel and robust strategy to guide the prediction of protein-protein complex structures. Of particular interest, it also provides a basic framework to efficiently track how protein surfaces could evolve while keeping their partners in contact.

Introduction

The modular assembly of proteins is a key determinant in the regulation of biological systems. Combinations of inter- and intra-molecular interactions govern the flow of information transmitted through cell signalling pathways. To unravel the complexity of cell responses, the first atlas of the physical interactome has been obtained for several model organisms (1, 2). To further elucidate the competitions and synergies ruling the molecular logic of these protein-protein interaction networks, a critical step relies on the structural characterization of the protein complexes. However, there is still a huge gap between the proteome-wide data accumulating and the available structural details of macromolecular complexes.

A number of studies have tackled the large scale analysis of protein-protein complexes from a structural perspective (3-6). They have emphasized that size, shape and the physico-chemical complementarities at the interfaces are key descriptors which could be used to develop computational methods able to predict protein binding sites from sequences or structures (7-13). In the context of evolution, seminal studies compared the binding modes of domain-domain interactions between homologous proteins and concluded that they tend to interact similarly, even if sequence identity has been maintained as low as 30 % (14, 15). Such a low conservation threshold suggests that interaction surfaces can evolve significantly while maintaining sufficient specificity between the binding partners. The paradox between sequence divergence and structural conservation of macromolecular assemblies has been recently related to the notion of superfamily and the existence of "heterodimer superfamilies" has been proposed (16). Although the hydrophobicity of buried positions is a major evolutionary constraint for the stable maintenance of a fold within a superfamily, it is not the only driving force at interfaces since these are usually made of heterogeneous physico-chemical textures. In that context, understanding and capturing how selection pressure is exerted at protein binding interfaces constitutes a fundamental challenge.

To address this issue, different evolutionary based methods have been examined, such as conservation and coevolution analyses. Sensitive conservation analyses have been proposed to detect functional sites within proteins (17-19) and have further been used for the identification of protein binding sites (20-25). The goal of coevolution analyses is to identify correlated or compensatory mutations between regions that are close in space (26-28). These analyses were shown to have high interest in predicting interacting sites within RNA (29) and were also used

to predict intra- or inter-molecular contacts between domains (22, 30). However, conservation or coevolution analyses (31) may still generate too many false positives to assist large-scale prediction of protein complex structures.

The high rate of mutations occurring at protein surfaces should severely challenge the cohesion of protein complexes across evolution. It is difficult to conceive how these perturbations can be accommodated with the delicate balance of elementary interactions that ensure affinity and selectivity. To address this issue we explored an alternative approach, distinct from conservation and covariation analyses, to inspect how the physico-chemical complementarities between contacting positions were constrained throughout evolution. Our analysis was first performed on a database of 86 complexes restricted to intra-molecular interactions between domains of multi-domain proteins, which constitute a suitable dataset to bypass the problems related to the definition of orthology.

A remarkable versatility was found in the way physico-chemical complementarities were maintained, likely accounting for the difficulty of capturing specific evolutionary traces at protein-protein interfaces. In many cases, complementarity is found to be preserved not only in a pairwise manner but also through clusters of neighboring residues. The complementarity appears strongly constrained only when mutations are considered in the framework of these clusters. Interactions belonging to different complementarities classes such as hydrophobic or electrostatic are then found minimally disrupted throughout evolution. Based on this analysis, a powerful predictive approach called SCOTCH as an acronym for Surface Complementarity Trace in Complex History was developed. The potency of the SCOTCH method for the prediction of intra-molecular interactions was tested and further assessed on a database of 11 inter-molecular complexes, highlighting its interest for docking applications. It provides a robust predictive method for the identification of native and near native interfaces which outperforms other evolutionary-based or statistically-based methods. The SCOTCH method provides an appealing strategy for the structural prediction of macromolecular assemblies and a basic framework to efficiently track how protein surfaces could evolve while keeping their partners in contact.

Results

Structural neighbors are key to capture evolutionary constraints at interfaces.

Electrostatic and hydrophobic interactions were found to be the physico-chemical properties that most contribute to complex stability (32, 33). Consistently, a specific analysis of the residue-residue preferences in inter-molecular interactions showed that most contacts occur between either hydrophobic or oppositely charged pairs of residues (34). Given that the relative positions of contacting residues are generally conserved even between remote homologs (14, 15), we considered how the physico-chemical complementarity could be maintained throughout evolution.

We first wondered whether the physico-chemical complementarity of interacting pairs was frequently disrupted through the evolution of intra-molecular interfaces. In that scope, a database of 86 intra-molecular complexes was built from the PSIMAP database (35) following the protocol described in Methods (Supp. Tables 2 and 4). We tested a simple model that restricts to three different types of complementarity between amino acids whose disruption is expected to be highly detrimental to the stability of a complex: (i) hydrophobic-hydrophobic, (ii) polar-polar, (iii) oppositely charged complementarities. Based on this model, about 62 % of the residues in an interface have at least one contact with a complementary residue (see distribution Supp. Figure 7). For each complex between domains A and B, we analyzed all pairs (i,j) formed by each residue i in interface A with all its contacting residues j in interface B (distance threshold for contact at 4.5 Å). For every pair (i,j), we defined a parameter called "complementarity ratio" as follows : the ratio of sequences inside a multiple sequence alignment for which the physico-chemical complementarity between two sites is observed (Figure 1) (see Methods). If this ratio is above a fixed threshold of 95%, the pair is considered as "significantly complementary" (the complementarity is disrupted in less than 5% of the sequences of the multiple sequence alignment). Considering our intra-molecular dataset, the proportion of pairs (i,j) at an interface exhibiting a given complementarity ratio was analyzed (Figure 2A, white bars). Only a weak proportion (11%) of significantly complementary pairs was found.

A limit of the former analysis is that the disruption of the physico-chemical properties between two residues may be rescued by the mutations of neighboring residues.

Complementarity may not be maintained in a strict pairwise manner. To account for that scenario, we further considered the *k* structural neighbors of each residue in the procedure (as in Figure 1 for the fish sequence). Using the structural neighbors of the contacting pairs (with *k*=2), a drastic shift of the distribution can be noticed (Figure 2A, black bars). As much as 50% of the pairs (i,j) found at the interface of domain-domain complexes are detected as significantly complementary by the method. Varying the number of structural neighbors from 1 to 3, the proportion of significantly complementary pairs detected reaches a plateau with 2 neighbors (Supp. Figure 5). The optimal number of structural neighbors was thus fixed to 2 in the rest of the study.

To test the specificity of the complementarity ratio analysis to predict protein binding sites, we generated a set of decoys for each complex of the intra-molecular database. 10 000 docking solutions for each of the 86 pairs of interacting domains were generated using the FTDOCK rigid body docking program (36). This program uses shape complementarity rules and fast-fourier transform algorithm to cover as thoroughly as possible the protein surface with physically realistic solutions. To be as specific as possible, we used a stringent distance criterion to define a solution as "near native" by setting the backbone root mean square deviation (RMSD) threshold to 3 Å from the structure of the native complex. On average, 7.4 "near native" solutions were generated for each set of 10000 decoys.

For each of the 860000 solutions, the proportion of pairs (i,j) detected as significantly complementary was calculated. A sharp discrimination was obtained between near native models and false alternative docking models (Figure 2B). As a matter of fact, the selection of all the models with a proportion of significantly complementary contacting pairs of residues above 20% can filter out 80% of the false models while losing only 15% of the near native solutions.

Discriminative power of evolutionary analysis considering intra-molecular cases

We have shown that the proportion of contacting pairs detected as significantly complementary is an efficient descriptor to discriminate between false and near native solutions. This proportion was further normalized to correct from two biases : (1) Based on the above principle, small and large interfaces exhibiting an equal proportion of pairs (i,j) are ranked the same. For small interfaces, the probability that this proportion arose by chance is

higher, the number of interacting pairs at the interface was thus taken into account as a correction factor. (2) Positions that are highly variable often correspond to misaligned regions. High scores observed in these regions are thus unreliable and high variability was also considered as a correction factor.

As detailed in the method section, two thirds of the intra-molecular database were randomly selected and used to adjust the weights of these correction factors by a logistic regression procedure. The logistic regression model estimates the probability that a given docking solution belongs to the class of near native or false complexes. In the following, the estimate of the class membership probability was denoted as the SCOTCH score. The remaining third of the intra-molecular database gathering 28 complexes was used as a validation dataset. For each of these cases, the docking models were scored using the SCOTCH score as well as scores based on conservation or coevolution analyses. The conservation analysis was performed using the rate4site algorithm which was shown to provide highly sensitive results and which infers the rate of evolution at each site of a multiple sequence alignment using a probabilistic-based evolutionary model (37). Two different methods were used to account for coevolutionary events : (i) detection of correlated mutations using substitution matrices as described in (26, 30) and (ii) detection of compensatory mutations which quantifies the compensatory events between pairs of residues considering the physico-chemical properties as described in (27, 38) (see methods and Supp. Text 3 for more details).

The quality of the predictions was evaluated for every method by counting the number of cases for which a near native complex (RMSD from native below 3 Å) has been selected within their best 10, 100, 1000 and 10000 solutions. Detection of a near native solution in these sets of solutions can be considered as very good, good, acceptable and bad, respectively. Following this evaluation procedure, the SCOTCH method was found to outperform all the others with 16 out of 28 complexes recognized in the very good and good categories, against 9 out of 28 complexes at best with the other approaches (Figure 3A). The performances of the different approaches were also evaluated using ROC curves which provide a powerful representation to compare predictive methods. The ROC curves plot the fraction of true positive against the fraction of false positive recovered for all possible threshold values. The largest area under the ROC curve indicates the best method. Again, the SCOTCH score exhibited a very good ability to discriminate between correct and false docking solutions (Figure 3B). To summarize, the classification of docking solutions as near native or false

based on the SCOTCH score appears very efficient : about 87% of the false solutions are filtered out, while 67% of the near native complexes are kept (Supp. Table 6).

Discriminative power of evolutionary analyses considering inter-molecular cases

To further probe the general applicability of the SCOTCH approach, a second test was carried out with a set of complexes comprising inter-molecular interactions. From the WENG benchmark 2.0 (39), 11 complexes whose structure involves only two chains and for which more than 10 homologous sequences can be aligned for each binding partner were considered (Supp. Tables 3 and 5). Applied on this set, the SCOTCH method performed better than for the intra-molecular dataset, since 10 out of 11 complexes were recognized in the very good and good categories (Figure 3C and 3D). The predictive power of the conservation analysis also increased with this dataset. In contrast, with the correlated and compensatory mutation analyses, only 1 out of 11 could be retrieved in the same categories. Overall, the classification of docking solution as near native or false based on the SCOTCH method filtered out about 93% of the false solutions, while keeping 86% of the near native complexes (Supp. Table 6).

SCOTCH highlights the versatile adaptation of complex interfaces

The specificity of the predictions obtained so far suggests that SCOTCH captures important constraints that apply to biological interfaces. The analysis of the complementarity ratio can thus be seen as a powerful tool to further probe the evolutionary events likely to occur at an interface. The native structures of the complexes correctly scored using SCOTCH were used for that analysis (48 complexes scored in the Top10 category in both intra- and inter-molecular validation dataset and in the intra-molecular learning dataset, see Supp. Tables 2 and 3). First, we analyzed to what extent structural neighbors are required to maintain strict complementarity between interacting pairs. On average, 54 % of the positions of an interface were found involved in at least one significantly complementary pair (Supp. Figure 8). Among these, a large majority (60 %) is detected significantly complementary only thanks to the structural neighbors, highlighting their importance to capture evolutionary process of interfaces (Supp. Figure 9).

We further wondered if it was possible that a site in an interface changed drastically its physico-chemical nature over evolution while maintaining a high complementarity. To do so,

we tracked the existence of unambiguous complementarity switches between hydrophobic-hydrophobic and charged-charged interacting clusters across evolution. To be significant we imposed that at least 10 % of the sequences in the multiple sequence alignment exhibit the switch event. Interestingly, such switches were observed at least once in about 30 % of the 48 analyzed structures (Supp. Table 4 and 5). These results illustrate the versatility of the elementary interactions that build up an interface and underscore the complexity of the underlying evolution history.

Optimization of SCOTCH using a statistical pairwise potential

An important feature of the SCOTCH score is that, contrary to previous approaches, it recognizes the near native interface by detecting the interface whose physico-chemical complementarity was the least disrupted across evolution. Other predictive methods based on statistical pair potentials rely on summing pairwise complementarity scores to recognize near native structures. Among them, the widely used RPScore was derived from a statistical analysis of pairwise interactions at complex interfaces and was reported to efficiently select near native structures from a set of decoys (40).

All the docking solutions generated from the intra-molecular and inter-molecular databases were scored with the RPScore. From the ROC curves (Figure 4), the RPScore was found to perform significantly better with the inter-molecular dataset (green curve Figure 4B) than with the intra-molecular dataset (green curve Figure 4A). In comparison, the SCOTCH score exhibited better performance, with similar predictive power on both datasets (red curves Figures 4A and 4B). Combination of the SCOTCH and the RPScore scores improved significantly the recognition of the near native solutions (black curve Fig. 4A and 4B) (the RPScore was added as an additional factor to the logistic regression procedure, see Supp. Text 2). The high predictive power of this score emphasizes that the evolutionary signals detected by SCOTCH can be strengthened by a statistical representation of the complementarity (Figure 4 and Supp. Table 6).

Article 2

Madaoui H, Becker E, Guerois R. (2006) Sequence search methods and scoring functions for the design of protein structures. *Methods Mol Biol;*340:183-206.

Article 3

Haubertin DY, Madaoui H, Sanson A, Guerois R, Orlowski S. (2006) Molecular dynamics simulations of E. coli MsbA transmembrane domain: formation of a semipore structure. *Biophysical Journal*;91(7):2517-31.

Molecular Dynamics Simulations of *E. coli* MsbA Transmembrane Domain: Formation of a Semipore Structure

David Y. Haubertin, Hocine Madaoui, Alain Sanson, Raphaël Guérois, and Stéphane Orlowski
Service de Biophysique des Fonctions Membranaires, Département de Biologie Joliot-Curie and URA 2096 CNRS,
Direction des Sciences du Vivant/Commissariat á l'Energie Atomique (CEA), Centre de Saclay, 91191 Gif-sur-Yvette cedex, France

ABSTRACT The human P-glycoprotein (MDR1/P-gp) is an ATP-binding cassette (ABC) transporter involved in cellular response to chemical stress and failures of anticancer chemotherapy. In the absence of a high-resolution structure for P-gp, we were interested in the closest P-gp homolog for which a crystal structure is available: the bacterial ABC transporter MsbA. Here we present the molecular dynamics simulations performed on the transmembrane domain of the open-state MsbA in a bilayer composed of palmitoyl oleoyl phosphatidylethanolamine lipids. The system studied contained more than 90,000 atoms and was simulated for 50 ns. This simulation shows that the open-state structure of MsbA can be stable in a membrane environment and provides invaluable insights into the structural relationships between the protein and its surrounding lipids. This study reveals the formation of a semipore-like structure stabilized by two key phospholipids which interact with the hinge region of the protein during the entire simulation. Multiple sequence alignments of ABC transporters reveal that one of the residues involved in the interaction with these two phospholipids are under a strong selection pressure specifically applied on the bacterial homologs of MsbA. Hence, comparison of molecular dynamics simulation and phylogenetic data appears as a powerful approach to investigate the functional relevance of molecular events occurring during simulations.

INTRODUCTION

The family of ABC (ATP-binding cassette) proteins is composed of an impressive, and still growing, number of proteins of various functions and is expressed at every level of the living kingdom, from bacteria and yeasts to plants and mammals (1). Most of them are membrane proteins, most generally devoted to ATP-dependent transports of substrates of highly various natures, and are thus often referred to as ''traffic ATPases''. In bacteria, ABC transporters are involved in either cell import of different metabolites or the efflux of antibiotics and exogenous molecules. In mammals, they are always expelling substrates out of cells, and in man they are involved in different pathophysiological disorders, such as Tangier disease (ABC A1) or cystic fibrosis (C7), and pharmacological processes, such as multidrug resistance (MDR) or absorption-distribution-elimination (ABC B1, C1, or G2) (2). The substrates handled by the human ABC transporters are, as far as they are known, often lipids (ABC A1, A4, B1, B4, D1, G1, G5, G8) (3) or drugs (B1, C1–6, G2) (4–6), and sometimes (B1, C1, G2) both of them (7).

Despite their biomedical importance, very limited insights have been provided into the molecular structures of these membrane proteins (8). Currently, high resolution structures have been obtained for a few isolated nucleotidic binding domains. The entire protein structure has only been obtained for the bacterial transporters BtuCD (9) and $(MsbA)_2$ (10). BtuCD is an importer of vitamin B_{12}, a large and hydrophilic molecule, and is rather atypical for an ABC transporter with its 20 transmembrane helices. In contrast, MsbA presents itself as a homodimer with a moderate sequence identity with the members of the B subfamily of the human ABC proteins, and it mediates the outward transport of the toxic lipid A but also of various lipids (11) and even drugs (12,13). This functional homology makes it an attractive bacterial model for the multidrug transporter P-glycoprotein (P-gp, ABC B1), especially considering that P-gp is suspected to work as a flippase for drugs (14), and this appears indeed the case as it can handle various lipids (15).

The interest in structural data about MsbA is reinforced by the report of crystal structures of three different isoforms from *Escherichia coli* (10), *Vibrio cholera* (16), and *Salmonella typhimurium* (17), which are proposed to be different conformations of the enzyme during its catalytic cycle. However, the *E. coli* MsbA structure has raised several concerns regarding the insertion of its transmembrane domain in the bilayer (18,19) since the bending of the intramembrane helices with respect to the membrane plane would result in an atypical geometry for the ''open'' conformation of the dimer. The investigation of the organization of the lipid molecules around this unique protein shape is thus of great interest. In particular, it is desirable to address the dynamic stability of a phospholipid bilayer in its vicinity, which conditions the functional relevance of this protein structure. If such an assembly is stable, the question to address is how phospholipid molecules can accommodate the bending of the transmembrane domain of the protein. In this context, we decided to tackle the question of the membrane perturbation induced at the molecular level by the

Submitted February 24, 2006, and accepted for publication May 31, 2006.

Address reprint requests to Dr. Stéphane Orlowski, SBFM/DBJC and URA 2096 CNRS, CEA Saclay, 91191 Gif-sur-Yvette cedex, France. Tel.: 33-1-69-08-95-77; Fax: 33-1-69-08-81-39; E-mail: orlowski@dsvidf.cea.fr.

David Y. Haubertin's present address is AstraZeneca, Centre de Recherche, Z.I. La Pompelle, BP 1050, 51689 Reims cedex 2, France.

0006-3495/06/10/2517/15 $2.00

doi: 10.1529/biophysj.106.084020

presence of MsbA in a model lipid bilayer in silico. This study was designed in relation to the lipid translocase function exhibited by MsbA, with the perspective of gaining some insights into its mechanisms of functioning.

Molecular dynamics (MD) simulations (20) offer a powerful way of studying biological macromolecules under approximately physiological conditions. These simulations have been extensively applied to the study of biological membranes (21–23) as well as various membrane-bound proteins (24–28). Recent advances in computer power have allowed the simulation of mixed micelles modeling human bile, with simulation times extending to 150 ns (29). Advances in simulation procedures and computer power mean than timescales exceeding 10 ns and large systems of ~100,000 atoms are becoming computationally tractable.

In this study, we report extended MD simulations of *E. coli* MsbA. A previous simulation of this protein has been published (19). However, whereas monomer simulations revealed themselves stable, the MsbA molecule was unstable in its dimeric form and showed large structural drifts over the course of the simulation. Moreover a membrane-mimetic octane slab was used. The interactions between membrane proteins and their lipid environment play important roles in the stability and function of proteins (30). Although an octane slab can represent a correct membrane-mimetic environment for cylindrical proteins, it might not be suited when lipid-protein interactions are critical for protein stability or when the protein does not possess a simple cylindrical shape. In this context, we present a 50-ns simulation of MsbA embedded in a palmitoyl oleoyl phosphatidylethanolamine (POPE) bilayer. For computational tractability reasons and since our main concern is the organization of lipids around the protein, only the transmembrane domain of the protein has been studied. As a result, it is shown that i), protein secondary/tertiary structure is reasonably stable; ii), polar headgroups of close lipids in the cytosolic leaflet partially invaginate, leading to a semipore formation; iii), two individualized lipid molecules can interact with the edge of the protein cavity with a clear specificity with respect to bulk phospholipids; and iv), the amino acid involved in this specificity has been strongly and specifically conserved during the evolution of bacterial MsbA-like ABC transporters, suggesting an important functional role for the observed structure.

METHODS

General strategy and characteristics of the simulation

The simulation of the transmembrane domain of *E. coli* MsbA inserted in a POPE bilayer is presented in this study (Table 1). Since the main components of the natural membrane in which MsbA is found are phosphatidylethanolamine lipids (31), POPE lipids represent a good approximation for the membrane environment. The MsbA cavity was filled with solvent and the protein was constrained during 16 ns; constraints were then removed and the system was further examined for 34 ns, therefore leading to a total simulation time of 50 ns. System preparation involved three steps: i), protein reconstruction, ii), phospholipid bilayer construction, and iii) protein and bilayer assembly.

All MD simulations were conducted using the GROMACS v3.2 MD simulation package (32) with parameter set ffgmx. Production simulations were performed in the isobaric-isothermal ensemble (NPT). The temperature of the system was coupled using the Berendsen algorithm (33) at 300 K with coupling constant $\tau_T = 0.1$ ps. Protein, phospholipids, solvent, and ions were coupled separately. The pressure was coupled using the Berendsen algorithm at 1 bar with coupling constant $\tau_P = 2$ ps and the compressibility was set to 4.5×10^5 bar^{-1} in all simulation box directions. Electrostatic interactions were calculated using particle mesh Ewald (34) with a real space cutoff of 0.9 nm. Cutoff for van der Waals interactions was set at 0.9 nm. Time step for integration was 2 fs; coordinates and velocities were saved every 0.5 ps. The LINCS algorithm was used to restrain bond lengths (35).

Protein reconstruction

The crystallographic structure of the MsbA protein from *E. coli* (Protein Data Bank (PDB) code: 1JSQ) was used as a starting structure for our simulations. The structure only contains an incomplete Cα trace; no side-chain or backbone atoms other than Cα are present. Several residues are missing (residues 1–9 and 208–237 for the first monomer). The transmembrane domain of *E. coli* MsbA, i.e., residues 1–104, 132–200, and 245–309, was reconstructed (Fig. 1). The nine N-terminal residue Cα atoms were reconstructed to form an α-helix with residues 10–20. The other missing residues (residues 208–237) were not part of the transmembrane domain and were therefore not reconstructed. Missing backbone and side-chain atoms were predicted using MAXSPROUT (36) and SCWRL (37). Residues lying inside the bilayer hydrophobic core were considered neutral. Ionization states of the protein's other side chains were the same as the standard at pH 7, except for buried residues. Buried residues were determined using WHATIF (38) to calculate accessible solvent area and with visual examination.

Phospholipid bilayer construction

The bilayer contained 662 POPE lipids; it was generated by extending a preequilibrated bilayer containing 340 lipids (26). The extension was achieved by replicating the 340 lipids box twice in both directions of the membrane plane, therefore leading to a 1360 POPE lipids system. Box dimensions were then reduced by 70% in the plane of the bilayer, yielding a system of 669 lipids. Seven lipids from the top layer were then randomly selected and removed to achieve an equal number of lipids in both layers. The system was then hydrated using a preequilibrated box of simple point charge-extended (SPC/E) water (39). The system was subjected to 1000 steps of steepest descent energy minimization. The coordinates of the lipids were then frozen, allowing the solvent molecules to relax and adapt their coordinates with the membrane during a 500-ps simulation in the canonical ensemble (NVT). The system was then equilibrated in a series of 250-ps simulations in the NPT ensemble with semiisotropic pressure coupling. The compressibility in the plane of the bilayer was fixed to 4.25×10^{-5} bar^{-1} and was discretely increased from 1.0×10^{-30} bar^{-1} along the bilayer normal to reach 4.25×10^{-5} bar^{-1} after 2.25 ns of simulation. The final equilibration run consisted of a 2.5-ns simulation in the NPT ensemble with a compressibility of 4.25×10^{-5} bar^{-1} in all directions.

TABLE 1 Simulation summary

Duration (ns)	Lipids	Waters and ions	Atoms	Restraints
50	591 POPE	18,601 waters and 14 Na^+, 34 Cl^-	91,271	Protein backbone: first 16 ns

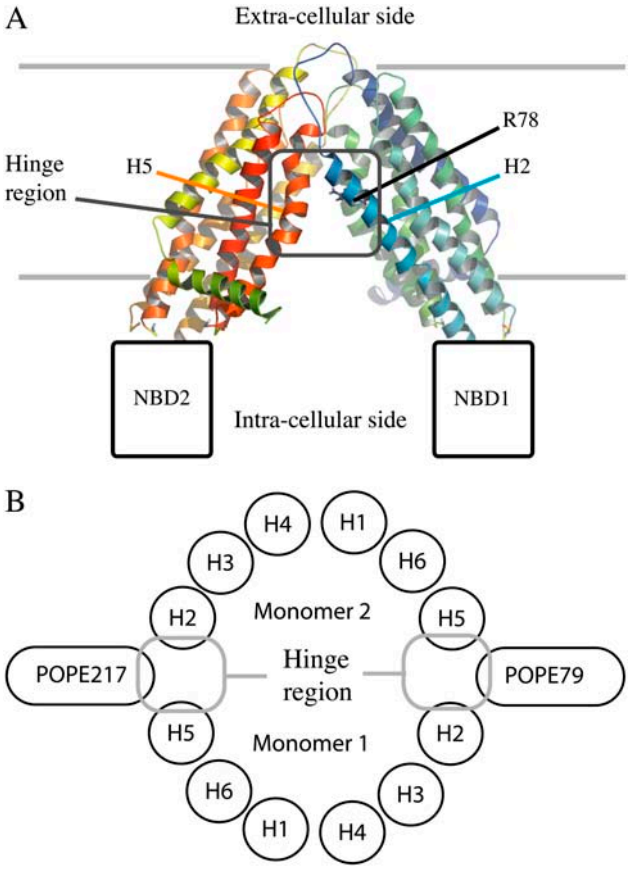


FIGURE 1 (*A*) General presentation of the open-state MsbA structure reconstruction, only considering its transmembrane domain. The ribbons are colored from blue to green from the N- to C-terminus of the M1 monomer and from green to red from the N- to C-terminus of the M2 monomer. (*B*) Schematic key phospholipids organization on the intracellular side of the protein.

Protein and bilayer assembly

The protein was inserted in a preequilibrated POPE bilayer in the fashion outlined by Sansom et al. (40). The protein was oriented such that its principal axis of inertia was aligned along the bilayer normal (z axis) and its other axes of inertia made an angle of 45° with the box axes in the plane of the bilayer. Positioning along the z axis was such that the bands of tryptophan residue side chains were located in the lipid headgroup region. To accommodate the protein, the box dimension along the z-direction was increased to 7.75 nm, and the system was subsequently hydrated using a preequilibrated box of SPC/E water (39). The cavity of the protein was filled with water. Randomly positioned potassium and sodium ions replaced 48 water molecules to balance the charges on the nonneutral protein residues, equivalent to a concentration of 17 mM. The size of the simulation box was $13.33 \times 13.21 \times 7.75$ nm. The resulting system was equilibrated following the procedure summarized in Table 2. After initial steepest descents energy minimization of the system, the protein and lipids coordinates were frozen to allow solvent molecules to relax their coordinates and optimize their interactions with the lipid-protein assembly in a 500-ps simulation in the NVT ensemble. The constraints on lipid molecules were then removed to optimize lipid-protein interactions in a 500-ps simulation in the NVT ensemble. The backbone atoms of the resulting system were then frozen, allowing side-chain atoms to relax their positions and optimize their interactions with lipid and solvent molecules in a 500-ps simulation in the NVT ensemble. The final equilibration step consisted of a 500-ps simulation in the NPT ensemble where backbone atoms of the protein were restrained harmonically using a force constant of 1000 kJ mol^{-1} nm^{-2}. The resulting system was used in a 50-ns simulation in the NPT ensemble. In the first 16 ns of this simulation, the protein backbone was restrained harmonically using a force constant of 1000 kJ mol^{-1} nm^{-2}; this restraint was removed for the remainder of the simulation.

Simulations were analyzed using GROMACS routines and locally written scripts. Secondary structure was analyzed using DSSP. All molecular graphics representations were generated using Pymol.

Conservation analysis

Given the sequence of the query protein, close homologous sequences were retrieved from the NR database using Psi-Blast (41). Divergent sequences leading to gaps or insertion in the transmembrane helices were removed from the resulting multiple sequence alignment (MSA) to be consistent with the structure of the protein. After removing any sequence with more than 90% sequence identity to any other as detected using Psi-Blast, a set of 96 sequences was extracted to form the MsbA-like data set.

A phylogenetic tree has been constructed (considering a neighbor joining method using a BLOSUM62 matrix) with the JALVIEW software (42). The 96 sequences were then divided into two major subgroups according to the phylogenetic tree: A1, including *E. coli* MsbA, which contains the closest homologs of this protein (49 sequences with an average identity of 42%), and A2, which contains remote homologs of *E. coli* MsbA (47 sequences with an average identity of 20%). Two MSAs were thus considered, MSA1, restricted to Msba functional homologs, and MSA2, gathering the remaining sequences of ABC transporters retrieved from the Blast procedure. For every position of MSA1 and MSA2, an evolutionary conservation score was computed. Conservation scores for each amino acid position in the resulting MSAs were computed using the rate4site program (43). The Bayesian method was applied for the calculation of the conservation scores using the Jones-Taylor-Thornton amino acid substitution model (44). The conservation scores computed by rate4site are a relative measure of evolutionary conservation at each position in the MSA: the lowest score represents the most conserved position in the MSA. Each score S has been rescaled between 0 and 99 using

$$Rscore = 99 - (99 \times (S - Lscore)/\Delta score),$$

where $Rscore$ is the rescaled score, $Lscore$ the lowest score, and $\Delta score$ the amplitude of all the scores. The rescaled conservation grades were then mapped onto the protein by replacing the B factors in the PDB file. The most conserved residues correspond to the highest rescaled scores.

The second goal of the phylogenetic analysis was to identify in the MSA the positions subjected to specific evolutionary pressure in the subfamily of MsbA-like bacterial lipid export ABC transporters. To obtain positions which are more specific to the A1 subgroup, conservation scores for each amino acid position in MSA1 were computed using rate4site and compared

TABLE 2 Equilibration procedure

Step	Length	Type*	Frozen groups	Restrained groups
1	1000 steps	SD	Protein, lipids	–
2	500 ps	NVT	Protein, lipids	–
3	1000 steps	SD	Protein	–
4	500 ps	NVT	Protein	–
5	1000 steps	SD	Protein backbone	–
6	500 ps	NVT	Protein backbone	–
7	500 ps	NPT	–	Protein backbone

*SD refers to steepest descents energy minimization. NVT refers to simulations in the NVT ensemble at 300 K using Berendsen algorithm and a coupling constant $\tau_T = 0.1$ ps. NPT refers to simulations in the NPT ensemble at 300 K, and 1 bar using a coupling constant $\tau_P = 2$ ps and a compressibility of 4.5×10^5 bar^{-1} in all box directions.

with those considering MSA2. The rate4site program assigns a confidence interval to each of the inferred evolutionary conservation score. For site specific comparison, nonoverlapping 95% confidence intervals (45) were used to indicate significant rate variations between both the MSA1 and the MSA2. One position was defined as specifically conserved in close homologs of MsbA-like proteins if the estimated confidence intervals between MSA1 and MSA2 did not overlap (the rate variation is defined as positive if the site is more conserved in the truncated MSA1). Only sites with a positive rate variation and a negative upper confidence limit considering MSA1 are defined as specifically conserved in the set of close homologous sequences (negative score indicates strong conservation).

RESULTS

To investigate the MsbA dimer stability as well as the lipid organization around such an assembly, a 50-ns simulation has been performed. Since lipid organization around the MsbA dimer cannot be predicted a priori, the simulation was started from an out of equilibrium structure in which the lipids close to the protein cavity entrance do not possess the equilibrium structure. The simulation has therefore been split up into two parts. In the first part (0–16 ns), the protein backbone was harmonically restrained, allowing lipids to organize themselves around the MsbA dimer and avoiding protein denaturation early in the simulation owing to inappropriate environment. In the second part of the simulation, the restraints were removed. The entire simulation provides information about the lipid-protein interactions, and the second part reports on the stability of these interactions and of the protein. Finally, conservation analysis enables us to examine those results in light of phylogenetic data.

A second ~20-ns simulation was performed in which the MsbA cavity was filled with POPE lipids (data not shown). The lipids quickly dived inside the cavity (~1 ns) and remained trapped for the remaining of the simulation. Calculation of the electrostatic potential of the MsbA inner cavity (data not presented) revealed its polar nature. It hence leads us to the conclusion that the cavity was more likely to interact with solvent in its steady state. The arrangement obtained might be an intermediate state or a local energy minimum. In both cases, the simulation time was not sufficient to observe a significant rearrangement of the system. The remainder of this work therefore only presents results referring to simulations where the protein cavity was filled with solvent.

Lipid organization around the protein

Before restraints on the lipid phosphorus along the z axis are removed, the bilayer is flat and the MsbA hinge region is in contact with both the cytosol region and the hydrocarbon chains of lipids in the vicinity. The hinge region corresponds to the interface of the two monomers that is located in the hydrophobic region of the membrane and in contact with the solvent. This region is defined by helix 2 of the first monomer and helix 5 of the second on one side and reciprocally on the other side (Fig. 1). This is a nonphysical arrangement since it involves contact between hydrophobic chains and polar residues of the protein as well as with the solvent (Fig. 2 *A*). During the last 1.5 ns of equilibration, where restraints on the POPE lipid phosphorus atoms are removed, a migration of the phospholipids headgroups toward the hinge region of MsbA can be observed (Fig. 2 *B*). This arrangement of POPE lipids around the protein early in the simulation before proper production is started balances out unfavorable interactions between protein, lipids, and solvent. The resulting organization, which can be described as a semipore-like structure, ensures that no hydrophobic groups are in contact with hydrophilic groups. The curvature of the cytosolic monolayer is indeed modified in the proximity of the protein hinge region. However, the membrane perturbation is only very local, not concerning the exoplasmic leaflet

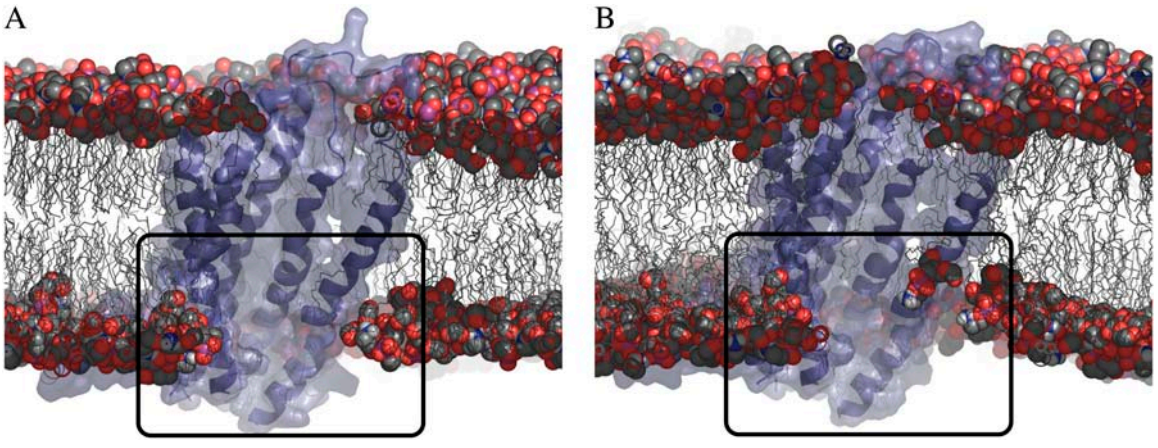


FIGURE 2 Lipid migration in the vicinity of the MsbA hinge region. Structure of the assembly at the beginning of the equilibration (*A*) ($t = -1.5$ ns) and at the end of the equilibration (*B*) ($t = 0$ ns). Protein is represented in blue, phospholipids chains are represented in gray lines, and headgroup atoms in spheres. For purpose of clarity, solvent and counterions are omitted. The extracellular region is located above the protein and the intracellular region below. The black frame highlights the very local perturbation of the cytosolic leaflet surrounding the protein.

and not propagated to long distances: outside a perimeter of 10 Å from the hinge region of MsbA, the bilayer is no longer perturbed and lipid headgroups regain their normal heights similar to that of lipids in the bulk of the membrane (Fig. 3).

Lipid-protein interaction

Two key phospholipids play an important role in the stability of the interface between the hinge region of MsbA and the membrane, POPE79 and POPE217. These phospholipids interact with the protein and remain in the hinge region during the integrality of the simulation. Quantifying lipid-protein interactions, as defined by atoms within 0.35 nm of each other, reveals that these phospholipids interact exclusively with helices 2 and 5 (see Fig. 1 *B* for a schematic representation). Indeed, interactions with these helices make up the integrality of the interactions with the protein. Both phospholipids have a similar number of interactions with helix 5, which stabilizes between 15 and 20 (Fig. 4). Most of these contacts are van der Waals contacts since helix 5 mostly contains hydrophobic and aromatic residues. On the other hand, POPE79 and POPE217 have a different behavior with respect to helix 2 (Fig. 4). Whereas POPE79 has a limited number of interactions with this helix (<2), POPE217 makes most of its interactions with the protein with helix 2 (~20–30 interactions). This suggests different conformations for these two POPE molecules.

The total number of protein-lipid interactions is steadily increasing during the unrestrained part of the simulation. There are ~1300 interactions after 16 ns and ~1500 interactions at the end of the simulation. Even though total simulation time reached 50 ns, the protein is still further stabilized by the membrane at the end of the study, and a final steady state is not completely reached.

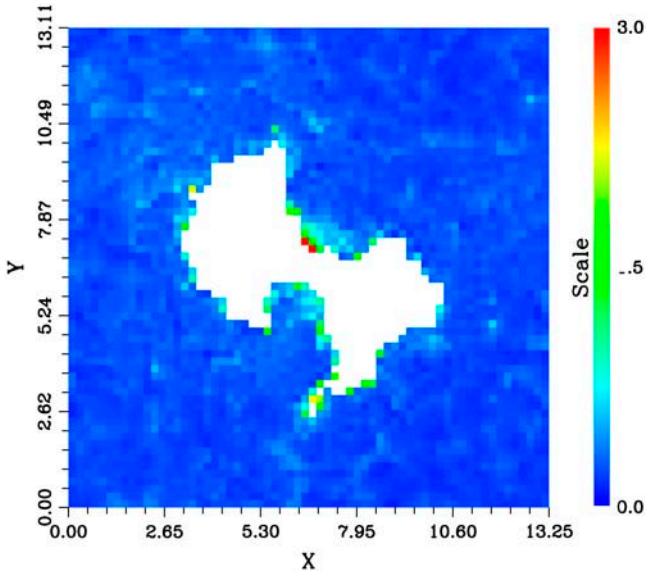


FIGURE 3 Average height of the phospholipid headgroup atoms in the intracellular layer over the 0–16-ns time range, in nanometers. Headgroup atoms considered in the calculations included those from the phosphatidyl ethanolamine group as well as those from the glycerol part. The grid size was 2.0 Å. The color scale is arbitrary.

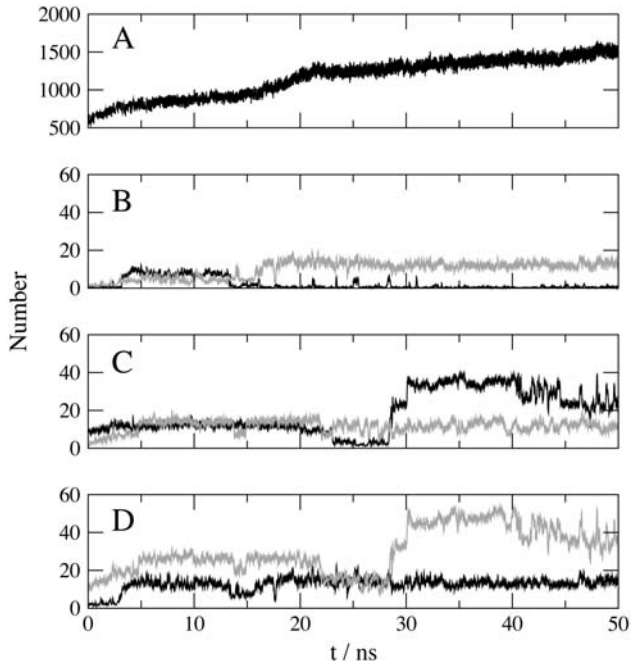


FIGURE 4 (*A*) Overall number of lipid-protein interactions (cutoff 0.35 nm). (*B*) Number of interactions between lipid POPE79 and helix 2 of monomer 2 (*black line*) and helix 5 of monomer 1 (*gray line*). (*C*) Number of interactions between lipid POPE217 and helix 2 of monomer 1 (*black line*) and helix 5 of monomer 2 (*gray line*). (*D*) Overall number of interactions between protein and lipid POPE79 (*black line*) and lipid POPE217 (*gray line*).

As a likely consequence for the fact that their locations and conformations are directly dependent upon the initial conformation, and indeed upon the process used to generate the protein-membrane complex, these two phospholipids adopt two different conformations. Whereas lipid POPE79 has its two hydrophobic chains outside of the protein cavity, lipid POPE217 has one chain half inside the cavity (SN2) and the other outside (SN1).

The examination of the average minimum distance of these two phospholipids with residues making helices 2 and 5 reveals the importance of several residues (Fig. 5). In helix 2, arginine Arg-78 and threonine Thr-81 are particularly important for the protein-lipid interface structure (see Fig. 1 *A*). Both residues are on average <0.2 nm away from POPE217, whereas with POPE79 only residue Thr-81 is at close distance (0.25 nm on average), testifying to the presence of two distinct conformations. In helix 5, most interactions are van der Waals contacts with hydrophobic residues Ile-258, Ser-260, Ala-262, and Phe-265 for POPE79 and Ile-258, Ala-262, and Phe-265 for POPE217.

The number of H-bonds (data not presented) also reveals two distinct behaviors. The total number of H-bonds stabilizes during the constrained part of the simulation and

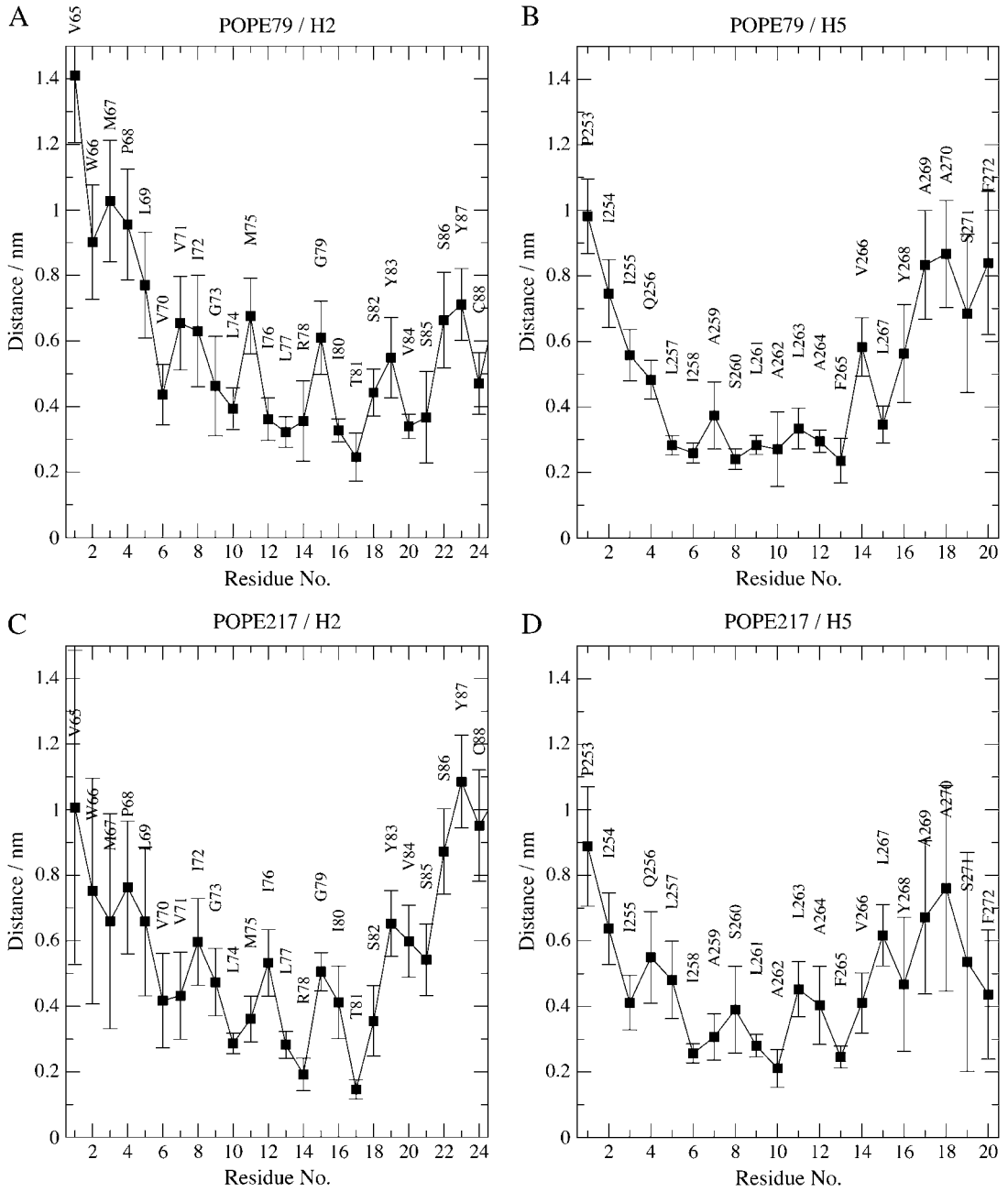


FIGURE 5 Average minimum distance of characteristic lipids POPE79 and POPE217 with each residue of the helices making the interface between the two protein monomers (helices 2 and 5). Profiles are calculated in 500-ps windows and averaged over all windows. (*A*) Minimum distance between lipid POPE79 and residues of helix 2 of monomer 2, (*B*) minimum distance between lipid POPE79 and residues of helix 5 of monomer 1, (*C*) minimum distance between lipid POPE217 and residues of helix 2 of monomer 1, and (*D*) minimum distance between lipid POPE217 and residues of helix 5 of monomer 2.

increases after constraints are removed to stabilize at around 80 H-bonds at the end of the simulation. During the unrestrained part of the simulation, local fluctuations of the protein enable new H-bonds to be created, further stabilizing the protein inside the membrane. The two individualized phospholipids have distinct behavior toward H-bonding. POPE217 interacts more with the protein than POPE79 and ~4 and ~1 H-bonds during the unrestrained part of the simulation, respectively. POPE79 has between one and two H-Bonds with helix 5 and occasionally one with helix 2. The other phospholipid, POPE217, has a greater number of H-bonds: up to four with helix 5 during the unrestrained part of the simulation. Interactions with helix 2 are more limited, between 1 and 2 H-bonds.

In agreement with these analyses, visual examination of these phospholipids (Fig. 6) shows the presence of one

H-bond between POPE79 and threonine Thr-81. The other phospholipid has two H-bonds with arginine Arg-78 and one H-bond with threonine Thr-81.

Lateral motion of lipids

In terms of short scale lateral diffusion, two classes of lipids can be distinguished: bound and free lipids (30). In our analysis, bound lipids were defined as those that had their phosphorus atom within 5 Å of the protein at the beginning of the simulation, whereas free lipids are defined as those not having their phosphorus atom within 5 Å of the protein. The selection was updated every 500 ps. To investigate the influence harmonic restraints might have on the lipid diffusion behavior, lateral diffusion coefficients have also been calculated for time range representative of the restrained and unrestrained part of the simulation, respectively, 1–15 ns and 20–50 ns. Characteristic lipids POPE79 and POPE217 have also been characterized. Values for short scale lateral diffusion coefficients are summarized in Table 3, where POPE44 is representative of a bulk lipid.

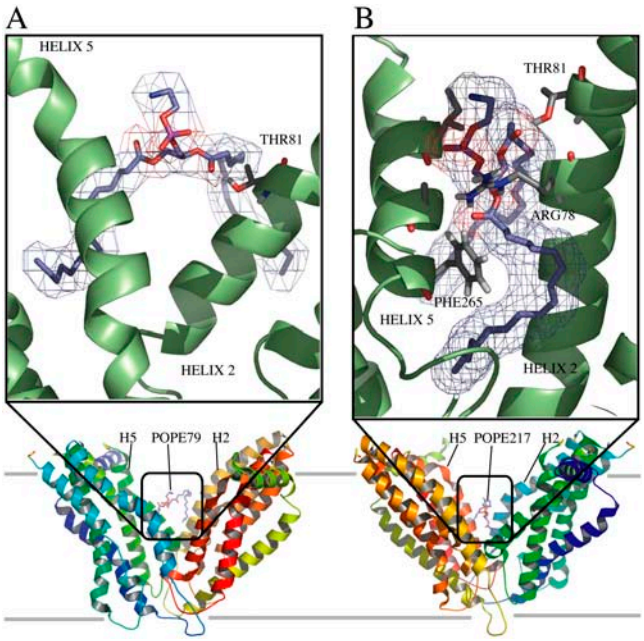


FIGURE 6 Snapshot of key protein-lipid interactions seen from the cavity. The extracellular side is located on the bottom and the intracellular side above. The bottom ribbon representation of the MsbA dimer indicates the orientation of the protein. The ribbons are colored from blue to green from the N- to C-terminus of the M1 monomer and from green to red from the N- to C-terminus of the M2 monomer. In both diagrams the protein is shown in ''cartoon'' format (in *green*). Selected side chains and lipids are shown in ''sticks'' format, with nitrogen and oxygen atoms colored blue and red, respectively. (*A*) Threonine (Thr-81) residue interacting with phospholipid POPE79. The threonine hydroxyl group is H-bonded with the POPE SN2 carbonyl group. (*B*) Threonine (Thr-81) and arginine (Arg-78) residues interacting with phospholipid POPE217. The threonine hydroxyl group is H-bonded with the POPE SN1 carbonyl group. The arginine amine group is H-bonded with the POPE SN2 carbonyl and ester group.

As frequently seen, free lipids have a larger diffusion coefficient than bound phospholipids but of the same order of magnitude. In contrast, POPE79 and POPE217 have smaller diffusion coefficients, as testified by the values obtained for a lipid located in the bulk, POPE44. The POPE79 diffusion coefficient is smaller than that of POPE217. This can be linked to the different conformations the two phospholipids are adopting, since POPE79 has its two hydrophobic chains outside of the protein cavity whereas POPE 217 has one chain inside the cavity and one chain outside. Lipids transported by MsbA are supposedly recruited inside the membrane before being flipped and expelled in the extracellular side membrane as suggested by Reyes and Chang (17) in the case of lipopolysaccharide (LPS). In our case, POPE217 might be in an intermediate state that would explain a significantly higher diffusion coefficient than POPE79. In light of these considerations, it is interesting to study the influence of restraints on these diffusion coefficients. Restraints have little effect on diffusion coefficients of the free, the bound, and the entire set of phospholipids. They have nonetheless a significant influence on the key phospholipids POPE79 and POPE217. Once restraints are removed, POPE217 diffusion is about two times larger, an observation consistent with the hypothesis of a transition state in which significant protein conformational change is required to allow the lipid to be transported. The POPE79 diffusion coefficient is, on the other hand, reduced by a factor of 2. Local protein movement observed in the unrestrained part of the simulation might allow this phospholipid to reach a stable state by enabling it to reach an energy minimum. These observations can be corroborated by the examinations of root mean square deviations (RMSDs) (see Fig. 8 *B*). Helix 5 of monomer 1, which is in contact with phospholipid POPE217, is prone to more fluctuations (final Cα RMSD 0.361 nm) than its counterpart of monomer 2 interacting with the other phospholipid (final Cα RMSD 0.316 nm). Helices 2 have similar RMSDs in both monomers.

Protein stability

Protein stability can be characterized by examining the RMSD of the protein structure as a function of time. RMSDs for the entire protein together with monomer 1 and 2 are presented in Fig. 7. The RMSD of the entire protein continuously increases during the restrained part of the simulation to reach 0.25 nm after 16 ns. After removal of the restraints, RMSD keeps on rising but stabilizes at 0.51 nm after 25 ns without any further significant drift. When plotting the contributions of the two monomers separately, it appears that the contribution from monomer 2 is larger than that of monomer 1, with final values of, respectively, 0.51 and 0.47 nm. These values, although large, are of the same order of magnitude as the one observed by Campbell et al. in their in silico study of MsbA (19). These numbers have to be related to the resolution of the crystal structure used (0.45 nm)

TABLE 3 POPE lipids short scale lateral diffusion coefficients ($\times 10^{-5}$ cm^2 s^{-1})

Time range (ns)	D*§ < 5 Å	D†‡ > 5 Å	Overall‡	POPE79§	POPE217§	POPE44§
0–50	0.0217 (±0.0080)	0.0334 (±0.0051)	0.0323 (±0.0048)	0.0034 (±0.0019)	0.0082 (±0.0090)	0.0104 (±0.0098)
0–15	0.0215 (±0.0086)	0.0340 (±0.0052)	0.0334 (±0.0048)	0.0070 (±0.0030)	0.0048 (±0.0044)	0.0094 (±0.0051)
20–50	0.0227 (±0.0075)	0.0334 (±0.0050)	0.0327 (±0.0047)	0.0033 (±0.0020)	0.0114 (±0.0107)	0.0134 (±0.0125)

*Lipids for which a phosphorus atom is located within 5 Å of the protein at the beginning of the simulation, updated every 500 ps.
†Lipids for which a phosphorus atom is located above 5 Å away from the protein at the beginning of the simulation, updated every 500 ps.
‡Diffusion coefficients calculated in 500-ps frames.
§Diffusion coefficients calculated in 5-ns frames.

and to the fact that it only contains a Cα trace. Furthermore, our study only considers the transmembrane domain of the protein: missing intracellular loops likely have a direct influence on the stability of the assembly.

The secondary structure reports on the protein stability over the course of the simulation (see Fig. 1 in Supplementary Data). Significant loss of helicity is observed for helix 3 in both monomers, helix 4 in monomer 1, and helix 5 in both monomers. Most of the loss of secondary structure happens on the extracellular side of the protein. This observation can be explained by the fact that intracellular loops joining helices 4 to 5 and helices 2 to 3 have been omitted for computational reasons. Consequently, amino acids located at the cut and in the vicinity of the membrane polar headgroups tend to interact with the latter, causing some helices to slide toward the intracellular side. Helices 3 and 4 are particularly exposed to the membrane headgroups, which could explain the loss of helicity observed on the extracellular side. The loss of helicity observed might be related to its interaction with the elbow helix. Interestingly, the elbow helix, which was model built, is conserved in the integrality of the simulation (especially in monomer 2). This feature was not observed in the study of Campbell et al. (19), which reveals the importance of the membrane choice for the protein stability.

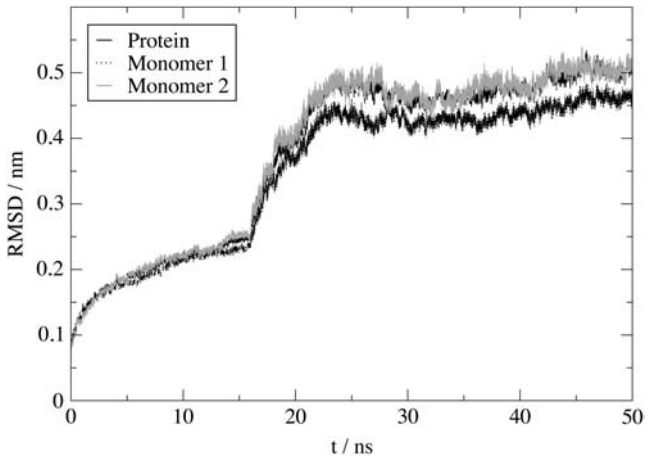


FIGURE 7 Time evolution of Cα RMSDs for protein (*black line*), monomer 1 (*black dotted line*), and monomer 2 (*gray line*). Intramonomer RMSDs are calculated by fitting each monomer on itself and measuring RMSDs relative to the fit over time.

Motions in the protein have been further characterized by examining tilt angles with respect to the membrane normal and the RMSDs of each helix relative to itself as a function of time (Fig. 8). Helices 2, 3, 4, and 6 show similar tilt angle fluctuations in both monomers with a final drift of ~5°, a value reflecting standard oscillations that might be observed in MD simulations exceeding tens of nanoseconds. There is a significant difference in the final tilt angles of helix 5 in monomers 1 and 2. Whereas the helix in monomer 2, which is in contact with phospholipid POPE79 does not show any significant final drift, helix 5 of monomer 1 shows a final drift of nearly 10°. This helix is in contact with phospholipid POPE217, which undergoes significant conformational change as noted above. We might therefore think that the observed tilt angle drift is part of the mechanism allowing phospholipid POPE217 migration toward the protein cavity and not just standard fluctuation. Concerning helix 1, there is a significant difference between the two monomers. The tilt angle is prone to more fluctuations in the first monomer; the question of whether this observation is significant remains to be elucidated. Intrahelix RMSDs are similar for both monomers except for helices 4 and 5. For the latter, the drift is more important in the first monomer, a logical observation considering the fact that this helix interacts with phospholipid POPE217, which undergoes significant conformational change. Helices 1, 2, and 6 are quite stable considering the overall protein stability, with final drifts of the order of 0.3 nm or less. The remaining helices show larger structural fluctuations with final drifts of the order of 0.4 nm.

Dimer dynamic behavior

Time evolution of the two monomers principal axes of inertia has been investigated. In Fig. 9 we report the angles made by the two monomers with the bilayer normal (z axis), together with the angle between the two monomers' axes of inertia. Both monomers are relatively stable. Monomers are initially oriented with an angle of 35° with respect to the bilayer normal; after restraints are removed, these angles oscillate with an amplitude of <10° to reach final values of 37° and

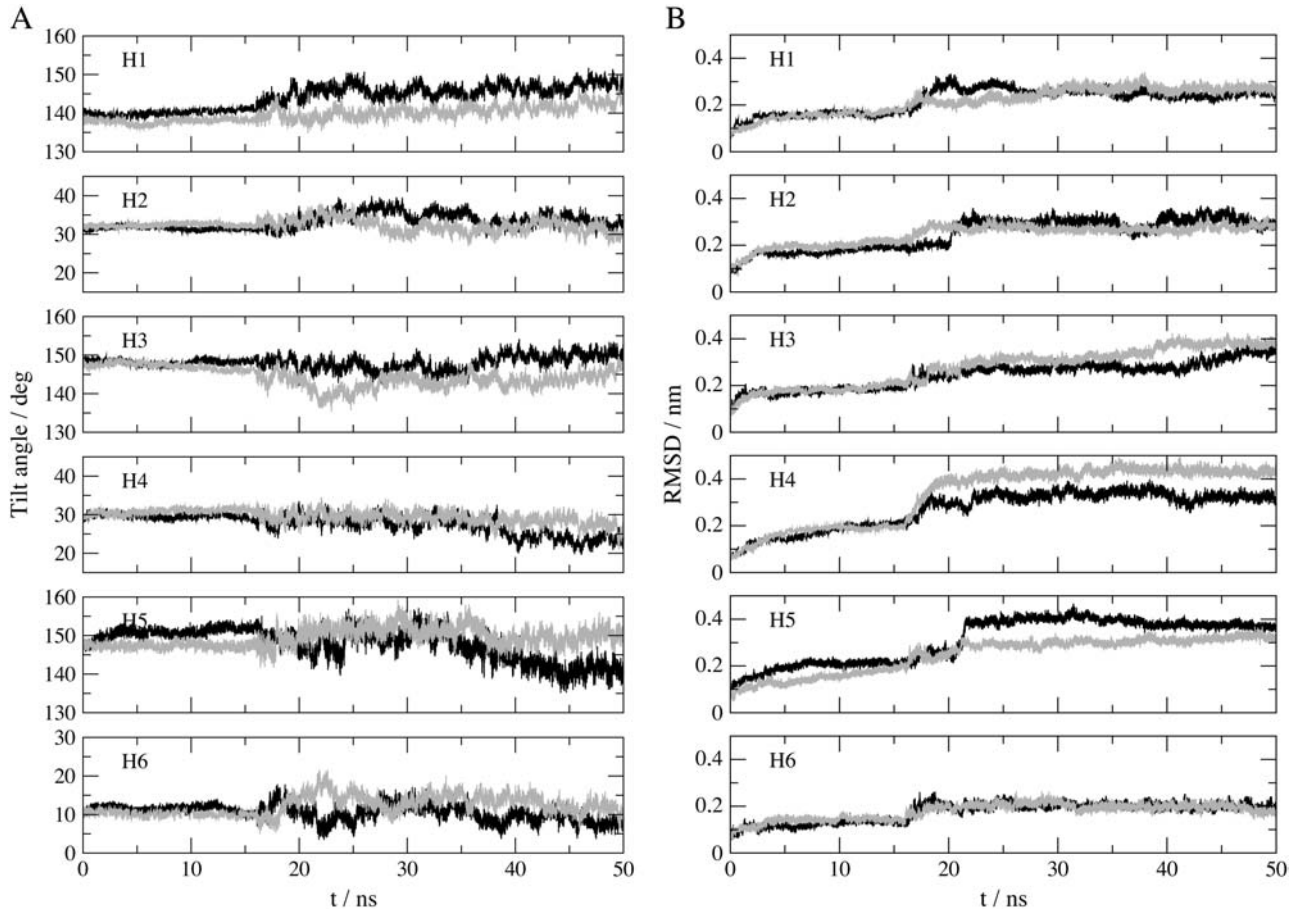


FIGURE 8 Time evolution of helices tilt with respect to the bilayer normal (*A*) and intrahelix Cα RMSDs (*B*) for monomer 1 (*black line*) and monomer 2 (*gray line*). Intrahelix RMSDs are calculated by fitting each helix on itself and measuring RMSDs relative to the fit over time.

29°, respectively, for monomers 1 and 2. Considering the length of the simulation, these drifts testify to the stability of the assembly. Observation of the angle between the two monomers' principal axes of inertia reveals a slight closing of the chamber. The angle has an initial value of 70°, which decreases during the simulation to reach a final value of 63°. Our protein model did not include the nucleotide binding domains, leaving helices of the intracellular domain free to interact with the membrane surface. These interactions affect the stability of the assembly and can partly explain the closing of the chamber. Indeed, the concerned amino acids tend to bend the helices to favor interactions with the membrane headgroups. These interactions do not affect the assembly stability on the timescale of this study (final drift of 7°) but might have more significant effects if longer timescales were considered.

The dimer stability has been further examined by investigating the time evolution of the number of H-bonds between the two monomers (Fig. 10). Since H-bonds between atoms can form and disappear quite quickly, the total number of H-bonds is subject to significant fluctuations. Hence, only the general trend of the curve is discussed here. The initial number of polar contacts is limited, with one to two H-bonds between the two monomers. This number continuously increases during the restrained part of the simulation (up to 16 ns) to reach a value of ~6 H-bonds. During the unrestrained part of the simulation (16 ns and beyond), the number of H-bonds initially decreases to about four interactions and finally ''stabilizes'' at about six interactions from 34 ns to the end of the simulation. This plateau testifies to the stability of the dimer interface.

The examination of polar contacts at different times reveals that most H-bonds are between extracellular domain 1 (ECD1) of one monomer and extracellular domain 3 (ECD3) of the other monomer, and reciprocally. Typical H-bonds are illustrated in Fig. 11. ECD1 of the first monomer, which is the loop joining helices 1 and 2, is drawn in blue; and ECD3 of monomer 2, the loop joining helices 5 and 6, is drawn in red. On this snapshot, amino acids Leu-52, Asp-53, Phe-56, Gly-57, Lys-58, and Arg-61 of ECD1 are H-bonded to amino acids Ser-274, Val-275, and Asp-276 of ECD3. Among notable interactions, the H-bond between the carboxyl group of Asp-276's side chain and the NH group of Arg-61's side chain as well as with the NH group of Lys-58's

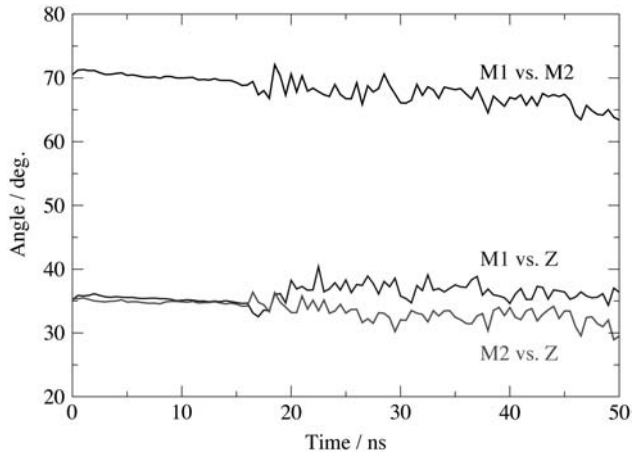


FIGURE 9 Time evolution of the principal axis of inertia of monomers 1 and 2. Angles of the principal axis of inertia of monomers 1 and 2 with the bilayer normal are reported together with the angle between the two monomers' principal axes of inertia.

backbone can be highlighted. H-bonds between Asp-53, Ser-274, and Val-275's backbone CO groups and Val-275, Phe-56, and Gly-57's backbone NH groups, respectively, together with H-bonds between Leu-52's backbone CO group and the Ser-274 hydroxyl group are relatively stable and can be observed during most of the second part of the simulation. These observations reveal the existence of a dimer interface which is stabilized by the appearance of numerous H-bonds during the course of the simulation and noticeably interactions involving backbone atoms.

Conservation analysis

We first analyzed the projection at the protein surface of the evolutionary conservation based on the MSA of 49 MsbA-like ABC transporters. Fig. 12 *A* shows that conserved

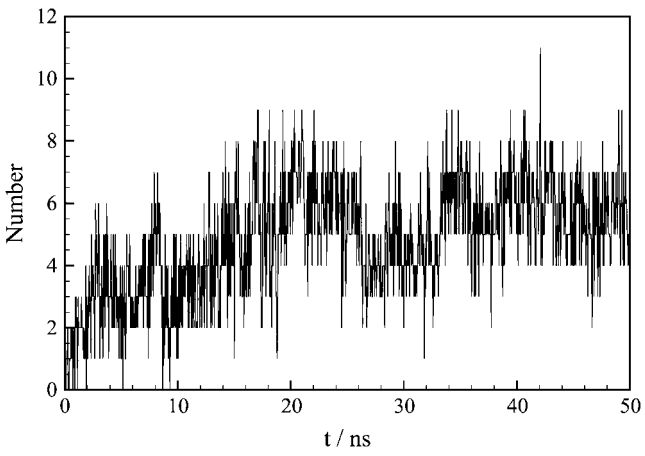


FIGURE 10 Time evolution of the number of H-bonds between the two monomers. The cutoffs used to define H-bonds are 0.25 nm for the hydrogen-acceptor distance and 60° for the donor-hydrogen-acceptor angle.

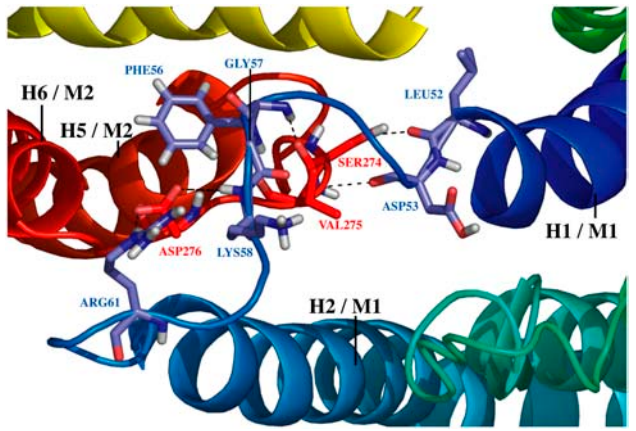


FIGURE 11 Snapshot of polar contacts between the two monomers at $t = 25$ ns. The protein is seen from the extracellular side. The protein is shown in ''cartoon'' format; selected side chains are shown in ''sticks'' format with nitrogen and oxygen atom colored blue and red, respectively. Helices 1 and 2 of monomer 1 are colored in dark and light blue, respectively, and helices 5 and 6 of monomer 2 are colored in orange and red, respectively. Illustrated H-bonds are those present between extracellular domains 1 and 3.

positions (colored *red*) can be found in different regions of the protein. The regions with the highest concentration of conserved positions are located inside the dimer cavity at the level of the cytosolic lipid interface and at the hinge between both monomers (shown in Fig. 12 *A* as hinge M1 and hinge M2 for monomers 1 and 2, respectively). The residues located at the membrane interface interacting with the lipid headgroups also appear relatively conserved, probably to ensure a proper positioning of the protein in the bilayer. In contrast, the external surface of the protein pointing toward the hydrophobic lipid environment contains more variable positions. Indeed, as long as the hydrophobic character of the residues is kept at the lipid-exposed surface, the structural integrity and function of the protein is likely to be maintained. The structure of the MsbA analyzed in this simulation only contains the transmembrane domain. Hence conservation data shown here only concern this region of the MsbA protein. It is interesting to note that in the entire structure the region containing the ICD region is also conserved to a similar extent as that of the internal cavity. This conservation probably demonstrates the importance of these regions in transmitting a proper signal transduction between the domains involved in the hydrolysis of the ATP and the transmembrane domain involved in the physical transport of substrates.

To further investigate the evolutionary constraints existing in the MsbA protein family, we decided to discriminate between two types of evolutionary pressures: i), those involved in the general transport mechanism of compounds through the membrane likely to be common to the entire ABC transporter superfamily, and ii), the conservation events that specifically occurred in the MsbA close homologs for reasons such as the specificity of the substrate

recognition. In that scope we further aligned 47 sequences of ABC transporters more distantly related to MsbA in the evolution. The clustering between the close and remote homologs was based on a rigorous phylogenetic tree analysis of the entire set of 96 sequences. We then analyzed among the conserved positions of MsbA homologs those positions significantly more conserved in the set of close homologs of MsbA than in the set of remote ones. This protocol corresponds to a differential conservation analysis based on the comparison of evolutionary rates and of their associated confidence intervals (see Methods). Interestingly, using a stringent confidence interval (probability above 95%) only two positions stand out, Arg-78 and Trp-165 (Fig. 12 *B*). Residue Arg-78 stands at the dimer hinge region, whereas residue Trp-165 is located in a loop pointing out toward the extracellular medium. Decreasing the confidence interval to 90%, three other positions are found to exhibit a differential conservation effect: Gln-134, Ala-142, and Ser-186 (see Figs. 2 and 3 in Supplementary Material). These residues are located in helices 3 and 4 close to the inner membrane interface. Although at this stage we cannot further interpret these evolutionary pressures, the conservation pattern suggests that these residues may also play a specific and important role in MsbA close homologs. Among the five positions pinpointed by the differential conservation protocol, Arg-78 remains the most conserved position in *E. coli* MsbA's close homologs. Arg-78 is precisely the residue found interacting with the two key lipids (POPE79 and POPE217 in our simulation) and stabilizing the semipore-like structure.

DISCUSSION

MsbA is a bacterial ABC transporter first recognized to export lipid A, a toxic metabolic intermediate involved in the biosynthesis of the outer membrane, and now reported to also handle various lipids, LPSs, and even cytotoxic drugs. As a lipid transporter, it is of interest to investigate the relationships between lipids and this protein, both from a general view of membrane biophysics and from the specific perspective of its molecular mechanism of functioning. As a multispecific active exporter of amphiphilic compounds, analyzing its molecular properties is highly desirable to gain some useful knowledge in the molecular enzymology of the mammalian multidrug transporter, P-gp, remarkable for its close structural and functional homologies, which would be valuable in various pharmacological applications. In this context, the evidencing by MD simulations of specific interaction of two phospholipid molecules at the hinge of MsbA dimer structure, in parallel with a very local invagination of the cytosolic leaflet of the membrane surrounding the protein as a consequence of its open-state conformation, can be considered a significant shedding of light on this model of ABC transporter. Moreover, these observations might be of great interest to suggest site-directed mutagenesis experiments.

Assessment of the model quality

MD is an adequate tool for analyzing molecular mechanisms determining protein-lipid interactions. However, the quality of the protein structure modeled is of prime importance for the reliability of the data generated. Actually, since we started from a rather moderate resolution crystal structure, we have chosen to take into account only the transmembrane domain of the protein, without considering the nucleotide binding domains and the other fragments clearly outside the membrane, because they would imply constraints that are too speculative when compared to their expected low repercussion

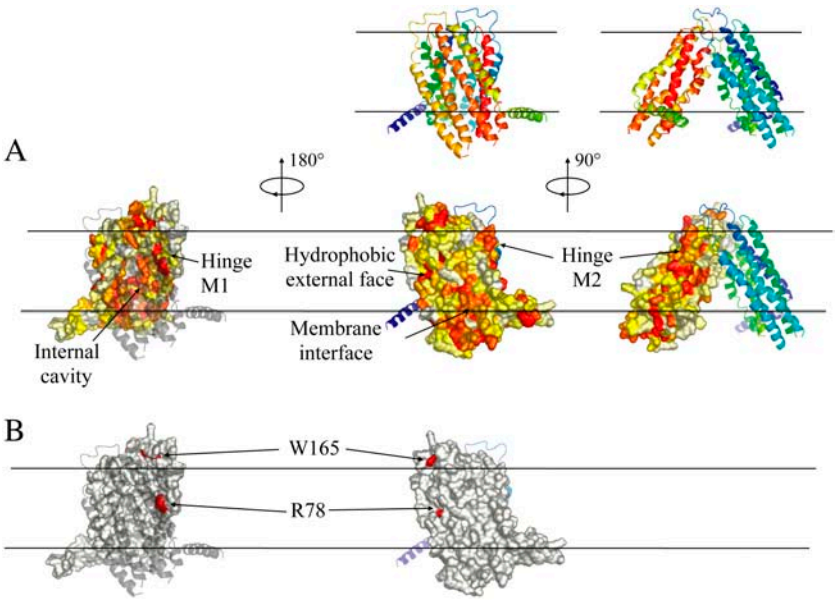


FIGURE 12 Conservation analysis calculated from the MSA of MsbA proteins and mapped on the surface of the dimer. The top ribbon representation of the MsbA dimer indicates the orientation of the corresponding surface representations shown below in panels *A* and *B*. The ribbons are colored from blue to green from the N- to C-terminus of the M1 monomer and from green to red from the N- to C-terminus of the M2 monomer. (*A*) The surface of monomer M2 is colored with respect to the conservation scores. Most to least conserved positions are colored from red to white, respectively, yellow corresponding to the average conservation score (*left*, *middle*, and *right panels* show the interior, the lipid-exposed, and the hinge regions of the dimer, respectively). (*B*) Positions that are specifically and significantly conserved in the MsbA close homologs family but not in the more general LPS ABC transporters family are colored in red at the surface of M2 monomer (*left* and *right panels* show the interior and the lipid-exposed regions, respectively).

on transmembrane domain dynamics. This has been validated a posteriori as shown by the reasonable stability which is obtained for the protein. In addition, this may be consistent with a report on the related protein P-gp according to which the transmembrane domain alone is sufficient for trafficking to cell surface and drug recognition (46). Missing intracellular domains make difficult a strong assumption of the model quality, and this point will be obviously further tested only after modeling the whole protein. Nevertheless, a PROCHECK (47) analysis performed on the initial structure revealed that all protein-health characteristics were within the acceptable range.

Structural stability of the protein

Even though intrahelical RMSDs denote important structural drifts for some transmembrane segments, the dimer structure is relatively stable. Previous studies performed using an octane slab (19) showed large structural drift of the transmembrane domains over just 1 ns. Our study shows a significant drift that stabilizes at 0.5 nm after 25 ns and remains constant until the end of the simulation. This highlights the critical role played by the nature of the membrane in the stability of the assembly. The structure used in this study was closely based upon the crystallographic MsbA Cα coordinates (PDB code, 1JSQ). The rather poor resolution of this structure partly explains the fluctuations observed in our simulation. The significant rearrangement of transmembrane helices that can be observed is not typical in membrane protein simulations. MsbA is a lipid transporter that supposedly undergoes significant transmembrane domain rearrangement during the transport cycle. As such, the transmembrane domains must consequently be very flexible, which could explain the structural drifts observed in this study. Anyway, this does not preclude the conclusions shown by the simulations about lipid dynamics in the protein environment.

Implication for lipid transport

The presented simulations show that the open-state MsbA structure can be inserted into a lipid membrane and that this structure may be physiologically viable. This (resting) conformation is in agreement with recent data obtained using electron paramagnetic resonance and chemical cross-linking techniques (48). Thanks to the dynamic vision they allow, the molecular simulations also give invaluable insights into the transport mechanism that might be adopted by MsbA to translocate lipids.

MsbA transports lipid A and also phospholipids and LPSs (7). Furthermore, it has been recently shown experimentally that MsbA is responsible for the export of lipid A by mediating its transbilayer translocation (i.e., it displays a ''flip-floppase activity'') (49). From the crystal structure obtained for *S. typhimurium* MsbA, a transport mechanism has been suggested by Reyes and Chang (17) from the sequestration of the polar sugar headgroup of LPS. The first two steps involve i), initial binding to the elbow helix of the apo structure, and ii), lipid headgroup insertion in the chamber of the apo structure. The last two steps involve transporter conformational change related to ATP binding and hydrolysis involving lipid headgroup flip while lipid chains are dragged through the lipid bilayer. In our simulation, analysis of protein-lipid interactions reveals the occurrence of two distinct conformations for phospholipids POPE79 and POPE217. These two singularized phospholipid molecules are actually the result of statistical sampling yielded by MD simulation. These two conformations might be in agreement with the transport mechanism suggested by Reyes and Chang in the sense that they show that the alkyl chains of these lipid molecules can adopt different conformations when interacting with this protein environment. This observation is also in good agreement with the differential conservation analysis that pinpointed the residue Arg-78 as one of the residues highly conserved in close homologs of MsbA although more variable in more distantly related organisms. All along the simulation, this residue is found interacting with the heads of the phospholipids POPE79 and POPE217 and stabilizing the semipore structure. Although lipid A was not included in the simulation, the evolutionary pressure found at position 78 may well reflect the functional specificity of the MsbA for the binding of lipid A.

Otherwise, we observed local and partial invagination in the cytosolic leaflet of lipid headgroups in contact with the protein, this being a direct consequence of the optimal filling of the periproteic space taking into account geometric constraint induced by the open state. This very local membrane structure perturbation might be the preliminary step of the transport mechanism. This would eventually imply translocation of the polar headgroup, coupled with energy-giving ATP hydrolysis, within the chamber of the observed semipore-like assembly. Indeed, it can be assumed that the subsequent energy-requiring translocation step will happen on a far larger timescale and will be accompanied by a significant rearrangement of the protein.

In addition, if we consider, in agreement with Chang's suggestion (16), as a fair possibility that MsbA alternates during its enzymatic cycle between open and closed states, it appears that the most altered part of the protein during such a transconformation is the hinge defined by the dimeric association. The fact that the hinge between both monomers is one of the most conserved regions in the transmembrane domain also supports the functional importance of this region in the transport process. It is then relevant, and even remarkable, that two phospholipids have been individualized on the hinge regions of the protein, located in the contact region of helices 2 and 5. These phospholipids establish specific interactions with the protein, when compared to other phospholipids in the bulk or in the ring surrounding the protein. They adopt two distinct conformations, resulting

from the initial setup of the simulation. Interestingly, their location is similar to the one suggested for initial LPS binding in the structure of *S. typhimurim* MsbA (17), proposed at the level of the hinge of the dimer, involving elbow and helix 1 but also helices 2 and 5 due to the size of the ligand. This constitutes a clear evocation of possible transport site, which would mainly involve six amino acids in each side of the hinge and could be a general characteristic of substrate transport by MsbA. Actually, this would be in full agreement with the fact that the drugs handled by MsbA are amphiphilic and likely come to interact with the protein via the membrane phase after partitioning into the lipids, as described for P-gp.

Connection with P-gp functioning

The mammal multidrug transporter P-gp is responsible for the MDR of some tumor cells against cytotoxic drugs used in anticancer chemotherapy, leading to treatment failures. P-gp presents enough sequence homology with MsbA to raise an interest in analyzing MD data to possibly shed light on its transport function. In particular, P-gp is now described as being rather a translocase, and more precisely a floppase, realizing the active transfer of its hydrophobic substrates, lipids, or drugs, from the cytosolic to the exoplasmic leaflet. As an ABC protein with a tandem structure, it looks like a possible MsbA dimer. Such a ''reminiscence'' of a dimeric structure is supported by data showing that the coexpression of the two halves of the protein is able to exhibit a drug-stimulated ATPase activity, a fair indication of a full functionality, whereas each separated half is only capable of a basal ATP hydrolysis (50). In the same line, it has been shown by molecular biology techniques that a good flexibility of the linker between the N- and C-terminus is required for P-gp activity (51). From a structural point of view, the low resolution image obtained by tridimensional reconstruction shows an inner chamber with a possible opening/communicating to the near lipid phase (52). This protein structure displays under the action of ATP hydrolysis a large amplitude transconformation (53) involving tertiary structure changes evidenced by various spectroscopic techniques (54–56). Furthermore, projection data from two-dimensional crystals of P-gp are consistent with the MsbA transmembrane domain structure (57), even if more recent structure refinements have shown that P-gp does not have a strictly symmetric transmembrane domain (58). As a whole, since the molecular mechanism of P-gp-mediated drug transport is currently very poorly understood, available data make MsbA a fair structural and functional model for P-gp in the frame of a first approach.

The data generated by molecular simulation of the interactions between MsbA and surrounding lipids have pinpointed few amino acids in the hinge region of the protein, involving helices 2 and 5 from each of the two MsbA monomers, for being able to specifically interact with particular lipid molecules. This is strikingly consistent with experimental data coming from cross-linking assays of the helix-helix neighborhood in P-gp, evidencing a close proximity between helices 2 and 11 as well as between helices 5 and 8 (59). In addition, helix 11 in P-gp has often been involved by directed point mutagenesis experiments in the specificity of drug recognition (60,61). Also, different photolabeling experiments indicated helices 4-5-6, 7-8, and 11-12 as targets for drug binding (62–64). Furthermore, it has been reported that among the 12 helices in P-gp, 7 (1-2-3-5-6-8-11) are rich in aromatic residues conserved among the various isoforms of P-gp, 4 of them (3-5-8-11) being suited for interacting with rhodamine 123, a typical polycyclic drug substrate, even if this is the result of a rather crude molecular modeling process considering each of the P-gp helices separately (65). As a whole, helices 5, 8, and 11 in P-gp thus appears to be consistently involved in the transported substrate recognition. In addition, among the six amino acids of MsbA specifically interacting with POPE lipids, three of them are conserved in the aligned P-gp sequence in each of its half-tandem: Ser-260, Ala-262, and Phe-265 respectively conserved in positions 309-311-314 and 952-954-957. According to available mutagenesis data (60), these residues are located close to some of the various amino acids involved in drug recognition by P-gp.

Finally, the molecular characteristics obtained here for MsbA interaction with membrane lipids, and that can be favorably compared with what is known for P-gp, could also be considered for the other ABC transporters from the B subfamily, which have the closest homology with MsbA. In particular, ABC B4 (also called MDR3) is known as a phosphatidylcholine translocase, mainly expressed in biliary canalicule where it participates to bile secretion (3), and it is conceivable that such a phospholipid transporter shares with MsbA (and P-gp) some similarities at the molecular level for transport mechanism. Also, ABC B2 and B3 (also known as TAP1/2) constitute as a dimer a transporter of hydrophobic peptides, formed by antigenic restriction and loaded into endoplasmic reticulum of macrophages (66), which have flexible structures and could thus be handled in a similar manner to phospholipids. Indeed, in this study, we showed that the coupling between simulation data and phylogenetic analyses could powerfully highlight residues likely to play a role in the transport of a specific class of substrates. The protocol developed here may be applied in the future to other classes of ABC transporters provided enough sequences can be aligned together with the corresponding knowledge on their transport specificities.

As a conclusion, based on the structure of the bacterial ABC transporter MsbA, molecular dynamic simulation of a large atomic system comprising a membrane protein inserted in its lipid membrane environment has allowed us to evidence the functional importance of the hinge region of the dimeric protein and the destabilization of the cytosolic leaflet, which may be proposed as a general mechanism for

other ABC transporters of lipids and/or amphiphilic drugs involved in mammalian physiology.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.

Our thanks go to the Centre de Calcul Recherche et Technologie (CCRT) of the Commissariat á l'Energie Atomique for access to computing resources. D.Y.H. thanks Massimo Marchi for helpful discussions.

H.M. is supported by a Délégation Générale pour l'Armement (DGA) fellowship.

REFERENCES

1. Higgins, C. F. 1992. ABC transporters: from microorganisms to man. *Annu. Rev. Cell Biol.* 8:67–113.
2. Gottesman, M. M., and S. V. Ambudkar. 2001. Overview: ABC transporters and human disease. *J. Bioenerg. Biomembr.* 33:453–458.
3. Borst, P., N. Zelcer, and A. van Helvoort. 2000. ABC transporters in lipid transport. *Biochim. Biophys. Acta.* 1486:128–144.
4. Bates, S. E., R. Robey, K. Miyake, K. Rao, D. D. Ross, and T. Litman. 2001. The role of half-transporters in multidrug resistance. *J. Bioenerg. Biomembr.* 33:503–511.
5. Kruh, G. D., H. Zeng, P. A. Rea, G. Liu, Z. S. Chen, K. Lee, and M. G. Belinsky. 2001. MRP subfamily transporters and resistance to anticancer agents. *J. Bioenerg. Biomembr.* 33:493–501.
6. Sauna, Z. E., M. M. Smith, M. Muller, K. M. Kerr, and S. V. Ambudkar. 2001. The mechanism of action of multidrug-resistance-linked P-glycoprotein. *J. Bioenerg. Biomembr.* 33:481–491.
7. Pohl, A., P. F. Devaux, and A. Herrmann. 2005. Function of prokaryotic and eukaryotic ABC proteins in lipid transport. *Biochim. Biophys. Acta.* 1733:29–52.
8. Schmitt, L., and R. Tampe. 2002. Structure and mechanism of ABC transporters. *Curr. Opin. Struct. Biol.* 12:754–760.
9. Locher, K. P., A. T. Lee, and D. C. Rees. 2002. The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science.* 296:1091–1098.
10. Chang, G., and C. B. Roth. 2001. Structure of MsbA from E-coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science.* 293:1793–1800.
11. Doerrler, W. T., M. C. Reedy, and C. R. Raetz. 2001. An Escherichia coli mutant defective in lipid export. *J. Biol. Chem.* 276:11461–11464.
12. Woebking, B., G. Reuter, R. A. Shilling, S. Velamakanni, S. Shahi, H. Venter, L. Balakrishnan, and H. W. van Veen. 2005. Drug-lipid A interactions on the Escherichia coli ABC transporter MsbA. *J. Bacteriol.* 187:6363–6369.
13. Reuter, G., T. Janvilisri, H. Venter, S. Shahi, L. Balakrishnan, and H. W. van Veen. 2003. The ATP binding cassette multidrug transporter LmrA and lipid transporter MsbA have overlapping substrate specificities. *J. Biol. Chem.* 278:35193–35198.
14. Higgins, C. F., and M. M. Gottesman. 1992. Is the multidrug transporter a flippase? *Trends Biochem. Sci.* 17:18–21.
15. Romsicki, Y., and F. J. Sharom. 2001. Phospholipid flippase activity of the reconstituted P-glycoprotein multidrug transporter. *Biochemistry.* 40:6937–6947.
16. Chang, G. 2003. Structure of MsbA from Vibrio cholera: a multidrug resistance ABC transporter homolog in a closed conformation. *J. Mol. Biol.* 330:419–430.
17. Reyes, C. L., and G. Chang. 2005. Structure of the ABC transporter MsbA in complex with ADP.vanadate and lipopolysaccharide. *Science.* 308:1028–1031.
18. Stenham, D. R., J. D. Campbell, M. S. Sansom, C. F. Higgins, I. D. Kerr, and K. J. Linton. 2003. An atomic detail model for the human ATP binding cassette transporter P-glycoprotein derived from disulfide cross-linking and homology modeling. *FASEB J.* 17:2287–2289.
19. Campbell, J. D., P. C. Biggin, M. Baaden, and M. S. Sansom. 2003. Extending the structure of an ABC transporter to atomic resolution: modeling and simulation studies of MsbA. *Biochemistry.* 42:3666–3673.
20. Karplus, M., and J. A. McCammon. 2002. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9:646–652.
21. Essex, J. W., M. M. Hann, and W. G. Richards. 1994. Molecular dynamics simulation of a hydrated phospholipid bilayer. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 344:239–260.
22. Tieleman, D. P., S. J. Marrink, and H. J. Berendsen. 1997. A computer perspective of membranes: molecular dynamics studies of lipid bilayer systems. *Biochim. Biophys. Acta.* 1331:235–270.
23. Tobias, D. J., K. C. Tu, and M. L. Klein. 1997. Atomic-scale molecular dynamics simulations of lipid membranes. *Curr. Opin. Colloid Interface Sci.* 2:15–26.
24. Sansom, M. S. 1998. Models and simulations of ion channels and related membrane proteins. *Curr. Opin. Struct. Biol.* 8:237–244.
25. Tieleman, D. P., I. H. Shrivastava, M. R. Ulmschneider, and M. S. Sansom. 2001. Proline-induced hinges in transmembrane helices: possible roles in ion channel gating. *Proteins.* 44:63–72.
26. Tieleman, D. P., and H. J. Berendsen. 1998. A molecular dynamics study of the pores formed by Escherichia coli OmpF porin in a fully hydrated palmitoyloleoylphosphatidylcholine bilayer. *Biophys. J.* 74:2786–2801.
27. Valadie, H., J. J. Lacapère, Y. H. Sanejouand, and C. Etchebest. 2003. Dynamical properties of the MscL of Escherichia coli: a normal mode analysis. *J. Mol. Biol.* 332:657–674.
28. Roux, B. 2005. Ion conduction and selectivity in K(+) channels. *Annu. Rev. Biophys. Biomol. Struct.* 34:153–171.
29. Marrink, S. J., and A. E. Mark. 2002. Molecular dynamics simulations of mixed micelles modeling human bile. *Biochemistry.* 41:5375–5382.
30. Deol, S. S., P. J. Bond, C. Domene, and M. S. Sansom. 2004. Lipid-protein interactions of integral membrane proteins: a comparative simulation study. *Biophys. J.* 87:3737–3749.
31. Burnell, E., L. van Alphen, A. Verkleij, and B. de Kruijff. 1980. 31P nuclear magnetic resonance and freeze-fracture electron microscopy studies on Escherichia coli. I. Cytoplasmic membrane and total phospholipids. *Biochim. Biophys. Acta.* 597:492–501.
32. Lindahl, E., B. Hess, and D. van der Spoel. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* [Online]. 7:306–317.
33. Berendsen, H. J., J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.
34. Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an N.log(N) method for Ewald sum in large systems. *J. Chem. Phys.* 98:10089–10092.
35. Hess, B., H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* 18:1463–1472.
36. Holm, L., and C. Sander. 1991. Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.* 218:183–194.
37. Canutescu, A. A., A. A. Shelenkov, and R. L. Dunbrack. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* 12:2001–2014.
38. Vriend, G. 1990. WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* 8:52–56.
39. Berendsen, H. J. C., J. R. Grigera, and T. P. Straatsma. 1987. The missing term in effective pair potentials. *J. Phys. Chem.* 91:6269–6271.
40. Bond, P. J., J. D. Faraldo-Gomez, and M. S. Sansom. 2002. OmpA: a pore or not a pore? Simulation and modeling studies. *Biophys. J.* 83:763–775.

41. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.

42. Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton. 2004. The Jalview Java alignment editor. *Bioinformatics.* 20:426–427.

43. Mayrose, I., D. Graur, N. Ben-Tal, and T. Pupko. 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* 21:1781–1791.

44. Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* 8:275–282.

45. Park, T., J. B. Loomis, and M. Creel. 1991. Confidence intervals for evaluating benefits estimates from dichotomous choice contingent valuation studies. *Land Econ.* 67:64–73.

46. Loo, T. W., and D. M. Clarke. 1999. The transmembrane domains of the human multidrug resistance P-glycoprotein are sufficient to mediate drug binding and trafficking to the cell surface. *J. Biol. Chem.* 274:24759–24765.

47. Laskowski, R. A., M. W. MacArthur, D. S. Moss, and J. M. Thornton. 1993. Procheck-a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* 26:283–291.

48. Buchaklian, A. H., A. L. Funk, and C. S. Klug. 2004. Resting state conformation of the MsbA homodimer as studied by site-directed spin labeling. *Biochemistry.* 43:8600–8606.

49. Doerrler, W. T., H. S. Gibbons, and C. R. Raetz. 2004. MsbA-dependent translocation of lipids across the inner membrane of Escherichia coli. *J. Biol. Chem.* 279:45102–45109.

50. Loo, T. W., and D. M. Clarke. 1994. Reconstitution of drug-stimulated ATPase activity following co-expression of each half of human P-glycoprotein as separate polypeptides. *J. Biol. Chem.* 269:7750–7755.

51. Hrycyna, C. A., L. E. Airan, U. A. Germann, S. V. Ambudkar, I. Pastan, and M. M. Gottesman. 1998. Structural flexibility of the linker region of human P-glycoprotein permits ATP hydrolysis and drug transport. *Biochemistry.* 37:13660–13673.

52. Rosenberg, M. F., R. Callaghan, R. C. Ford, and C. F. Higgins. 1997. Structure of the multidrug resistance P-glycoprotein to 2.5 nm resolution determined by electron microscopy and image analysis. *J. Biol. Chem.* 272:10685–10694.

53. Rosenberg, M. F., G. Velarde, R. C. Ford, C. Martin, G. Berridge, I. D. Kerr, R. Callaghan, A. Schmidlin, C. Wooding, K. J. Linton, and C. F. Higgins. 2001. Repacking of the transmembrane domains of P-glycoprotein during the transport ATPase cycle. *EMBO J.* 20:5615–5625.

54. Sonveaux, N., A. B. Shapiro, E. Goormaghtigh, V. Ling, and J. M. Ruysschaert. 1996. Secondary and tertiary structure changes of reconstituted P-glycoprotein. A Fourier transform attenuated total reflection infrared spectroscopy analysis. *J. Biol. Chem.* 271:24617–24624.

55. Sonveaux, N., C. Vigano, A. B. Shapiro, V. Ling, and J. M. Ruysschaert. 1999. Ligand-mediated tertiary structure changes of reconstituted P-glycoprotein. A tryptophan fluorescence quenching analysis. *J. Biol. Chem.* 274:17649–17654.

56. Liu, R., A. Siemiarczuk, and F. J. Sharom. 2000. Intrinsic fluorescence of the P-glycoprotein multidrug transporter: sensitivity of tryptophan residues to binding of drugs and nucleotides. *Biochemistry.* 39:14927–14938.

57. Lee, J. Y., I. L. Urbatsch, A. E. Senior, and S. Wilkens. 2002. Projection structure of P-glycoprotein by electron microscopy. Evidence for a closed conformation of the nucleotide binding domains. *J. Biol. Chem.* 277:40125–40131.

58. Rosenberg, M. F., R. Callaghan, S. Modok, C. F. Higgins, and R. C. Ford. 2005. Three-dimensional structure of P-glycoprotein: the transmembrane regions adopt an asymmetric configuration in the nucleotide-bound state. *J. Biol. Chem.* 280:2857–2862.

59. Loo, T. W., M. C. Bartlett, and D. M. Clarke. 2004. Disulfide cross-linking analysis shows that transmembrane segments 5 and 8 of human P-glycoprotein are close together on the cytoplasmic side of the membrane. *J. Biol. Chem.* 279:7692–7697.

60. Ambudkar, S. V., S. Dey, C. A. Hrycyna, M. Ramachandra, I. Pastan, and M. M. Gottesman. 1999. Biochemical, cellular, and pharmacological aspects of the multidrug transporter. *Annu. Rev. Pharmacol. Toxicol.* 39:361–398.

61. Ambudkar, S. V., C. Kimchi-Sarfaty, Z. E. Sauna, and M. M. Gottesman. 2003. P-glycoprotein: from genomics to mechanism. *Oncogene.* 22:7468–7485.

62. Wu, Q., P. Y. Bounaud, S. D. Kuduk, C. P. Yang, I. Ojima, S. B. Horwitz, and G. A. Orr. 1998. Identification of the domains of photoincorporation of the 3′- and 7-benzophenone analogues of taxol in the carboxyl-terminal half of murine mdr1b P-glycoprotein. *Biochemistry.* 37:11272–11279.

63. Demeule, M., A. Laplante, G. F. Murphy, R. M. Wenger, and R. Beliveau. 1998. Identification of the cyclosporin-binding site in P-glycoprotein. *Biochemistry.* 37:18110–18118.

64. Safa, A. R. 2004. Identification and characterization of the binding sites of P-glycoprotein for multidrug resistance-related drugs and modulators. *Curr. Med. Chem. Anticancer Agents.* 4:1–17.

65. Pawagi, A. B., J. Wang, M. Silverman, R. A. Reithmeier, and C. M. Deber. 1994. Transmembrane aromatic amino acid distribution in P-glycoprotein. A functional role in broad substrate specificity. *J. Mol. Biol.* 235:554–564.

66. Abele, R., and R. Tampe. 1999. Function of the transport complex TAP in cellular immune recognition. *Biochim. Biophys. Acta.* 1461:405–419.

Article 4

Becker E, Meyer V, Madaoui H, Guerois R. (2006) Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response. *Bioinformatics*;22(11):1289-92.

BIOINFORMATICS DISCOVERY NOTE Vol. 22 no. 11 2006, pages 1289–1292
doi:10.1093/bioinformatics/btl075

Structural bioinformatics

Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response

Emmanuelle Becker[1,†], Vincent Meyer[2,†], Hocine Madaoui[1] and Raphaël Guerois[1,*]

[1]Service de Biophysique des Fonctions Membranaires, URA CNRS 2096, Département de Biologie Joliot-Curie and [2]Département d'Etude et d'Ingénierie des Protéines, CEA Saclay, 91191 Gif-Sur-Yvette, Cedex, France

Received on January 30, 2006; revised and accepted on February 27, 2006
Advance Access publication March 7, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Human Nbs1 and its homolog Xrs2 in *Saccharomyces cerevisiae* are part of the conserved MRN complex (MRX in yeast) which plays a crucial role in maintaining genomic stability. NBS1 corresponds to the gene mutated in the Nijmegen breakage syndrome (NBS) known as a radiation hyper-sensitive disease. Despite the conservation and the importance of the MRN complex, the high sequence divergence between Nbs1 and Xrs2 precluded the identification of common domains downstream of the N-terminal Fork-Head Associated (FHA) domain.

Results: Using HMM–HMM profile comparisons and structure modelling, we assessed the existence of a tandem BRCT in both Nbs1 and Xrs2 after the FHA. The structure-based conservation analysis of the tandem BRCT in Nbs1 supports its function as a phosphoserine binding domain. Remarkably, the 5 bp deletion observed in 95% of NBS patients cleaves the tandem at the linker region while preserving the structural integrity of each BRCT domain in the resulting truncated gene products.

Contact: guerois@cea.fr

Supplementary information: http://www-spider.cea.fr/Groups/si6661/view.html

1 INTRODUCTION

Nbs1 in human (or Xrs2 in yeast) is an essential component of the so-called MRN complex associating Mre11, Rad50 and Nbs1 (Petrini and Stracker, 2003; van den Bosch *et al*., 2003) and plays a crucial role in DNA repair pathways (Kobayashi *et al*., 2004). The human Nbs1 protein is a 754 amino acid long protein composed of several functional domains identified from sequence analysis and biochemical experiments (Fig. 2A). At the N-terminus, a Fork-Head Associated (FHA) domain (Durocher and Jackson, 2002) followed by a single BRCA1 C-terminal (BRCT) domain (Bork *et al*., 1997; Callebaut and Mornon, 1997) can be detected from sequence to profile searches. The C-terminus of Nbs1 contains a Mre11 binding region (Desai-Mehta *et al*., 2001) and an ATM recruitment motif (Falck *et al*., 2005). In Xrs2, the *Saccharomyces cerevisiae* functional homolog of Nbs1, the FHA domain together with the Tel1 (ATM homologue) and Mre11 binding regions are conserved but the existence of a BRCT domain was never detected from sequence analysis. As a matter of fact, the sequences of Xrs2 and Nbs1 are highly divergent in the 250 amino acids following the FHA domain (10% sequence identity). Using a specific strategy, new sequences of Xrs2 homologs not present in databases such as GenBank or EMBL could be retrieved and aligned to human sequences. From the resulting multiple sequence alignment, we show that in fact two BRCT domains are present in both human Nbs1 and yeast Xrs2 right behind the FHA domain.

Tandem BRCT have been recently recognized as major mediators of phosphorylation-dependent protein–protein interactions in processes related to cell-cycle checkpoint and DNA repair functions (Glover *et al*., 2004). The ability of the tandem BRCT of Nbs1 to bind phospho-peptides was never probed before since the existence of the second BRCT was not suspected. The model-based analysis of the tandem BRCT of Nbs1 strongly suggests that it is a phosphoserine binding module. The 5 bp deletion observed in 95% of NBS patients splits up the tandem at position 218. Remarkably, this mutation preserves the structural integrity of the second BRCT at plus or minus one residue. Altogether, our findings suggest that the NBS disease could be partly linked to a disruption of the interaction properties of the tandem BRCT: cleavage of the tandem BRCT may alter the selectivity of target recognition by Nbs1 and hence affect the signaling network required for efficient DNA damage responses.

2 METHODS

An initial profile containing close homologs of Nbs1 was built from searches of the non-redundant database using PSI-BLAST (Altschul *et al*., 1997) on the MPI server (Soding *et al*., 2005). For Xrs2, the initial profile gathered three sequences retrieved from tblastn searches on the *Saccharomyces* comparative genomic database (Kellis *et al*., 2003). The profiles were enriched by aligning profiles of more divergent sequences using the profile–profile alignment method HHalign (Soding, 2005).

Iteratively, the profile–profile alignment procedure led to a global multiple sequence alignment gathering 25 sequences from human Nbs1 to *S.cerevisiae* Xrs2 (see Supplementary information). The profile consisting of 25 sequences was compared against a database of profiles built from the PDB using the HHpred server (Soding *et al*., 2005). Three structures of

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

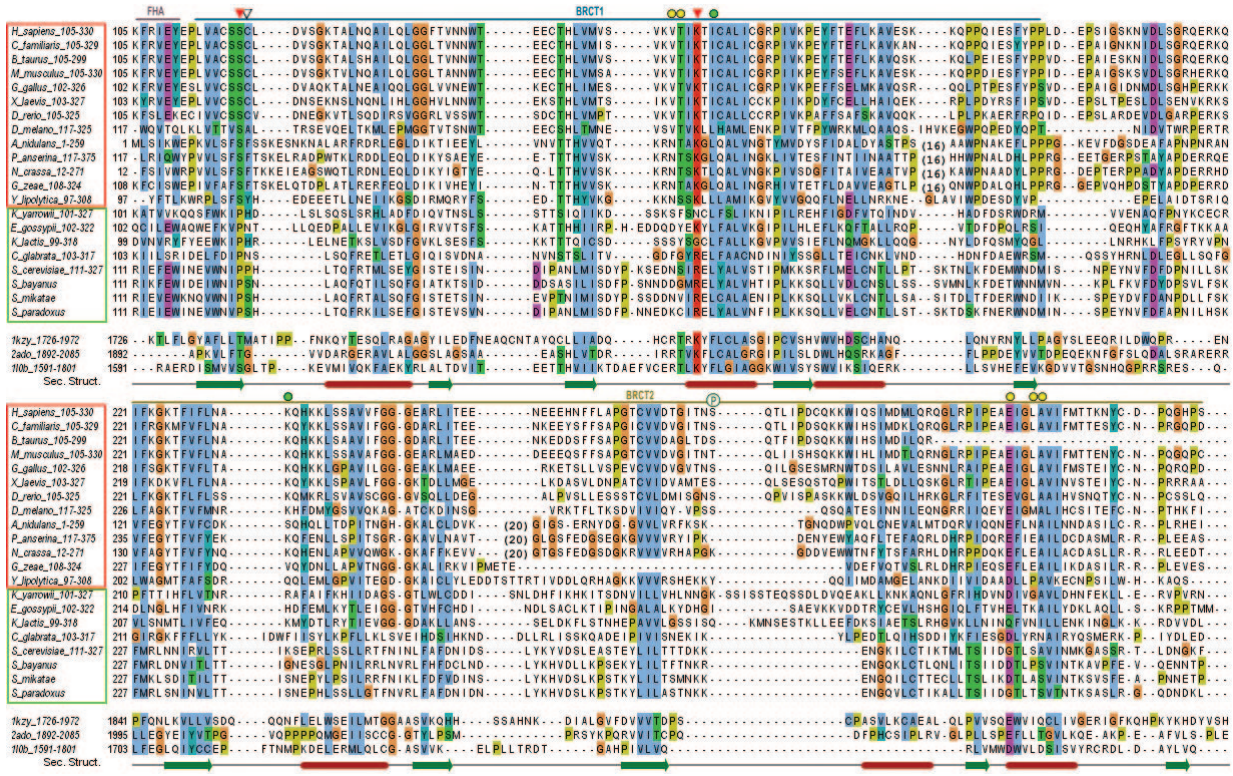


Fig. 1. Multiple sequence alignment of the tandem BRCT in homologs of Nbs1 and Xrs2 aligned with three structures of tandem BRCT (1kzy: 53BP1, 2ado: MDC1, 1l0b: BRCA1) represented with Jalview (Clamp *et al*., 2004). Helices and strands are noted by red sticks and green arrows below. Domains boundaries are shown by an horizontal line above. Red and green boxes group the sequences names with respect to their patterns at the pSer binding positions. Positions contacting the pSer residue with sidechain or backbone in tandem BRCT complex are indicated by red and white triangles, respectively. Positions contacting the pSer+3 positions are shown by green circles. Other positions found in direct contact with the phosphopeptide atoms are shown by yellow circles.

tandem BRCT were detected with significant scores and confidence levels >96% (PDB codes 1l0b, 2ado, 1kzy). Over the major length of the profile, the built alignment was consistent with the structural alignment of the templates. Yet, significant divergence could be observed at the N-terminus (first strand) and C-terminus (last $\alpha/\beta/\alpha$ motif). At the N-terminus, only the alignments with 1l0b and 2ado were compatible with the presence of an upstream FHA domain. At the C-terminus, only the alignment with the 1kzy template suggested the existence of a long insertion in the $\beta 3/\alpha 2$ loop of the second BRCT, consistent with the conservation profile in the whole Nbs1 family. A global sequence to structure alignment between the 25 sequences and the structural alignment of the three templates (1l0b, 2ado and 1kzy) was created based on these features.

Models were generated for both human Nbs1 and *S.cerevisiae* Xrs2 with Modeller 8v2 (Sali and Blundell, 1993) using the three structures 1kzy, 2ado and 1l0b as templates (max. Seq. ID: 13.2%). The quality of the models was assessed using Verify3D (Luthy *et al*., 1992), Prosa2003 (Sippl, 1993), ProQ and MaxSub (Wallner and Elofsson, 2003). The profile–profile alignment between the tandem BRCT of the Nbs1/Xrs2 family and that of the structural alignment of the three templates was iteratively refined in order to reduce the alignment errors pinpointed by the four evaluation scores.

To further assess the physical relevance of the model built for the tandem BRCT of Nbs1, a 5 ns molecular dynamic simulation was performed at 300 K in explicit solvent using GROMACS 3.2 (Van Der Spoel *et al*., 2005) (see Supplementary information for details). Conservation analyses were carried out using the Rate4site algorithm (Pupko *et al*., 2002). Possible arrangements of the FHA domain with respect to the tandem BRCT were explored using the HADDOCK program (Dominguez *et al*., 2003) by docking models of the FHA domain onto models of the tandem BRCT while constraining the distance between their C- and N-termini in respect of the Nbs1 sequence.

3 RESULTS

Models of the Nbs1 and Xrs2 tandem BRCT were built from the multiple sequence alignment in Figure 1 and assessed using standard evaluation tools. The scores of Nbs1 model are (Prosa2003: −1.92), (Verify3D: 0.395), (ProQ : 3.51) and (MaxSub : 0.348) and those of Xrs2 (Prosa2003: −1.16), (Verify3D: 0.332), (ProQ : 3.75) and (MaxSub : 0.338). The absence of residues with Verify3D scores below 0.1 together with ProQ and MaxSub scores significantly above 1.5 and 0.1, respectively, ensures the absence of major issues in the models of both tandem BRCT (Wallner and Elofsson, 2003). The physical quality of the Nbs1 model was further assessed by running a 5 ns simulation of molecular dynamics in an explicit solvent. The Cα rmsd stabilizes around 4 Å from the initial model structure (3.2 Å excluding the long loops 201–216 and 273–291) and secondary structures are overall preserved after a 5 ns of simulation as illustrated in Figure 2B (see also Supplementary information).

3.1 Functional insights from the tandem BRCT model

3.1.1 Clues for phosphoserine binding in Nbs1 So far, the tandem BRCT repeats of MDC1, PTIP, BARD1, 53BP1, RAD4, Ect2, TOPBP1, DNA ligase IV, *S.pombe* Crb2 and *S.cerevisiae* Rad9 have been shown to have phospho-serine (pSer) binding properties *in vitro* (Manke *et al*., 2003; Yu *et al*., 2003). The consensus signature for the pSer binding property was described as [S/T-G] in $\beta 1/\alpha 1$ loop

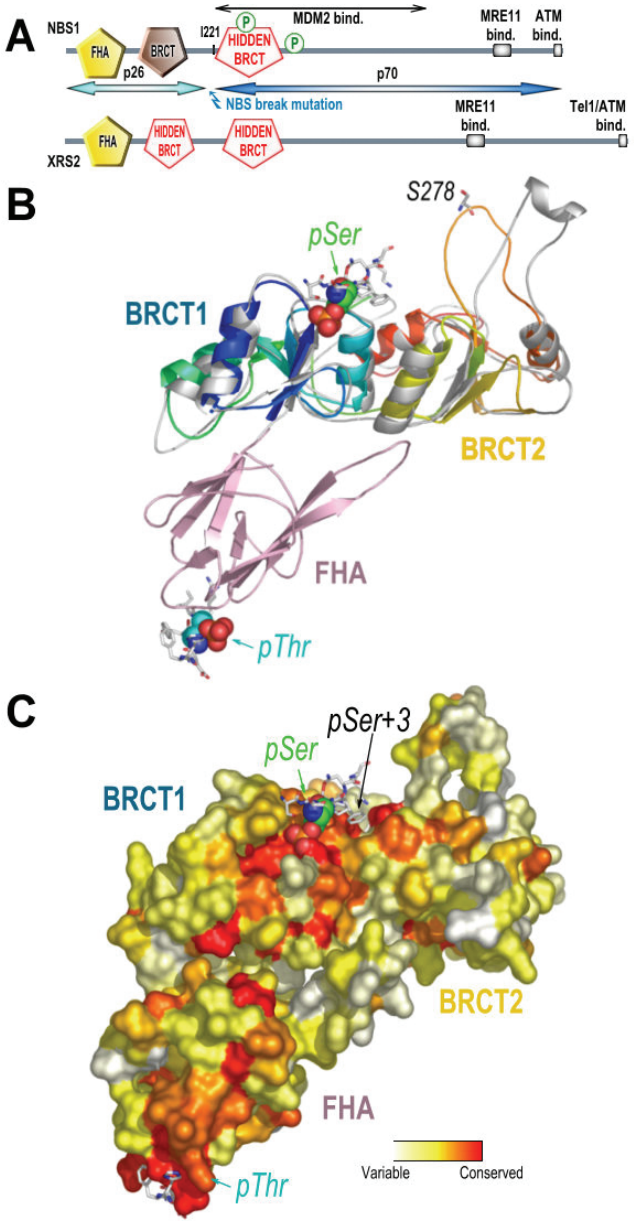


Fig. 2. (**A**) Domain organization of Nbs1 and Xrs2 with newly identified BRCT shown in red. (**B**) Ribbon representation of the model of the tandem BRCT before (rainbow colors) and after a 5 ns simulation of molecular dynamics (gray). In pink, a model of the FHA domain in a putative orientation with respect to the tandem. Phospho-peptides as found in known complexes of FHA and tandem BRCT with their ligands are shown as sticks. Phospho-Ser and -Thr residues are shown as spheres. (**C**) Surface projection of the evolutionary rates as calculated by the rate4site algorithm (Pupko *et al*., 2002). Colors from red to white report from the most conserved to the most variable positions. Drawn with pymol (DeLano, 2002).

and [S/T-X-K] in α2 helix of the first BRCT (Glover *et al*., 2004). The Gly in β1/α1 loop is quite versatile probably because only the backbone atoms at that position interacts with the pSer residue [53BP1 (1kzy in Fig. 1) has a Met and binds pSer *in vitro*]. In the tandem BRCT of Nbs1, the most conserved region localizes in the sites shown to bind the pSer residue in the structures of the tandem BRCT complexes (Stucki *et al*., 2005). The positions that directly contact the pSer with their sidechain or backbone are indicated by red and white triangles, respectively (Fig. 1). In sequences from *Homo sapiens* Nbs1 to *Yarrowia lipolytica* fungus (red box, Fig. 1), the consensus motif [S-C/F] in β1/α1 loop and [T/S-X-K] in α2 is strictly conserved supporting the function of this module as a pSer binding domain. In species ranging from *Kluyveromyces yarrowii* to *Solenodon paradoxus* (green box, Fig. 1), including Xrs2, positions binding pSer with their sidechains (red triangles) are conserved but do not match the consensus phospho-binding signature (Glover *et al*., 2004). The corresponding motif in *S.cerevisiae* Xrs2 is [P-P] in β1/α1 loop and [S-X-R] in α2 helix. Yet, several clues support that the tandem BRCT of Xrs2 might still be a pSer-binding module: (1) the BRCT domain of the ligase III shown to bind pSer peptides *in vitro* contains a Pro instead of a Ser in β1/α1 loop, as in Xrs2, (2) an Arg in α2 instead of a Lys is found in the tandem BRCT of ligase IV, also shown to be a pSer-binding module *in vitro* (Yu *et al*., 2003).

The groove at the interface between the BRCT domains is involved in the specific recognition of the residues flanking the pSer amino acid and is significantly conserved in the Nbs1 family (Fig. 2C). In structures of tandem BRCT/phosphopeptide complexes, the pSer+3 position was shown to hold much of the binding selectivity (Glover *et al*., 2004). Positions whose sidechain were shown to directly contact the position pSer+3 are indicated by green circles in Figure 1. In contrast to known structures where hydrophobic residues are often found at those positions, a Lys is quite conserved in one position of the Nbs1 multiple alignment (K233 in *H.sapiens* Nbs1).

3.1.2 Location of the phosphorylated sites in Nbs1 In response to ionizing radiation, Nbs1 is phosphorylated at Ser278 and Ser343 by the ATM kinase, and this event is required for activation of the intra S phase checkpoint (Kobayashi *et al*., 2004). From the structural model, Ser278 is located in the long β3/α2 loop of the second BRCT (Fig. 2B) and Ser343 is found 13 residues after the last residue of the tandem BRCT. Interestingly, the flexible linkers surrounding Ser278 and Ser343 are not long enough to allow for an intramolecular recognition of the pSer by the tandem BRCT.

3.1.3 Tandem BRCT and disease related mutations Of the NBS patients, 95% carry a 5 bp deletion in exon 6 of the *NBS*1 gene, which results in the expression of two truncated proteins of 26 (p26) and 70 kDa (p70) (Fig. 2A). The mutation splits the tandem precisely in the linker between the two BRCT domains. P26 moiety includes the region 1–218 spanning the FHA and the integrality of the first BRCT domain. P70 corresponds to the C-terminal half of Nbs1 and is produced by an alternative initiation of translation upstream of the 5 bp deletion. After a 18 residue extension at the N-terminus, the sequence of p70 is identical to that of the wild-type Nbs1 from I221 to the end (Williams *et al*., 2002).

I221 sharply corresponds to the beginning of the second BRCT and is the first residue fully buried in its hydrophobic core. Several structures of well-folded single C-terminal BRCT domains isolated from a tandem support that each BRCT domain can adopt its structure independently (Gaiser *et al*., 2004; Zhang *et al*., 1998). Hence, despite the severe sequence variations induced by the mutation in the linker, elements crucial for the structural integrity of the second BRCT have been preserved. It suggests that the second BRCT may not only fold independently but also hold a function important for

viability in NBS patients. Regarding the first BRCT, it has been shown that the FHA/BRCT could bind *in vitro* the histone H2AX phosphorylated by ATM (Kobayashi *et al*., 2002). Phosphorylation of H2AX at Ser129 is among the first events of the repair of double strand breaks (Lowndes and Toh, 2005). Our data suggest that the p26 fragment (Fig. 2A) may still be able to bind pSer residues in NBS cells but with a loss of binding selectivity due to the truncation of the second BRCT. This novel hypothesis would be interesting to test in the light of the results obtained on animal models of the NBS pathology (Difilippantonio *et al*., 2005; Williams *et al*., 2002).

3.1.4 Nbs1 and Mdm2 interaction Mdm2 has been extensively studied as a negative regulator of p53 tumor suppressor (Vousden and Prives, 2005). Mdm2 overexpression was recently shown to inhibit the DNA repair function of the MRN complex and this effect required the binding of Mdm2 to Nbs1 (Alt *et al*., 2005). The region 198–314 of Mdm2 was shown to associate with the MRN complex through the central region of Nbs1 221–540. This region encompasses the newly identified second BRCT domain 221–330 but not the first one. Downstream of the second BRCT, the region 330–540 is predicted to be largely unfolded (see Supplementary information). We hypothesize that the second BRCT of Nbs1 by itself may be involved in the interaction with Mdm2.

3.2 Functional implications from the FHA-tandem BRCT structural model

A striking feature of the domain organization among all Nbs1 homologs is the absence of a linker between the FHA and the tandem BRCT modules. Despite the high versatility in position and length of the insertions inside the FHA or the BRCT and between the two BRCT, not even a single amino acid was ever added at the hinge between the two modules. A structural model of the ensemble composed by the FHA and the tandem BRCT domains was built to probe the potential organization of the modules (Fig. 2B). Owing to steric hindrance, the phospho-binding sites of both domains are constrained on opposite sides of the whole assembly and could hardly be closer than 45 Å (see Supplementary information). It excludes the possibility to bind simultaneously a pThr neighboring a pSer at <15 residues. The structural constraint between the domains may originate from a specific evolutionary constraint coupling both pThr and pSer binding functions. Interestingly, the FHA and the BRCT were shown to be both required for optimal chromatin association of the MRN complex (Kobayashi *et al*., 2002; Zhao *et al*., 2002). Moreover, a mutation disrupting the FHA pThr binding site revealed that this domain is involved in a signal amplification step crucial for DNA repair after low doses of irradiation (Difilippantonio *et al*., 2005). The coupling between pThr and pSer binding functions suggested from the model might as well contribute to this amplification process.

ACKNOWLEDGEMENTS

The authors are grateful to F. Ochsenbein, M.-C. Marsolier-Kergoat and S. Zinn-Justin for their useful comments about the manuscript. This work is partly funded by the ACI IMPBIO 2004. V.M. is supported by an AFM fellowship (Association Française contre les Myopathies). H.M. is supported by a DGA fellowship. Funding to pay the Open Access publication charges was provided by the CEA Saclay.

Conflict of Interest: none declared.

REFERENCES

Alt,J.R. *et al*. (2005) Mdm2 binds to Nbs1 at sites of DNA damage and regulates double strand break repair. *J. Biol. Chem*., **280**, 18771–18781.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bork,P. *et al*. (1997) A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J*., **11**, 68–76.

Callebaut,I. and Mornon,J.P. (1997) From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. *FEBS Lett*., **400**, 25–30.

Clamp,M. *et al*. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

DeLano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.

Desai-Mehta,A. *et al*. (2001) Distinct functional domains of nibrin mediate Mre11 binding, focus formation, and nuclear localization. *Mol. Cell. Biol*., **21**, 2184–2191.

Difilippantonio,S. *et al*. (2005) Role of Nbs1 in the activation of the Atm kinase revealed in humanized mouse models. *Nat. Cell Biol*., **7**, 675–685.

Dominguez,C. *et al*. (2003) HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc*., **125**, 1731–1737.

Durocher,D. and Jackson,S.P. (2002) The FHA domain. *FEBS Lett*., **513**, 58–66.

Falck,J. *et al*. (2005) Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. *Nature*, **434**, 605–611.

Gaiser,O.J. *et al*. (2004) Solution structure, backbone dynamics, and association behavior of the C-terminal BRCT domain from the breast cancer-associated protein BRCA1. *Biochemistry*, **43**, 15983–15995.

Glover,J.N. *et al*. (2004) Interactions between BRCT repeats and phosphoproteins: tangled up in two. *Trends Biochem. Sci*., **29**, 579–585.

Kellis,M. *et al*. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kobayashi,J. *et al*. (2002) NBS1 localizes to gamma-H2AX foci through interaction with the FHA/BRCT domain. *Curr. Biol*., **12**, 1846–1851.

Kobayashi,J. *et al*. (2004) NBS1 and its functional role in the DNA damage response. *DNA Repair (Amst)*, **3**, 855–861.

Lowndes,N.F. and Toh,G.W. (2005) DNA repair: the importance of phosphorylating histone H2AX. *Curr. Biol*., **15**, R99–R102.

Luthy,R. *et al*. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Manke,I.A. *et al*. (2003) BRCT repeats as phosphopeptide-binding modules involved in protein targeting. *Science*, **302**, 636–639.

Petrini,J.H. and Stracker,T.H. (2003) The cellular response to DNA double-strand breaks: defining the sensors and mediators. *Trends Cell Biol*., **13**, 458–462.

Pupko,T. *et al*. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** ((Suppl. 1)), S71–S77.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.

Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.

Soding,J. (2005) Protein homology detection by HMM–HMM comparison [Erratum, (2005), *Bioinformatics*, **21**, 2144.]. *Bioinformatics*, **21**, 951–960.

Soding,J. *et al*. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*., **33**, W244–W248.

Stucki,M. *et al*. (2005) MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. *Cell*, **123**, 1213–1226.

van den Bosch,M. *et al*. (2003) The MRN complex: coordinating and mediating the response to broken chromosomes. *EMBO Rep*., **4**, 844–849.

Van Der Spoel,D. *et al*. (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem*., **26**, 1701–1718.

Vousden,K.H. and Prives,C. (2005) P53 and prognosis: new insights and further complexity. *Cell*, **120**, 7–10.

Wallner,B. and Elofsson,A. (2003) Can correct protein models be identified? *Protein Sci*., **12**, 1073–1086.

Williams,B.R. *et al*. (2002) A murine model of Nijmegen breakage syndrome. *Curr. Biol*., **12**, 648–653.

Yu,X. *et al*. (2003) The BRCT domain is a phospho-protein binding domain. *Science*, **302**, 639–642.

Zhang,X. *et al*. (1998) Structure of an XRCC1 BRCT domain: a new protein–protein interaction module. *EMBO J*., **17**, 6404–6411.

Zhao,S. *et al*. (2002) Functional analysis of FHA and BRCT domains of NBS1 in chromatin association and DNA damage responses. *Nucleic Acids Res*., **30**, 4815–4822.

Article 5

Becker E, Cotillard A, Meyer V, Madaoui H, Guerois R. (2007) HMM-Kalign: a tool for generating sub-optimal HMM alignments. Bioinformatics. [Epub ahead of print] PMID: 17921492

Bioinformatics Advance Access published October 6, 2007

Application note

HMM-Kalign: a tool for generating sub-optimal HMM alignments

Emmanuelle Becker*, Aurélie Cotillard, Vincent Meyer, Hocine Madaoui and Raphaël Guérois*

CEA, iBiTecS, URA 2096, SB^{2}SM, Laboratoire de Biologie Structurale et Radiobiologie, Gif sur Yvette, F-91191 France.

Associate Editor: Prof. Anna Tramontano

ABSTRACT

Summary: Recent development of strategies using multiple sequence alignments (MSA) or profiles to detect remote homologies between proteins has led to a significant increase in the number of proteins whose structures can be generated by comparative modelling methods. However, prediction of the optimal alignment between these highly divergent homologous proteins remains a difficult issue. We present a tool based on a generalized Viterbi algorithm that generates optimal and sub-optimal alignments between a sequence and a HMM. The tool is implemented as a new function within the HMMER package called *hmmkalign*.

Availability:http://www-spider.cea.fr/Groups/hk3039/view.html

Contacts: raphael.guerois@cea.fr, emmanuelle.becker@cea.fr.

The present work aims at automatically exploring the alignment space in the neighborhood of the optimal sequence alignment (OSA) in order to find an alignment closer to the structural alignment than the OSA.

The sequence alignment space in the neighborhood of the OSA has been quite extensively explored in the context of pairwise sequence alignments. Waterman (Waterman, 1983) proposed an algorithm derived from the standard Sellers algorithm to determine all the pairwise alignments whose scores are within a range ε of the OSA's score. Later and still dealing with pairwise sequence alignments, Saqi and Sternberg (Saqi and Sternberg, 1991) proposed a heuristic known as the Iterative Elimination Method, based on the progressive perturbation of the distance matrix. Another method to generate alternative pairwise sequence alignments has been introduced by Zucker (Zuker, 1991).

With the rise of sequence-profile, sequence-HMM, and more recently profile-profile and HMM-HMM alignments, these algorithmic developments were less studied. However, although progress have been made especially for the detection of remote homology, the alignment of sequences sharing less than 25% of sequence identity is still problematic in the context of comparative modelling. Based on this observation, some articles (Chivian and Baker, 2006; Jaroszewski et al., 2002; John and Sali, 2003) reintroduced the idea of generating alternative alignments by using heuristics such as a parametric approach (Chivian and Baker, 2006) coupled with Saqi and Sternberg's Iterative Elimination Method (Jaroszewski et al., 2002), or a genetic algorithm (John and Sali, 2003).

In this work, we explore the possibility of generating alternative alignments in the context of alignments obtained using Hidden Markov Models, such as HMMER (Eddy, 1996) or SAM (Karplus et al., 2005). Instead of heuristics, HMM-Kalign generates the exact neighborhood of the OSA.

* To whom correspondence should be addressed.

The Viterbi algorithm is classically used to align a sequence s_{obs} to a profile HMM and consists in finding the sequence of states that maximizes the emission probability of s_{obs} (Viterbi, 1967). To generate alternative alignments in the neighborhood of the OSA, one solution is to use a generalized Viterbi algorithm that precisely determines the k-best sequences of states that maximize the emission of s_{obs}. This generalization of the Viterbi algorithm has been used in the field of speech recognition and elegant variants have been developed recently that fasten the process (Huang and Chiang, 2005). We implemented and included the generalized Viterbi algorithm in the program HMMER (Eddy, 1996).

1 GENERATING SUB-OPTIMAL ALIGNMENTS

To use the *hmmkalign* command, two files are required :

- <*MSA*>, that contains a multiple sequence alignment (derived for example from the alignment of structural templates);
- <*sequences*>, that contains two sequences in fasta format: (i) the sequence to be aligned, (ii) one sequence from the <MSA> file that may be used as a template to further build a model of the first sequence.

To build the HMM, it is possible to use the classical command :

$ *./hmmbuild <hmm file > <MSA>* (command 1)

although our results show that within highly divergent families, it is more effective to drive explicitly the HMM architecture with respect to the conservation of the secondary structures (details in supplementary data 1). This is possible via the command :

$ *./hmmbuild --hand <hmm file > <MSA>* (command 2)

where the <*MSA*> file contains an additional line with symbols '-' and 'x' encoding for the positions of insertions and match states, respectively. After building the HMM, the command to generate *k* alignments is :

$ *./hmmkalign k <hmm file> <sequences>*

The OSA classically generated with HMMER corresponds to the alignment with the best score (K=1) (*cf.* command 1).

Exploration can be targeted to specific regions. For a sequence $s_{obs}=s_1...s_T$ in which only the region $s_i...s_j$ is to be sampled, add a hybrid sequence in the <*MSA*> file, that contains the "anchors" $s_1...s_{i-1}$ and $s_{j+1}...s_T$ and insertions '-' symbols instead of $s_i...s_j$.

2 TESTING PROCEDURE

We studied 115 alignments from 22 highly divergent protein families, sharing on average less than 25% identity (supp. data 2). These alignments were extracted from the HOMSTRAD database (Stebbings and Mizuguchi, 2004). The following procedure was applied : (1) exclude the test sequence from the multiple structural alignment; (2) build two distinct HMMs with commands 1 and 2, (3) get 20 sub-optimal alignments of the excluded sequence

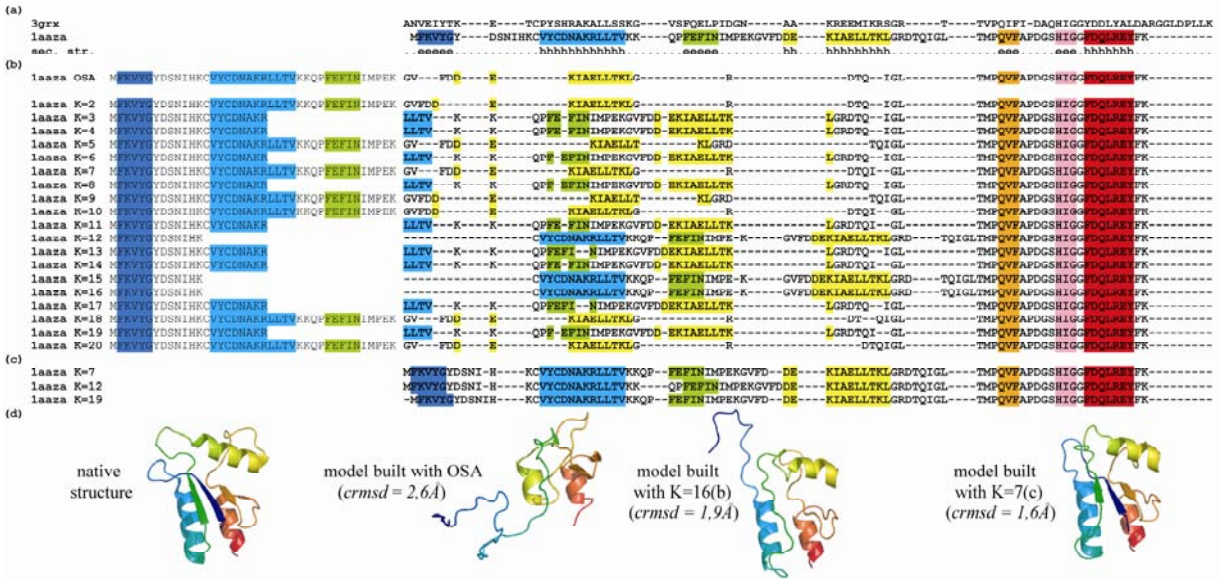


Figure 1 : The K best pairwise alignments generated with *hmmkalign* between 1aaza (test sequence) and 3grx (template) are shown with reference to the sequence of 3grx (supp. data 3). Location of secondary structures in the 1aaz structure is highlighted in color. **(a)** Structural alignment between 3grx and 1aaza and secondary structure assignment of 1aaza (HOMSTRAD annotations). **(b)** The 20 sub-optimal alignments generated by *hmmkalign* between 3grx and 1aaza. Unaligned amino acids are not bold. Alignment K=1 corresponds to the OSA. **(c)** Alignments generated when the HMM architecture explicitly restrained by using the command 2. The 20 sub-optimal alignments were computed but only 3 of them are presented (the 7th , 12th and 19th). **(d)** Native x-ray structure versus models produced by comparative modelling with Modeller6v2 using the OSA, and two sub-optimal alignments, K=16 (b) and K=7 (c).

on both HMMs, (4) evaluate with respect to the structural alignment.

3 RESULTS FOR THE 115 TEST CASES.

In 95 of the 115 test cases, at least one sub-optimal alignment had a Q_{mod} better than the OSA. For 26 of them, the Q_{mod} increased by more than 0.10 (supp. data 2). The alternative alignments generated by *hmmkalign* were also found to be of greater interest than the ones generated with heuristic approaches (supp. data 4). These results highlight that targeted sampling of the sequence alignment space in the neighborhood of the OSA by *hmmkalign* is efficient in generating optimized alignments and thereby better models.

4 EXAMPLE WITHIN THE THIOREDOXIN FAMILY.

The thioredoxin family contains small enzymes that are involved in redox reactions. Their sequences are on average 100 amino acids long and highly divergent (17% sequence identity on average), while their 3-layers sandwich fold is conserved. Aligning the sequence of the oxidized bacteriophage T4 glutaredoxin with the other members of the family is a difficult task. As a matter of fact, the OSA (figure 1b) is far from the structural alignment (ratio of correctly aligned positions Q_{mod} = 0.50).

First, we studied the 20 sub-optimal alignments produced when the HMM is built with command 1 (figure 1b). The alignment can be divided in two parts: the first 63 amino acids, whose positions are extremely variable, and the last 24 amino acids that are not shifted. Not surprisingly, the least varying positions along the sampled alignments correlate with the correctly aligned ones. Within the sub-optimal alignments, alignments K=12, K=15 and K=16, are substantially better than the OSA (Q_{mod} = 0.79).

We then studied the 20 sub-optimal alignments produced when HMM architecture is explicitly driven by secondary structure conservation (*cf.* command 2). Alignments with Q_{mod} reaching 0.89 were obtained (3 of them are shown in figure 1c).

Homology models of the oxidized bacteriophage T4 glutaredoxin were constructed with the OSA and all the sub-optimal alignments. As illustrated in figure 1d, the root mean square deviation between the native structure and the models are much smaller with models produced with the sub-optimal alignments (K=16 or K=7) than with models produced with the OSA.

ACKNOWLEDGEMENTS

This work is partly funded by the ACI IMPBIO SpIDER.

REFERENCES

Chivian, D. and Baker, D. (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res*, **34**, e112.

Eddy, S.R. (1996) Hidden Markov models. *Curr Opin Struct Biol*, **6**, 361-365.

Huang, L. and Chiang, D. (2005) Better k-best Parsing. *Proceedings of the 9th International Workshop on Parsing Technologies (IWPT)*, Vancouver, BC.

Jaroszewski, L., Li, W. and Godzik, A. (2002) In search for more accurate alignments in the twilight zone. *Protein Sci*, **11**, 1702-1713.

John, B. and Sali, A. (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res*,**31**,3982-3992.

Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M. and Hughey, R. (2005) SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins*, **61 Suppl 7**, 135-142.

Saqi, M.A. and Sternberg, M.J. (1991) A simple method to generate non-trivial alternate alignments of protein sequences. *J Mol Biol*, **219**, 727-732.

Stebbings, L.A. and Mizuguchi, K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res*,**32**,203-207.

Viterbi, A.J. (1967) Error bounds for convolutional codes. *IEEE Transactions on Information Theory*, **13**, 260–269.

Waterman, M.S. (1983) Sequence alignments in the neighborhood of the optimum. *Proc Natl Acad Sci U S A*, **80**, 3123-3124.

Zuker, M. (1991) Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J Mol Biol*, **221**, 403-420.

Article 6

Martin G, Heyd B, Sun Y, Kan E, Joly P, Ulmer J, Madaoui H, Combes O, Vita C, Barnett S, Srivastava I, Kessler P, Martin L. (2007) Cross-linking strategies to generate stable gp120-CD4 mimic complexes for immunogen design. *Journal Of Virology*. Submitted.

Bibliographie

Abhiman, S., and Sonnhammer, E. L. (2005). FunShift: a database of function shift analysis on protein subfamilies. Nucleic Acids Res *33*, D197-200.

Abhiman, S., and Sonnhammer, E. L. (2005). Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins *60*, 758-768.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Afonnikov, D. A., Oshchepkov, D. Y., and Kolchanov, N. A. (2001). Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics *17*, 1035-1046.

Agez, M., Chen, J., Guerois, R., van Heijenoort, C., Thuret, J. Y., Mann, C., and Ochsenbein, F. (2007). Structure of the histone chaperone ASF1 bound to the histone H3 C-terminal helix and functional insights. Structure *15*, 191-199.

Aloy, P., Ceulemans, H., Stark, A., and Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. J Mol Biol *332*, 989-998.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002). The Pfam protein families database. Nucleic Acids Res *30*, 276-280.

Becker, E., Meyer, V., Madaoui, H., and Guerois, R. (2006). Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response. Bioinformatics *22*, 1289-1292.

Berendsen, H. J. C., van der Spoel, D., and van Drunen, R. (1995). GROMACS: A message-passing parallel molecular dynamics implementation. Comp Phys Comm *91*, 43-56.

Berezovsky, I. N., Kilosanidze, G. T., Tumanyan, V. G., and Kisselev, L. L. (1999). Amino acid composition of protein termini are biased in different manners. Protein Eng *12*, 23-30.

Berger, J. M., Gamblin, S. J., Harrison, S. C., and Wang, J. C. (1996). Structure and mechanism of DNA topoisomerase II. Nature *379*, 225-232.

Bernauer, J., Aze, J., Janin, J., and Poupon, A. (2007). A new protein-protein docking scoring function based on interface residue properties. Bioinformatics *23*, 555-562.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol *112*, 535-542.

Bhattacharyya, R. P., Remenyi, A., Yeh, B. J., and Lim, W. A. (2006). Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits. Annu Rev Biochem.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.* (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453-1474.

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

Boter, M., Amigues, B., Peart, J., Breuer, C., Casais, C., Moore, G., Kleanthous, C., Ochsenbein, F., Shirasu, K., and R., G. (2007). HSP90, SGT1, and RAR1 form a ternary chaperone complex to regulate R-gene dependent plant immunity. Plant Cell *in press*.

Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O., and Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol *5*, R35.

Bradford, J. R., and Westhead, D. R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. Bioinformatics *21*, 1487-1494.

Byeon, I. J., Li, H., Song, H., Gronenborn, A. M., and Tsai, M. D. (2005). Sequential phosphorylation and multisite interactions characterize specific target recognition by the FHA domain of Ki67. Nat Struct Mol Biol *12*, 987-993.

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

Camacho, C. J., Gatchell, D. W., Kimura, S. R., and Vajda, S. (2000). Scoring docked conformations generated by rigid-body protein-protein docking. Proteins *40*, 525-537.

Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci *12*, 2001-2014.

Carugo, O., and Argos, P. (1997). Protein-protein crystal-packing contacts. Protein Sci *6*, 2261-2263.

Carvalho, A. L., Dias, F. M., Prates, J. A., Nagy, T., Gilbert, H. J., Davies, G. J., Ferreira, L. M., Romao, M. J., and Fontes, C. M. (2003). Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. Proc Natl Acad Sci U S A *100*, 13809-13814.

Chandonia, J. M., and Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. Science *311*, 347-351.

Chang, G., and Roth, C. B. (2001). Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. Science *293*, 1793-1800.

Chaudhury, S., Sircar, A., Sivasubramanian, A., Berrondo, M., and Gray, J. J. (2007). Incorporating biochemical information and backbone flexibility in RosettaDock for CAPRI rounds 6-12. Proteins.

Chelliah, V., Blundell, T. L., and Fernandez-Recio, J. (2006). Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. J Mol Biol *357*, 1669-1682.

Chen, R., Li, L., and Weng, Z. (2003). ZDOCK: an initial-stage protein-docking algorithm. Proteins *52*, 80-87.

Chen, R., Mintseris, J., Janin, J., and Weng, Z. (2003). A protein-protein docking benchmark. Proteins *52*, 88-91.

Chen, X. W., and Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. Bioinformatics *21*, 4394-4400.

Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat, R. J., and Stoddard, B. L. (2002). Design, activity, and structure of a highly specific artificial endonuclease. Mol Cell *10*, 895-905.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. Embo J *5*, 823-826.

Cole, S. P., and Deeley, R. G. (2006). Transport of glutathione and glutathione conjugates by MRP1. Trends Pharmacol Sci *27*, 438-446.

Craig, R. A., and Liao, L. (2007). Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. BMC Bioinformatics *8*, 6.

Dahiyat, B. I., and Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. Science *278*, 82-87.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci *23*, 324-328.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol *332*, 449-460.

Date, S. V., and Marcotte, E. M. (2003). Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. Nat Biotechnol *21*, 1055-1062.

de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. Science *307*, 724-727.

De, S., Krishnadev, O., Srinivasan, N., and Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. BMC Struct Biol *5*, 15.

de Vries, S. J., van Dijk, A. D., and Bonvin, A. M. (2006). WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins *63*, 479-489.

Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics *1*, 349-356.

Desai-Mehta, A., Cerosaletti, K. M., and Concannon, P. (2001). Distinct functional domains of nibrin mediate Mre11 binding, focus formation, and nuclear localization. Mol Cell Biol *21*, 2184-2191.

Devico, A., Silver, A., Thronton, A. M., Sarngadharan, M. G., and Pal, R. (1996). Covalently crosslinked complexes of human immunodeficiency virus type 1 (HIV-1) gp120 and CD4 receptor elicit a neutralizing immune response that includes antibodies selective for primary virus isolates. Virology *218*, 258-263.

Doerrler, W. T., Reedy, M. C., and Raetz, C. R. (2001). An Escherichia coli mutant defective in lipid export. J Biol Chem *276*, 11461-11464.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc *125*, 1731-1737.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., *et al.* (2003). PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics *4*, 11.

Duan, Y., Reddy, B. V., and Kaznessis, Y. N. (2005). Physicochemical and residue conservation calculations to improve the ranking of protein-protein docking solutions. Protein Sci *14*, 316-328.

Dueber, J. E., Yeh, B. J., Bhattacharyya, R. P., and Lim, W. A. (2004). Rewiring cell signaling: the logic and plasticity of eukaryotic protein circuitry. Curr Opin Struct Biol *14*, 690-699.

Dueber, J. E., Yeh, B. J., Chak, K., and Lim, W. A. (2003). Reprogramming control of an allosteric signaling switch through modular recombination. Science *301*, 1904-1908.

Dunbrack, R. L., Jr., and Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci *6*, 1661-1681.

Dutheil, J., Pupko, T., Jean-Marie, A., and Galtier, N. (2005). A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol *22*, 1919-1928.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

English, C. M., Adkins, M. W., Carson, J. J., Churchill, M. E., and Tyler, J. K. (2006). Structural basis for the histone chaperone activity of Asf1. Cell *127*, 495-508.

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

Falck, J., Coates, J., and Jackson, S. P. (2005). Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. Nature *434*, 605-611.

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

Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2002). Soft protein-protein docking in internal coordinates. Protein Sci *11*, 280-291.

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

Filikov, A. V., Hayes, R. J., Luo, P., Stark, D. M., Chan, C., Kundu, A., and Dahiyat, B. I. (2002). Computational stabilization of human growth hormone. Protein Sci *11*, 1452-1461.

Fouts, T., Godfrey, K., Bobb, K., Montefiori, D., Hanson, C. V., Kalyanaraman, V. S., DeVico, A., and Pal, R. (2002). Crosslinked HIV-1 envelope-CD4 receptor complexes elicit broadly cross-reactive neutralizing antibodies in rhesus macaques. Proc Natl Acad Sci U S A *99*, 11842-11847.

Fouts, T. R., Tuskan, R., Godfrey, K., Reitz, M., Hone, D., Lewis, G. K., and DeVico, A. L. (2000). Expression and characterization of a single-chain polypeptide analogue of the human immunodeficiency virus type 1 gp120-CD4 receptor complex. J Virol *74*, 11427-11436.

Fryxell, K. J. (1996). The coevolution of gene family trees. Trends Genet *12*, 364-369.

Gabb, H. A., Jackson, R. M., and Sternberg, M. J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol *272*, 106-120.

Gardiner, E. J., Willett, P., and Artymiuk, P. J. (2001). Protein docking using a genetic algorithm. Proteins *44*, 44-56.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141-147.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., *et al.* (2003). A protein interaction map of Drosophila melanogaster. Science *302*, 1727-1736.

Glaser, F., Steinberg, D. M., Vakser, I. A., and Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. Proteins *43*, 89-102.

Glover, J. N., Williams, R. S., and Lee, M. S. (2004). Interactions between BRCT repeats and phosphoproteins: tangled up in two. Trends Biochem Sci *29*, 579-585.

Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. Proteins *18*, 309-317.

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. J Mol Biol *299*, 283-293.

Goh, C. S., and Cohen, F. E. (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol *324*, 177-192.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol *331*, 281-299.

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

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol *320*, 369-387.

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

Halperin, I., Wolfson, H., and Nussinov, R. (2006). Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins *63*, 832-845.

Haubertin, D. Y., Madaoui, H., Sanson, A., Guerois, R., and Orlowski, S. (2006). Molecular dynamics simulations of E. coli MsbA transmembrane domain: formation of a semipore structure. Biophys J *91*, 2517-2531.

He, Y., D'Agostino, P., and Pinter, A. (2003). Analysis of the immunogenic properties of a single-chain polypeptide analogue of the HIV-1 gp120-CD4 complex in transgenic mice that produce human immunoglobulins. Vaccine *21*, 4421-4429.

Higgins, C. F., and Gottesman, M. M. (1992). Is the multidrug transporter a flippase? Trends Biochem Sci *17*, 18-21.

Hillier, B. J., Christopherson, K. S., Prehoda, K. E., Bredt, D. S., and Lim, W. A. (1999). Unexpected modes of PDZ domain scaffolding revealed by structure of nNOS-syntrophin complex. Science *284*, 812-815.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., *et al.* (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180-183.

Hofmann, K. (1999). The modular nature of apoptotic signaling proteins. Cell Mol Life Sci *55*, 1113-1128.

Hu, Z., Ma, B., Wolfson, H., and Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. Proteins *39*, 331-342.

Huse, M., and Kuriyan, J. (2002). The conformational plasticity of protein kinases. Cell *109*, 275-282.

Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics *5*, 299-314.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A *98*, 4569-4574.

Jackson, R. M., Gabb, H. A., and Sternberg, M. J. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol *276*, 265-285.

Jensen, L. J., Jensen, T. S., de Lichtenberg, U., Brunak, S., and Bork, P. (2006). Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature *443*, 594-597.

Jimenez, J. L. (2005). Does structural and chemical divergence play a role in precluding undesirable protein interactions? Proteins *59*, 757-764.

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

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A *89*, 2195-2199.

Kim, C. A., Gingery, M., Pilpa, R. M., and Bowie, J. U. (2002). The SAM domain of polyhomeotic forms a helical polymer. Nat Struct Biol *9*, 453-457.

Kim, W. K., Bolser, D. M., and Park, J. H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics *20*, 1138-1150.

Kim, W. K., and Ison, J. C. (2005). Survey of the geometric association of domain-domain interfaces. Proteins *61*, 1075-1088.

Knudsen, B., and Miyamoto, M. M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci U S A *98*, 14512-14517.

Kobayashi, J., Antoccia, A., Tauchi, H., Matsuura, S., and Komatsu, K. (2004). NBS1 and its functional role in the DNA damage response. DNA Repair (Amst) *3*, 855-861.

Korkegian, A., Black, M. E., Baker, D., and Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. Science *308*, 857-860.

Kortemme, T., and Baker, D. (2004). Computational design of protein-protein interactions. Curr Opin Chem Biol *8*, 91-97.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., *et al.* (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637-643.

Kuffner, J. J. (2004). Effective Sampling and Distance Metrics for 3D Rigid Body Path Planning, Paper presented at: Proc. IEEE Int'l Conf. on Robotics and Automation.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science *302*, 1364-1368.

Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y., and Baker, D. (2001). Conversion of monomeric protein L to an obligate dimer by computational protein design. Proc Natl Acad Sci U S A *98*, 10687-10691.

Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y., and Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. J Mol Biol *315*, 471-477.

Kumar, A., and Snyder, M. (2002). Protein complexes take the bait. Nature *415*, 123-124.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J., and Hendrickson, W. A. (1998). Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature *393*, 648-659.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol *307*, 1487-1502.

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

Lazaridis, T., and Karplus, M. (2000). Effective energy functions for protein structure prediction. Curr Opin Struct Biol *10*, 139-145.

Lee, J. H., and Paull, T. T. (2005). ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. Science *308*, 551-554.

Lee, Y. T., Jacob, J., Michowski, W., Nowotny, M., Kuznicki, J., and Chazin, W. J. (2004). Human Sgt1 binds HSP90 through the CHORD-Sgt1 domain and not the tetratricopeptide repeat domain. J Biol Chem *279*, 16511-16517.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., *et al.* (2004). A map of the interactome network of the metazoan C. elegans. Science *303*, 540-543.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol *257*, 342-358.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J. F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., *et al.* (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell *125*, 801-814.

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

Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E., and Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol *316*, 139-154.

Madaoui, H., Becker, E., and Guerois, R. (2006). Sequence search methods and scoring functions for the design of protein structures. Methods Mol Biol *340*, 183-206.

Madaoui, H., and Guerois, R. (2007). Surface Complementarity Trace in Complex History (SCOTCH), a robust method to discriminate native protein complex interfaces. Proc Natl Acad Sci U S A *in revision*.

Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Ten Eyck, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. Protein Eng *14*, 105-113.

Manke, I. A., Lowery, D. M., Nguyen, A., and Yaffe, M. B. (2003). BRCT repeats as phosphopeptide-binding modules involved in protein targeting. Science *302*, 636-639.

Mann, M., Hendrickson, R. C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. Annu Rev Biochem *70*, 437-473.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science *285*, 751-753.

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

Martin, G., Heyd, B., Sun, Y., Kan, E., Joly, P., Ulmer, J., Madaoui, H., Combes, O., Vita, C., Barnett, S., *et al.* (2007). Cross-linking strategies to generate stable gp120-CD4 mimic complexes for immunogen design. *submitted*.

Martinon, F., Burns, K., and Tschopp, J. (2002). The inflammasome: a molecular platform triggering activation of inflammatory caspases and processing of proIL-beta. Mol Cell *10*, 417-426.

Mathivanan, S., Periaswamy, B., Gandhi, T. K., Kandasamy, K., Suresh, S., Mohmood, R., Ramachandra, Y. L., and Pandey, A. (2006). An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics *7 Suppl 5*, S19.

Matsuoka, S., Ballif, B. A., Smogorzewska, A., McDonald, E. R., 3rd, Hurov, K. E., Luo, J., Bakalarski, C. E., Zhao, Z., Solimini, N., Lerenthal, Y., *et al.* (2007). ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. Science *316*, 1160-1166.

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res *11*, 2120-2126.

Mayrose, I., Graur, D., Ben-Tal, N., and Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol *21*, 1781-1791.

McCoy, A. J., Chandana Epa, V., and Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. J Mol Biol *268*, 570-584.

Mendez, R., Leplae, R., De Maria, L., and Wodak, S. J. (2003). Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins *52*, 51-67.

Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A *102*, 10930-10935.

Moont, G., Gabb, H. A., and Sternberg, M. J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins *35*, 364-373.

Moreau, C., Jacquet, H., Prost, A. L., D'Hahan, N., and Vivaudou, M. (2000). The molecular basis of the specificity of action of K(ATP) channel openers. Embo J *19*, 6644-6651.

Mousson, F., Lautrette, A., Thuret, J. Y., Agez, M., Courbeyrette, R., Amigues, B., Becker, E., Neumann, J. M., Guerois, R., Mann, C., and Ochsenbein, F. (2005). Structural basis for the interaction of Asf1 with histone H3 and its functional implications. Proc Natl Acad Sci U S A *102*, 5975-5980.

Munoz, V., and Serrano, L. (1995). Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides. J Mol Biol *245*, 275-296.

Murphy, J., Gatchell, D. W., Prasad, J. C., and Vajda, S. (2003). Combination of scoring functions improves discrimination in protein-protein docking. Proteins *53*, 840-854.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.

Natsume, R., Eitoku, M., Akai, Y., Sano, N., Horikoshi, M., and Senda, T. (2007). Structure and function of the histone chaperone CIA/ASF1 complexed with histones H3 and H4. Nature *446*, 338-341.

Nauli, S., Kuhlman, B., and Baker, D. (2001). Computer-based redesign of a protein folding pathway. Nat Struct Biol *8*, 602-605.

Neher, E. (1994). How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci U S A *91*, 98-102.

Neuvirth, H., Raz, R., and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol *338*, 181-199.

Nilges, M., and O'Donoghue, S. I. (1998). Ambiguous NOEs and automated NOE assignment. Progress in Nuclear Magnetic Resonance Spectroscopy *32*, 107-139.

Nooren, I. M., and Thornton, J. M. (2003). Diversity of protein-protein interactions. Embo J *22*, 3486-3492.

Ofran, Y., and Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. FEBS Lett *544*, 236-239.

Palma, P. N., Krippahl, L., Wampler, J. E., and Moura, J. J. (2000). BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins *39*, 372-384.

Pawson, T. (2007). Dynamic control of signaling by modular adaptor proteins. Curr Opin Cell Biol *19*, 112-116.

Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science *300*, 445-452.

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

Pazos, F., Ranea, J. A., Juan, D., and Sternberg, M. J. (2005). Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol *352*, 1002-1015.

Pazos, F., and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. Protein Eng *14*, 609-614.

Pazos, F., and Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins *47*, 219-227.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A *96*, 4285-4288.

Petrini, J. H. (2007). Cell signaling. A touching response to damage. Science *316*, 1138-1139.

Petrini, J. H., and Stracker, T. H. (2003). The cellular response to DNA double-strand breaks: defining the sensors and mediators. Trends Cell Biol *13*, 458-462.

Phizicky, E. M., and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. Microbiol Rev *59*, 94-123.

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics *18 Suppl 1*, S71-77.

Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins *63*, 490-500.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., *et al.* (2001). The protein-protein interaction map of Helicobacter pylori. Nature *409*, 211-215.

Reina, J., Lacroix, E., Hobson, S. D., Fernandez-Ballester, G., Rybin, V., Schwab, M. S., Serrano, L., and Gonzalez, C. (2002). Computer-aided design of a PDZ domain to recognize new target sequences. Nat Struct Biol *9*, 621-627.

Reuter, G., Janvilisri, T., Venter, H., Shahi, S., Balakrishnan, L., and van Veen, H. W. (2003). The ATP binding cassette multidrug transporter LmrA and lipid transporter MsbA have overlapping substrate specificities. J Biol Chem *278*, 35193-35198.

Ritchie, D. W., and Kemp, G. J. (2000). Protein docking using spherical polar Fourier correlations. Proteins *39*, 178-194.

Romsicki, Y., and Sharom, F. J. (2001). Phospholipid flippase activity of the reconstituted P-glycoprotein multidrug transporter. Biochemistry *40*, 6937-6947.

Rouse, J., and Jackson, S. P. (2002). Interfaces between the detection, signaling, and repair of DNA damage. Science *297*, 547-551.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., *et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., and Ranganathan, R. (2005). Natural-like function in artificial WW domains. Nature *437*, 579-583.

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779-815.

Salwinski, L., and Eisenberg, D. (2003). Computational methods of analysis of protein-protein interactions. Curr Opin Struct Biol *13*, 377-382.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A *95*, 5857-5864.

Schymkowitz, J. W., Rousseau, F., Martins, I. C., Ferkinghoff-Borg, J., Stricher, F., and Serrano, L. (2005). Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc Natl Acad Sci U S A *102*, 10147-10152.

Seet, B. T., Dikic, I., Zhou, M. M., and Pawson, T. (2006). Reading protein modifications with interaction domains. Nat Rev Mol Cell Biol *7*, 473-483.

Sharan, R., and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. Nat Biotechnol *24*, 427-433.

Sheinerman, F. B., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol *10*, 153-159.

Shifman, J. M., and Mayo, S. L. (2002). Modulating calmodulin binding specificity through computational protein design. J Mol Biol *323*, 417-423.

Shilling, R. A., Venter, H., Velamakanni, S., Bapna, A., Woebking, B., Shahi, S., and van Veen, H. W. (2006). New light on multidrug binding by an ATP-binding-cassette transporter. Trends Pharmacol Sci *27*, 195-203.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins *17*, 355-362.

Smith, G. R., Sternberg, M. J., and Bates, P. A. (2005). 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.

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., and Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. Nature *437*, 512-518.

Songyang, Z., Fanning, A. S., Fu, C., Xu, J., Marfatia, S. M., Chishti, A. H., Crompton, A., Chan, A. C., Anderson, J. M., and Cantley, L. C. (1997). Recognition of unique carboxyl-terminal motifs by distinct PDZ domains. Science *275*, 73-77.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., *et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Stricher, F., Martin, L., and Vita, C. (2006). Design of miniproteins by the transfer of active sites onto small-size scaffolds. Methods Mol Biol *340*, 113-149.

Takahashi, A., Casais, C., Ichimura, K., and Shirasu, K. (2003). HSP90 interacts with RAR1 and SGT1 and is essential for RPS2-mediated disease resistance in Arabidopsis. Proc Natl Acad Sci U S A *100*, 11777-11782.

Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. J Mol Evol *44*, 66-73.

Tarjan, R. (1972). Depth first search and linear graph algorithms. SIAM J Comput *1*, 146-160.

Taylor, J. S., and Burnett, R. M. (2000). DARWIN: a program for docking flexible molecules. Proteins *41*, 173-191.

Teichmann, S. A., and Babu, M. M. (2002). Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends Biotechnol *20*, 407-410; discussion 410.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res *22*, 4673-4680.

Tress, M., de Juan, D., Grana, O., Gomez, M. J., Gomez-Puertas, P., Gonzalez, J. M., Lopez, G., and Valencia, A. (2005). Scoring docking models with evolutionary information. Proteins *60*, 275-280.

Tschochner, H., and Hurt, E. (2003). Pre-ribosomes on the road from the nucleolus to the cytoplasm. Trends Cell Biol *13*, 255-263.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Vakser, I. A. (1995). Protein docking for low-resolution structures. Protein Eng *8*, 371-377.

van den Bosch, M., Bree, R. T., and Lowndes, N. F. (2003). The MRN complex: coordinating and mediating the response to broken chromosomes. EMBO Rep *4*, 844-849.

van der Burgt, I., Chrzanowska, K. H., Smeets, D., and Weemaes, C. (1996). Nijmegen breakage syndrome. J Med Genet *33*, 153-156.

van der Sloot, A. M., Mullally, M. M., Fernandez-Ballester, G., Serrano, L., and Quax, W. J. (2004). Stabilization of TRAIL, an all-beta-sheet multimeric protein, using computational redesign. Protein Eng Des Sel *17*, 673-680.

van Dijk, A. D., Boelens, R., and Bonvin, A. M. (2005). Data-driven docking for the study of biomolecular complexes. Febs J *272*, 293-312.

Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., *et al.* (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. Science *303*, 844-848.

Vidal, M., and Legrain, P. (1999). Yeast forward and reverse 'n'-hybrid systems. Nucleic Acids Res *27*, 919-929.

Vita, C., Drakopoulou, E., Vizzavona, J., Rochette, S., Martin, L., Menez, A., Roumestand, C., Yang, Y. S., Ylisastigui, L., Benjouad, A., and Gluckman, J. C. (1999). Rational engineering of a miniprotein that reproduces the core of the CD4 site interacting with HIV-1 envelope glycoprotein. Proc Natl Acad Sci U S A *96*, 13091-13096.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J Mol Graph *8*, 52-56, 29.

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., and Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. Science *287*, 116-122.

Wang, B., Chen, P., Huang, D. S., Li, J. J., Lok, T. M., and Lyu, M. R. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate. FEBS Lett *580*, 380-384.

Wang, C., Bradley, P., and Baker, D. (2007). Protein-protein docking with backbone flexibility. J Mol Biol *373*, 503-519.

Wang, C., Schueler-Furman, O., and Baker, D. (2005). Improved side-chain modeling for protein-protein docking. Protein Sci *14*, 1328-1339.

Woebking, B., Reuter, G., Shilling, R. A., Velamakanni, S., Shahi, S., Venter, H., Balakrishnan, L., and van Veen, H. W. (2005). Drug-lipid A interactions on the Escherichia coli ABC transporter MsbA. J Bacteriol *187*, 6363-6369.

Wollacott, A. M., Zanghellini, A., Murphy, P., and Baker, D. (2007). Prediction of structures of multidomain proteins from structures of the individual domains. Protein Sci *16*, 165-175.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M., Kavraki, L., and Lichtarge, O. (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol *326*, 255-261.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome Res *14*, 1107-1118.

Yu, X., Chini, C. C., He, M., Mer, G., and Chen, J. (2003). The BRCT domain is a phospho-protein binding domain. Science *302*, 639-642.

Zhang, Y., Zhou, J., and Lim, C. U. (2006). The role of NBS1 in DNA double strand break repair, telomere stability, and cell cycle checkpoint control. Cell Res *16*, 45-54.

Zhong, H., and Carlson, H. A. (2005). Computational studies and peptidomimetic design for the human p53-MDM2 complex. Proteins *58*, 222-234.