

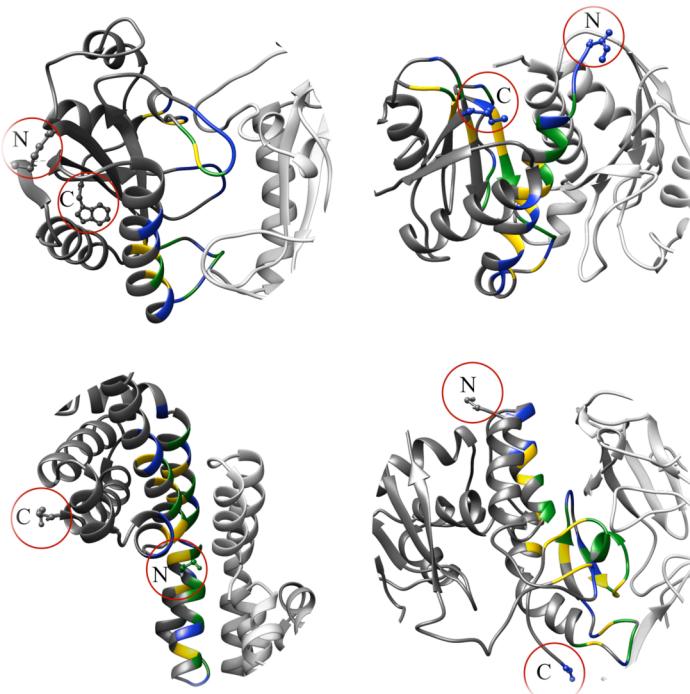
Master 2 Recherche Biochimie Structurale et Fonctionnelle
Université Claude Bernard Lyon 1
2013 – 2014

**Interactions Protéine-Protéine
Implication des Résidus Terminaux**

Tutrice pédagogique et Responsable de la Formation : **Pr. Sylvie RICARD-BLUM**

Tutrice professionnelle : **Dr. Juliette MARTIN**

Etudiant : **Olivier MARTIN**



Bases Moléculaires et Structurales des Systèmes Infectieux (BMSSI) - UMR 5086
Bioinformatique : Structures et Interactions (BISI)
7, passage du Vercors 69367 Lyon

Sommaire

1. Introduction.....	3
2. Matériels et méthodes.....	5
2.1 Sources des données	5
2.1.1 InterEvol.....	5
2.1.2 Protein-Protein Docking Benchmark 4.0	7
2.1.3 IntAct.....	8
2.2 Surface accessible au solvant.....	8
2.3 Définition des régions structurelles.....	9
2.4 Modèle stochastique	9
2.5 Statistique.....	10
2.5.1 Bootstrap	10
2.5.2 Test d'indépendance du χ^2	10
2.5.3 Prise en compte des tests multiples.....	11
2.5.4 Index de variation qualitative.....	12
2.6 Informatique	12
3. Résultats.....	12
3.1 Importance de la prise en compte de la taille des protéines	12
3.2 Implication des terminus au niveau des interfaces protéine-protéine.....	13
3.2.1 Proportion de protéines ayant un ou deux terminus à l'interface.....	14
3.2.2 Modèle stochastique : proportion de protéines ayant un terminus à l'interface	15
3.2.3 Co-interfaçage des terminus.....	17
3.3 Caractérisation dynamique des terminus au niveau des interfaces protéine-protéine	18
3.3.1 Conformères des complexes protéine-protéine	18
3.3.2 Conformères libres et liés.....	19
3.4 Terminus et étiquettes.....	19
3.4.1 Terminus privilégié pour la fixation d'une étiquette.....	19
3.4.2 Biais de méthodologie : une conséquence interactionnelle des étiquettes	20
4. Discussion.....	22
5. Conclusion	24
6. Références Bibliographiques.....	24
7. Hyperliens	26
7.1 Bases de données.....	26
7.2 Logiciels et Langages de Programmation	26

Abréviations

IPP : Interaction Protéine-Protéine

IVQ : Index de Variation Qualitative

PDB : Protein Data Bank

RMN : Résonance Magnétique Nucléaire

SAS : Surface Accessible au Solvant

SAS_r : Surface Accessible au Solvant Relative

SAS_{rc} : Surface Accessible au Solvant Relative de la protéine Complexée

SAS_{rm} : Surface Accessible au Solvant Relative de la protéine Monomérique

Remerciements

J'adresse tout d'abord mes remerciements à Richard Lavery pour m'avoir accepté au sein de son laboratoire. Je souhaite remercier tout particulièrement Juliette Martin pour sa confiance et son encadrement tout au cours de ce stage. Je souhaite également remercier l'ensemble de l'équipe BISI et en particulier les doctorants Jonathan Barnoud et Loïc Ethève pour leur gentillesse et leurs conseils.

1. Introduction

Les protéines sont dotées d'activités biochimiques particulièrement variées ce qui leur confère un rôle fonctionnel primordial au sein de la cellule. Afin de mener à bien leur fonction, les protéines agissent rarement seules. Dans environ 80% des cas, elles interagissent entre elles au niveau d'une interface pour former des complexes (Berggård *et al.*, 2007). Les interactions protéine-protéine (IPP) ont des rôles dans de nombreux processus cellulaires dont la transduction du signal, le métabolisme cellulaire, le transport membranaire et l'immunité (Berggård *et al.*, 2007). Les IPP peuvent aussi avoir un rôle dans des processus de pathogénèse comme par exemple dans la drépanocytose. Celle-ci est provoquée par la formation d'une nouvelle interface par mutation au niveau d'une région exposée d'un résidu chargé par un résidu apolaire. La polymérisation de l'hémoglobine mutante est à l'origine de la maladie (Gabriel et Przybylski, 2010). L'ensemble des IPP au sein d'une cellule peut être représenté sous la forme d'un réseau, c'est-à-dire un ensemble de nœuds (les protéines) reliés par des arrêtes (les interactions). Ce réseau porte le nom d'interactome. Son analyse peut permettre d'élucider la fonction de protéines mal caractérisées et de déterminer des cibles thérapeutiques potentielles (Bonetta, 2011).

Plusieurs méthodes expérimentales permettent de détecter les IPP et à terme permettront de résoudre l'interactome de diverses espèces (Berggård *et al.*, 2007). Plusieurs critères permettent de distinguer ces méthodes : le milieu de caractérisation (*in vitro*, *in vivo*), le débit d'information qu'elles génèrent, la possibilité d'obtenir des données quantitatives, la capacité à mettre en évidence des co-complexes, la nécessité d'utiliser des protéines purifiées ainsi que la nécessité d'utiliser une étiquette (en anglais, *tag*). Des exemples de méthodes expérimentales sont la co-immunoprecipitation, le double hydride en levure, le *pull down* et la résonance plasmonique de surface. Il est à noter que ces méthodes sont grandement dépendantes du protocole employé et que les résultats issus de méthodes différentes ne se chevauchent pas toujours (Bonetta, 2011).

La compréhension des interactions au niveau atomique requiert la connaissance de la structure tridimensionnelle des complexes protéines-protéines. Les approches les plus utilisées sont la cristallographie aux rayons X, la résonance magnétique nucléaire (RMN) et la microscopie électronique (Berman *et al.*, 2000). L'ensemble de ces structures sont déposées dans la Protein Data Bank (PDB) (Berman *et al.*, 2000). Toutefois, les protéines sont des objets flexibles (Goh *et al.*, 2004). Des mouvements peuvent se dérouler à des échelles de temps et d'espace différent. Ceci va de la variation de la longueur d'une liaison covalente à l'ouverture d'un canal ionique. Ainsi, le modèle clef-serrure introduit par Fischer n'est pas toujours adapté pour la compréhension des IPP. Deux autres modèles prenant en compte les changements conformationnels sont plus appropriés. Il s'agit de l'ajustement induit et la sélection conformationnelle (Goh *et al.*, 2004). L'ajustement induit introduit par Koshland apporte une notion de plasticité permettant aux protéines d'ajuster leur forme au cours de l'interaction. Le plus récent modèle de la sélection conformationnelle suppose que les protéines existent sous la forme d'un ensemble de conformations différentes en équilibre. Le partenaire interagit sélectivement avec une de ces conformation (c'est-à-dire la conformation complexée ou active) ce qui déplace l'équilibre vers celle-ci. Des données dynamiques sont accessibles par la RMN et la microscopie électronique précédemment citées ainsi que d'autres méthodes moins couteuses en temps de préparation comme la spectroscopie de fluorescence ou le dichroïsme circulaire.

La structure des complexes protéiques apporte des informations concernant leurs interfaces. Bien que les protéines interagissant soient diverses, ces interfaces possèdent des propriétés communes entre elles. La distinction des complexes en fonction de leur composition (hétérodimères et homodimères) et leur affinité (obligatoires ou non-obligatoires) permet de former des groupes d'interfaces homogènes. Les interfaces ont été étudiées en regard de propriétés structurales : taille, planarité, circularité, complémentarité, composition en acide aminé et hydrophobie (Jones et Thornton, 1996). La partition des complexes en plusieurs régions structurelles (par exemple, surface, intérieur et interface) permet de faciliter l'étude des interfaces. Cette partition se réalise classiquement sur la base de l'accessibilité au solvant des résidus et permet de mettre en évidence des différences entre l'interface et le reste de la protéine, notamment en termes de composition en acides aminés (Chakrabarti et Janin, 2002).

Les terminus possèdent des propriétés différentes l'un de l'autre et du reste de la chaîne qui pourraient être exploitées par la protéine notamment afin d'interagir. Ces résidus sont le plus souvent situés à la surface des protéines (Jacob et Unger, 2007). Cette accessibilité confère aux terminus un contact direct avec leur environnement et donc la possibilité d'interagir avec un partenaire protéique. On explique en partie cette accessibilité par la charge des terminus (positive pour le N-terminus et négative pour le C-terminus) qui interagirait avec l'eau. Bien que les autres résidus chargés soient le plus souvent dans des régions exposées, leur surface accessible reste moins importante (Jacob et Unger, 2007). De plus, le désordre structurel intrinsèque des protéines est prédominant aux niveau des extrémités (Uversky *et al.*, 2013). Les terminus héritent de nombreuses propriétés spécifiques aux régions des protéines intrinsèquement désordonnées. Celles-ci sont associées à une multitude de partenaires et fonctions biologiques de part de leur grande plasticité conformationnelle. Les résidus terminaux ont été l'objet de plusieurs études visant à déterminer leurs propriétés particulières. Il a ainsi été montré que certains résidus sont plus fréquents que d'autres aux extrémités (Pal et Chakrabarti, 1999). L'existence de cette distribution non uniforme a une implication biologique du moins pour le N-terminus. La durée de vie des protéines *in vivo* est en effet dépendante de la nature de l'acide aminé en position N-terminal (Varshavsky, 1997). Finalement, il a été avancé que certains biais structuraux au niveau des terminus comme la proximité des terminus (Christopher et Baldwin, 1996) ou la compacité du N-terminus (Alexandrov, 1996) contenaient des informations concernant les mécanismes de repliement des protéines.

Dans ce travail, nous nous sommes intéressés aux terminus dans le contexte des IPP. Il est à noter que dans le cadre expérimental, les terminus sont souvent modifiés par la fixation d'une étiquette. C'est-à-dire une séquence peptidique de taille variable. Elles permettent de grandement faciliter certains protocoles et leur utilisation est devenue presque incontournable pour la purification à grande échelle de protéines. Approximativement 75% des protéines exprimées pour des études cristallographiques seraient basées sur l'utilisation d'étiquette (Derewenda, 2004). De même, les méthodes de détection des IPP à haut débit nécessitent presque systématiquement la présence d'au moins une étiquette. L'étiquette peut cependant altérer certaines propriétés de la protéine comme son taux d'expression (Berggård *et al.*, 2007), sa croissance cristalline (Bucher *et al.*, 2001), sa localisation (Palmer et Freeman, 2007) ou sa fonction (Christensen *et al.*, 2012). Bien que certaines publications essaient de sonder ces conséquences en utilisant des protocoles standardisés, la majorité de ces résultats sont anecdotiques empêchant l'élaboration d'une méthode pouvant prédire les conséquences des étiquettes. A notre connaissance, aucune étude ne s'est intéressée aux effets des étiquettes sur la détection des IPP.

L'objectif de cette étude est d'étudier les relations entre les terminus et les interactions protéine-protéine. Notre étude sera divisée en trois parties :

1. Nous montrerons l'importance de la prise en compte de la taille des protéines lors de l'analyse de paramètres structuraux liés aux interfaces.
2. Nous analyserons la présence des terminus dans les interfaces protéine-protéine sur un grand ensemble de structures protéiques.
3. Nous étudierons les implications de la localisation des terminus pour la détection des IPP à l'aide de méthodes nécessitant des étiquettes.

2. **Matériels et méthodes**

2.1 Sources des données

Cette étude s'est limitée aux dimères protéine-protéine dont la structure tridimensionnelle est connue et déposée dans la PDB.

2.1.1 InterEvol

Une liste non-redondante de 17658 dimères protéiques a été extraite de la base de données InterEvol (Faure *et al.*, 2012). Cette liste a les propriétés suivantes :

- Elle a été construite à partir des unités biologiques de la PDB.
- Elle favorise les protéines entières aux fragments de protéines ce qui nous permet d'étudier les terminus des protéines natives. Toutes les séquences des protéines ont un équivalent dans la base de données UniProt.
- Elle utilise la notion de redondance entre interfaces. Cela permet de ne pas surreprésenter une interface et ainsi biaiser les résultats. La redondance est définie comme suit : une identité de séquence primaire de plus de 70% sur plus de 70% de sa longueur et plus de 40% d'identité entre les résidus à l'interface.
- Elle distingue les interfaces natives des complexes biologiques (interfaces biologiques) des interfaces constituant des artefacts de la cristallisation (interfaces cristallines). La probabilité d'avoir une interface cristalline est calculée par NOXclass. Il s'agit d'un algorithme d'apprentissage informatique ayant une précision de 92% (Zhu *et al.*, 2006). Nous avons considéré les interfaces comme cristallines à partir d'une probabilité de 50%.
- Elle distingue les complexes obligatoires des complexes non-obligatoires. Les complexes obligatoires sont ceux dont les composants ne sont pas observables à l'état isolé. La probabilité d'avoir une interface obligatoire est calculée par NOXclass. Nous avons considéré les interfaces comme obligatoires à partir d'une probabilité de 50%.

Deux fichiers issus de cette base de données ont été utilisés : INTER70_REFINFO et INTER70. Le fichier INTER70_REFINFO compte 17658 complexes protéine-protéines dimériques ayant des interfaces non redondantes. On dénombre 35316 monomères dont 5350 issues hétérodimères biologiques, 2254 de hétérodimères cristallins, 19408 de homodimères biologiques et 8304 de homodimères cristallins. La majorité de ces complexes sont des structures cristallographiques. Nous avons également cherché à étudier des structures résolues par RMN. Le fichier INTER70 contient un ensemble d'identifiants renvoyant à des complexes dimérique groupés selon la redondance des interfaces. Une unique structure RMN a été sélectionnée au sein de chaque groupe. On dénombre au final 194 complexes dimériques résolus par RMN. De part la faible quantité de données, les homodimères et les hétérodimères n'ont pas été distingués. Les terminus ont été définis comme le premier et le dernier résidu de la séquence native.

Les dimères contenant une structure d'anticorps ont été retirés afin de ne pas biaiser l'étude, les complexes anticorps-antigène étant particulièrement nombreux dans la PDB. Les mots clés suivants ont été recherchés dans la liste initiale et les dimères correspondants ont été retirés : « fab », « antibody », « antigen » et « igg ».

Il n'est pas toujours possible de déterminer des coordonnées pour tous les résidus d'une protéine. On retrouve dans certains fichiers PDB, des résidus sans coordonnées. Ces résidus sont présents dans la partie SEQRES (séquence complète de la protéine étudiée) mais absents de la partie ATOM (coordonnées des résidus). Un alignement de ces deux séquences est déjà présent dans la partie « pdbx_poly_seq_scheme » du fichier Macromolecular Crystallographic Information. Cet alignement permet d'automatiser rapidement la détection des coordonnées manquantes. Les auteurs d'InterEvol nous ont fourni les documents nécessaires pour faire la correspondance entre leur base de données et la PDB. Les monomères ayant au moins un résidu terminal sans coordonnées ont été retirés du jeu de données initial.

Les protéines recombinantes sont souvent exprimées avec une étiquette. Cette étiquette n'est pas systématiquement excisée avant des études structurelles. Il existe donc un risque de confondre un résidu de l'étiquette pour un résidu terminal de la protéine native. La détection de ces étiquettes est plus délicate car il n'existe pas de manière standardisée de les indiquer. Pour cela, une recherche d'expression régulière a été réalisée : « ^X{0,5}ETIQUETTE » pour le N-terminus et « ETIQUETTE(X{0,5})\$ » pour le C-terminus. « X » est un résidu quelconque et « ETIQUETTE » est la séquence de l'étiquette (Tableau 1). L'expression « ^ » recherche en début de séquence, « \$ » recherche en fin de séquence, « {m,n} » recherche entre m et n fois l'expression placée en amont et « () » est utilisé pour former une sous-expression. Nous considérons que le résidu en aval (N-terminus) ou en amont (C-terminus) de l'étiquette est le terminus de la protéine native.

On retrouve six cas possibles pour chaque résidu terminal en fonction de la présence ou absence des coordonnées et d'une étiquette (Figure 1). De manière à étudier les résidus terminaux de la protéine native, certains monomères ont dû être retirés du jeu de données. Ce sont les monomères ayant au moins un résidu terminal répondant au cas 2 ou 6. Les fréquences d'apparition sont différentes selon la méthode employée pour résoudre la structure.

Nom	Taille	Séquence
Poly-Arginine	5 - 6	(R) _n
Poly-Histidine	5 - 10	(H) _n
FLAG	8	DYKDDDDK
(FLAG) ₃	22	KRRWKKNFIAVSAANRFKKISSLGAL
Hemagglutinin	19	YPYDVPDYA
c-myc	19	EQKLISEEDL
Histidine Affinity Tag	19	KDHЛИHNВHKEFHAHAHNK
Calmodulin-Binding Peptide	26	KRRWKKNFIAVSAANRFKKISSLGAL
Streptavidin-Binding Peptide	38	MDEKTTGWRGGHVVEGLAGELEQLRARLEH HPQQQREP

Tableau 1. Noms, tailles et séquences des étiquettes recherchées. Cette liste a été conçue à partir de données issues de la littérature scientifique (Terpe et al., 2003).

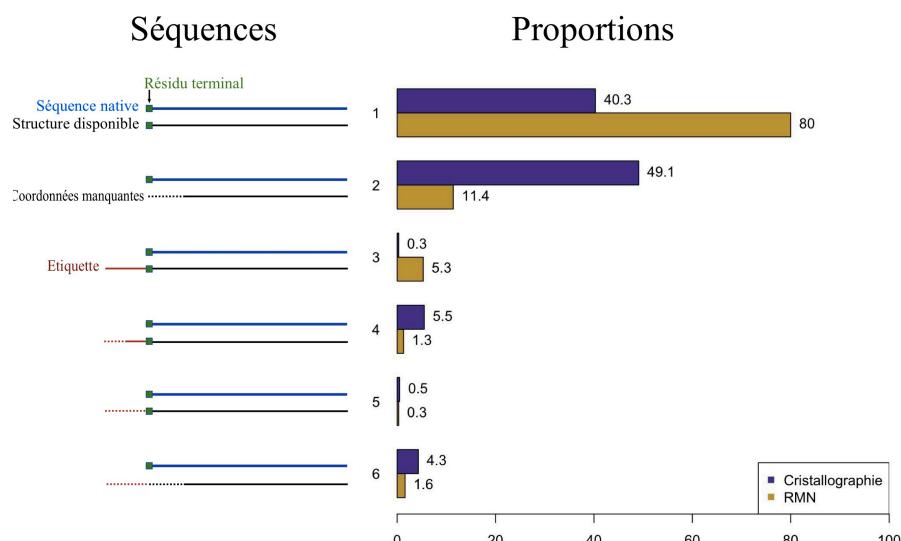


Figure 1. Résidus terminaux sans coordonnées et étiquettes. Cas n°1 : on ne retrouve pas coordonnées manquantes ou d'étiquette. Cas n°2 : les coordonnées manquantes concernent le résidu terminal. Cas n°3 : on retrouve une étiquette. Cas n°4 : les coordonnées manquantes concernent partiellement l'étiquette. Cas n°5 : les coordonnées manquantes concernent l'intégralité de l'étiquette. Cas n°6 : les coordonnées manquantes concernent l'intégralité de l'étiquette et le résidu terminal.

2.1.2 Protein-Protein Docking Benchmark 4.0

La base de données Protein-Protein Docking Benchmark 4.0 (Hwang *et al.*, 2010) contient 176 hétéro-oligomères de protéines aussi bien sous leur conformation libre que leur conformation liée. Elle a permis d'étudier les résidus terminaux dans ces différentes conformations. Seuls les dimères ont été maintenus ce qui réduit le nombre de complexes à 119. La base de données a été construite de manière à ce que les structures aient le minimum de coordonnées manquantes et le maximum de correspondance de séquence entre la conformation libre et liée. Cependant, les séquences ne sont pas toujours de même longueur ce qui implique que les résidus terminaux ne correspondent pas toujours. Un alignement de séquence global a donc été réalisé entre les séquences de la forme libre et liée. Des résidus aux extrémités ont été retirés de manière à pouvoir étudié le même résidu. Seuls les monomères pour lesquelles moins de 15 résidus ont été retirées ont été maintenues. Ceci restreint le jeu de données à 66 hétérodimères biologiques soit 264 résidus terminaux.

Le fait que certains résidus terminaux fluctuent dans une gamme de distance inférieure à la résolution a dû être pris en compte. Nous avons retenu la résolution la plus faible parmi les différentes structures (c'est-à-dire le complexe et les deux monomères). Certains complexes possèdent sous leur conformation libre une structure résolue par RMN. Nous avons fixé leur résolution arbitrairement à 2Å. On dénombre alors 74 résidus terminaux.

2.1.3 IntAct

Les données concernant les méthodes de détection des interactions protéine-protéine ont été obtenues par la base de données IntAct (Orchard et al., 2014). Elle contient des informations détaillées et standardisées concernant les IPP pour lesquelles il existe des preuves expérimentales. Les données utilisées ici ont été extraites du fichier compressé pmidMIF25.zip. Celui-ci contenait 12397 fichiers XML correspondants aux publications dont sont issues les données. La correspondance entre les chaines PDB et les numéros d'accession Uniprot a été possible grâce à la base de données PDB/Uniprot Mapping (Martin, 2005). On dénombre 27858 expériences pour 262415 interactions concernant 62915 protéines. Environ 97% de ces interactions sont de type hétérodimère. Le dernier niveau nous indique parfois qu'une étiquette a été utilisée pour l'expérience et parfois même la localisation de cette étiquette. Cette information n'a été annotée que dans une minorité des cas (environ 3%) et ne peut donc pas être utilisée.

2.2 Surface accessible au solvant

Les calculs de surface accessible au solvant (SAS) ont été réalisés avec Naccess 2.1.1 (Hubbard et Thornton, 1993). Ce logiciel correspond à une implémentation de la méthode de Lee et Richards. Il consiste à générer autour de chaque atome une distribution uniforme d'un nombre fini de points (N_{tot}) assimilables à une sphère. Le rayon R de cette sphère est la somme du rayon de Van der Waals de l'atome ainsi que d'un rayon sonde représentant le solvant (ici, une molécule d'eau de rayon 1.4Å). Les points non enfouis sous d'autres sphères (N_{acc}) vont constituer la surface accessible au solvant (Figure 2). La ratio N_{acc} / N_{tot} permet d'approximer l'aire de la SAS à partir de l'aire de la sphère initialement définie.

Cette aire est normalisée afin d'être comparable entre résidus. Pour cela, on utilise l'aire du résidu X à l'intérieur du tripeptide Ala-X-Ala dans une conformation étendue. Celui-ci est supposé représenter l'aire maximale que la SAS peut avoir pour le résidu X. On parle dans ce cas de surface accessible au solvant relative (SAS_r). Pour une protéine faisant partie d'un complexe, il est possible de calculer la SAS_r d'un résidu dans deux contextes différents : la forme momomérique (SAS_{rm}) et la forme complexée (SAS_{rc}). A partir de ces deux valeurs, il est également possible de calculer la variation de la SAS_r provoquée par la complexation (ΔSAS_r).

$$\Delta SAS_r = SAS_{rm} - SAS_{rc} \geq 0$$

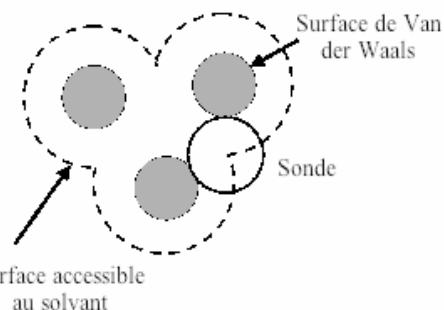


Figure 2. Surface accessible au solvant. Figure d'origine en anglais.
http://www ccp4.ac.uk/newsletters/newsletter38/03_figs/03_fig1.gif

2.3 Définition des régions structurelles

Les régions structurelles sont déduites des valeurs de SAS_r des résidus selon la définition proposée par Levy (Levy *et al.*, 2010) (Figure 3) :

- Interface : $\Delta SAS_r > 0$
 - Rebord (*Rim*) : $SAS_{rc} > 25\%$
 - Cœur (*Core*) : $SAS_{rc} < 25\%$ et $SAS_{rm} > 25\%$
 - Support (*Support*) : $SAS_{rm} < 25\%$
- Non interface $\Delta SAS_r = 0$
 - Surface (*Surface*) : $SAS_r > 25\%$
 - Intérieur (*Interior*) : $SAS_r < 25\%$

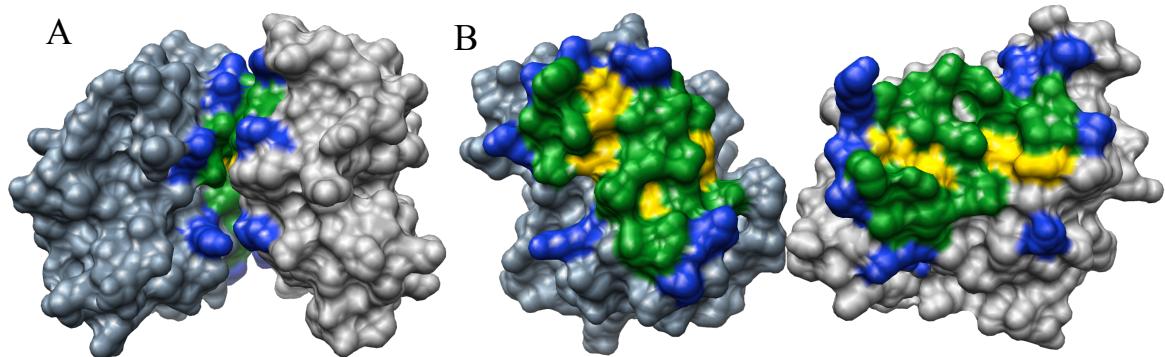


Figure 3. Régions structurelles d'interface selon la définition de Levy

Exemple de l'hémoglobine humaine (PDB : 1HHO). Seul un hétérodimère $\alpha\beta$ est représenté pour plus de clarté. Les régions d'interfaces sont colorées tandis que la surface et l'intérieur sont en gris. Le rebord est coloré en bleu, le cœur en vert et le support en jaune. La chaîne α est en gris plus foncé que la chaîne β . (A) Vue sous sa forme liée. (B) Vue éclatée du complexe : les monomères ont été distancés et ont subit une rotation de 90°.

2.4 Modèle stochastique

Nous avons cherché à connaître la distribution du nombre de protéines ayant un terminus à l'interface attendu par hasard. Dans notre modèle, nous associons à chaque protéine une variable aléatoire X . Celle-ci peut prendre deux valeurs selon la réalisation de l'événement aléatoire d'interfaçage : « 0 » si le terminus n'est pas à l'interface et « 1 » si le terminus est à l'interface. Nous sommes donc face à une suite d'épreuves de Bernoulli. Pour chaque variable aléatoire définie, il est alors nécessaire de définir une probabilité de « succès » $P(X = 1)$ ainsi qu'une probabilité « d'échec » $P(X = 0)$. Pour une protéine, cela correspond à calculer la probabilité qu'un résidu quelconque soit à l'interface. Celle-ci peut être estimée par la proportion de résidus à l'interface. Cependant, les terminus sont majoritairement situés dans des régions exposées. On omettra donc la contribution des régions non exposées dans la forme monomérique ($ASA_{rm} < 25\%$), c'est-à-dire le support de l'interface et l'intérieur de la protéine.

$$P(X = 1) = \frac{\text{Rebord} + \text{Cœur}}{\text{Rebord} + \text{Cœur} + \text{Surface}}$$

Il est possible de simuler, sous loi hypergéométrique, pour chaque protéine individuellement, la réalisation de cet événement aléatoire. Nous pouvons compter le nombre de protéines pour lesquelles il y a eu succès ($X = 1$, le terminus est à l'interface). Nous répétons alors cette simulation un grand nombre de fois. Nous obtenons ainsi une distribution théorique du nombre de protéines ayant un terminus à l'interface.

On compare le nombre observé de protéines avec notre distribution théorique. Cela peut se faire à travers le calcul d'une valeur p comme dans un test statistique classique. On parle dans ce cas de valeur p empirique. La valeur p correspond à l'aire sous la courbe de la densité de probabilité (Figure 4). Cette aire peut être estimée par la formule suivante. N_{SE} correspond au nombre de fois que des valeurs simulées étaient supérieures ou égales à la valeur observée, N_{IE} au nombre de fois que celles-ci étaient inférieures ou égales et N_T au nombre total de simulation. Nous avons réalisé 10000 simulations. En prenant le minimum, on détectera à la fois les effets sur- et sous-représentés.

$$p = \min \left(\frac{N_{SE}}{N_T}, \frac{N_{IE}}{N_T} \right)$$

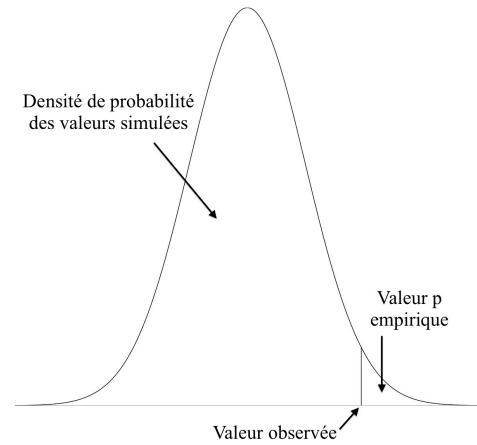


Figure 4. Représentation graphique de la valeur p empirique.

2.5 Statistique

2.5.1 Bootstrap

La distribution exacte de nos variables nous est inconnue. Cela rend difficile le calcul de paramètres comme l'erreur standard. Afin d'estimer cette distribution, nous pouvons utiliser la méthode du bootstrap. Cette méthode consiste à générer un grand nombre de nouveaux échantillons de même taille que l'échantillon initial et d'en faire la synthèse. Ces échantillons sont construits en tirant au sort avec remise parmi les valeurs de l'échantillon initial. On parle de ré-échantillonnage. Pour chaque échantillon, il devient alors possible d'estimer un paramètre (par exemple, la moyenne). Cette étape est répétée plusieurs fois et nous nous retrouvons avec un ensemble d'estimations du paramètre. L'ensemble de ces estimations constitue sa distribution statistique.

2.5.2 Test d'indépendance du χ^2

Le test d'indépendance du χ^2 permet de mesurer l'indépendance entre deux variables qualitatives. La première étape consiste à définir l'hypothèse nulle (H_0) : l'indépendance entre les deux variables. L'hypothèse alternative (H_1) est définie comme la dépendance entre les variables. Le risque de première espèce α est alors fixé. Celui-ci constitue une règle de décision car il définit les statistiques du tests pour lesquelles l'hypothèse nulle est maintenue ou rejetée. Les effectifs observés (O_i) sont placés dans un tableau de contingence. Ce tableau est caractérisé par un degré de liberté (k) calculé par la formule suivant où n_c est le nombre de colonnes et n_r est le nombre de rangées.

$$k = (n_c - 1)(n_r - 1)$$

Les effectifs attendus (E_i) sont les valeurs qui auraient été observées en cas d'indépendance. Ils sont calculés par la formule suivante où S_{ci} représente la somme de la colonne correspondante, S_{ri} représente la somme de la rangée correspondant et S la somme totale du tableau.

$$E_i = \frac{S_{Ci} \times S_{Ri}}{S}$$

Lorsque les variables sont indépendantes, la somme des écarts aux carrés normalisés par les valeurs attendues suit une loi du χ^2 à k degrés de liberté ($\chi^2(k)$).

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i} \in \chi^2(k)$$

A partir de cette statistique calculée et de la distribution du χ^2 , la valeur p peut être extraite. Elle représente la probabilité d'obtenir une statistique au moins aussi extrême que celle observée. L'hypothèse nulle est rejetée si la valeur p est inférieure au seuil α . Afin de savoir quelles cases contribuent à une valeur de χ^2 significative, il est possible de faire une analyse des résidus standardisés (R_i). Un résidu standardisé correspond à l'écart normalisé par la racine carrée de la valeur attendu. Ces résidus standardisés suivent une loi normale centrée réduite ($N(0,1)$). Il est donc possible de calculer une valeur p pour chacune des cases et savoir si sa contribution au χ^2 est significative.

$$R_i = \frac{O_i - E_i}{\sqrt{E_i}} \in N(0,1)$$

2.5.3 Prise en compte des tests multiples.

La probabilité de faire une erreur de type I (faux positif) grandit avec le nombre de tests statistiques employés. C'est le problème des comparaisons multiples. Afin de limiter le taux de fausses découvertes, la valeur p a été ajustée par la méthode de Benjamini – Hochberg (Benjamini et Hochberg, 1995). Considérons N hypothèses (H_1, \dots, H_N) ayant respectivement leur valeur p (p_1, \dots, p_N). Les valeurs p sont classées dans l'ordre croissant tel que : $p_{(1)} \leq \dots \leq p_{(N)}$. Notons alors les n hypothèses classées ($H_{(1)}, \dots, H_{(N)}$) et leur valeur p correspondant ($p_{(1)}, \dots, p_{(N)}$). Les valeurs p sont alors corrigées en utilisant ce système d'équation.

1. $\tilde{p}_N = p_N$
2. $\tilde{p}_{(i)} = \min \left(\tilde{p}_{(i+1)}, \frac{N}{i} p_i \right), i = N - 1, \dots, 1$

Ces formules correspondent à appliquer la méthode suivante :

1. La valeur p la plus élevée $p_{(N)}$ est laissée inchangée.
2. On prend ensuite la deuxième valeur p la plus élevée $p_{(n-1)}$. Celle-ci est multipliée par N et divisée par son index ($i = N-1$). Si cette valeur est inférieure à la valeur p précédemment calculée, alors c'est cette valeur qui est retenue. Autrement, c'est la valeur p précédemment calculée qui est retenue. Ce deuxième calcul est alors itéré jusqu'à la plus petite valeur p .

2.5.4 Index de variation qualitative

Les mesures de dispersion classiques ne sont applicables qu'aux variables quantitatives. Pour cette raison, un grand nombre de mesures analogues applicables aux variables qualitatives ont été développées. Toutes ces mesures varient toutes entre 0 et 1. Elles sont égales à 0 lorsque la dispersion est minimale, c'est-à-dire que toutes les variables ont la même valeur. Elles sont égales à 1 lorsque la dispersion est maximale, c'est-à-dire que toutes les valeurs possibles sont uniformément distribuées. La mesure que nous avons choisie se nomme l'index de variation qualitative (IVQ) (Wilcox, 1967). Sa formule est suivante où N est la taille de l'échantillon, K le nombre de valeurs et f_i est la fréquence d'une catégorie.

$$IVQ = 1 - \frac{K}{N^2(K-1)} \sum_{i=1}^K (f_i - \frac{N}{K})^2$$

2.6 Informatique

Toute cette étude a été réalisée sur un Mac Mini 2.1 avec Mac Os X 10.6.8. Trois langages de programmation ont été utilisés : Python, R et Shell script. Python a permis l'extraction des données des différentes bases de données et d'y appliquer des opérations de manière automatique (par exemple, lancer un calcul de surface accessible au solvant avec Naccess). R a été utilisé pour analyser les données générées : réalisation de tests statistiques et création de graphiques. Un Shell script a été rédigé dans certains cas afin d'automatiser le lancement de plusieurs scripts à la suite. Le logiciel Chimera (Pettersen *et al.*, 2004) a été utilisé pour la visualisation de structures de protéines, réaliser des calculs de distance ainsi que créer les images de structures.

3. Résultats

3.1 Importance de la prise en compte de la taille des protéines

La proportion de résidus à l'interface est souvent considérée comme étant constante. Or, de nombreux paramètres structuraux, dont la taille des interfaces, dépendent de la taille des protéines. Cela est illustré dans la Figure 5 sur les structures du jeu de données InterEvol initial (17658 complexes). Les interfaces cristallines sont en moyenne plus petites que les interfaces biologiques et les interfaces des hétérodimères sont en moyenne plus petites que les interfaces des homodimères. Compte tenu de cette dépendance à la taille, il est important pour la suite de l'analyse de tenir compte de la taille des protéines étudiées.

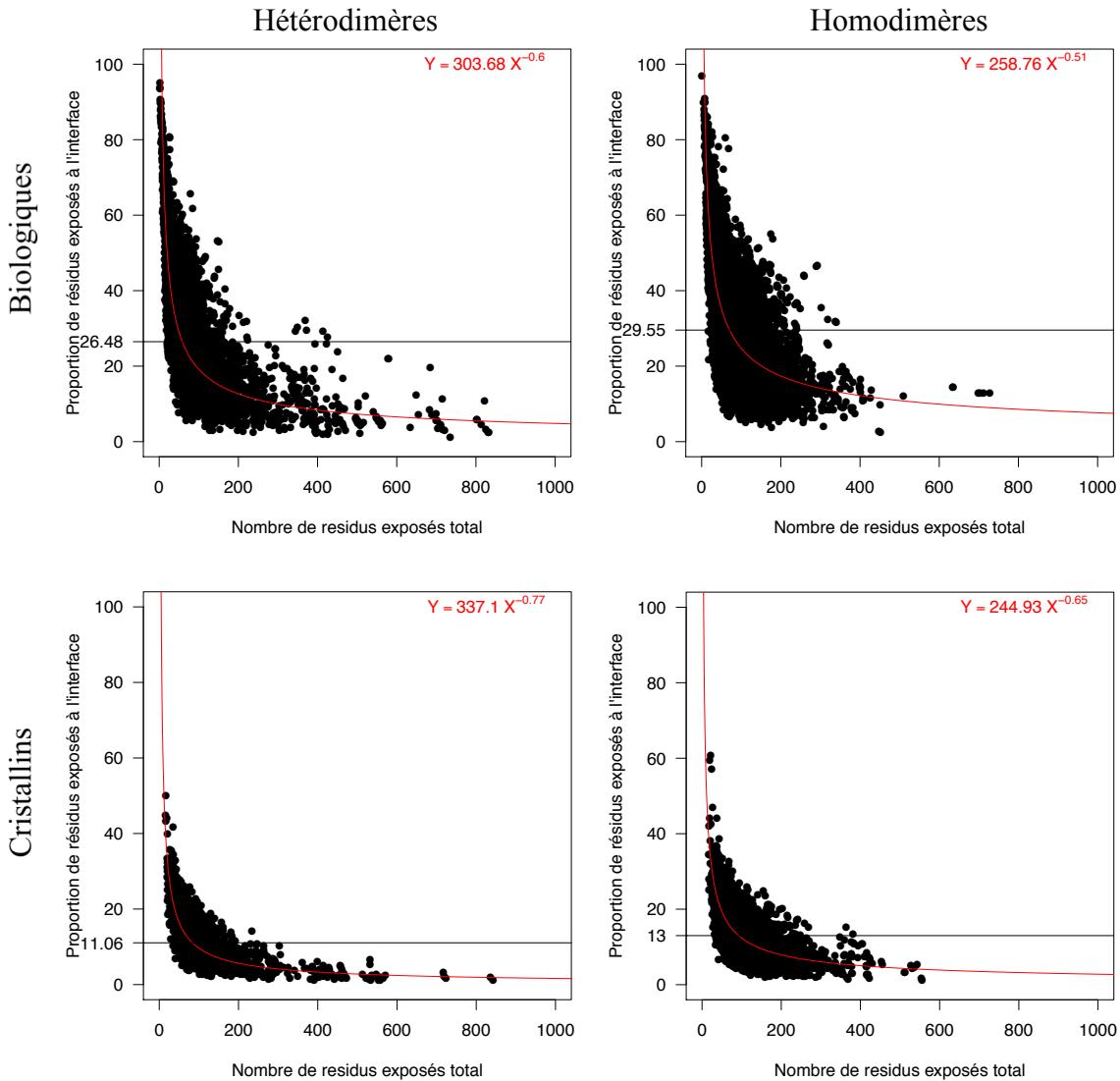


Figure 5. Relation entre le nombre de résidus exposés (rebord, cœur et surface) et la proportion de résidus exposés à l'interface (rebord et cœur). La courbe rouge est issue d'un ajustement de courbe utilisant une loi de puissance. L'équation correspondante est exposée en haut à droite. La droite noire représente la moyenne de l'ensemble de ces valeurs.

3.2 Implication des terminus au niveau des interfaces protéine-protéine

Une unique conformation des complexes dimériques a été analysée : conformation obtenue par cristallographie ou structure représentative obtenue par RMN. Le fichier INTER70_REFINFO a été utilisé. La présence d'étiquette et de coordonnées manquantes a été prise en compte comme expliqué dans les matériaux et méthodes. L'analyse compte 6655 dimères dont 1332 issues de hétérodimères biologiques, 434 de hétérodimères cristallin, 3295 de homodimères biologiques et 1594 de homodimères cristallins. Les terminus ont été définis comme le premier et le dernier résidu de la protéine. La fréquence d'apparition des terminus au niveau des différentes régions des protéines définies par Levy a d'abord été étudiée (Tableau 2).

Les terminus sont préférentiellement au niveau de la surface et du rebord caractérisés par une surface accessible au solvant relative supérieure à 25% même à l'état complexé. Les répartitions sont considérablement différentes entre les complexes biologiques et cristallins. La proportion de terminus à l'interface (rebord, cœur et support) est environ deux fois plus importante pour les complexes biologiques. De même, les terminus issus des homodimères sont plus souvent à l'interface que les terminus des hétérodimères. Ces différences sont sans doute dues à des différences de proportion moyenne de résidus à l'interface. Celle-ci est plus importante pour les complexes biologiques par rapport aux complexes cristallins ainsi que plus importante pour les homodimères par rapport aux hétérodimères (Figure 5).

	Hétérodimères Biologiques	Homodimères Biologiques	Hétérodimères Cristallins	Homodimères Cristallins
Surface	72,1%	67,7%	84,6%	79,1%
Intérieur	5,1%	5,0%	4,8%	7,4%
Rebord	18,5%	22,9%	8,9%	11,8%
Cœur	3,6%	3,7%	1,3%	1,2%
Support	0,8%	0,7%	0,5%	0,5%

Tableau 2. Fréquence d'apparition des terminus au niveau des régions structurelles. Les sommes peuvent ne pas être égale à 100% de part de l'arrondissement à la première décimale.

3.2.1 Proportion de protéines ayant un ou deux terminus à l'interface

Chaque monomère formant un complexe possède deux terminus susceptibles d'être à l'interface (Figure 6). On dénombre quatre catégories selon la présence des terminus à l'interface et les opérateurs logiques « OU » et « ET » : le N-terminus est à l'interface (Figure 6B et C), le C-terminus est à l'interface (Figure 6B et D), le N-terminus OU le C-terminus sont à l'interface (Figure 6B, C et D) et le N-terminus ET le C-terminus sont à l'interface (Figure 6B). Les proportions de monomères de ces catégories ont été calculées en prenant en compte la taille des protéines à travers des intervalles de taille d'une étendue de 50 résidus contenant au moins 20 monomères (Figure 7). La proportion d'interfaces biologiques impliquant au moins un terminus (N- OU C-) est loin d'être négligeable. Elle est en moyenne de 38% pour les hétérodimères biologiques et 45% pour les homodimères biologiques. Ces proportions sont dépendantes de la taille des monomères. Cette dépendance semble cependant complexe et difficile à interpréter. On remarque également une différence dans le comportement du N- par rapport au C-terminus. Pour certains intervalles, la proportion de l'un semble significativement différente de l'autre. On retrouve en moyenne plus de C-terminus que de N-terminus à l'interface. Pour les hétérodimères biologiques, on retrouve 25,1% de N-terminus contre 29,6% de C-terminus à l'interface. Pour les homodimères biologiques, on retrouve 22,3% de N-terminus contre 23,3% de C-terminus à l'interface.

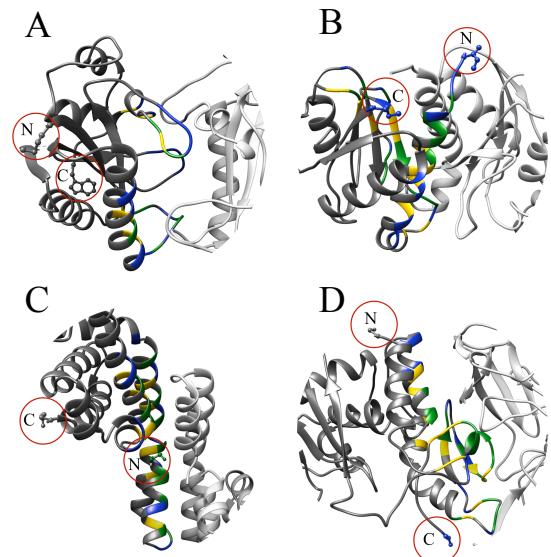


Figure 6. Terminus aux interfaces protéine-protéine. Le rebord est coloré en bleu, le cœur en vert et le support en jaune. (A) Aucun des terminus n'est à l'interface (PDB : 2QX0). (B) Les deux terminus sont à l'interface (PDB : 1TO6). (C) Seul le N-terminus est à l'interface (PDB : 1KN1). (D) Seul le C-terminus est à l'interface (PDB : 1GGP)

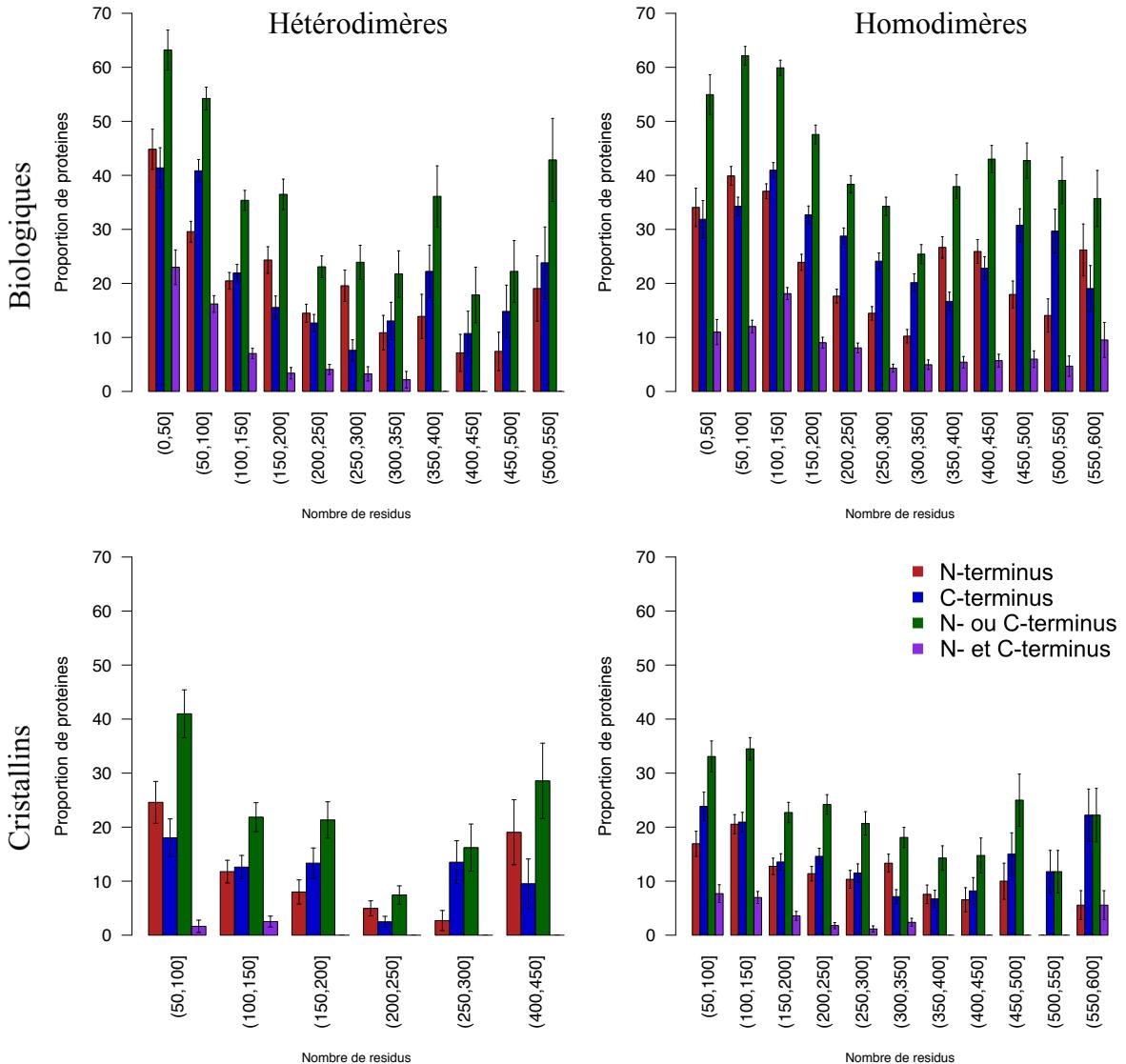


Figure 7. Proportion de protéines ayant son ou ses terminus à l'interface. Les barres correspondent à deux erreurs standard obtenues par bootstrap (10000 ré-échantillonnages).

3.2.2 Modèle stochastique : proportion de protéines ayant un terminus à l'interface

Les proportions de protéines observées sont loin d'être négligeables pour les interfaces biologiques. De plus, on retrouve plus de C-terminus que de N-terminus à l'interface. La significativité statistique a été évaluée en comparant les nombres de protéines observés ayant un terminus à l'interface avec les nombres attendus par hasard. Un modèle stochastique de l'interfaçage d'un terminus a été construit. Celui-ci est construit à partir de la proportion de résidus à l'interface et est décrit en détail dans les matériels et méthodes. Il permet d'estimer la distribution statistique du nombre de protéines attendues (Figure 8). Les interfaces cristallines représentent des associations transitoires qui peuvent se dérouler entre n'importe quelle paire de protéine (Zhu *et al.*, 2006). Elles n'ont pas été sélectionnées au cours de l'évolution pour des propriétés structurelles ou fonctionnelles particulières. Ces interfaces constituent donc un contrôle pour le modèle stochastique. Celui-ci semble particulièrement bien prédire les nombres observés de protéines ayant un terminus à l'interface pour les protéines à interfaces cristallines. Aucune valeur p ne descend en dessous 0,01. Le modèle construit est donc valide. On remarquera également que le modèle est valide pour les homodimères non-obligatoires.

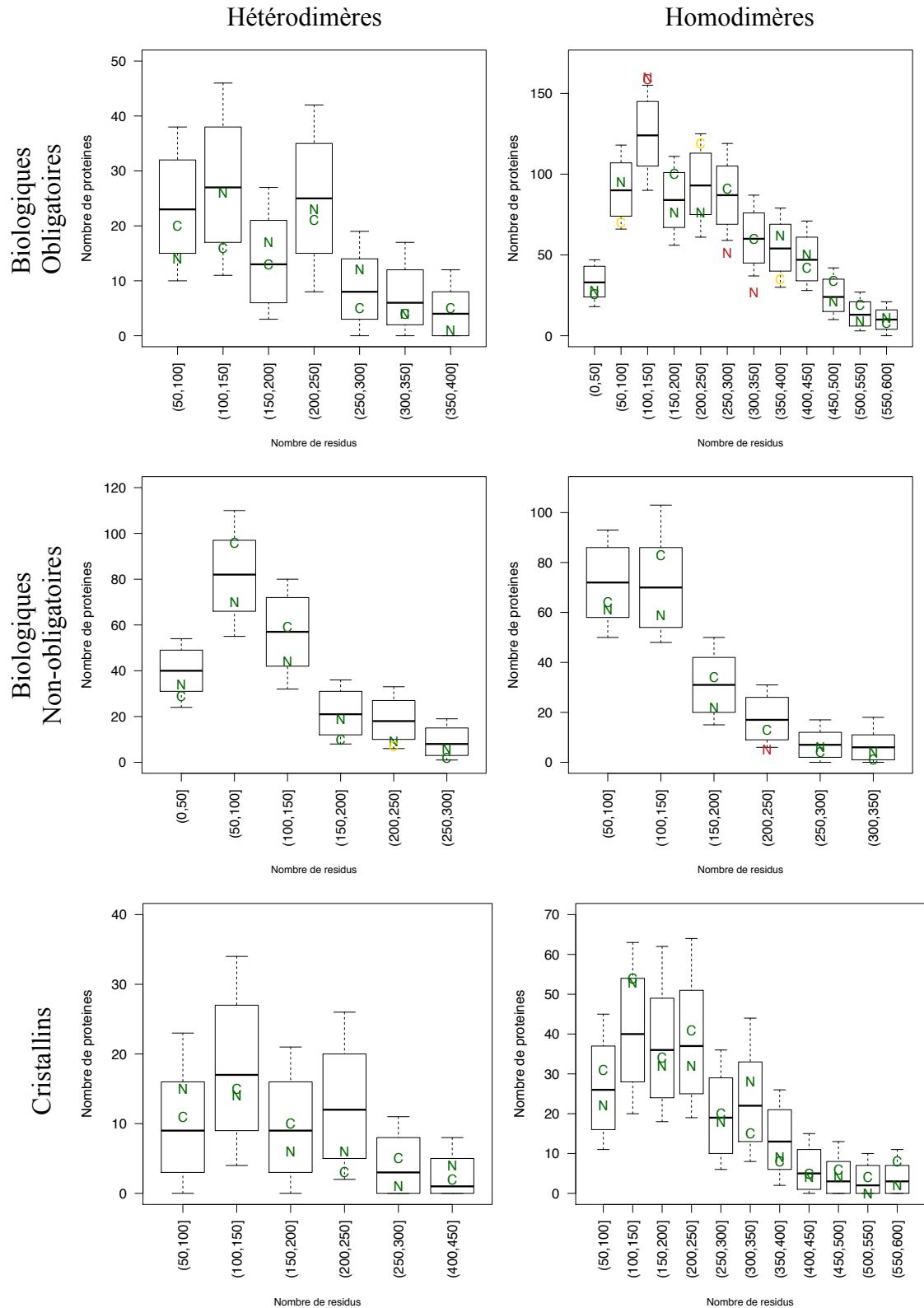


Figure 8. Modèle stochastique et nombres observés de protéines ayant un terminus à l'interface. Les distributions statistiques calculées par le modèle stochastique sont représentées par des boîtes à moustaches. Le trait noir central est la médiane. La boite est l'intervalle de confiance à 99%. Les moustaches sont les valeurs en dehors de cet intervalle. Les nombres observés de protéines sont représentés par un « N » si elles possèdent leur N-terminus à l'interface ou par un « C » si elles possèdent leur C-terminus à l'interface. Ces lettres sont colorées en fonction de la valeur p empirique : $p > 0.01$ (vert), $0.01 \geq p > 0.001$ (jaune) et $0.001 \geq p$ (rouge). Les valeurs p ont été corrigées par Benjamini-Hochberg.

Au contraire, il paraît ne pas être complètement valide pour les hétérodimères biologiques et les homodimères biologiques obligatoires. Les différences touchent aussi bien le N- que le C-terminus. Pour les hétérodimères, il semble y avoir un évitement des interfaces. Pour les homodimères biologiques obligatoires, selon les intervalles de taille, la distribution théorique sous-estime ou surestime le nombre de terminus attendus. L'interfaçage des terminus de ces complexes n'est donc pas assimilable à un processus stochastique. Ceci pourrait impliquer la présence d'un avantage fonctionnel à présenter ou non un terminus à l'interface.

3.2.3 Co-interfaçage des terminus

Nous définissons un co-interfaçage comme la présence à l'interface d'au moins un terminus par monomère formant un complexe. Quatre catégories sont dénombrés pour des complexes dimériques : présence des deux N-terminus à l'interface (NN), présence des deux C-terminus à l'interface (CC) et deux cas pour la présence d'un N-terminus et d'un C-terminus (NC et CN), chacun provenant d'un monomère. Pour chaque complexe, les quatre terminus doivent être définis. Cela réduit le nombre de complexes total employables (Tableau 3). Les co-interfaçages ne se répartissent pas de manière homogène entre ces différentes catégories. Les cas NC et CN semblent particulièrement sous-représentés. Le cas le plus fréquent pour l'ensemble des complexes est celui où chaque monomère présente un C-terminus à l'interface (CC).

	Hétérodimères Biologiques	Hétérodimères Cristallins	Homodimères Biologiques	Homodimères Cristallins
NN	29	1	295	61
CC	34	4	346	72
NC et CN	30	1	173	22
Valeurs attendues	23,25	1,5	203,5	38,75
Co-interfaçages	65	5	563	118
Complexes	384	133	1379	683

Tableau 3. Nombre observé de co-interfaçage de terminus. Quatre catégories sont dénombrés pour des complexes dimériques : présence des deux N-terminus à l'interface (NN), présence des deux C-terminus à l'interface (CC) et deux cas pour la présence d'un N-terminus et d'un C-terminus (NC et CN), chacun provenant d'un monomère. La somme des co-interfaçages n'est pas égale au nombre de complexes où ils sont présents : un même complexe possède plusieurs cas de co-interfaçage s'il possède plus de deux terminus à l'interface.

3.3 Caractérisation dynamique des terminus au niveau des interfaces protéine-protéine

Les protéines existent sous la forme d'un ensemble de conformations distinctes, chacune respectant le repliement général. La position des différents résidus n'est donc pas constante. Un résidu à l'interface dans une conformation peut ne pas l'être pas dans une autre. Les terminus ont donc été examinés au niveau de différentes conformations de complexes protéine-protéine ainsi dans les conformations libres et liées des protéines.

3.3.1 Conformères des complexes protéine-protéine

La majorité des complexes protéines-protéines d'InterEvol ont été résolus par cristallographie aux rayons X. Nous cherchons à étudier des structures résolues par RMN. Le fichier INTER70 a été utilisé comme expliqué dans les matériaux et méthodes. La présence d'étiquette et de coordonnées manquantes a été prise en compte comme expliqué dans les matériaux et méthodes. On dénombre 194 complexes dimériques soit 388 monomères. De part la faible quantité de données, les homodimères et les hétérodimères n'ont pas été distingués. Les intervalles de taille contiennent au moins 10 monomères.

La position des terminus a été définie en utilisant les cinq régions de Levy. De manière à avoir une mesure objective de la variation du positionnement des terminus, l'index de variation qualitative (IVQ) a été utilisée. La proportion de terminus ayant un IVQ donné a été établie (Figure 9A). Seulement la moitié des terminus reste fidèle à une même région de Levy ($IVQ = 0$) ce qui pourrait traduire le caractère intrinsèquement désordonné des terminus. La fluctuation la plus fréquemment observée est celle entre le rebord de l'interface et la surface de la protéine (résultats non présentés). Cependant, la RMN n'est applicable que pour les petites protéines. La moyenne de l'IVQ par intervalle de nombre de résidus a été étudiée (Figure 9B). L'IVQ semble être fortement dépendant de la taille de la protéine : il décroît rapidement avec la taille des protéines. Il est donc possible de parler d'une certaine stabilité dans le positionnement des terminus pour les protéines de taille élevée. Il est également intéressant de remarquer que le C-terminus semble caractérisé par une plus grande variabilité.

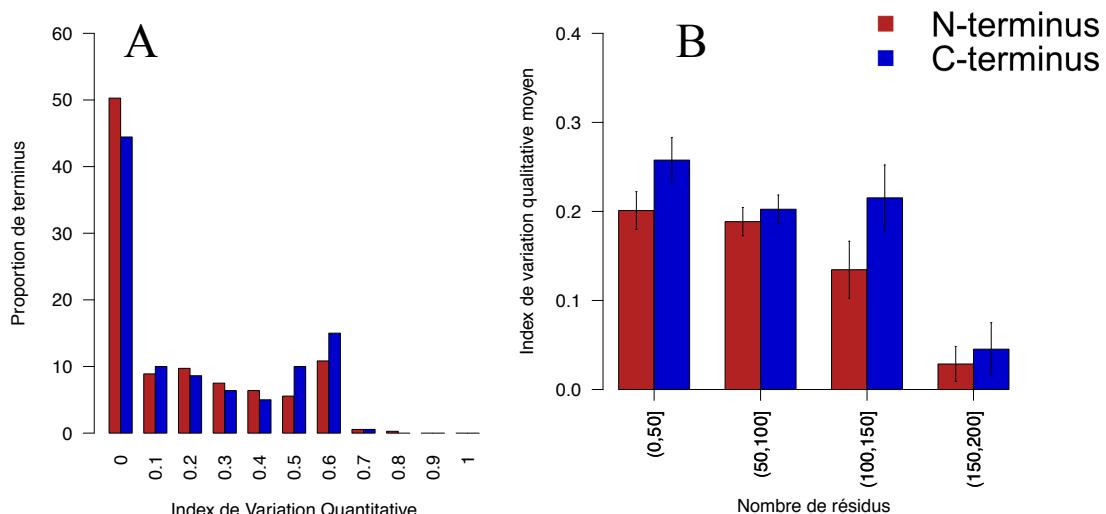


Figure 9. Variation du positionnement des terminus dans différentes conformations de complexes protéine-protéine. (A) Proportion de terminus ayant un index de variation qualitative (IVQ) donné. L'IVQ a été arrondi à la décimale près. (B) Moyenne de l'IVQ par intervalle de taille protéique.

3.3.2 Conformères libres et liés

Afin de considérer le mouvement des terminus au cours de la complexation, la base de données Protein-Protein Docking Benchmark 4.0 (Hwang *et al.*, 2010) a été utilisée. Il a été possible de recueillir des données concernant 66 hétérodimères biologiques (264 terminus).

Nous définissons une transition comme étant un changement de région entre le complexe et la configuration obtenue par superposition des conformations libres sur le complexe comme illustré dans la Figure 10. On dénombre 22 transitions parmi les 264 terminus (8%). Certains terminus changent de région après un déplacement très faible (c'est-à-dire en dessous de la résolution). Comme décrit dans les matériaux et méthodes, nous avons retiré ces terminus de l'analyse ce qui réduit le nombre de terminus à 74. On dénombre parmi ceux-ci 8 transitions de régions (11%). Ceci dépend encore une fois de la taille des protéines : aucune transition n'a été mise en évidence pour des protéines de plus de 250 résidus (résultats non présentés). Ces données semblent encore indiquer une certaine stabilité des terminus en regard des différentes régions.

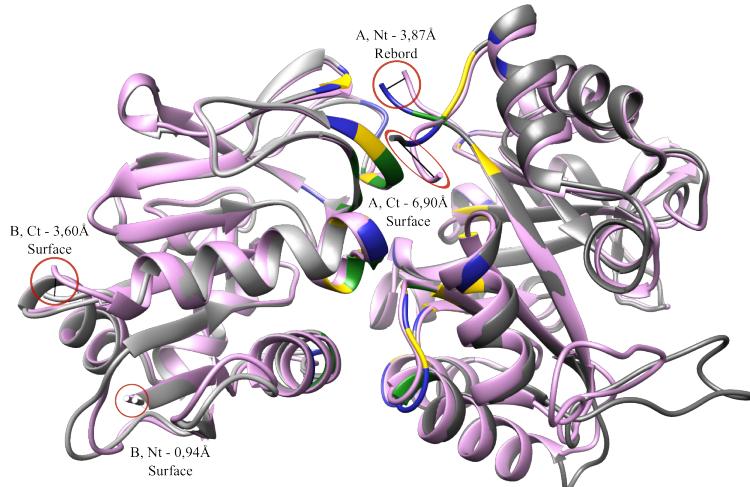


Figure 10. Mouvement des terminus au cours de la complexation. Exemple de l'imidazole glycerol phosphate synthase de *Thermotoga maritima*. Le complexe sous sa conformation libre (PDB : 1THF et 1K9V) est aligné structurellement avec le complexe sous sa conformation liée (PDB : 1GPW). Le complexe libre est coloré en violet. Le complexe lié a son interface colorée tandis que la surface et l'intérieur sont en gris. Le rebord est coloré en bleu, le cœur en vert et le support en jaune. Les distances indiquées correspondent à la distance entre les terminus des conformations libres et liées. Le terminus B Nt a été exclu de l'analyse car son déplacement en inférieur à la résolution (2,40 Å). Aucune transition n'est présente dans cet exemple.

3.4 Terminus et étiquettes

3.4.1 Terminus privilégié pour la fixation d'une étiquette

Nous avons cherché à déterminer si les étiquettes étaient plus souvent fixées sur un terminus que sur un autre. Pour cela, les séquences (SEQRES) de la PDB et IntAct ont été analysées. Il est important de remarquer que les expériences que ces bases de données répertorient utilisent des étiquettes très différentes. On retrouvera des étiquettes d'affinité dans la PDB alors que IntAct concerne une variété d'étiquettes allant de l'étiquette épitopique au domaine protéique.

Une recherche d'expression régulière comme décrit dans les matériaux et méthodes a été réalisée sur l'ensemble des séquences de la PDB (Figure 11). A l'exception de c-myc, il semblerait que le N-terminus soit légèrement privilégié. Ces différences ne sont cependant pas significatives statistiquement en réalisant un test d'indépendance du χ^2 ($p = 0,13$). De plus, il est important de remarquer que cette analyse sonde les étiquettes qui n'ont pas été excisées avant l'étude structurelle et non pas l'ensemble des étiquettes utilisées pour la purification. Certaines entrées de IntAct signalent l'utilisation d'une étiquette pour détection des IPP. A partir des données recueillies, on dénombre 720 (55,6%) étiquettes en C-terminal contre 575 (44,4%) étiquettes en N-terminal. Ces données semblent nous indiquer une légèrement préférence pour le C-terminus. Ces résultats semblent en contradiction. Les différences d'utilisation des terminus sont cependant très faibles. Cela pourrait simplement traduire l'absence d'une position privilégiée par les expérimentalistes pour l'étiquetage.

3.4.2 Biais de méthodologie : une conséquence interactionnelle des étiquettes

Les différentes méthodes de détection des IPP peuvent être distinguées selon si elle nécessite ou non une étiquette (Tableau 4). La présence d'une étiquette à l'interface pourrait avoir une influence sur ces méthodes. Cette influence constituera un biais de méthodologie. Nous faisons l'approximation que les régions occupées par les résidus de l'étiquette sont identiques à la région occupée par le résidu terminal. Les méthodes de détection concernant les dimères biologiques d'InterEvol ont été extraites d'IntAct (Orchard *et al.*, 2014). La correspondance entre InterEvol et IntAct est décrite dans les matériaux et méthodes. Les méthodes de détection de 662 interactions entre 315 monomères ont ainsi été extraites. Le biais a été étudié en fonction de la nature des terminus présents à l'interface ainsi que de la taille des protéines composants les complexes dimériques (Figure 12). Un test d'indépendance du χ^2 met en évidence un biais global dans les deux cas ($p < 0,001$). La présence d'un terminus à l'interface a donc une influence sur les méthodes de détection des IPP. L'analyse en fonction de la nature des terminus montre une augmentation de l'utilisation des méthodes sans étiquettes pour les protéines ayant au moins un terminus à l'interface à l'exception de deux cas : N0 00 et N0 N0. L'analyse en fonction de la taille des protéines montre une augmentation de la proportion des protéines ayant au moins un terminus à l'interface détectées par des méthodes nécessitant au moins une étiquette. Cependant, l'analyse des résidus standardisés ne permet d'attribuer à aucune case, une contribution significative ($p > 0,01$). Il est donc impossible de formuler un profil type de protéine concernée par ce biais.

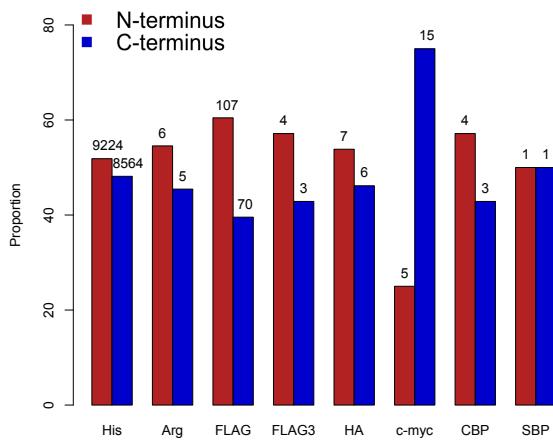


Figure 11. Etiquettes de la PDB. La proportion relative de fixation sur le N-terminus et C-terminus est tracée pour chaque étiquette. Le nombre de séquences correspondantes est présenté en haut de chaque barre. Les abréviations utilisées sont les suivantes : His pour His-tag, Arg pour Arg-tag, HA pour Hémagglutinine, CBP pour Calmodulin Binding Peptide et SBP pour Streptavidin Binding Peptide. Aucune séquence contenant la Histidine Affinity Tag n'a été détectée.

Ne nécessitent pas d'étiquette	Nécessitent au moins une étiquette
Anti bait coimmunoprecipitation	Adenylate cyclase complementation
Circular dichroism	Affinity technology
Classical fluorescence spectroscopy	Anti tag coimmunoprecipitation
Differential scanning calorimetry	Beta galactosidase complementation
Dynamic light scattering	Beta lactamase complementation
Electron microscopy	Bimolecular fluorescence complementation
Isothermal titration calorimetry	Cross-linking study
Molecular sieving	Dihydrofolate reductase reconstruction
Nuclear magnetic resonance	Fluorescent resonance energy transfer
Surface plasmon resonance	Lex-a dimerization assay
Surface plasmon resonance array	One hybrid
Transmission electron microscopy	Phage display
X-ray crystallography	Pull down
	Tandem affinity purification
	Tox-r dimerization assay
	Transcriptional complementation assay
	Two hybrid (et dérivés)

Tableau 4. Classification des méthodes de détection des IPP selon la nécessité d'une étiquette. Les noms utilisés sont issus de l'ontologie *Protein Structure Initiative Molecular Interaction* 2.5. L'ordre est alphabétique.

- Méthode ne nécessitant pas d'étiquette
- Méthode nécessitant au moins une étiquette

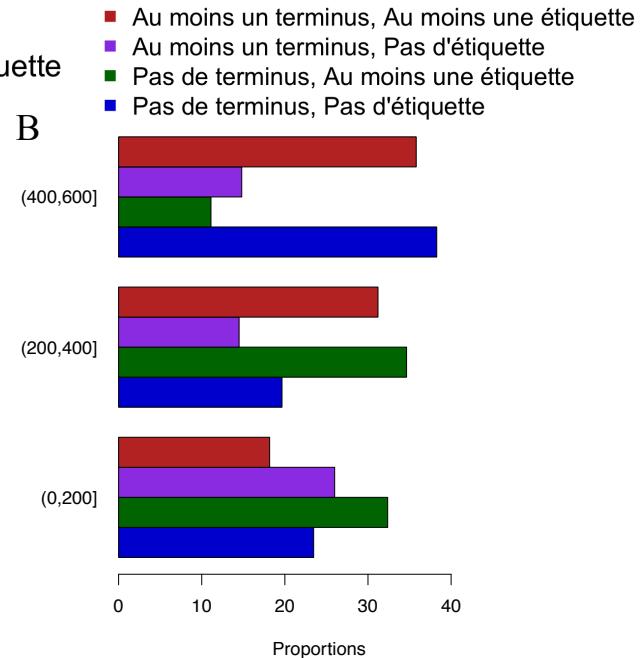
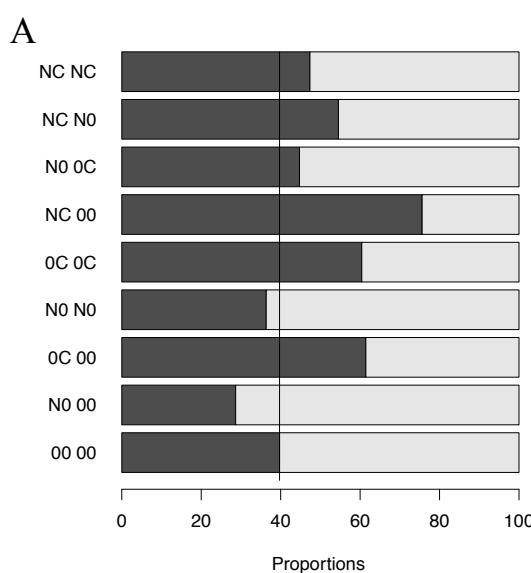


Figure 12. Proportion des méthodes utilisées pour détectées les IPP. Les proportions ont été normalisées à 100%. (A) Proportions en fonction des résidus terminaux présent à l'interface. « N » signifie N-terminus à l'interface, « C » signifie C-terminus à l'interface et « 0 » signifie absence de terminus à l'interface (par exemple, « N0 00 » signifie qu'une seule protéine a son N-terminus à l'interface. La droite représente la proportion du cas sans aucun terminus à l'interface (00 00). (B) Proportions en fonction du nombre de résidus total des monomères issus des complexes dimériques. Les proportions sont catégorisées en fonction de la présence ou absence de terminus à l'interface ainsi que de la nécessité ou non d'une étiquette dans la méthode de détection. Les monomères ont été comptés individuellement.

4. Discussion

Nous avons voulu étudier les relations entre les terminus et les interactions protéine-protéine. Nous avons d'abord procédé à une analyse structurale à partir d'une grande collection de structures de dimères extraits de la PDB. Un ensemble de précautions a été pris de manière à étudier les résidus terminaux des protéines natives. Nous avons retiré les anticorps qui possèdent systématiquement leur N-terminus à l'interface et les fichiers PDB qui ne possèdent pas de coordonnées pour les résidus terminaux. De plus, la présence d'étiquettes a été prise en compte de manière à ne pas confondre les résidus des étiquettes avec les résidus terminaux.

Nous montrons que les terminus occupent principalement des régions structurelles ayant maintenues une accessibilité au solvant même à l'état complexé. Les terminus se répartissent donc entre la surface de la protéine et une région périphérique de l'interface nommée rebord. Ceci est en accord avec des données de la littérature concernant les monomères qui indiquent que les terminus sont des résidus exposés (Jacob et Unger, 2004). La répartition de ces résidus dans les complexes implique une proportion non négligeable de terminus à l'interface. D'après nos observations, pour des monomères issus de complexes à interface biologique, plus de deux monomères sur dix ont un terminus particulier à l'interface et environ quatre monomères sur dix ont au moins un de leurs terminus à l'interface.

Nous avons ensuite cherché à savoir si ces proportions étaient le résultat d'un processus stochastique. Un modèle stochastique permettant d'estimer la distribution du nombre de protéines ayant un terminus à l'interface a donc été conçu. Les nombres observés de protéines pour les hétérodimères biologiques et les homodimères biologiques non-obligatoires étaient significativement différents de ces distributions ce qui pourrait traduire une implication fonctionnelle. Le phénomène mis en évidence est complexe et ces données sont difficiles d'interprétation. Il semble avoir un évitement des interfaces par les terminus des hétérodimères biologiques. Les terminus des homodimères obligatoires ont un comportement beaucoup moins homogène, et selon la taille de la protéine, les terminus préfèrent ou évitent les interfaces. Il est possible que les protéines profitent ou non des spécificités des terminus afin d'interagir selon un ensemble de facteurs qui caractérisent le terminus et l'interface de la protéine en question. Par exemple, le mécanisme *fly-casting* postule que le désordre intrinsèque constitue un avantage cinétique pour interagir avec d'autres partenaires protéiques (Shoemaker *et al.*, 2000). Le désordre intrinsèque étant fréquemment localisé au niveau des extrémités (Uversky, 2013), il ne serait pas étonnant d'observer une surreprésentation des terminus aux interfaces. A l'inverse, une étude de dynamique moléculaire sur trois complexes est arrivée à la conclusion que les interfaces sont composées de régions rigides (Kuttner et Engel, 2011). Cette rigidité pourrait constituer une pénalité entropique en contraignant conformationnellement les terminus flexibles. Un évitement des interfaces serait ainsi attendu.

Les résultats obtenus indiquent que les monomères constituants des complexes présentent chacun un C-terminus à l'interface plus souvent que n'importe quelle autre paire de terminus. Ceci est aussi bien valide pour les hétérodimères que pour les homodimères signifiant l'absence de lien avec la symétrie des constituants. De plus, cette situation ne semble pas être favorable sur le plan énergétique de part des charges identiques des terminus. Nous ne savons pas pour l'instant comment expliquer cette observation mais il se pourrait que cette paire de C-terminus signe un mécanisme interactionnel.

De manière à prendre en compte la flexibilité des protéines, nous avons également analysé les terminus dans différentes conformations de complexes protéine-protéine ainsi dans les conformations libres et liées des protéines. Cette flexibilité pourrait rendre l'assignation d'une région du complexe à un terminus peu pertinente car il pourrait exister une fluctuation non négligeable. Néanmoins, notre analyse montre une certaine fidélité à une région, en particulier pour les protéines les plus grandes. Le mouvement des terminus est donc contraint et les cas d'intercalation à l'interface sont rares.

Suite à notre caractérisation structurelle et dynamique, nous avons étudié si la localisation des terminus avait des conséquences sur la détection des interactions. Nous réussissons à mettre en évidence la présence un biais de méthodologie global pour les protéines ayant son ou ses terminus à l'interface. Les méthodes nécessitant des étiquettes sont moins performantes pour les complexes impliquant des terminus à l'interface à l'exception des cas où seul le ou les N-terminus sont présents. Cependant, aucune des catégories construites n'apporte une contribution statistique significative. Ceci pourrait être expliqué en partie par la classification des méthodes de détection. Celles étant classées comme « ne nécessitant pas d'étiquette » n'excluent pas la possibilité d'en utiliser une. De plus, la région occupée par le résidu terminal n'est pas nécessairement la même que celles des résidus de l'étiquette. Bien que la majorité des résidus des étiquettes mises en évidence dans InterEvol étaient au niveau de la surface, certains terminus auxquelles ces étiquettes étaient fixées étaient localisés à l'interface. La considération d'un segment de résidus terminaux à la place d'un unique résidu pourrait peut être mieux sonder l'environnement du terminus et ainsi mieux corrérer à la position de l'étiquette. Ceci est important car la région occupée par l'étiquette peut influencer la réussite du protocole. En effet, toutes les étiquettes d'affinité présentes dans InterEvol ont la majorité de leurs résidus en dehors de l'interface. Ces étiquettes doivent être considérées comme des cas de réussites car elles n'ont pas gêné la formation du complexe. Il est cependant important de remarquer que le choix du terminus à étiqueter ne peut se faire sur l'unique argument structural : les étiquettes peuvent avoir des conséquences au niveau de plusieurs propriétés de la protéine. Le choix du terminus doit donc prendre en compte un ensemble de contraintes qui dépendent de la méthodologie employée et des propriétés de la protéine. En absence d'informations permettant de choisir rationnellement, il semble préférable de toujours essayer les deux terminus.

Au cours de cette étude, les modifications post-traductionnelles n'ont pas été considérées. Cependant, les terminus sont des résidus accessibles ce qui pourrait faciliter ces modifications (Jacob et Unger, 2007) et ainsi avoir un effet sur les IPP en créant des sites d'interaction. L'utilisation des données issues d'une base de données comme dbPTM pourrait se révéler particulièrement intéressant. Notre analyse comporte également un biais structural provenant des structures cristallographiques. En effet, près de la moitié de ces terminus n'ont pas de coordonnées ce qui implique que notre analyse concerne principalement des terminus suffisamment contraint conformationnellement pour avoir leur position déterminée. Notre analyse du mouvement des terminus au cours de la complexation pourrait alors sous-estimer leur véritable flexibilité. Néanmoins, les résultats obtenus à partir des terminus dans différents modèles RMN semble rassurante à ce niveau. En outre, les différentes bases de données comportent peu d'information concernant l'utilisation des étiquettes. Dans le cas de la PDB, ces données ne sont pas standardisées. Ceci complique l'extraction automatisée des données. L'annonce systématique et standardisée de l'utilisation d'une étiquette dans une expérience à l'aide d'un vocabulaire contrôlé pourrait être grandement utile pour mieux étudier les conséquences des étiquettes.

5. Conclusion

Nous avons caractérisé les terminus au sein des complexes protéine-protéine et nous avons montré que leur localisation peut influencer la détection des interactions. Ces résidus sont en effet fréquemment au niveau d'une région périphérique de l'interface. Nous avons également conçu un modèle stochastique de l'interfaçage des terminus qui indique que ces résidus ne préfèrent ou n'évitent pas systématiquement les interfaces. Néanmoins, nous obtenons un biais vers un évitement dans les hétérodimères et une préférence dans les homodimères ainsi que des comportements différents pour le N-terminus et le C-terminus. Nous révélons que les monomères constituants des complexes présentent chacun un C-terminus à l'interface plus souvent que n'importe quelle autre paire de terminus. Par ailleurs, nous montrons que les terminus, malgré leur flexibilité, restent fidèles à une même région du complexe aussi bien au cours de la complexation qu'au sein du complexe. Cette information est notable car elle indique qu'il est possible de prédire la localisation des résidus terminaux par rapport aux interfaces d'après les structures isolées. De plus, elle autorise à assigner une région particulière du complexe à un terminus. Finalement, nous démontrons que la présence d'un terminus à l'interface peut influencer les méthodes de détection des interactions nécessitant des étiquettes. L'ensemble de ces données fournit un jeu de données préparé à l'étude des terminus et permet de mieux définir les conditions d'application des étiquettes. Elles sont d'intérêts à toute personne souhaitant développer un algorithme prédisant les conséquences des étiquettes et permettant de choisir rationnellement le terminus à étiqueter.

6. Références Bibliographiques

- Alexandrov N. Structural argument for N-terminal initiation of protein folding. *Protein Sci.* (1993) 2:1989–1991
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B (Methodological)* (1995) 57:289-300
- Berggård T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. *Proteomics.* (2007) 7:2833-2842.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research* (2000) 28:235-242
- Bonetta L. Protein-protein interactions: Interactome under construction. *Nature.* (2010) 468:851-854
- Bucher MH, Evdokimov AG, Waugh DS. Differential effects of short affinity tags on the crystallization of Pyrococcus furiosus maltodextrin-binding protein. *Acta Crystallogr D Biol Crystallogr.* (2002) 58:392-397
- Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. *Proteins* (2002) 47:334-343
- Christensen B, Kläning E, Nielsen MS, Andersen MH, Sørensen ES. C-terminal modification of osteopontin inhibits interaction with the $\alpha_V\beta_3$ -integrin. *J Biol Chem.* (2012) 287:3788-3797

- Christopher JA, Baldwin TO. Implications of N and C-terminal proximity for protein folding. *J Mol Biol.* (1996) 257:175-187
- Derewenda ZS. The use of recombinant methods and molecular engineering in protein crystallization. *Methods.* (2004) 34:354-363.
- Faure G, Andreani J, Guerois R. InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic Acids Res* (2012) 40:D847-856.
- Gabriel A, Przybylski J. Sickle-cell anemia: A Look at Global Haplotype Distribution. *Nature Education* (2010) 3:2
- Goh CS, Milburn D, Gerstein M. Conformational changes associated with protein-protein interactions. *Curr Opin Struct Biol.* (2004) 14:104-109
- Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. *Proteins.* (2010) 78:3111-3114
- Hubbard SJ, Thornton JM. 'NACCESS'. Computer Program. Department of Biochemistry and Molecular Biology, University College London. (1993)
- Jacob E, Unger R. A tale of two tails: why are terminal residues of proteins exposed? *Bioinformatics* (2007) 23:225–230.
- Jones S, Thornton JM. Principles of protein-protein interactions. *Proc Natl Acad Sci USA* (1996) 93:13-20
- Kuttner YY, Engel S. Protein hot spots: the islands of stability. *J Mol Biol.* (2012) Jan 415:419-428.
- Levy ED. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J Mol Biol* (2010) 403:660–670
- Martin AC. Mapping PDB chains to UniProtKB entries. *Bioinformatics.* (2005) 21:4297-4301
- Orchard S, Ammari M, Aranda B, Breuza L, Brigandt L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, Del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, van Roey K, Cesareni G, Hermjakob H. The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* (2014) 42:358-63
- Pal D, Chakrabarti P. Terminal residues in protein chains: Residue preference, conformation, and interaction. *Biopolymers* (1999) 53:467-47
- Palmer E, Freeman T. Investigation into the use of C- and N-terminal GFP fusion proteins for subcellular localization studies using reverse transfection microarrays. *Comp Funct Genomics.* (2004) 5:342-353.

- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* (2004) 25:1605-1612
- Shoemaker BA, Portman JJ, Wolynes PG. Speeding molecular recognition by using the folding funnel: The fly-casting mechanism. *Proc Natl Acad Sci USA* (2000) 97:8868-8873
- Terpe K. Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl Microbiol Biotechnol* (2003) 60:523–533
- Uversky VN. The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. *FEBS Lett* (2013) 587:1891-1901
- Varshavsky A. The N-end rule pathway of protein degradation. *Genes Cells.* (1997) 2:13-28.
- Wilcox AR. Indices of qualitative variation. (1967)
- Zhu H, Domingues FS, Sommer I, Lengauer T. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* (2006) 7:27

7. Hyperliens

7.1 Bases de données

IntAct : <http://www.ebi.ac.uk/intact/>

InterEvol : <http://biodev.cea.fr/interevol/>

PDB : <http://www.rcsb.org/pdb/>

PDB/Uniprot Mapping : <http://www.bioinf.org.uk/pdbsws/>

Protein-Protein Docking Benchmark : <http://zlab.umassmed.edu/benchmark/>

7.2 Logiciels et Langages de Programmation

Chimera : <http://www.cgl.ucsf.edu/chimera/>

Python : <http://www.python.org/>

R : <http://www.R-project.org/>

Master 2 Recherche Biochimie Structurale et Fonctionnelle
Université Claude Bernard Lyon 1
2013 – 2014

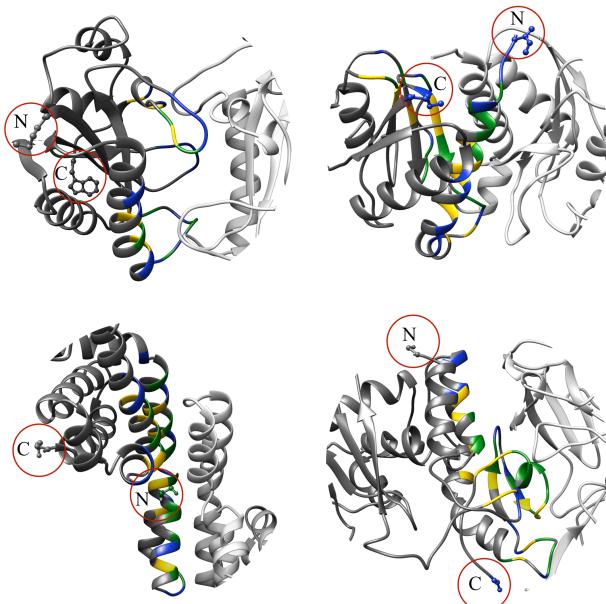
Interactions Protéine-Protéine - Implication des Résidus Terminaux
Olivier MARTIN

Bases Moléculaires et Structurales des Systèmes Infectieux - UMR 5086
Bioinformatique : Structures et Interactions
Juliette MARTIN

Les protéines agissent rarement seules. Les interactions protéine-protéine permettent la formation de complexes macromoléculaires qui sont essentiels pour les fonctions cellulaires comme la transduction du signal. Les résidus terminaux des protéines possèdent des propriétés particulières par leur charge et leur flexibilité structurale. Pour les besoins expérimentaux, ils sont souvent modifiés par la fixation d'une séquence peptidique nommée étiquette qui peut avoir des conséquences sur les propriétés des protéines.

Nous avons caractérisé les résidus terminaux dans les complexes dimériques, leur présence dans les interfaces protéine-protéine ainsi que les conséquences interactionnelles des étiquettes. Les caractéristiques structurales des résidus terminaux ont été étudiées à partir d'une grande collection de dimères issus de la Protein Data Bank. Les caractéristiques dynamiques ont été étudiées sur un ensemble de dimères obtenus par RMN, ainsi qu'un ensemble de structures de dimères pour lesquelles les structures des protéines isolées sont disponibles. Les conséquences interactionnelles des étiquettes ont été étudiées en croisant les données structurales avec les données issues d'expériences de détection des interactions protéine-protéine de la base de données IntAct.

Nos résultats indiquent que quatre protéines sur dix ont au moins un résidu terminal au niveau de l'interface. Les résidus terminaux ne semblent pas fortement préférés ou évités dans les interfaces. Néanmoins, nous obtenons un biais vers un évitement dans les hétérodimères et une préférence dans les homodimères ainsi que des comportements différents pour les extrémités N-terminus et C-terminus. De plus, malgré la flexibilité des terminus, les résidus terminaux restent généralement fidèles à une même région de la protéine. Finalement, nous montrons que la présence d'un résidu terminal à l'interface peut influencer l'efficacité des méthodes de détection nécessitant des étiquettes. L'ensemble de ces données fournit un jeu de données préparé à l'étude des terminus et permet de mieux définir les conditions d'application des étiquettes.



Résidus terminaux aux interfaces protéine-protéine