Développement d'une méthode in silico pour caractériser le potentiel d'interaction des surfaces protéiques dans un environnement encombré

Hugo Schweke

université PARIS-SACLAY

NNT: 2018SACLS554

Thèse de doctorat

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Développement d'une méthode *in silico* pour caractériser le potentiel d'interaction des surfaces protéiques dans un environnement encombré

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud

École doctorale n°577
SDSV : Structure et Dynamique des Systèmes Vivants
Spécialité de doctorat: biochimie et biologie structurale

Thèse présentée et soutenue à Orsay, le 13 décembre 2018, par

Hugo Schweke

Composition du Jury :

Philippe Minard Professeur, Univ. Paris-Sud (I2BC)	Président du jury
Frédéric Cazals Directeur de recherche, INRIA	Rapporteur
Emmanuel Levy Chercheur, Institut Weizmann	Rapporteur
Marc Baaden Directeur de recherche, CNRS (IBPC)	Examinateur
Alexandre de Brevern Directeur de recherche, INSERM (DSIMB)	Examinateur
Raphaël Guerois Chercheur CEA, CEA (I2BC)	Examinateur
Marie-Hélène Mucchielli-Giorgi Maître de conférence, Univ. Paris-Sud (I2BC)	Directrice de thèse
Anne Lopes Maître de conférence, Univ. Paris-Sud (I2BC)	Co-directrice de thèse

Remerciements

Ce travail est le fruit de nombreuses semaines de travail, de plusieurs dizaines de litres de café, ainsi que d'un certain nombre de samedis et de dimanches après-midi passés au laboratoire. Énormément de temps passé devant un écran. Beaucoup de stress aussi, mais au final une expérience très stimulante.

Après avoir passé plus de trois ans dans cette équipe, cela va faire bizarre de partir. Je me dois de remercier de nombreuses personnes. Pour commencer, tout particulièrement et en toute évidence, je remercie du fond du cœur Anne et Marie-Hélène. Elles ont vraiment tout fait pour que l'on en soit là ! Depuis mon arrivée en stage de master jusqu'à la fin de ma thèse j'ai toujours pu compter sur elles. Nous avons toujours pu échanger en toute confiance et en toute sincérité et elles ont toujours su montrer une grande disponibilité, mais aussi une grande patience à mon égard. Je me souviendrai toujours des longues discussions que j'ai eues avec Anne, que ce soit à propos de mon sujet de thèse, sur la science en général, ou même d'autre chose. Être sous la direction de quelqu'un d'aussi passionné, ce n'est pas toujours de tout repos, mais ça n'a pas de prix ! Honnêtement je ne pourrai jamais les remercier assez toutes les deux pour tout ce qu'elles m'ont apporté pendant ces trois ans. J'ai énormément appris scientifiquement mais aussi humainement et j'ai probablement vécu l'une des expériences les plus importantes de ma vie avec cette thèse.

Je remercie aussi tout les membres de l'équipe Bioinformatique Moléculaire, les anciens aussi bien que les nouveaux pour les moments passés pendant ces trois ans. Au final je retiens énormément de bons moments, notamment les nombreux (mais très irréguliers) footings !

Je remercie également les membres de mon jury de thèse pour avoir accepté d'évaluer mon travail. Je remercie Emmanuel Levy et Frédéric Cazals d'avoir accepté d'être rapporteurs de ma thèse. Je remercie Alexandre de Brevern et Philippe Minard pour avoir accepté d'être respectivement examinateur et président du jury. Et je remercie Raphaël Guerois et Marc Baaden pour avoir accepté d'être examinateurs de ma thèse mais aussi membres de mon comité de thèse, et donc d'avoir suivi l'avancée de mes travaux pendant ces trois ans.

Enfin, pour terminer, je remercie mes parents ainsi que l'ensemble de ma famille pour leur soutien, ainsi que tous mes amis, les anciens aussi bien que les nouveaux, pour tout ce qu'ils m'ont apporté.

Sommaire

Liste des abréviations

2D : Deux Dimensions

3D : Trois Dimensions

ACC : *Accuracy*

AUC : *Area Under the Curve*

ADN : Acide DésoxyRibonucléique

ARN : Acide RiboNucléique

ASA : *Accessible Surface Area*

CM : Centre de Masse

CV : *Circular Variance*

DSH : Différence Significative Honnête

FFT : *Fast Fourier Transform*

FN : *False Negative*

FP : *False Positive*

IP : *Interaction Propensity*

IPOPS : *Interaction Propensity of Protein Surface*

NIP : *normalized interaction propensity*

PDB : *Protein Data Bank*

PPV *: Positive Predictive Value*

RMN : Résonance Magnétique Nucléaire

ROC : *Receiver Operating Characteristics*

RSA : *Relative Solvent Accessibility*

SENS : Sensibilité

SPE : Spécificité

TAP : *Tandem Affinity Purification*

TN : *True Negative*

TP : *True Positive*

Y2H : *Yeast Two-Hybrid*

1 La biologie des protéines

1.1 Les protéines: historique et premières découvertes

L'étude des protéines date de plus de deux siècles. La première mise en évidence de protéines en tant que molécules biologiques fut réalisée en 1789 par Antoine Fourcroy, comte de Fourcroy. Il distingua trois types de protéines : l'albumine, la fibrine et la gélatine. Le terme "protéine" fut proposé en 1838 par le chimiste suédois Jöns Jacob Berzelius dans une lettre au chimiste hollandais Gerardus Johannes Mulder, puis employé et popularisé par ce dernier (1). Ce mot dérive du grec ancien *prôteíon* (*πρωτεῖον*) que l'on peut traduire par "prééminent" ou "tout premier", Berzelius étant convaincu que ces molécules possédaient un rôle majeur dans l'organisation du vivant.

La mise en évidence des 20 acides aminés constitutifs des protéines (excluant les acides aminés spéciaux que sont la sélénocystéine et la pyrrolysine) mit presque 130 ans, entre la découverte de l'asparagine en 1806 par Louis-Nicolas Vauquelin et Pierre Jean Robiquet (2), et celle de la thréonine en 1935 par Curtis Meyer et William Cumming Rose (3). Parallèlement à ces découvertes, le modèle des protéines comme étant des molécules constituées d'acides aminés reliés entre eux par des liaisons peptidiques fut proposé de manière indépendante par Emil Fisher et Franz Hofmeister en 1902. La notion de séquence primaire des protéines était ainsi née. Cependant il fallut attendre 1949 et le séquençage de la structure primaire de l'insuline bovine par Frederick Sanger (4) pour qu'il soit possible de déterminer les séquences protéiques.

Après ces premiers résultats les découvertes s'accélérèrent. Grâce à l'essor de la cristallographie la première structure cristallographique d'une protéine, la myoglobine de cachalot, fut élucidée en 1958 par John Kendrew et al (5). Cette avancée fut suivie de près par l'élucidation de la structure de l'hémoglobine par Max Perutz en 1959 (6). Dans les années qui suivirent plusieurs structures de protéines furent obtenues par cristallographie : le lysozyme de poulet en 1965 (7), les ribonucléases A et S en 1967 (8,9), la papaïne (10) et l'hémoglobine de cheval en 1968 (11), l'insuline en 1971 (12). Une base de donnée appelée la PDB (*Protein Data Bank*) (13) fut créée en 1971 dans le but de répertorier les structure de macromolécules biologiques élucidées. À

l’origine elle ne comptait que sept structures, mais ce nombre augmenta rapidement, pour finalement atteindre plus de 143 000 structures moléculaires biologiques en août 2018.

1.2 Des protéines aux interactions protéiques

Malgré les progrès réalisés dans la compréhension des protéines en tant que molécules biologiques, leurs rôles dans la cellule n’était pas encore compris (14). Le fait que les protéines interagissent entre elles a été observé bien avant la découverte des premières structures protéiques. Par exemple l’inhibition de la trypsine (même si la nature de cette dernière n’était pas encore connue) est connue depuis 1906 avec les travaux de Hedin (15). A la fin des années 20, Svedberg développa l’ultracentrifugation. A l’aide de cette méthode révolutionnaire, il put calculer les masses moléculaires de nombreuses protéines en déterminant leur vitesse de sédimentation (16). En travaillant sur l’hémocyanine, il observa plusieurs masses moléculaires pour cette même protéine (16,17), chacune de ces masses moléculaires étant un multiple de la plus petite des masses observées. Il fit alors l’hypothèse que chacune de ces masses moléculaires représentait un complexe formé par l’assemblage de plusieurs protéines d’hémocyanine. Ces résultats sont ainsi souvent considérés comme la découverte de l’existence de la structure quaternaire des protéines (18).

Dans les décennies qui suivirent plusieurs expériences permirent de mettre en évidence l’importance des interactions protéiques dans certains mécanismes cellulaires (14), par exemple la nature multimérique de nombreuses enzymes impliquées dans des réactions clés du métabolisme cellulaire, comme dans le cas du complexe α-cétoglutarate déshydrogénase (19). Celui-ci est en effet constitué de trois enzymes, une décarboxylase, une acyltransférase et une oxydo-réductase, chacune étant impliquée dans une étape précise de la voie métabolique catalysant la transformation de l’acide α-cetoglutarique en Succinyl-CoA, une étape clé du cycle de Krebs (20). Durant cette période d’autres découvertes mirent en évidence le rôle des interactions protéiques dans les processus cellulaires, montrant à chaque fois leur rôle primordial dans la biologie de la cellule. Cependant toutes ces expériences et observations ne permettaient d’observer le rôle des interactions protéiques que dans des cas spécifiques. Celles-ci étaient étudiées dans le cadre de machineries protéiques bien délimitées, comme dans les cas de complexes enzymatiques ou de l’hémoglobine. Il fallut encore du temps pour mesurer le rôle des interactions protéiques dans le fonctionnement global de la cellule.

Les années 70 et 80 virent l'essor de nouvelles technologies de biologie moléculaire telles que l'électrophorèse sur gel de polyacrylamide (21) ou le clonage moléculaire (22). Ces avancées méthodologiques facilitèrent grandement l'étude des interactions protéiques, et permirent ainsi de démontrer leur rôle essentiel dans un grand nombre de processus biologiques, tels que les cascades de phosphorylations (23,24) ou encore la division cellulaire (14). Quelques soient les processus étudiés les protéines et leurs interactions semblaient jouer un rôle important. À cette époque une vision plus générale des interactions protéiques et de leur caractère ubiquitaire dans la cellule commença à émerger (25).

La fin des année 80 vit d'importants progrès dans le domaine de la spectrométrie de masse ainsi que la publication d'une méthode de détection des interactions protéine-protéine qui allait révolutionner la prédiction d'interactions protéiques: le criblage par méthode double-hybride (Y2H) (26). Avec ces méthodes, les biologistes disposaient d'outils pour évaluer les interactions protéiques à grande échelle. Mais ce fut à partir du début des années 1990, avec l'apparition des méthodes de séquençage haut-débit que les travaux de détection d'interactions protéiques à grande échelle furent possibles. L'essor de ces méthodologies eut un effet primordial sur l'étude des interactions protéiques. Avec ces nouveaux outils, il devenait possible d'obtenir le séquençage de génomes entiers, et par la même occasion, la quasi-totalité des gènes codants pour les protéines d'un organisme. Ces innovations ont permis le perfectionnement des méthodes de prédiction d'interactions à haut débit telles que les méthodes de Y2H et de spectrométrie de masse et ont rendu possible leur utilisation à grande échelle. C'est au début des années 2000 que furent réalisés les premiers travaux de caractérisation de l'interactome complet d'un organisme modèle chez *S. cerevisiae,* par méthode de double-hybride (27–29) et par spectrométrie de masse (30,31). Ces interactomes ont permis de recenser des milliers d'interactions protéiques mais de manière inattendue, seul un recouvrement faible était observé entre les différentes expériences (32). Depuis lors, d'autres réseaux d'interactions ont été réalisés avec des méthodes plus efficaces, comme par exemple dans (33–37). Un grand nombre de méthodes expérimentales (38) et bioinformatiques (39–41) ont été développées et perfectionnées pour la cartographie des réseaux d'interaction.

Ces résultats montrent toute la complexité de la biologie de la cellule. Les difficultés rencontrées sont dues à la nature même des interactions protéiques, qui ont lieu dans un milieu spécifique (l'environnement cellulaire) à une concentration donnée, chose qu'il est difficile de représenter dans un réseau d'interactions binaires. De plus l'information structurale, c'est à dire de quelle manière les protéines interagissent, est essentielle pour avoir une vue d'ensemble des interactions protéiques à l'échelle d'un organisme (42).

Il ressort de ces deux siècles de travaux sur les protéines et leurs interactions que celles-ci sont présentes dans quasiment tous les processus de la cellule et sont indispensables à son fonctionnement. Les protéines représentent une classe de macromolécules extrêmement diverse, avec des fonctions bien distinctes dans tous les mécanismes cellulaires : structuration de la cellule, signalisation ou encore transcription de l'ADN et traduction de l'ARN. De plus les interactions protéiques sont dynamiques et dépendent de leur localisation et de leur concentration cellulaire. Cette grande complexité multifactorielle rend leur caractérisation difficile à appréhender à grande échelle. Comprendre les mécanismes survenant dans les interactions protéine-protéine est donc une étape nécessaire pour comprendre leur rôle dans la cellule. Dans la partie suivante nous allons caractériser plus en détail les propriétés structurales des interactions protéine-protéine.

1.3 Caractéristiques des interactions protéiques fonctionnelles

Les interactions protéiques se font à travers des interactions non covalentes entre deux protéines ou plus. La stabilité de ces interactions dépend des propriétés biologiques des protéines mises en jeu ainsi que des propriétés de leur environnement (pH, température...).

Nous allons nous intéresser plus précisément aux propriétés physiques et chimiques des interactions protéiques. Du fait de leur importance majeure, il existe une riche bibliographie sur les caractéristiques des interfaces protéiques. Je rappellerai ici les grandes lignes concernant les différents types d'interfaces protéiques, ainsi que leurs propriétés physico-chimiques et évolutives.

1.3.1 Les types d'interactions protéiques

Plusieurs types d'interactions protéiques ont été définis dans la littérature. On peut catégoriser les interactions selon plusieurs critères. Il existe par exemple des classifications basées sur la stabilité des complexes (43,44). On fait ainsi fréquemment la distinction entre interactions permanentes et interactions transitoires, ou encore entre interactions obligatoires ou non-obligatoires. On peut aussi classer les interactions protéiques selon la nature de leurs composants. On fait souvent la distinction entre interactions homo-oligomériques et hétéro-oligomériques. Cette

classification fait la distinction entre une interaction impliquant des monomères identiques et une interaction entre monomères différents.

Il est important de noter que ces catégories ne sont pas mutuellement exclusives. Par exemple les interactions permanentes sont généralement considérées comme obligatoires, mais peuvent aussi être non obligatoires (43). De plus la stabilité d'une interaction protéique dépend des conditions dans lesquelles celle-ci s'opère : Une interaction peut être considérée comme transitoire sous certaines conditions et permanente sous d'autres (y compris différentes conditions cellulaires).

Nous allons voir un peu plus précisément quels sont les détails et les particularités de ces différentes catégories.

1.3.1.1 Classifications basées sur la stabilité des interactions

Interactions permanentes et transitoires

Les protéines engagées dans des interactions permanentes présentent une affinité de liaison très forte. Ces interactions sont généralement très stables et irréversibles. Les monomères (sous-unités de base d'un complexe protéique multimérique) liés par ce type d'interactions ne sont généralement jamais retrouvés sous forme monomérique *in vivo (43)*. Ces interactions sont souvent nécessaires à la fonction des protéines concernées. Elles sont très courantes dans la cellule dans le cas de machineries protéiques complexes telles que le protéasome ou le ribosome.

Les interactions transitoires recouvrent une gamme très variable d'affinités de liaison (45,46). De ce fait on fait généralement une distinction entre les interactions transitoires faibles et les interactions transitoires fortes (43,46). Les premières consistent en des interactions très brèves avec une constante de dissociation de l'ordre de la micromole et une durée de quelques secondes (46). Les secondes sont plus stables avec une constante de dissociation de l'ordre de la nanomole et une durée de vie généralement plus longue (43,46).

Il est important de souligner que les affinités de liaison des milliers d'interactions protéiques ayant lieu dans la cellule peuvent couvrir 12 ordres de grandeur, avec tout un éventail de valeurs entre les deux extrêmes (45).

Interactions obligatoires et non-obligatoires

Les monomères engagés dans une interaction obligatoire ne sont pas retrouvés sous forme stable et repliée sous forme monomérique. Ces interactions sont généralement nécessaires à la fonction des protéines concernées. Il existe cependant des cas d'interactions permanentes non-obligatoires. A l'opposé, les protéines formant des interactions non-obligatoires sont stables sous forme monomérique dans la cellule. Ce type d'interaction est très courant, elles sont notamment très fréquentes dans les voies de signalisations de la cellule qui nécessitent des signaux temporaires (47). Les interactions non-obligatoires sont souvent synonymes d'interactions transitoires, mais pas toujours.

1.3.1.2 Catégories basées sur la nature des protéines formant les interactions

Ces catégories sont basée sur le type de protéines formant les interactions. On distingue les interactions hétéro-oligomériques des interactions homomériques. Les interactions hétéro-oligomériques mettent en jeu des monomères de types différents. Un exemple connu est celui de la molybdoptérine synthase de *E. coli* (48) Figure 1.1A. Les interactions homo-oligomériques mettent en jeu des monomères de composition identique. Il s'agit d'un type d'interaction très répandu : en se basant sur les structures déterminées expérimentalement, on estime que 45 % des protéines d'organismes eucaryotes et 60 % des protéines d'organismes procaryotes forment des homomères dans des conditions physiologiques (49,50). Un exemple type d'interaction homo-oligomérique est celui de l'alcool déshydrogénase de classe 3 de *H. sapiens*, qui forme un dimère (51) (Figure 1.1B).

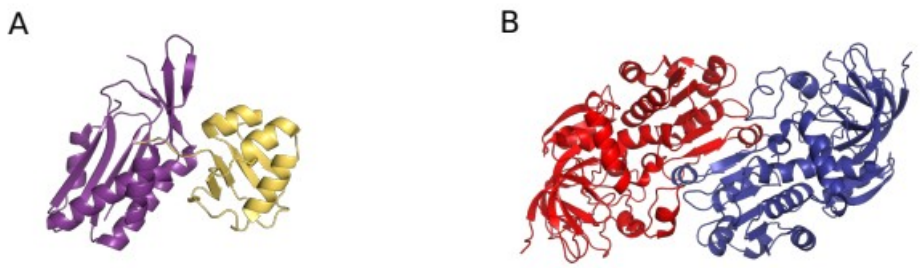


Figure 1.1. Exemple d'interaction hétéromérique et homomérique. (A) La molydoptérine synthase de *E. coli* (code pdb 1fm0) forme un hétérodimère. (B) L'alcool déshydrogénase de classe 3 forme un homodimère (code pdb 1m6h).

Les interfaces impliquées dans ces différents assemblages possèdent des propriétés chimiques et structurales particulières qui les différencient du reste de la surface des protéines. Nous allons brièvement décrire quelles sont les caractéristiques de ces interfaces dans la partie suivante.

1.3.2 Les interfaces protéiques

Avec la mise en évidence de l'importance des interactions protéiques dans les processus biologiques, de nombreuses études visant à caractériser les interfaces protéine-protéine ont été réalisées. Dès le milieu des années 70, Chothia et Janin analysaient dans des travaux pionniers (52,53) les structure de plusieurs complexes : l'association entre la trypsine bovine et l'inhibiteur de trypsine pancréatique, la trypsine porcine liée à l'inhibiteur de trypsine du soja, l'insuline et l'hémoglobine. Ces premières analyses révélèrent que l'enfouissement de résidus hydrophobes est le facteur majeur de la stabilité des complexes protéiques. Depuis ces premiers travaux, il y a eu une multitude d'études visant à caractériser les interfaces protéine-protéine. Pour les travaux majeurs se référer à (53–61), ainsi qu'aux revues et chapitres de livres suivants (43,49,62–65). Ces analyses ont mis en évidence plusieurs propriétés structurales et biophysiques des interfaces protéiques.

1.3.2.1 Comment définir les interfaces protéiques ?

Il existe plusieurs définitions des surfaces protéiques. Parmi les critères couramment employés on distingue ceux basés sur les distances atomiques ou sur l'accessibilité au solvant :

- distance inter-atomique : soit deux protéines A et B formant un complexe. Un résidu de la protéine A est considéré comme faisant partie d'une interface avec la protéine B si au moins un de ses atomes se trouve à un distance inférieure à un seuil (souvent 5Å) d'un atome d'un résidu de la protéine B (66,67). Ce critère de distance peut-être appliqué sur n'importe quelle paire d'atomes, ou être employé en ne prenant en compte que les paires de Cα ou de Cβ (le seuil de distance variant en conséquence) (67). Il est aussi possible de calculer les distances interatomiques en prenant en compte les rayons de van der Waals des atomes considérés. On peut par exemple fixer un seuil de distance de 0.5Å ajouté à la somme des rayons de van der Waals des deux atomes considérés (67).

- changement d'accessibilité au solvant : un résidu est considéré comme faisant partie d'une interface si son accessibilité au solvant diminue entre la forme complexée et la forme libre de la protéine. L'ASA représente l'aire de la surface d'une molécule qui est accessible au solvant et est

exprimé en Å^2. Ce concept a été introduit par Lee et Richards en 1971 (68). L'accessibilité au solvant est généralement calculée par l'algorithme de Shrake et Rupley (69) qui consiste à explorer la surface d'une biomolécule à l'aide d'une sphère de 1.4Å de rayon (approximativement le rayon d'une molécule d'eau), afin de déterminer l'accessibilité au solvant des résidus.

1.3.2.2 Caractéristiques structurales des interfaces

Les interfaces des complexes protéiques présentent une surface totale généralement comprise entre 1200 et 2000 Å^2 (même si dans de nombreux complexes la surface enfouie est supérieur à 2000 Å^2), pour une moyenne à environ 1600 Å^2 (59). Elles impliquent en moyenne entre 50 et 60 résidus répartis également entre les deux protéines (58,59). Ces interfaces sont très compactes au niveau atomique, avec généralement une grande complémentarité de forme entre les surfaces des protéines qui interagissent (59,70). Leur compaction est souvent comparable à celle de l'intérieur des protéines. Elles sont généralement relativement plates, bien que leur morphologie soit très variable (59). Les interfaces mises en jeu dans les interactions permanentes sont souvent plus grandes que les interfaces trouvées dans les interactions transitoires (64).

1.3.2.3 Caractéristiques physico-chimiques des interfaces

Les sites d'interaction se distinguent du reste de la surface des protéines au niveau de leur composition chimique, c'est-à-dire de par le type d'acides aminés qui les composent (acides aminés polaires, apolaires ou chargés). Les interfaces protéiques ont un caractère plus hydrophobe que le reste de la surface (58,71,72). L'hydrophobicité est l'une des propriétés majeures de la reconnaissance entre protéines, permettant à celles-ci d'engager des interactions entre elles. Les interfaces protéiques sont en moyenne enrichies en Met, Tyr et Trp tandis que l'on observe une forte déplétion en Glu, Lys et Asp (64).

Cependant il est aussi possible de séparer les interfaces en plusieurs catégories en fonction de leurs propriétés. Par exemple les interfaces d'interactions transitoires sont généralement moins hydrophobes et plus polaires que celles des interactions permanentes (43,56,60,62), bien que leur propriétés restent différentes du reste de la surface (62). Mintseris et Weng (61) ont montré que les résidus appartenant à des interfaces permanentes subissent une pression de sélection plus forte que ceux appartenant à des interfaces transitoires, et évoluent donc en moyenne plus lentement. Ils ont

tendance à coévoluer, ce qui est beaucoup moins le cas pour les résidus appartenant à des interfaces transitoires. Ofran et Rost (60) ont déterminé six types d'interfaces protéine-protéine en fonction de leurs compositions en acides aminés, et des fréquences de contacts entre types d'acides aminés : interfaces intra-domaine, domaine-domaine, homomériques obligatoires, homomériques non-obligatoires, hétéromériques obligatoires, hétéromériques non-obligatoires. Bien que d'après leur analyse chacun de ces quatre types d'interfaces possèdent des propriétés particulières, toutes ont en commun d'être clairement différentes par rapport au reste de la surface des protéines.

1.3.2.4 Les interfaces, des objets non homogènes

Tous les résidus d'une interface ne contribuent pas de la même manière à la stabilité de l'interaction. Il a été montré que seuls quelques résidus de l'interface contribuent généralement à la majeure partie de l'énergie d'interaction (73). Ces résidus essentiels à l'interaction sont appelés *hotspots*. On définit généralement comme *hotspots* tout résidu ayant un $\Delta\Delta G$ supérieur à 1 ou 2 kcal.mol^{-1}, c'est à dire tout résidu dont la mutation en alanine induit une perte de 1 à 2 kcal.mol^{-1} sur l'énergie totale d'interaction. Ces résidus sont généralement très conservés (74). Ils sont souvent localisés au centre de l'interface, où ils sont protégés du solvant par d'autres résidus (75). Les résidus jouant le plus souvent le rôle de *hotspots* sont Trp, Arg et Tyr tandis que Leu, Ser, Thr et Val jouent rarement ce rôle (73,76).

Les caractéristiques des interfaces protéiques ne sont donc pas uniformes sur toute leur surface. Lo Conte et al (59) ont proposé un modèle structural dans lequel ils faisaient la distinction entre le cœur et la bordure des interfaces (modèle *core/rim*). La composition du cœur est proche de celle de l'intérieur des protéines tandis que la composition de la bordure est plus proche de celle de la surface des protéines. Levy (77) a par la suite introduit dans ce modèle une partie appelée support dont la composition est similaire à celle du cœur hydrophobe des protéines (modèle *core/rim/support*, voir figure 1.2). Ce modèle à trois composantes maximise les spécificités chimiques de chacune de ces régions :

- le support (*support*) possède une composition similaire à celle de l'intérieur des protéines.

- la bordure (*rim*) possède une composition similaire à celle de la surface des protéines.

- le cœur (core) possède une composition plus proche de l'intérieur des protéines que de la surface, tout en restant intermédiaire entre les deux.

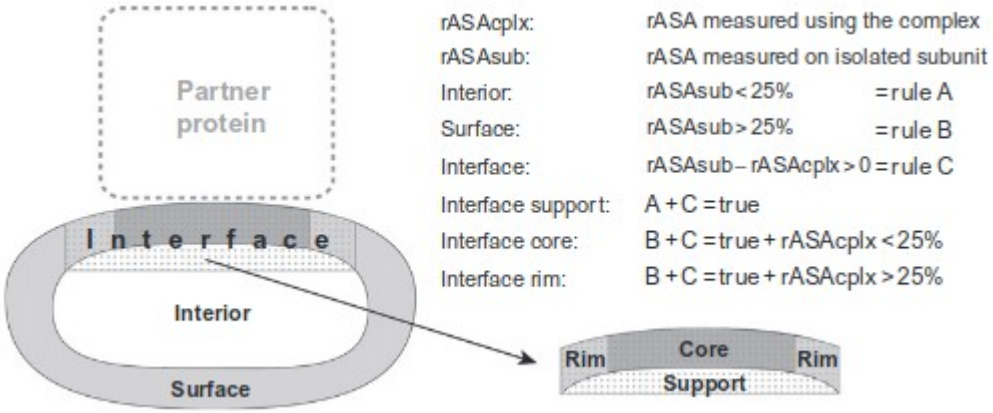


Figure 1.2. Modèle structurale à trois régions des interfaces protéiques. Selon ce modèle, les interfaces sont représentées par trois régions structurales bien définies : le *core*, le *support* et le *rim*. Ces régions sont définies en fonction de leur changement d'accessibilité au solvant après complexation (voir section 1.3.2.1). Figure extraite du chapitre de livre de Levy et Teichmann (49).

1.3.3 Evolution des interfaces protéiques

1.3.3.1 La conservation des interfaces protéiques

Il a été montré que deux protéines homologues partageant au moins 30% à 40% d'identité de séquence possèdent généralement une structure similaire (78). Aloy et al (79) et d'autres (80) ont analysé la conservation des modes d'assemblages des complexes formés de protéines homologues. On appelle ce type de complexes homologues des interologues, deux complexes AB et A'B' étant dits interologues si respectivement A et A' et B et B' sont homologues structuraux. Aloy et al ont montré que les couples de complexes interologues partageant plus de 30 % d'identité de séquence ont généralement des modes d'assemblage similaires. Ce résultat est très important et forme la base de la modélisation de complexes par homologie (42), comme nous l'évoquerons dans la partie 2.2.2. Leur analyse structurale de ces modes d'assemblage est basée sur le critère du rmsd, qui est une mesure de similarité globale.

Andreani et al. (81) sont allés plus loin et ont analysé dans le détail les caractéristiques et la conservation des différents types de contacts atomiques et des propriétés physico-chimiques des interfaces de complexes interologues. De manière surprenante, ils ont mis en évidence une grande plasticité au niveau des contacts atomiques, avec en moyenne seulement 60 % de conservation entre

complexes interologues. En plus de cela, en moyenne un quart des résidus présents dans l'interface d'un complexe ne font pas partie de l'interface d'un complexe interologue. Les résidus appartenant au *core* et au *support* (voir section 1.3.2.4) des interfaces sont beaucoup plus conservés dans les interfaces d'interologues que les résidus appartenant à la bordure des interfaces. Les auteurs montrent aussi que les ponts salins, les liaisons hydrogènes, les contacts entre résidus chargés sont peu conservés entre interfaces de complexes interologues.

Leur analyse montre que, bien que la structure globale des interfaces soit largement conservées entre interologues, il existe une étonnante plasticité des interfaces de ces complexes.

1.3.3.2 La co-évolution et la co-adaptation

Un autre point important à évoquer est la co-évolution et la co-adaptation des interfaces protéiques. La co-évolution réfère à l'évolution coordonnée entre deux organismes ou molécules. La co-adaptation réfère plus spécifiquement aux interactions entre deux partenaires protéiques (82). Elle relève du fait qu'une ou plusieurs mutations impactant la spécificité et/ou la stabilité d'un complexe peut, afin de maintenir l'interaction fonctionnelle, être compensée par une ou des mutations compensatoires chez le partenaire (83). Ce type de phénomène est très important car il permet aux protéines de maintenir des interactions fonctionnelles au cours de l'évolution (84).

1.3.3.3 Émergence de nouveaux sites d'interaction

La composition en résidus de la bordure des interfaces est similaire quelque soit le type d'interface considéré (77). C'est donc le cœur des interfaces qui leur confère leur spécificité physico-chimique. De plus, Levy a montré qu'en théorie il suffit de deux mutations ponctuelles en moyenne pour transformer une région de surface protéique d'une trentaine de résidus en une région avec les même propriétés physico-chimiques qu'un site d'interaction (77). Les implications de cette étude sont extrêmement importantes puisqu'elle montre que la barrière pour passer d'une surface protéique à un site d'interaction est en fait très mince. Un travail théorique récent a été réalisé par Tonddast-Navaei et Skolnick (85). Les auteurs montrent qu'une grande partie de la surface des protéines (les trois-quarts selon leurs estimations) possède des propriétés géométriques semblables à des sites d'interaction connus, et sont donc en théorie capable de former des interactions protéiques.

Jusqu'ici nous avons traité les protéines comme des objets biologiques isolés exerçant une fonction dans la cellule. Mais les protéines évoluent dans un environnement, que celui-ci soit le cytosol, la membrane cellulaire ou un organelle, ou encore un environnement extracellulaire. Comprendre les protéines et leurs interactions ne peut donc se faire sans comprendre comment celles-ci évoluent et s'organisent dans leur environnement. C'est ce que nous allons aborder dans le chapitre suivant.

1.4 Les protéines dans l'environnement cellulaire

1.4.1 La cellule, un environnement encombré

La cellule est composée en grande partie d'eau: entre 60 et 70% de sa masse totale chez la levure de boulanger *S. cerevisiae (86)*. Toujours chez la levure de boulanger, les protéines constituent entre 40 et 50% du poids sec de la cellule. Il s'agit du type de macromolécule le plus présent devant les ARN (environ 10%) et les lipides (environ 10%) (87). Ces quantités varient selon l'état de la cellule mais restent toujours de cet ordre.

Il a été calculé qu'au sein d'une cellule la densité de protéines pourrait atteindre 2 à 4 millions par μm^3 (88). Pour un organisme modèle comme la levure de boulanger, dont la cellule représente en moyenne un volume d'environ 42 μm^3 (89), cela représente entre 80 et 160 millions de protéines par cellule. Par ailleurs on estime que le protéome de la levure de boulanger est constitué d'environ 6000 protéines, ce qui signifie qu'en moyenne une protéine est représentée par 20000 copies par cellule. Cependant il a été montré que les niveaux d'expressions des protéines dans les cellules sont extrêmement variables selon le type de protéine et l'état de la cellule (90). Dans (90), les auteurs montrent par exemple que, chez la levure de boulanger, les deux tiers des protéines codées par l'organisme ont un nombre de copies par cellule compris entre 1000 et 10000, alors que les protéines les plus représentées peuvent atteindre $7.5x10^5$ copies par cellules, et les moins représentées aucune copie par cellule. Il y a donc une très grande diversité aussi bien dans le type des protéines que dans leur nombre d'exemplaires qui recouvre quatre ordres de grandeur. En plus des protéines viennent s'ajouter les nombreuses molécules non protéiques, telles que les acides nucléiques, les petits métabolites, les ions etc…

1.4.2 Les interactions fonctionnelles et non-fonctionnelles

La cellule est donc un environnement encombré dans lequel les protéines entrent constamment en contact entre elles, ainsi qu'avec d'autres molécules biologiques. Parmi ces nombreuses interactions certaines sont fonctionnelles et d'autres non-fonctionnelles. Ici nous considérons comme interaction fonctionnelle toute interaction ayant un effet positif sur le *fitness* de l'organisme. C'est au milieu d'un grand nombre d'interactions non-fonctionnelles que les protéines intracellulaires, et tout particulièrement les protéines cytosoliques, doivent assurer leurs fonctions à travers des interactions fonctionnelles. Pour le maintien de la cellule il est nécessaire que les protéines engagent des interactions avec leurs partenaires en quantité suffisante pour assurer leur fonction, bien que la plupart (ou au moins une grande partie) des interactions qu'elles engagent en circulant dans leur environnement se fassent de manière non-fonctionnelle. La compétition entre interactions fonctionnelles et non-fonctionnelles est donc omniprésente.

Plusieurs travaux visant à déterminer le comportement des protéines dans la cellule ont été réalisés. Par exemple Levy et al. (91) ont développé une méthode pour mesurer les concentrations de protéines en interaction avec une protéine rapportrice, en se basant sur le nombre et la stabilité de ces interactions non-fonctionnelles. Grâce à leur nouvelle méthodologie, les auteurs ont montré qu'une protéine cytosolique peut engager des interactions non-fonctionnelles de faible affinité avec plusieurs milliers d'autres protéines cytosoliques. Cela montre l'importance de l'encombrement mais aussi de la diffusion cellulaire : les protéines n'engagent pas des interactions uniquement avec leur partenaires fonctionnels ; elles sont constamment en contact avec d'autres protéines et engagent des interactions non-fonctionnelles de faible affinité avec ces dernières.

Afin de rendre compte de la dynamique de l'environnement cellulaire, plusieurs simulations par dynamique moléculaire d'un cytoplasme bactérien ont été réalisées ces dernières années (92,93). Ces travaux visent à modéliser l'environnement cellulaire, en simulant la dynamique des molécules biologiques présentes dans la cellule. En 2010, McGuffee et Elcock (92) ont réalisé la simulation par dynamique moléculaire d'une portion du cytoplasme de *E. coli* en incluant les 50 macromolécules biologiques les plus abondantes (représentant 85% du poids sec du cytoplasme et donc la majorité des macromolécules cytoplasmiques) pour un total d'environ un millier de molécules. Il ressort de leur travaux qu'une protéine change de "voisins", c'est à dire de macromolécules situées à proximité immédiate, à un rythme extrêmement rapide, de l'ordre de la microseconde, du fait de la diffusion très rapide des molécules dans le cytoplasme. Une macromolécule a donc la possibilité d'engager des interactions transitoires non-fonctionnelles avec

un grand nombre de protéines. Ces résultats concordent avec ceux de Levy et al (91) qui montrent la grande diversité d'interactions protéiques ayant lieu dans la cellule.

Plus récemment Yu et al (93) ont réalisé la simulation du cytoplasme de *M. genitalium*. Ils ont inclus des protéines, acides nucléiques, métabolites, ions et l'eau en solvant implicite pour environ 2000 macromolécules et plusieurs dizaines de milliers de solutés et de molécules d'eau. Ils ont représenté les molécules par des modèles tout atomes flexibles. Leur simulation a mis en évidence le fait que les conditions *in vivo* et *in vitro* sont très différentes du fait des nombreuses interactions non-fonctionnelles ayant lieu *in vivo*. Ils montrent notamment que, comme suggéré dans des travaux précédents (94), les interactions protéiques non-fonctionnelles pourraient avoir comme conséquence la déstabilisation des structures protéiques natives. Ce résultat concorde avec plusieurs expériences qui montrent que les interactions non-fonctionnelles de faibles affinité peuvent aussi jouer un rôle dans la stabilité des protéines *in vivo* (95,96). Comme déjà souligné par McGuffee et Elcock (92) la diffusion des protéines dans le cytoplasme est ralentie par les nombreuses interactions non-fonctionnelles et est dépendante des autres molécules auxquelles elles sont confrontées dans leur environnement direct. Elles ont par exemple une forte tendance à lier les métabolites environnants.

Tous ces résultats montrent que les conditions dans la cellule jouent un rôle très important sur la diffusion des protéines et sur leur disponibilité pour engager des interactions fonctionnelles. Il n'est donc pas possible d'aborder les interactions fonctionnelles entre protéines sans prendre en compte les interactions non-fonctionnelles qui se déroulent constamment dans leur environnement cellulaire.

1.4.3 Conséquences des interactions non-fonctionnelles sur l'évolution des protéines

En plus des travaux expérimentaux ou de modélisation présentés plus haut, plusieurs travaux théoriques ont montré l'importance des contraintes exercées par l'environnement sur les protéomes. En effet une double contrainte s'exerce sur les protéines :

- il est nécessaire pour celles-ci de pouvoir engager des interactions avec leurs partenaires.
- il est aussi nécessaire de minimiser la fréquence des interactions non-fonctionnelles avec les autres protéines présentes dans le même environnement cellulaire.

En partant de ces deux postulats, Zhang et al (97) ont proposé un modèle théorique suggérant que minimiser la perte de « ressources » due aux interactions non-fonctionnelles conduit à limiter la diversité du protéome ainsi que les concentrations des protéines co-exprimées et co-localisées (c'est à dire entrants potentiellement en contact). La diversité protéique ainsi que les concentrations individuelles des protéines du cytoplasme, de la mitochondrie et du noyau de *S. cerevisiae* sont proches des limites au-delà desquelle celles-ci deviennent nocives pour la cellule. Il s'agit d'un résultat conceptuel important car il montre que les contraintes inhérentes à l'environnement cellulaire influence la diversité de protéomes entiers.

Un autre travail de modélisation des interactions fonctionnelles et non-fonctionnelles a été réalisé par Johnson et Hummer (98). Ils ont utilisé une représentation très simple pour modéliser la compétition entre interactions fonctionnelles et non-fonctionnelles. Ils ont créé un modèle de réseau d'interaction dans lequel chaque protéine interagit avec un unique partenaire, toutes les autres interactions étant considérées comme non-fonctionnelles. Afin de favoriser les interactions fonctionnelles, celles-ci doivent avoir une énergie d'interaction plus favorable entre elles qu'avec les autres. En optimisant les séquences des sites d'interactions à l'aide de simulation de Monte Carlo, ils ont essayé de minimiser l'énergie d'interaction des interactions fonctionnelles tout en augmentant celles des interactions non-fonctionnelles. Leur modèle suggère que plus la diversité protéique est forte dans un environnement donné, plus il sera difficile de favoriser les interactions fonctionnelles par rapport aux non-fonctionnelles.

1.4.4 Conséquences des interactions non-fonctionnelles sur l'évolution des surfaces protéiques

Dans les parties 1.4.2 et 1.4.3 nous avons montré l'importance des interactions non-fonctionnelles dans la cellule ainsi que les contraintes qu'elles exercent à l'échelle des protéomes. Nous allons maintenant étudier plus en détail comment ces contraintes se manifestent sur la structure et la composition chimique des protéines, et tout particulièrement des surfaces protéiques.

Plusieurs études (99–101) ont montré sur différents organismes que la vitesse d'évolution des protéines est inversement corrélée à leur abondance. Des travaux (102–106) ont montré que ces différences de vitesse d'évolution sont liées aux contraintes exercées par l'environnement cellulaire. En effet les capacités cytotoxiques d'une protéine sont grandement dépendantes de sa concentration cellulaire. Les protéines les plus abondantes peuvent potentiellement se montrer nocives pour la

cellule en cas de mauvais repliements (causant des agrégats), ou encore si elles engagent un grand nombre d'interactions non-fonctionnelles. A l'inverse des protéines présentes en faible quantité dans une cellule auront un effet beaucoup moins nocif pour la cellule. En effet leur faible concentration rend leur capacité à engager des interactions non-fonctionnelles et à former des agrégats plus faible. Plusieurs travaux ont justement montré que cette pression de sélection sur les protéines les plus abondantes permettait d'une part de limiter la surface des protéines à engager un grand nombre d'interactions non-fonctionnelles (105,106) et de limiter les propriétés cytotoxiques dues à de mauvais repliements des protéines (102–105). Par exemple, Levy et al (106) ont montré que les interactions non-fonctionnelles entre protéines influencent de manière globale la composition de la surface des protéines. Pour démontrer cela ils ont développé une échelle de propension à l'interaction des acides aminés (ou échelle de *stickiness*) qui reflète la propension des acides aminés à être impliqué dans des surfaces protéiques (voir section 5.3.2). De manière intéressante cette échelle a été définie de manière purement statistique à partir de structures de complexes répertoriées dans la PDB (13), mais présente une forte corrélation avec l'échelle d'hydrophobicité de Wimley-White (107), déterminée de manière expérimentale. Les résultats obtenus dans cette étude suggèrent que les séquences protéiques subissent des contraintes évolutives afin de limiter les interactions non-fonctionnelles entre protéines repliées. Ces contraintes s'expriment de manière plus forte sur les protéines les plus abondantes, étant donné que ce sont ces dernières qui ont la capacité d'engager le plus grand nombre d'interactions non-fonctionnelles. Ces observations ont été réalisées sur trois organismes modèles: *E. coli*, *S. cerevisiae* et *H. sapiens*. De manière intéressante les conclusions faites sont globalement les mêmes pour ces trois organismes bien que les tendances soient plus prononcées pour *E. coli*, la compartimentation observée chez les eucaryotes permettant d'alléger ces contraintes (97).

Un autre travail remarquable a été réalisé par Schavemaker et al. (108). Les auteurs suggèrent que les propriétés de surface des ribosomes orientent la charge globale des protéines cytosoliques. En effet les ribosomes sont chargés négativement et présents en grand nombre dans le cytoplasme. Schavemaker et al. ont observé chez *E. coli* que les protéines positivement chargées ont un coefficient de diffusion jusqu'à 100 fois inférieur à ceux des protéines chargées négativement ou neutres. Les auteurs ont supposé que les protéines chargées positivement sont séquestrées beaucoup plus facilement dans des interactions non-fonctionnelles avec les ribosomes que les autres protéines. Ces observations vont de pair avec le fait que le protéome de *E. coli* est significativement enrichi en protéines possédant une charge globale de surface négative. De plus les auteurs ont montré que des variants de la GFP (*Green fluorescent protein*) arborant des charges globales largement positives étaient effectivement maintenus dans des interactions non-fonctionnelles avec des ribosomes. Ces

résultats suggèrent donc que les protéines cytoplasmiques de *E. coli* subissent une pression de sélection afin de limiter les charges positives à leurs surfaces, dans le but de limiter les interactions non-fonctionnelles entre les protéines cytosoliques et les ribosomes. Les interactions non-fonctionnelles entre les ribosomes et les protéines influencent donc la composition chimique des surfaces des protéines.

Les conséquences des interactions non-fonctionnelles s'exercent donc sur l'ensemble de la surface des protéines. Cette contrainte s'applique sur plusieurs niveaux tels que (i) la charge des protéines (108) (ii) les propriétés de *stickiness* (propension à l'interaction de façon non-spécifique) des surfaces protéiques et (iii) l'évolution des surfaces protéiques (99).

Tout ces travaux montrent que la cellule est un environnement complexe dans lequel l'évolution des protéines est guidée par un compromis permanent entre interactions fonctionnelles et non-fonctionnelles. En définitive, la pression évolutive s'exerce sur les protéines à plusieurs niveaux :

- les protéines, pour garantir le maintien de leurs fonctions, doivent maintenir les interactions fonctionnelles avec leurs partenaires via leurs sites d'interaction qui subissent une pression de sélection (*positive design*).
- les protéines doivent limiter les interactions non-fonctionnelles avec les protéines et autres molécules biologiques qu'elles peuvent rencontrer dans leur environnement. L'ensemble de leur surface est donc aussi contraint pour limiter ces interactions (*negative design*).

Ces contraintes sont inter-dépendantes : une protéine piégée dans des interactions non-fonctionnelles sera limitée pour réaliser des interactions fonctionnelles par exemple.

1.5 Conclusion

Tout au long de cette introduction j'ai présenté quelques-une des grandes découvertes relatives aux protéines, en particulier sur les relations existant entre celles-ci et leur environnement cellulaire. Nous avons vu que les protéines sont des objets biologiques nécessitant généralement d'interagir avec d'autres molécules pour réaliser leurs fonctions. En particulier les interactions protéiques se font à travers des sites d'interactions possédant des propriétés qui les distinguent du reste de la surface des protéines.

Les protéines sont souvent considérées comme des objets biologiques ayant une surface, une région de cette surface correspondant au(x) site(s) d'interaction qui va/vont permettre à la protéine

d’interagir avec son/ses partenaire(s) pour jouer son rôle fonctionnel. Cependant la réalité est plus complexe : les protéines évoluent dans un environnement cellulaire encombré avec une forte concentration en molécules biologiques, et cet environnement exerce des contraintes sur l’ensemble de leur surface (*positive design* et *negative design*). En effet, le reste de la surface est constamment en compétition avec les sites d’interaction. Pour pouvoir comprendre le fonctionnement des protéines dans un tel environnement, il est donc nécessaire de prendre en compte la compétition entre interactions fonctionnelles et non-fonctionnelles et en particulier d’étudier comment elle opère sur l’ensemble de la surface protéique. Ainsi, l’objectif de ma thèse était de définir et mettre en place un nouveau cadre théorique pour explorer le potentiel de l’ensemble de la surface d’une protéine à interagir avec un ensemble de protéines d’intérêt (les protéines du même environnement cellulaire par exemple).

En effet, caractériser les interactions non-fonctionnelles de manière expérimentale est très difficile. Il est notamment difficile d’obtenir une information structurale ou énergétique de ces interactions. Sauf exception (109) nous ne savons pas par exemple quelles sont les zones des surfaces protéiques les plus ciblées par ces interactions non-fonctionnelles. Dans ce contexte général, les moyens fournis par la biologie computationnelle permettent de contourner ces limites. Elles permettent de plus d’étudier les interactions mises en jeu à l’échelle atomique. Elles permettent également de simuler des interactions n’ayant pas lieu ou ayant lieu de manière sporadique dans la cellule. Je vais développer ces aspects dans la partie suivante.

2 La modélisation des protéines et de leurs interactions

2.1 La bioinformatique structurale

Les approches de bioinformatique structurale ont pour but de modéliser la structure et le comportement des molécules biologiques. La bioinformatique structurale a débuté il y a plus de 50 ans avec les travaux pionniers de Lifson, Allinger, Karplus, Scheraga, Levitt et Warshel entre autres (110,111). Depuis lors ces méthodes se sont largement démocratisées. Elles ont donné lieu à un grand nombre de méthodes répondant à des objectifs variés tels que la dynamique moléculaire, qui vise à représenter l'évolution d'un système moléculaire au cours du temps, l'étude du processus de repliement des protéines, le dessin computationnel qui a pour but de développer de nouvelles protéines avec des propriétés particulières, la prédiction de repliement de protéines à partir de leurs séquences d'acides aminés ou encore l'amarrage moléculaire (*docking* moléculaire) qui vise à prédire la structure tridimensionnelle d'assemblages moléculaires.

Ces approches ont un rôle complémentaire aux approches expérimentales. Leurs avantages sont multiples : elles permettent de rendre compte de phénomènes qu'il n'est pas possible ou très difficile d'observer par des méthodes expérimentales, tels que l'évolution de systèmes moléculaires à travers le temps (par exemple la dynamique d'assemblage de certaines molécules, ou le passage d'un ion à travers un canal). Elles permettent d'autre part de prédire des structures de protéines et de leurs assemblages avec les méthodes d'amarrage moléculaire.

Pour caractériser le potentiel d'interaction de la surface d'une protéine à interagir avec ses partenaires mais aussi des protéines non-partenaires, il faut un outil rapide et capable de modéliser n'importe quelle paire de protéines, qu'elles interagissent effectivement ou non en réalité. Les méthodes d'amarrage moléculaire répondent à ces impératifs. Je vais maintenant présenter plus en détail ces méthodes, puis je vais expliquer la procédure que j'ai employée.

2.2 L’étude des interactions protéiques par *docking* moléculaire

2.2.1 Le *docking* moléculaire : définition et historique

L’amarrage moléculaire (plus connu sous son anglicisme *docking* moléculaire, que j’utiliserai dans la suite de ce manuscrit) regroupe un ensemble de méthodes développées dans le but de prédire la structure d’un complexe formé par deux molécules connues pour interagir. L’objectif est simple : étant donné deux molécules A et B, on cherche à prédire la structure tridimensionnelle du complexe A-B qu’elles forment lors de leur interaction (Figure 2.1). On distingue plusieurs type de *docking* en fonction de la nature des molécules mises en jeu. Les plus communs sont le *docking* protéine-protéine et le *docking* protéine-ligand, le *docking* protéine-acides nucléiques étant aussi employé. On distingue le *docking* protéine-protéine du *docking* protéine-ligand, chaque méthode disposant de caractéristiques propres. Le premier vise à reconstruire la structure des complexes protéiques à partir de leurs sous-unités. Le second est largement employé en recherche pharmaceutique dans le cadre de criblages virtuels permettant de sélectionner des molécules candidates contre des cibles thérapeutiques. Le *docking* protéine-protéine et protéine-ligand sont donc deux champs d’application du *docking* avec des problématiques et des méthodologies différentes. Dans le cadre de ce travail, l’emploi du terme *docking* fera uniquement référence, sauf précision, au *docking* protéine-protéine. Ce dernier a connu un vif intérêt depuis plusieurs décennies, étant donné le nombre limité d’interactions protéine-protéine pouvant être déterminées expérimentalement. En effet élucider la structure des interactions protéine-protéine de manière expérimentale est difficile, la cristallisation étant souvent l’étape limitante. De plus un réel problème se pose pour l’élucidation de structures de complexes transitoires à faible affinité de liaison, qui ne sont généralement pas assez stables pour être déterminés de manière expérimentale à une résolution atomique, bien qu’ils représentent une part très importante des interactions protéiques.

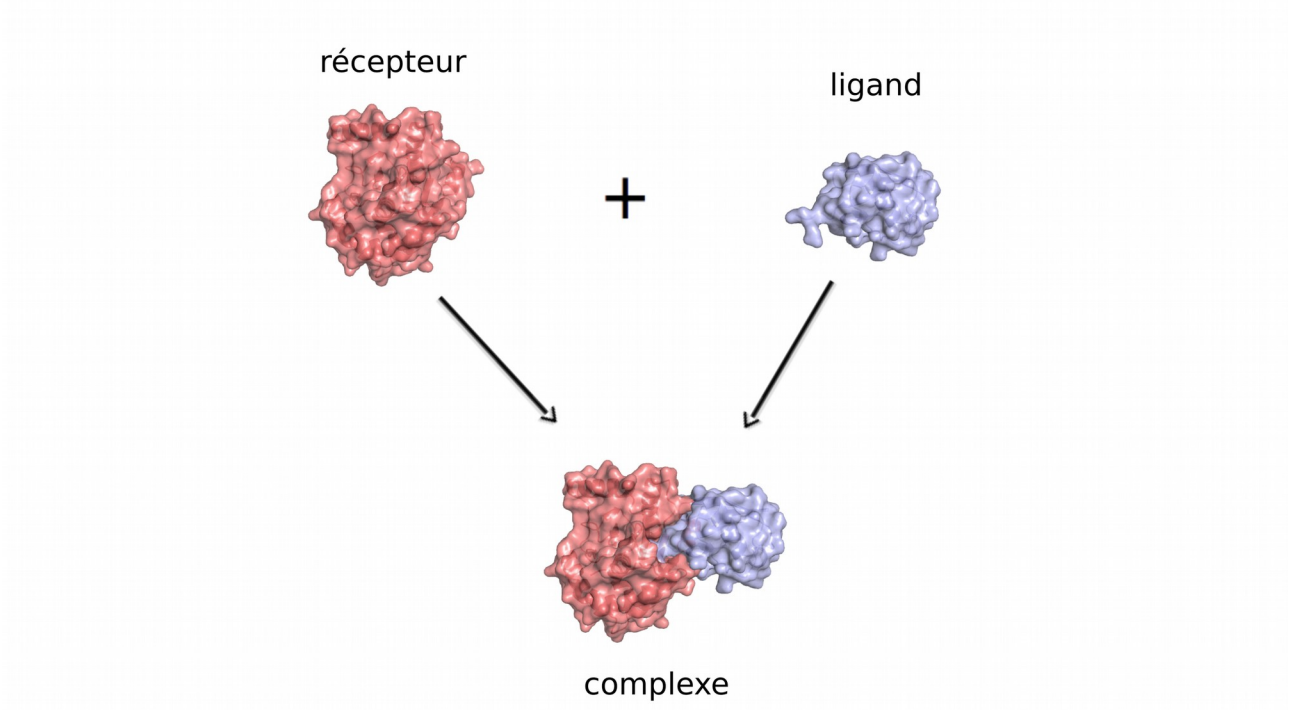


Figure 2.1. Le *docking* protéine-protéine. Le *docking* protéine-protéine est généralement employé pour déterminer la structure d'un complexe protéique à partir de celles de ses sous-unités.

Les premiers algorithmes de *docking* ont vu le jour au début des années 1970. ils ont été développés notamment dans l'équipe de Harold Scheraga qui développa le premier algorithme de *docking* protéine-ligand. L'espace de recherche de solution était limité aux sites d'interactions connus mais les méthodologies utilisées représentaient déjà les ligands comme des molécules flexibles (112). À ma connaissance, les premières procédures de *docking* protéine-protéine sont arrivées quelques années plus tard, les premiers travaux ayant été publiés en 1978 par respectivement Greer et Bush (113) et Wodak et Janin (114).

À l'époque, Greer et Bush (113) avaient développé une méthode pour calculer la surface moléculaire d'une protéine. Étant donné que la complémentarité de surface est l'une des caractéristiques majeures des interfaces protéiques, ils avaient appliqué leur méthode à la prédiction de la structure formée par deux monomères de la methémoglobine de cheval. Leur critère d'évaluation des modes d'interaction était basé sur la complémentarité de surface entre les deux sous-unités (113).

De leur côté, Janin et Wodak avaient développé un procédure utilisant une représentation simplifiée des protéines (chaque résidu étant remplacé par un centre d'interaction assimilable à un

pseudo-atome), qui prenait en compte la complémentarité de surface à travers un potentiel de Lennard-Jones, et modélisant les interactions électrostatique par un potentiel de Coulomb (114). Du fait des capacités de calcul de l'époque et du nombre limité de structures protéiques, cet algorithme avait été testé sur un unique complexe formé par l'association entre la trypsine pancréatique bovine et l'inhibiteur la BPTI (*bovine pancreatic trypsin inhibitor*).

Depuis ces travaux pionniers, de nombreuses avancées ont vu le jour. Du fait des enjeux importants que représente la prédiction de la structure de complexes protéiques, la communauté scientifique du *docking* moléculaire est très active. De nombreux algorithmes de *docking* ont été développés et de nombreuses méthodologies ont été mises au point. La communauté du *docking* protéine-protéine dispose notamment d'un concours annuel, CAPRI (*Community Assessment of Predicted Interactions*) qui fourni chaque année aux participants des structures monomériques de protéines, le but étant de prédire la structure des complexes qu'elles forment (115,116). Les participants doivent fournir leurs propositions sans avoir connaissance des véritables structures des complexes, les propositions étant ensuite évaluées selon des critères préétablis. Cette initiative permet d'évaluer les progrès réalisés chaque année par les équipes de recherche participantes.

Nous allons maintenant expliquer les méthodologies de *docking* les plus utilisées. En pratique deux types de stratégies conceptuellement différentes prédominent: le *template-based docking* et *le docking* libre, ou global (plus connu sous le nom de *free docking*).

2.2.2 Le *template-based docking*

Le principe des algorithmes *template-based* est de modéliser la structure d'un complexe à partir de la structure connue d'un complexe impliquant des protéines homologues à celles dont on veut prédire la structure de l'assemblage (117). Cette méthode se base sur les principes de la modélisation de structures protéiques par homologie. Il a été en effet montré que deux protéines homologues ayant au moins 30% à 40% d'identité de séquence ont généralement une structure similaire (78). Il est donc possible de créer un modèle d'une structure d'une protéine inconnue à partir de la structure connue d'une protéine homologue. De plus Aloy et al ont montré dans (79) que les couples de complexes interologues partageant plus de 30 % d'identité de séquence ont généralement des modes d'assemblage similaires (voir section 1.3.3.1). Il est donc possible de modéliser le complexe formé par deux protéines A et B à partir d'un complexe interologue A'B'. Les structures utilisées pour modéliser le complexe d'intérêt peuvent être sélectionnées en se basant par exemple sur des alignements de séquences (118) ou de structure (119). Cette méthode permet en

principe de générer des modèles proches de la solution native, qu'il est possible d'affiner par des approches *post-docking* (voir section 2.2.3.3) prenant en compte par exemple la flexibilité des protéines.

Le principe de cette méthodologie consiste donc à s'appuyer sur l'information extraite de la structure de complexes interologues pour réduire de manière drastique l'espace des conformations à évaluer et ainsi éviter d'explorer un grand nombre de modes d'interactions putatifs. Ces méthodes ont reçu récemment un grand intérêt du fait du nombre toujours croissant de structures de complexes répertoriées dans la PDB (120) et de nombreux développements ont été réalisés (119,121–123) en suivant cette méthodologie, avec de très bons résultats (117).

Les principales faiblesses de cette catégorie de méthode de *docking* résident dans le principe même qui fait leur force: la nécessité de posséder un modèle structural fiable pour pouvoir modéliser le complexe d'intérêt. De plus, il existe des cas de figure de couples d'interologues ayant des modes d'assemblages différents malgré une forte identité de séquence (80).

2.2.3 Les algorithmes de *docking* libre

Contrairement aux algorithmes de *template-based docking*, les algorithmes de *docking* libre ne modélisent pas la structure des complexes à partir de structures connues. Ils doivent répondre à deux problèmes : (1) générer un ensemble de modes d'interaction entre les deux protéines, que nous appelons étape d'échantillonnage, (2) évaluer les conformations, que nous appelons étape d'évaluation. Ces deux étapes peuvent être couplées ou réalisées l'une à la suite de l'autre. Par convention une des protéines, généralement la plus grosse, est appelée récepteur et est maintenue fixe pendant la procédure de *docking*. L'autre protéine est appelée ligand et est mobile pendant la procédure. Il existe de nombreuses méthodologies (124). Je vais traiter plus en détail les plus employées.

2.2.3.1 L'échantillonnage

L'étape d'échantillonnage est une étape de recherche qui consiste à explorer l'espace conformationnel du récepteur et du ligand pour générer un ensemble de conformations de *docking*, c'est-à-dire un ensemble de modes d'interaction, entre les deux protéines d'intérêts. Pour pouvoir prédire correctement la structure d'un complexe, l'étape d'échantillonnage doit parvenir à générer

au moins une conformation proche de la conformation native. L'étape d'évaluation devra ensuite être capable de la sélectionner parmi les conformations générées. En effet si aucun mode d'interaction proche de la solution native n'est générée lors de cette étape, alors il ne sera pas possible de générer un modèle de bonne qualité du complexe à modéliser.

Il existe de nombreuses méthodologies pour parvenir à ce but. Cette recherche peut-être réalisée sans a priori, c'est à dire en n'ayant aucune connaissance préalable sur les sites d'interaction potentiels des deux protéines d'intérêt. Cette étape peut aussi se faire avec en intégrant des données pour limiter le nombre de conformations à générer. En effet connaître par avance la localisation du site d'interaction d'une ou des deux protéines mises en jeu permet de limiter de manière significative le nombre de modes d'interaction à évaluer. Certains logiciels de *docking* sont donc conçus de manière à prendre en compte des informations permettant de réduire l'espace de recherche. Ces informations peuvent être de tout type: données d'évolution de séquences protéiques, données issues d'expériences de mutagenèse, contraintes issues d'expériences de RMN. Par exemple HADDOCK (125) intègre des données de RMN pour définir des contraintes de distances entre des résidus, permettant de limiter considérablement l'espace conformationnel à analyser.

L'exploration de l'espace conformationnel sans a priori est plus coûteuse et l'échantillonnage s'effectue généralement en considérant les protéines comme des structures rigides afin de limiter le nombre de degrés de liberté du système. De nombreux logiciels effectuent ce type de recherche sans a priori. Parmi les plus connus on peut citer ATTRACT (126), HEX (127), FRODOCK (128) ou ZDOCK (129). L'échantillonnage est réalisé de manière exhaustive dans les six degrés de liberté du système (trois degrés de libertés translationnels et trois degrés de liberté rotationnels).

Méthodes basées sur des FFT

L'emploi des FFT (*Fast Fourier Transform*) dans un algorithme de *docking* fut réalisé pour la première fois en 1992 par Katchalski-Katzir et al (130). Ce fut une révolution à l'époque car cela permis de réaliser un échantillonnage exhaustif de millions de conformations différentes dans les six degrés de liberté du système à un coût calculatoire abordable.

Pour réaliser ces calculs la première étape consiste à discrétiser l'espace de recherche. Chaque protéine est ainsi représentée dans une grille en trois dimensions. Chaque cellule de la grille se voit assigner une valeur distinguant les cellules localisées à l'intérieur, à la surface ou à l'extérieur des protéines. Le score de *docking* est réalisé en calculant, pour chaque mode d'assemblage entre les deux protéines, le degré de complémentarité entre les deux protéines à l'aide

d'une fonction de corrélation. Utiliser cette méthodologie dans les six degrés de liberté translationnels et rotationnels nécessite de tester plusieurs millions de conformations différentes dans une complexité $O(n^6)$, ce qui est prohibitif en terme de temps de calcul. Cependant l'emploi de FFT permet de réaliser la recherche dans l'espace translationnel en $O(\ln(n^3))$ au lieu de $O(n^3)$, résultant en une complexité totale de $O(n^3 \times \ln(n^3))$, largement abordable même sur des ordinateurs personnels. Cette méthodologie possède cependant aussi quelques inconvénients :

1- elle ne permet pas d'introduire la flexibilité des protéines, même si celle-ci peut-être rajoutée dans une étape de post-traitement sur les meilleures conformations.

2- il est nécessaire de réaliser les calculs basés sur les FFT pour toutes les orientations possibles entre le récepteur et le ligand, ce qui rend difficile l'intégration de données visant à réduire l'espace de recherche des solutions.

Il est aussi important de noter que les FFT peut aussi s'appliquer sur l'espace rotationnel au lieu de translationnel comme c'est le cas dans FRODOCK (128). Dans ce cas là c'est la recherche dans l'espace rotationnel qui est réalisée avec une complexité en $O(\ln(n^3))$ (au lieu de $O(n^3)$). Mais dans tous les cas il s'agit d'une méthode peu coûteuse et efficace, et de nombreux algorithmes de *docking*, tels que ZDOCK (129) ou ClusPRO (131) sont basés sur l'utilisation de FFT pour explorer de manière exhaustive les modes d'interaction entre les protéines mises en jeu.

Méthodes basées sur la recherche dans l'espace cartésien

Il existe un grand nombre de méthodes basées sur une exploration de l'espace conformationnel entre deux protéines dans l'espace cartésien. Je vais expliquer brièvement les principes de certaines d'entre elles. Pour plus de précisions et d'exhaustivité, voir la revue de Huang (124).

Méthodes basées sur la complémentarité de surface

Ces méthodes peuvent être basées sur les principes de la reconnaissance d'images utilisée en informatique. Il existe plusieurs méthodologies différentes. Par exemple le logiciel PatchDock (132) représente les protéines par une surface de Conolly (133), qui permet de représenter les surfaces des protéines par des points dans l'espace. Puis l'algorithme va segmenter les surfaces protéiques en fonction de leurs propriétés : planéité, concavité ou convexité. Les surfaces des protéines sont ensuite représentées par des graphes, chaque nœud ayant une propriété distincte. L'algorithme va ensuite rechercher dans les graphes de chacune des deux protéines les portions complémentaires. Pour cela ils appliquent une technique hybride reposant sur un hachage géométrique (*geometric*

hashing) et un partitionnement de données afin de détecter les surfaces possédant des propriétés complémentaires, comme par exemple des surfaces concaves et convexes.

Méthodes de recherche par minimisation d'énergie

Ces méthodes utilisent des approches d'échantillonnage par méthode de Monte Carlo, dynamique brownienne ou encore en procédant par minimisation d'énergie à partir de points de départs générés autour du récepteur. Classiquement le récepteur est maintenu fixe, tandis que plusieurs points de départ du ligand sont placés autour de celui-ci. Ceux-ci peuvent être placés de manière aléatoire autour de la surface du récepteur ou générés autour de points d'intérêts, comme par exemple un site d'interaction connu (134). Dans le cas du logiciel ATTRACT (126) les points de départ sont générés de manière équidistante autour de la surface du récepteur. Pour chaque point de départ plusieurs orientations du ligand sont générées. Chacune de ces orientations donne lieu à une recherche dans l'espace cartésien par minimisation d'énergie (Figure 2.2). D'autres procédures sont possibles. Par exemple pour le cas du logiciel HADDOCK (125) les deux protéines sont placées à 150Å l'une de l'autre, puis l'échantillonnage des conformations se fait par la génération d'orientations aléatoires entre les deux protéines, suivi d'une série de minimisations d'énergie rotationnelles afin d'optimiser l'orientation des deux protéines, puis d'une série de minimisations rotationnelles et translationnelles. Les conformations ainsi générées sont le point de départ d'une procédure d'affinage (voir section 2.2.3.3). Dans ces méthodes, les protéines peuvent être représentées par des atomes (ou pseudo-atomes) mais aussi par des grilles (135). Parmi les algorithmes de *docking* les plus connus utilisant cette méthodologie, on retrouve ATTRACT (136), HADDOCK (125) RosettaDock (137) et SwarmDock (138).

Il est important de noter qu'avec ce type de méthodes, l'échantillonnage n'est pas réalisé de manière exhaustive comme dans les méthodes utilisant les FFT. Cependant le nombre de conformations générées peut tout de même atteindre plusieurs dizaines de milliers pour un logiciel comme ATTRACT, qui utilise une représentation simplifiée des protéines (chaque résidu étant représenté comme trois ou quatre pseudo-atomes). Toujours en prenant l'exemple d'ATTRACT, en théorie chaque mode d'interaction final obtenu à partir de chaque point de départ est un minimum local, ce qui rend les modes d'interactions obtenus plus pertinents que dans le cadre des millions de conformations testées par les algorithmes faisant usage de FFT. Zacharias a montré que réaliser un échantillonnage plus dense, et donc en théorie plus exhaustif que celui employé par défaut dans ATTRACT, ne permet pas d'obtenir de meilleures solutions (126). La couverture de l'échantillonnage est donc suffisante pour détecter les minimums locaux les plus bas.

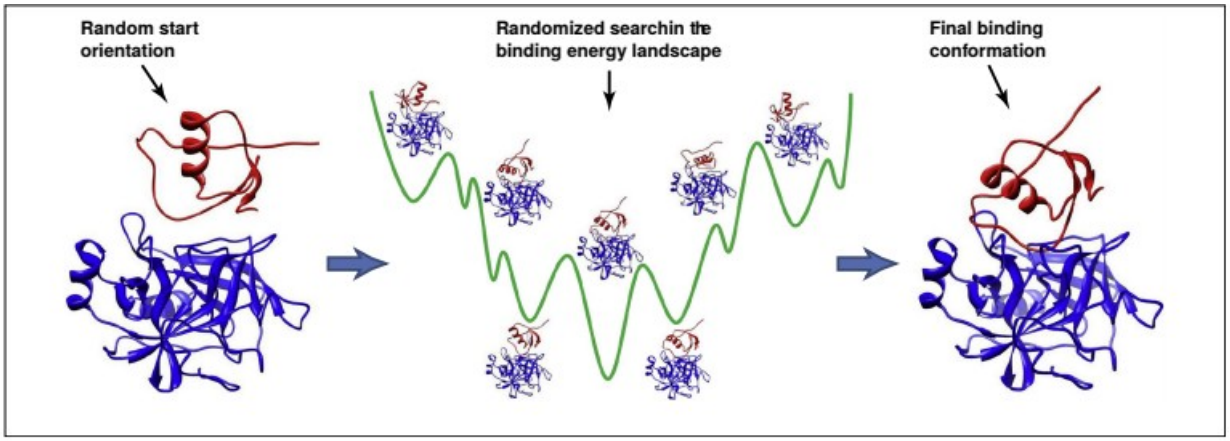


Figure 2.2. Minimisation d'énergie d'une position de départ de *docking*. La recherche a généralement lieue en plaçant de manière aléatoire les deux protéines l'une par rapport à l'autre. Le paysage énergétique est ensuite exploré en optimisant les positions relatives des deux protéines, et, en cas de *docking* flexible, en optimisant les positions des chaînes latérales des résidus et/ou de la chaîne principale jusqu'à atteindre un minimum local d'énergie**.** Figure extraite de l'article de Huang (124).

Représenter la flexibilité des protéines

Les protéines sont des molécules flexibles en solution. De nombreuses protéines subissent des réarrangements conformationnels lors de leurs interactions avec d'autres protéines. Ces réarrangement peuvent avoir lieu au niveau des chaînes latérales des résidus impliqués dans l'interaction mais peuvent aussi être de plus grande ampleur avec des réarrangements de domaines, ou encore le repliement de régions entières des protéines mises en jeu. Dans de nombreux cas, il est donc nécessaire de prendre en compte ces changements conformationnels pour pouvoir prédire la structure native du complexe. Dans le cas de petits réarrangements conformationnels à l'échelle de résidus, la flexibilité peut être modélisée de manière implicite en autorisant un certain degré d'interpénétration atomique (typiquement en atténuant le terme répulsif du potentiel de Lennard-Jones, dont nous parlerons plus loin). Dans le cas de changements conformationnels plus importants, simuler ces réarrangements est actuellement un véritable défi pour le *docking* moléculaire. En effet le nombre très important de modes d'assemblages testés couplé au nombre de conformations de chaque protéine qu'induit la prise en compte de leur flexibilité rend la procédure prohibitive du point de vue du temps de calcul.

Pour pallier à cela, de nombreux algorithmes de *docking* réalisent leur procédure en deux étapes : une première étape en considérant les protéines comme des corps rigides, avec éventuellement une représentation implicite de la flexibilité des protéines pour les petits

réarrangements conformationnels, puis une deuxième étape d'affinement des meilleures solutions proposées lors de la première étape (139). Cependant il est aussi possible d'introduire la flexibilité durant la première étape d'évaluation. C'est par exemple le cas du logiciel ATTRACT (126) qui peut employer des modes normaux pré-calculés (modes dans lesquels tous les atomes de la molécule vibrent à la même fréquence et dans la même phase dans des directions différentes passant simultanément par leurs positions d'équilibre) (140) pour simuler des changements conformationnels de grande ampleur sans augmenter de manière considérable le nombre de degrés de liberté. D'autres méthodes existent, telles que celle employée par HADDOCK (125) qui introduit notamment la flexibilité au niveau des résidus de l'interface des poses de *docking* (ces derniers pouvant modifier les positions de leurs chaînes latérales), et qui fait aussi usage de dynamique moléculaire dans une étape de raffinement des conformations obtenues.

2.2.3.2 L'évaluation des conformations

L'étape d'évaluation a pour but d'identifier les modes d'interactions les plus favorables, donc avec l'énergie la plus basse, parmi les conformations générées pendant l'étape d'échantillonnage. Cette évaluation des complexes est réalisée par une fonction de score optimisée de manière à discriminer des autres solutions les conformations correspondant aux solutions natives. Il existe de nombreuses méthodes pour évaluer les conformations issues de calculs de *docking*. Les fonctions de score peuvent-être classée en plusieurs catégories selon leur propriétés et la méthodologie employée pour les mettre au point. On distingue :

- les fonctions de score basées sur des modèles physiques. Ces modèles représentent les forces qu'engagent les atomes entre eux lors de l'interaction. Elles font usage de champs de forces dont les paramètres représentant les contributions énergétiques sont spécifiquement optimisés pour le *docking*. Typiquement elles emploient un terme pour modéliser les interactions de van des Waals, un terme de Coulomb pour modéliser les interactions électrostatiques et éventuellement un terme pour modéliser la désolvatation. Parmi les exemples les plus connus on retrouve les fonctions de score de la première version d'ATTRACT (126), de RosettaDock (137), de HADDOCK (125) et de PyDock (141).
- les fonctions de score basées sur des potentiels statistiques. Ces fonctions sont déduites de l'analyse statistique réalisée sur des interfaces de structures de complexes protéiques déterminées par des méthodes expérimentales. Elles sont donc basées sur la connaissance que l'on a des interactions protéiques, et permettent de modéliser des interactions protéiques à partir des propriétés des interactions protéiques connues. Cependant étant basées sur ce qui est connu, elles ne peuvent

modéliser de nouveaux comportements absents des bases de données. Il existe des potentiels statistiques basés sur des potentiels de paires comme dans le cas d'ATTRACT (142) ou encore de triplets, comme dans le cas de InterEvScore (143–145). Il existe aussi des fonctions de score basées sur de l'apprentissage automatique (146). Depuis le début des années 2000, plusieurs méthodes d'apprentissage pour l'évaluation des solutions de *docking* ont été développées. Celles-ci reposent sur l'apprentissage par forêts aléatoires (146), machines à vecteurs de support (147) ou réseaux de neurones (148,149) des propriétés des interfaces : composition en acides aminés, fréquences de paires de résidus, score de conservation et score de complémentarité de forme pour chaque conformation de *docking*. A mon sens, leur principal défaut réside dans le fait que les informations sur le poids des différents descripteurs fournis en entrée et sur la façon dont ils sont combinés pour évaluer les solutions de *docking* sont difficilemnt interprétables.

A l'heure actuelle, il existe un très grand nombre de fonctions de scores. Par exemple, dans leur travail de 2013. Moal et al (150) ont évalués les performances de 115 fonctions de score. Il est important de garder en tête que les catégorisations de ces fonctions ne sont pas si claires, car de nombreuses fonctions sont composées de plusieurs termes appartenant à des catégories évoquées plus haut.

2.2.3.3 Étape d'affinage *post-docking*

De nombreux logiciels de *docking* emploient une étape de traitement post-*docking* dans le but d'affiner les meilleures solutions obtenues lors de l'étape d'évaluation. Le principe de cette procédure consiste à sélectionner les meilleures solutions (par exemple le top 100 ou le top 1000) pour les réévaluer en utilisant une méthode d'évaluation plus sophistiquée, donc plus performante mais aussi plus coûteuse en temps de calcul (et donc inapplicable sur l'ensemble des conformations). Cette méthode implique que la fonction de score plus grossière utilisée lors de l'étape précédente soit suffisamment performante pour classer les modes d'interactions natifs parmi les meilleures conformations. Généralement pendant cette étape la flexibilité des chaînes latérales des résidus, voir des protéines entières, est simulée. ZRANK (151) est un exemple de fonction de score utilisée pour l'affinage des solutions de *docking*.

2.2.4 Champ d'application du *docking* protéique

2.2.4.1 Prédiction de structure de complexes

Comme je l'ai expliqué plus haut, le *docking* protéine-protéine a été créé à l'origine dans le but de reconstruire la structure tridimensionnelle de complexes protéiques à partir des structures tridimensionnelles de leurs composants. Il s'agit de son rôle majeur, et la plupart des algorithmes de *docking* libre sont utilisés à cet effet. Grâce aux progrès réalisés au niveau des temps de calcul ils peuvent maintenant être utilisés à grande échelle. Par exemple, Mosca et al (152) ont cherché chez *S. cerevisiae* à fournir des modes d'assemblages putatifs de complexes binaires formés par des protéines détectées comme interagissantes par des méthodes expérimentales, mais dont la structure tri-dimensionnelle du complexe qu'elles forment n'est pas connue. Les auteurs avaient donc pour objectif de fournir une information structurale aux réseaux d'interactions protéine-protéine binaires (évoqués dans la section 1.2). Ils ont pour cela utilisé les structures expérimentales partielles ou complètes disponibles de ces protéines, ainsi que des modèles par homologie. Ils ont appliqué leur méthode sur environ 3000 interactions binaires, et ont donc réalisé autant de calculs de *docking*, un nombre important pour l'époque. Il s'agissait d'un projet ambitieux, mais qui reste toujours dans le cadre classique du *docking* protéique, à savoir prédire des complexes binaires à partir de leurs sous-unités. Les progrès réalisés sur la vitesse des algorithmes de *docking* ont ouvert la voie à de nouvelles applications ces dernières années. En effet, plusieurs travaux pionniers permirent d'étendre le champ d'application des algorithmes de *docking* à la prédiction de sites d'interaction et même à la prédiction de partenaires protéiques.

2.2.4.2 Prédiction de sites d'interaction

Dans leur travail datant de 2004, Fernandez-Recio et al (153) établirent une méthode de prédiction de sites d'interaction protéiques qu'ils appliquèrent sur un jeu de données de 21 protéines. Ils observèrent que les solutions de *docking* de basse énergie avaient tendance à s'accumuler au niveau des sites d'interaction caractérisés expérimentalement, et ce quelque soit la nature du partenaire (connu pour interagir avec la protéine récepteur ou non). A partir de cette observation ils réalisèrent des calculs de *docking* entre un récepteur et plusieurs protéines choisies arbitrairement et développèrent un indice – la NIP pour *Normalized Interaction Propensity* - permettant de prédire les sites d'interfaction protéiques à partir de ces calculs de *docking*. Les

résidus appartenant le plus souvent à une interface de *docking* étant prédits comme appartenant au site d'interaction du récepteur.

A la suite de ces résultats Grosdidier et Fernandez-Recio ont appliqué cette méthodologie pour le prédiction de *hotspots (154)*, des résidus contribuant de manière importante dans l'énergie de liaison entre partenaires (voir section 1.3.2.4). Là encore le *docking* s'est révélé performant, avec des résultats similaires à ceux obtenues avec d'autres méthodes de prédiction de *hotspots (155)*. L'avantage du *docking* arbitraire pour la prédiction de *hotspots* est qu'il ne repose pas sur la connaissance de la structure du complexe ni même du partenaire impliqué.

Les travaux de Martin et Lavery reprenant aussi le *docking* arbitraire (156) font écho à ceux présentés plus haut. Les auteurs montrent qu'en combinant les scores de NIP avec ceux de JET (157), un logiciel de prédiction de sites d'interaction basé sur la recherche d'îlots de résidus conservés à la surface des protéines, ils améliorent la qualité de la prédiction de sites d'interaction par rapport aux prédictions NIP utilisées seules. Néanmoins les performances obtenues restent inférieures à celle d'un algorithme spécialisé dans la prédiction d'interfaces protéiques tel que VORFFIP (158). Dans cette lignée, les travaux de Vamparys et al (159) et de Lagarde et al (160) montrent que le *docking* entre protéines choisies arbitrairement permet de révéler des sites d'interaction alternatifs tels que des sites de fixation à l'ADN.

L'ensemble de ces travaux montrent que l'application d'algorithmes de *docking* à la prédiction de sites d'interaction est une méthodologie viable et encore peu explorée.

2.2.4.3 Prédiction de partenaires protéiques

Un autre enjeu majeur de la biologie structurale est de prédire quelles protéines interagissent dans la cellule et quels sont leurs modes d'assemblage. Ainsi une question récurrente en biologie est : "qui interagit avec qui, et comment ?". Ces deux aspects ont généralement été traités séparément et par des approches très différentes. Les méthodes expérimentales haut-débit telles que les méthodes de double hybride (Y2H) (161), la méthode TAP (162) et la spectrométrie de masse (35), ainsi que les méthodes *in silico* haut-débit telles que les approches systémiques, d'analyse de données de coexpression, la génomique à grande échelle (profils phylogénétiques par exemple (163,164)) permettent d'identifier massivement des partenaires protéines, et donc de répondre à la question « qui interagit avec qui ?». Les méthodes biophysiques telles que les méthodes de cristallographie, RMN, cryo-microscopie électronique pour les méthodes expérimentales, et de *docking* pour les méthodes *in silico*, permettent quant-à-elles de caractériser la structure tridimensionnelle des assemblages formés par des partenaires protéiques. Elles répondent donc à la

question « comment les partenaires interagissent ?». Cependant, ces méthodes biophysiques ne sont pas applicables à la même échelle que les méthodes haut-débit de type Y2H, TAP, etc. Jusqu'à il y a encore peu de temps, le *docking* restait limité lui aussi à la prédiction de la structure de couples protéiques à relativement petite échelle. Or avec les progrès récents en *docking* protéine-protéine, celui-ci est maintenant applicable à grande échelle, ce qui a permis à de nouvelles applications de voir le jour.

Ainsi dans des travaux pionniers, Sacquin-Mora et al (165) furent les premiers à appliquer les calculs de *docking* à la prédiction de partenaires. Pour se faire ils ont développé un algorithme de *docking* gros grain, MAXDo, très similaire au logiciel ATTRACT (126). Le but de leur expérience était de déterminer s'il était possible de retrouver les vrais partenaires protéiques, parmi de faux partenaires, à partir des scores de *docking* de chaque paire. L'hypothèse testée était de savoir si les vrais partenaires obtenaient des scores de *docking* plus favorables que les couples choisis arbitrairement. Pour chaque protéine, ils comparèrent les meilleures énergies de *docking* obtenues pour le vrai partenaire et les 11 autres faux partenaires. De manière intéressante ils montrèrent que le vrai partenaire ne conduisait pas nécessairement à de meilleures scores d'énergie que les 11 autres protéines. Ainsi l'information extraite des calculs de *docking* ne semblait pas suffisante pour prédire des partenaires protéiques. Cependant ils montrèrent qu'en restreignant l'algorithme de *docking* à explorer les régions correspondant au site d'interaction avec le partenaire, alors les vrais couples se distinguaient clairement des autres. Ainsi la question de la prédiction de partenaires se résumait à la prédiction de sites d'interaction. Ce résultat est très intéressant car il déplace le problème : il montre que si on est capable de prédire les sites d'interaction des protéines, on est capable de prédire les partenaires en interaction. Finalement ils essayèrent de prédire les sites d'interaction des 12 protéines en utilisant une variante de la NIP (153) pour prédire les sites d'interaction des protéines, puis en restreignant l'exploration conformationnelle à ces sites d'interaction prédits. Les résultats obtenus furent meilleurs que ceux obtenus en utilisant l'ensemble des solutions de *docking*, mais moins bons qu'en restreignant les calculs de *docking* aux sites d'interaction identifiés expérimentalement.

Ainsi ces travaux montrent qu'une connaissance parfaite des sites d'interaction mis en jeu, alliée aux calculs de *docking* arbitraire permettrait en théorie d'obtenir une excellente capacité de prédiction des partenaires protéiques parmi un ensemble de faux partenaires.

En 2013 Lopes et al ont réalisé un travail important (166) dans la continuité de celui réalisé par Sacquin-Mora et al (165), toujours dans le but de prédire des vrais partenaires protéiques à partir d'un jeu de données composé de vrais et faux partenaires. Ils réalisèrent un calcul de *docking*

croisé complet (168x168 calculs de *docking*) à l'aide de l'algorithme MAXDo sur l'ensemble de la base de données PPI benchmark 2.0 (167) en utilisant uniquement des protéines sous forme non-liée. Utiliser des formes non-liées augmente la difficulté de l'exercice dès que les protéines subissent des changements conformationnels (même mineurs) lors de l'assemblage. Ils évaluèrent leur capacité prédictive en calculant l'aire sous la courbe ROC (l'AUC) lors de la prédiction des vrais partenaires parmi le jeu de faux partenaires. En restreignant l'exploration conformationnelle lors du *docking* autour des sites d'interaction identifiés expérimentalement, ils obtinrent une AUC de 85%. Ce résultat très prometteur confirme que connaître le site d'interaction peut permettre de prédire les partenaires d'une protéine à partir des formes non liées des monomères. Ils développèrent ensuite un nouvel indice numérique pour classer les partenaires et identifier les plus probables. Cet indice reposait sur une combinaison de l'énergie de *docking*, d'une nouvelles version de la NIP normalisée et d'un score reflétant la conservation des résidus de surface (157). L'hypothèse sous-jacente était que les interfaces entre paires fonctionnelles de protéines (i) doivent avoir une énergie acceptable et (ii) subissent une pression de sélection pour maintenir l'assemblage fonctionnel. On devrait trouver une trace de cette pression évolutive (conservation de résidus) au niveau des interfaces fonctionnelles et les distinguer ainsi des paires non-fonctionnelles. Avec ce nouvel index exempt d'information expérimentale et satisfaisant le compromis stabilité/fonction (le caractère fonctionnel étant indirectement relié au signal de conservation), ils obtinrent une AUC de 61% sur l'ensemble de la base de données et 72% sur les 23 couples enzymes-inhibiteurs de la base. Ces résultats sont très encourageants pour la suite car ils montrent qu'en améliorant la prédiction des sites d'interaction, on devrait considérablement améliorer la prédiction des partenaires protéiques à partir de la seule information de leurs structures monomériques (rappelons que ces calculs ont été réalisés avec les formes non liées (c'est-à-dire monomériques) des partenaires). Les auteurs ont ensuite distingué trois catégories de protéines. (i) les « compétitrices fortes » qui avaient un fort index d'interaction avec un grand nombre de partenaires, (ii) les « non-compétitrices » qui avaient un index d'interaction très faible avec l'ensemble des protéines de la base et (iii) les « intermédiaires » qui avaient un index fort avec seulement un nombre limité de partenaires. Ces résultats rappellent les observations faites par Levy et al (106) qui montraient que toutes les protéines n'avaient pas toutes la même propension d'interaction.

Maheshwari et Brylinski ont très récemment développé une méthode de prédiction de partenaires à l'aide de calculs de *docking* couplés à de l'intégration de données hétérogènes et des méthodes d'apprentissage automatique (168). Ils ont pour cela réalisé des calculs de *docking* avec ZDOCK3.0.2 (129) sur le protéome de *E. coli* en utilisant toutes les structures expérimentales disponibles et en modélisant par homologie un grand nombre de structures. Ils ont couplé leurs

calculs de *docking* avec des calculs de prédiction de sites d'interaction. Ils ont ensuite affiné les meilleurs modèles avec une étape de *post-docking* puis suivi d'une étape de forêt d'arbres décisionnels afin d'estimer la probabilité qu'un complexe généré représente une vraie interaction. Ils ont obtenu une AUC de 0.72 sur leur jeu de test, montrant qu'une méthodologie couplant *docking* et intégration de données est viable pour la prédiction d'interactions protéine-protéine.

Tous ces résultats montrent que la prédiction de partenaires protéiques à l'aide de calculs de *docking* est une approche prometteuse. Ce champ d'application reste encore peu développé, à l'heure actuelle moins d'une dizaine de travaux ont été publiés dans ce domaine (165,166,168–172).

2.3 Conclusion

Le *docking* moléculaire est traditionnellement utilisé dans le but de prédire la structure d'un complexe à partir de ses sous-unités. Cependant depuis plus d'une décennie maintenant, des travaux ont été réalisés en étendant le champ d'application du *docking* protéique à la prédiction de sites d'interaction mais aussi à la prédiction de partenaires protéiques. Il s'avère que ces voies sont toutes prometteuses et encore peu explorées. Par ailleurs, le *docking* moléculaire permet (i) de simuler l'interaction de n'importe quelle paire de protéines, et cela à une résolution atomique et (ii) de simuler tous les modes d'interaction possibles (incluant modes d'interaction énergétiquement favorables mais aussi défavorables – ces derniers sont très difficiles à caractériser expérimentalement) entre deux protéines et donc d'explorer les modes d'assemblages impliquant l'ensemble des surfaces des deux protéines en interaction. De plus la rapidité de certains d'algorithmes de *docking* les rend aujourd'hui applicables pour des expériences haut-débit. Ce sont pour toutes ces raisons que le *docking* me semble être l'outil de choix pour étudier le potentiel de l'ensemble de la surface d'une protéine à interagir avec un jeu de protéines d'intérêt.

3 Mon travail de thèse

L'objectif de ma thèse était de mettre en place un cadre théorique permettant d'investiguer comment l'environnement encombré de la cellule contraint les surfaces protéiques et leurs paysages d'énergie d'interaction. En d'autres termes, l'objectif était de développer une méthode pour caractériser l'effet de l'environnement cellulaire sur le protéome et en particulier sur le potentiel d'interaction d'une protéine avec ses partenaires mais aussi avec des protéines arbitraires. Pour cela, il fallait (i) mettre en place une méthode pour caractériser des interactions entre paires de protéines fonctionnelles mais aussi entre des paires non-fonctionnelles et (ii) mettre en place une méthode permettant de caractériser le potentiel d'interaction de l'ensemble de la surface d'une protéine.

Ainsi, le cœur de ma thèse a consisté à mettre en place un cadre théorique reposant sur (i) une procédure de *docking* arbitraire impliquant des paires fonctionnelles et non-fonctionnelles et (ii) une représentation originale des paysages d'énergie d'interaction à l'aide de cartes d'énergie en deux dimensions qui représentent de façon synthétique le potentiel d'interaction de la surface d'une protéine à interagir avec un partenaire d'intérêt. Enfin, j'ai développé une nouvelle représentation intégrant l'information extraite de sous-ensembles de cartes d'énergie qui permet de caractériser le potentiel d'interaction d'une protéine avec un sous-ensemble de partenaires d'intérêt. En jouant sur la nature des partenaires mis en jeu, ces cartes en deux dimensions que je nommerai cartes IPOPS (*Interaction Propensity Of Protein Surfaces*) fournissent une information systémique en permettant de refléter de façon synthétique le potentiel d'interaction d'une protéine dans un environnement donné.

- La première étape de ma thèse a donc consisté à développer SURFMAP, qui permet de représenter en 2D différentes propriétés de surface d'une protéine (hydrophobicité, relief de la surface, potentiel électrostatique, conservation des résidus ou potentiel d'interaction par exemple). SURFMAP sera mis à disposition de la communauté à l'aide d'un serveur web.

- la seconde étape a consisté à mettre en place le cadre théorique permettant de caractériser le potentiel d'interaction d'une protéine avec un jeu de partenaires d'intérêt depuis la procédure de *docking* arbitraire jusqu'à la production d'une carte IPOPS. J'ai testé ma

méthode sur un sous-ensemble de 348 protéines extraites de la base de données PPI4DOCK (173). Cette dernière présente l'avantage de contenir de nombreuses structures de couples d'interologues sous forme non-liée. Lors de cette étape, j'ai pu évaluer la « robustesse » de ma méthode et j'ai pu caractériser de façon systématique les propriétés énergétiques, physico-chimiques mais aussi évolutives des différentes régions de surface. Cette étude m'a permis d'élaborer un nouveau modèle de surface où les surfaces protéiques peuvent être vues comme des vecteurs de nouvelles interactions protéiques.

- J'ai ensuite appliqué ma méthode sur différents jeux de données :

    - le jeu de complexes homomériques présenté dans (174). Ce jeu de données était très intéressant car il comprenait des complexes présentant des symétries différentes. L'objectif était d'étudier les propriétés de surface (évolutives, hydrophobicité mais aussi le potentiel d'interaction) des homomères présentant différents types de symétrie pour voir si les contraintes qui s'exerçaient sur ces différents assemblages étaient les mêmes.

    - j'ai constitué un jeu de 12 familles d'homologues (comprenant 74 protéines) pour étudier la conservation du potentiel d'interaction au sein de ces familles et ainsi investiguer la conservation du paysage énergétique d'interaction d'une paire de protéines (fonctionnelle ou non-fonctionnelle) au cours de l'évolution.

    - J'ai réalisé le *docking*-croisé complet (cas de *docking* arbitraire où toutes les paires possibles du jeu de données sont testées. Cette procédure est utilisée lorsque il n'y a aucun a priori sur les paires mises en jeu ou lorsque l'on veut réaliser une exploration exhaustive des interactions entre les protéines du jeu de données) des protéines du cytosol de *S. cerevisiae* dont la structure tridimensionnelle avait été caractérisée expérimentalement ou modélisée par homologie (400000 calculs de *docking* pour un total de 2 millions d'heures environ). L'objectif à long terme de ce travail étant de caractériser l'effet de l'encombrement cellulaire sur le protéome et l'interactome du cytosol de *S. cerevisiae*.

4 Représenter les surfaces protéiques : cartographie moléculaire

Au cours de ma thèse j'ai réalisé une méthode permettant d'analyser et de visualiser les propriétés de la surface des protéines. Cette méthode repose sur l'utilisation d'une projection cartographique de la surface des protéines. A l'origine ce type de méthodes a été développé dans le domaine de la géodésie (science de l'étude des dimensions et de la forme de la Terre) mais est applicable dans de nombreux domaines. Elles reposent sur la projection de la surface d'une sphère (telle que le globe terrestre) sur une carte en deux dimensions (figure 4.1). Dans le cas d'une protéine, ces cartes sont réalisées par l'approximation de sa surface par une sphère puis par la projection des propriétés de la surface sur une carte.

Il existe de nombreuses méthodes pour réaliser une projection cartographique, chacune possédant des spécificités qui lui sont propres (175). Ces particularités viennent du fait qu'il n'est pas possible de projeter la surface d'une sphère sur un plan en deux dimensions sans entraîner de distorsions. De ce fait chaque type de projection possède des caractéristiques spécifiques. Par exemple certaines projections ont la propriété de conserver localement les angles, et donc les formes, mais ne permettent pas de conserver les surfaces (projections conformes). A l'opposé d'autres projections vont permettre de conserver les surfaces au prix d'une distorsion des angles (projections équivalentes).

Pour l'étude de la surface des protéines, le choix des projections cartographiques va dépendre de plusieurs contraintes : (i) il est préférable d'utiliser une projection permettant de représenter l'ensemble de la surface de la protéine d'intérêt sur un même plan. (ii) Dans le but d'appréhender de manière correcte la distribution des propriétés de l'ensemble de la surface des protéines, il est aussi important d'utiliser une projection équivalente (projections dites *equal-area*), c'est à dire que ce type de projection a pour propriété que, quelques soient deux aires de même surface sélectionnées sur la projection, ces deux aires ont également la même surface sur la sphère ayant servie à la projection. Une projection très connue réunissant ces propriétés est la projection sinusoïdale (figure 4.1B), c'est celle que j'ai choisi d'utiliser au cours de ma thèse. Il est cependant important de noter que d'autres projections, telle que les projections de Mollweide, de Hammer ou

encore la projection azimutale équivalente de Lambert, remplissent également les conditions évoquées plus haut et auraient aussi pu être utilisées (175).

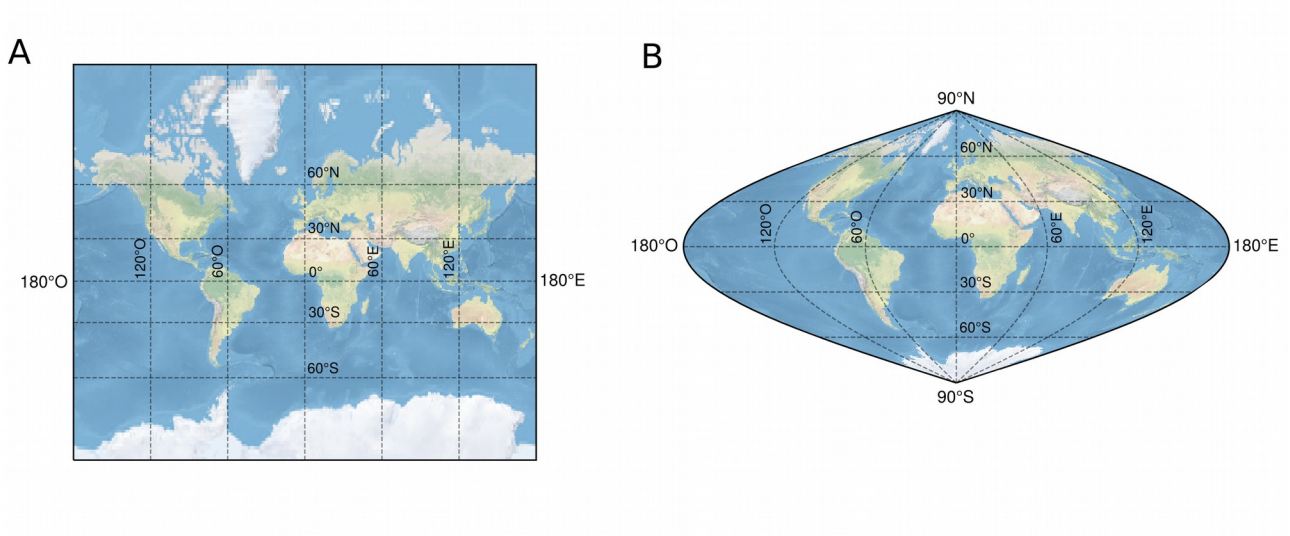


Figure 4.1. Exemples de projections cartographiques du globe terrestre. (A) Projection de Mercator. Il s'agit d'une projection conforme : les angles sont conservés mais pas les aires, celles-ci s'accroissant au fur et à mesure que l'on s'éloigne de l'équateur. (B) Projection sinusoïdale. Il s'agit d'une projection équivalente : les surfaces sont conservées mais pas les angles, les pôles étant notamment déformés. N, S, E et O désignent respectivement le nord, le sud, l'est et l'ouest. Les méridiens et parallèles sont représentés par des lignes noires en pointillés allant respectivement du sud au nord et de l'est à l'ouest. Figures réalisées à l'aide du *package* python cartopy.

Bien qu'étant un outil évident et simple à mettre en œuvre pour pouvoir analyser de manière rapide les propriétés de la surface des protéines, l'utilisation d'une projection cartographique possède aussi des limites et des inconvénients :

- Le fait d'approximer la surface d'une protéine par une sphère a pour résultat d'effacer toute irrégularité de sa surface, telles que les protubérances et les cavités. De plus les surfaces correspondant à celles-ci seront sous-échantillonnées sur la projection. Cependant il est à noter qu'il est tout de même possible de conserver la trace de ces propriétés sur une carte, par exemple en cartographiant la variance circulaire (qui est une mesure du degré d'exposition d'un atome, voir section 5.3.3) des atomes de la surface des protéines.

- Approximer une protéine par une sphère peut être considéré comme une approximation acceptable dans le cas de protéines globulaires, cependant dans le cas de protéines possédant des formes non globulaires, tel que la forme en fer à cheval de l'inhibiteur de ribonucléase (176) (figure 4.2A) ou la forme très allongée de la protéine de réparation de l'ADN RecN (177) (figure 4.2B), alors une telle approximation peut être trop importante pour obtenir un modèle réaliste de la surface de la protéine.

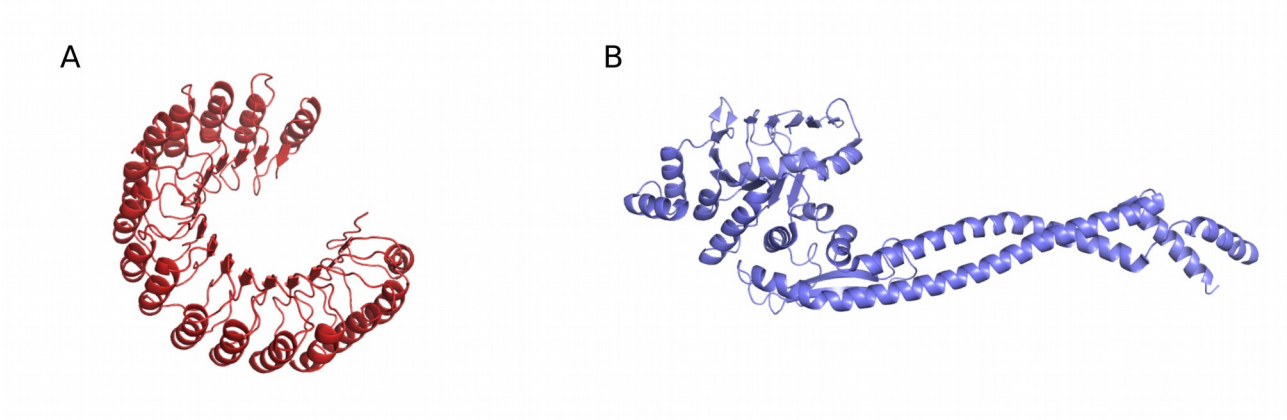


Figure 4.2. Structures de l'inhibiteur de ribonucléase et de la protéine de réparation de l'ADN RecN. (A) L'inhibiteur de ribonucléase (code pdb 2bnh) possède une structure non globulaire en forme de fer à cheval. Dans le cas de cette protéine la projection cartographique n'est pas un outil adapté pour étudier les propriétés de sa surface. (B) RecN (code pdb 4ad8) possède une structure non globulaire de forme très allongée. Là aussi la projection cartographique n'est pas forcément un outil adapté pour étudier les propriétés de sa surface.

Même si la représentation de protéines en deux dimensions est utilisée depuis longtemps (Wodak et Janin y ont eu recours en 1978 par exemple (114)), à ma connaissance les premiers travaux de cartographie moléculaire systématique de la surface des protéines datent du milieu des années 80. Fanning et al (178) avaient appliqué leur méthode à l'étude de la surface de la myoglobine et du lysozyme, et tout particulièrement à la topographie de leurs régions antigéniques. La même année Barlow et Thornton avaient réalisé un logiciel permettant l'analyse de la distribution des résidus chargés et non chargés sur la surface des protéines (179). Leur méthodologie repose sur la projection cartographique des coordonnées des résidus de surface par une projection de Hammer. La carte obtenue est ensuite analysée par un algorithme de partitionnement en k-moyennes afin de mettre en évidence les régions de la surface les plus chargées, permettant d'obtenir une analyse quantitative de la distribution des charges à la surface des protéines.

Depuis ces premiers travaux, plusieurs méthodes faisant usage d'une projection cartographique des surface protéiques ont été réalisées. Par exemple Pawlowski et Godzik (180) ont réalisé un protocole de prédiction de fonction de protéines homologues basé sur la comparaison des propriétés physico-chimiques de leurs surfaces. Ils ont pour cela calculé les propriétés des résidus de surface (résidus positivement chargés, négativement chargés, hydrophobes ou hydrophiles) puis ont projeté ces informations sur une carte réalisée à l'aide d'une projection sinusoïdale. Ils ont aussi fait usage de cartes cumulatives d'homologues structuraux qui représentent sur une même carte les propriétés sommées de plusieurs protéines homologues. Ces cartes permettent de mettre en avant les régions de la surface possédant des propriétés conservées entre homologues structuraux, et celles au contraire possédant des propriétés physico-chimiques variables. Ils ont testé leur méthodologie sur plusieurs familles de protéines homologues (hémoglobines, domaines TRAF et domaines de mort (*death domains*)) et ont montré que la comparaison de propriétés de la surface de ces protéines est un prédicteur efficace de leurs fonctions. Leur protocole a ensuite été mis en place sur un serveur web (181). Cependant, bien que très intéressante, leur méthodologie repose sur l'analyse d'un unique descripteur de la surface des protéines, les propriétés de charge et d'hydrophibicité/ hydrophilicité des résidus, alors que plusieurs autres descripteurs pourraient aussi être comparés.

C'est dans l'optique de permettre l'analyse systématique de descripteurs des surfaces protéiques que le logiciel Structuprint a été réalisé (182). Ce logiciel est spécialisé dans la cartographie des propriétés de surface des protéines. Il permet de cartographier 328 descripteurs (classé en plusieurs grandes catégories) des surface protéiques. Ce logiciel réalise la cartographie de la surface des protéines par une projection cylindrique de Miller. Malheureusement cette projection ne permet pas de conserver les surfaces : les distances entre atomes situés vers l'équateur sont correctement représentées, mais sont graduellement surestimées au fur et à mesure de l'éloignement de l'équateur, jusqu'à une surestimation maximale aux pôles. Il est donc difficile de comparer les propriétés de surface de différentes parties d'une même carte. De plus la méthodologie des cartes cumulatives réalisée par Pawlowski et Godzik (180) est difficile à mettre en place avec ce type de projection.

Koromyslova et al (183) ont réalisé un logiciel permettant de cartographier la distribution du potentiel électrostatique et du potentiel hydrophobe moléculaire des surfaces protéiques. Par rapport à l'approche développé par Pawlowski et Godzik (180), leur méthode permet une représentation plus fine des propriétés de la surface des protéines. En effet Pawlowski et Godzik se sont intéressés à la distribution des résidus chargés ou hydrophobes/hydrophiles, tandis que Koromyslova et al ont

étudié la distribution des potentiels électrostatique et hydrophobe, ce qui permet de mieux appréhender les caractéristiques physico-chimiques des régions étudiées.

Les travaux présentés plus haut visent tous à analyser un certain nombre de propriétés des surfaces protéiques. Les auteurs se sont concentrés sur l'analyse des caractéristiques topographiques des régions protéiques, de la distribution du potentiel électrostatique/hydrophobe, ainsi que de la distribution et de la conservation entre homologues des résidus chargés, hydrophobes ou hydrophiles. Dans le cadre de ma thèse je me suis intéressé au potentiel d'interaction protéine-protéine des surfaces protéiques. Plusieurs travaux modélisant ce potentiel ont été réalisé mais peu ont mis en avant cet aspect là dans leurs études.

4.1 Étudier le potentiel d'interaction des surfaces protéiques

A ma connaissance deux travaux de *docking* ont déjà fait usage de cartes d'énergie. Le travail pionnier de Fernandez-Recio et al (153) fut le premier à représenter ces cartes. Dans ce travail leur objectif était de développer une méthode de prédiction d'interface, les cartes d'énergie ne leur servant que de support visuel pour localiser les solutions de *docking* les plus favorables à la surface du récepteur. De manière similaire, dans Sacquin-Mora et al (165) l'emploi de carte d'énergie avait une fonction de support visuel pour valider l'emploi de leur algorithme de *docking*, et non pas un but d'analyse. Ces deux travaux n'avaient donc pas pour but d'analyser les propriétés du paysage énergétique d'interaction des protéines.

Pour analyser l'ensemble du paysage énergétique d'interaction entre une protéine et des partenaires protéiques, j'ai donc choisi d'employer des cartes d'énergie en deux dimensions. Celles-ci représentent une approximation du potentiel de la surface d'une protéine à interagir avec une autre protéine.

Cette question a été abordée dans un travail réalisé en 2012 par Ravikumar et al (184) sur un jeu de 5 complexes protéiques. Ils ont mis au point un protocole spécifiquement calibré pour pouvoir explorer l'ensemble de la surface du récepteur à l'aide de courtes simulations de dynamique moléculaire démarrant à de multiple points de départ autour du récepteur (méthode qui peut rappeler la méthode d'échantillonnage de solutions de *docking* décrite dans la section 2.2.3.1). Afin

de rendre cette méthodologie abordable en terme de temps de calculs, ils ont utilisé un modèle gros grain très simplifié des protéines: chaque résidu est représenté par un unique pseudo-atome localisé au niveau de son Cα (185). Les auteurs ont représenté le paysage énergétique d'interaction entre deux protéines par une sphère centrée sur le centre de masse du récepteur, approximant la surface de ce dernier. Leur simulation leur a permis de retrouver les modes d'interactions natifs des complexes protéiques mis en jeu, ce qui valide leur méthode, mais aussi, chez certains complexes, des solutions de basse énergie ne correspondant pas aux modes d'interaction natifs (Figure 4.3). Ces travaux sont très intéressants et prometteurs. Le fait que certains complexes possèdent plusieurs minimums locaux très attractifs dans leur paysage énergétique d'interaction mériterait une analyse approfondie. Cependant leur méthodologie reste difficilement applicable à grande échelle. La dynamique moléculaire, même avec un modèle très simplifié comme le leur, ne permet pas de réaliser des simulations à grande échelle comme nous avions l'ambition de le faire dans cette thèse.

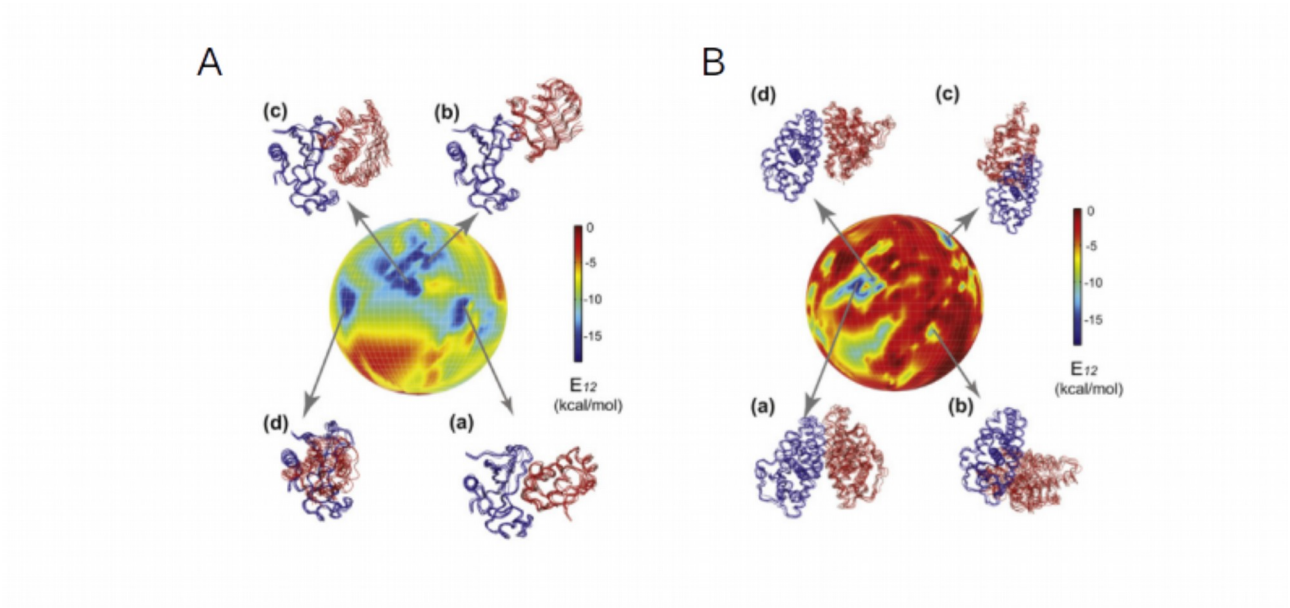


Figure 4.3. Paysage énergétique d'interaction entre les protéines barnase et barstar et les protéines RXR et LBD. Le paysage énergétique est représenté sous la forme d'une sphère centrée sur le centre de masse de la protéine récepteur (en bleu dans les schémas). Les régions favorables à l'interaction sont colorées en bleu, tandis que les régions défavorables sont colorées en rouge. (A) Dans le cas de l'interaction entre la barnase et la barstar, on distingue trois régions du récepteur favorables à l'interaction. (B) Dans le cas de l'interaction entre la protéine RXR et la protéine LBD, on distingue deux régions du récepteur favorables à l'interaction. Les figures ont été extraites de l'article de Ravikumar et al. (184).

5 Méthodologies

Dans cette section sont présentées les procédures et méthodes que j'ai utilisées et mises au point durant ma thèse et qui sont communes aux trois grandes sections de ma thèse. Le matériel et méthodes spécifique à chaque section, tels que les jeux de données de structures utilisées, sont présentés à l'intérieur des sections correspondantes.

5.1 Manipulation des structures protéiques

5.1.1 Définition des résidus d'interface

Dans ce travail les résidus appartenant à une interface sont définis à partir des formes complexées des protéines. Pour une protéine donnée, on définit comme résidu de l'interface tout résidu dont l'accessibilité au solvant diminue ($\Delta rsa > 0$) entre la forme libre de la protéine et sa forme complexée. Tout les calculs d'accessibilité au solvant on été réalisés avec le logiciel FreeSASA (186).

5.1.2 Préparation des structures protéiques

J'ai utilisé trois jeux de données de structures protéiques que je présenterai dans chacune des sections qui suivent. Toutes les structures protéiques ont été préparées avec le logiciel DOCKPREP (187) afin de supprimer les ions, molécules d'eau et de compléter les chaînes latérales incomplètes. Tous les alignements structuraux ont été réalisés avec le logiciel TM-align (188). Tous les alignements de séquences ont été réalisés avec le logiciel MUSCLE (189).

5.2 Procédure de *docking* : ATTRACT

Au cours de ma thèse, j'ai réalisé les calculs de *docking* moléculaire avec le logiciel ATTRACT, développé en 2003 par Martin Zacharias (126). Cet algorithme est basé sur une représentation simplifiée, dite gros grain des protéines, et utilise une fonction de score composée d'un terme de Lennard-Jones pour modéliser les interactions de van der Waals, et d'un terme de Coulomb pour représenter les interactions électrostatiques. Le *docking* moléculaire est réalisé par minimisation d'énergie dans les six degrés de liberté rotationnels et translationnels (voir section 2.2.3.1) à partir de milliers de conformations de départ générées de manière homogène autour de la surface du récepteur.

5.2.1 Représentation des protéines

La représentation des protéines se fait avec un modèle gros grain des résidus. L'utilisation d'un tel modèle permet de diminuer la complexité du système, et donc de réduire de manière conséquente le temps de calculs, qui est l'une des limites principales du *docking*. Cette représentation simplifiée se traduit par un nombre plus faible de minimum locaux par rapport à un modèle de représentation protéique tout-atome, et permet donc une convergence beaucoup plus rapide lors des étapes de recherche de minimum d'énergie (du fait du plus petit nombre de minima locaux).

Les acides aminés sont représentés par trois ou quatre pseudo-atomes selon leur type. Pour chaque résidu, la chaîne principale est représentée par deux pseudo-atomes (un est positionné sur l'atome d'azote, l'autre sur l'atome d'oxygène) (Figure 5.1), les propriétés de ces deux pseudo-atomes sont les mêmes quelque soit la nature de l'acide aminé. La chaîne latérale des petits acides aminés (Ala, Asp, Asn, Cys, Ile, Leu, Pro, Ser, Thr, et Val) est représentée par un unique pseudo-atome situé au centre de masse des atomes lourds la composant. Les chaînes latérales des autres acides aminés (Tyr, Phe, Gln, Glu, Lys, Arg, Met, His) sont représentées par deux pseudo-atomes, ce qui permet de prendre en compte leur forme et propriétés chimiques spécifiques (136). La Glycine est un cas particulier, ne possédant pas de chaîne latérale elle est représentée par les seuls deux pseudo-atomes du squelette.

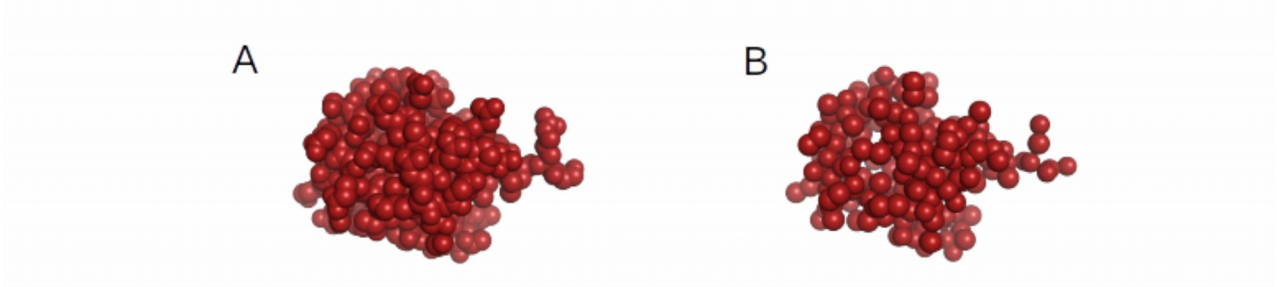


Figure 5.1. Représentation de l'ubiquitine avec un modèle tout atome et avec le modèle gros grain d'ATTRACT. (A) Modèle protéique de l'ubiquitine (pdb 1xd3, chaîne B) à une résolution atomique. Chaque sphère représente un atome lourd (les atomes d'hydrogène ne sont pas représentés). (B) Représentation de la même protéine avec le modèle gros grain d'ATTRACT. Chaque sphère représente un pseudo-atome. Les deux figures ont été réalisées avec le logiciel PyMOL (190).

5.2.2 Fonction de score

La fonction de score d'ATTRACT est composée d'un terme de Lennard-Jones pour modéliser les forces de van der Waals et un terme de Coulomb pour modéliser les interactions électrostatiques. ATTRACT a la particularité d'employer un champ de forces permettant des interactions attractives entre paires d'atomes (voir équation (1)) mais aussi purement répulsives (voir équations (2) et (3)).

Les interactions entre deux pseudo-atomes A et B sont modélisées par les équations suivantes :

- Cas de paires attractives :

$$V = \varepsilon_{AB}\left[\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right] + \frac{q_i q_j}{\varepsilon r_{ij}} \quad (1)$$

- Cas de paires répulsives :

$$V = -\varepsilon_{AB}\left[\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right] + \frac{q_i q_j}{\varepsilon r_{ij}} \qquad \text{si} \quad r_{ij} > r_{min} \quad (2)$$

$$V = 2e_{min} + \varepsilon_{AB}\left[\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right] + \frac{q_i q_j}{\varepsilon r_{ij}} \quad \text{si} \quad r_{ij} \leqslant r_{min} \quad (3)$$

avec R_{AB} et ε_{AB} des paramètres spécifiques à chaque couple de pseudo-atomes, ε une constante diélectrique (ici $\varepsilon = 15r$), r_{ij} la distance entre les deux pseudo-atomes, q_i la charge du pseudo-atome

A, q_j la charge du pseudo-atome B. r_{min} correspond à la distance entre les deux atomes à laquelle un minimum d'énergie est atteint pour les paires attractives, ou un point de selle pour les paires répulsives (figure 5.2). e_{min} correspond à l'énergie à la distance r_{min} entre les deux atomes A et B (figure 5.2). Pour plus de détails, se référer à la publication originelle de Fiorucci et Zacharias (142)).

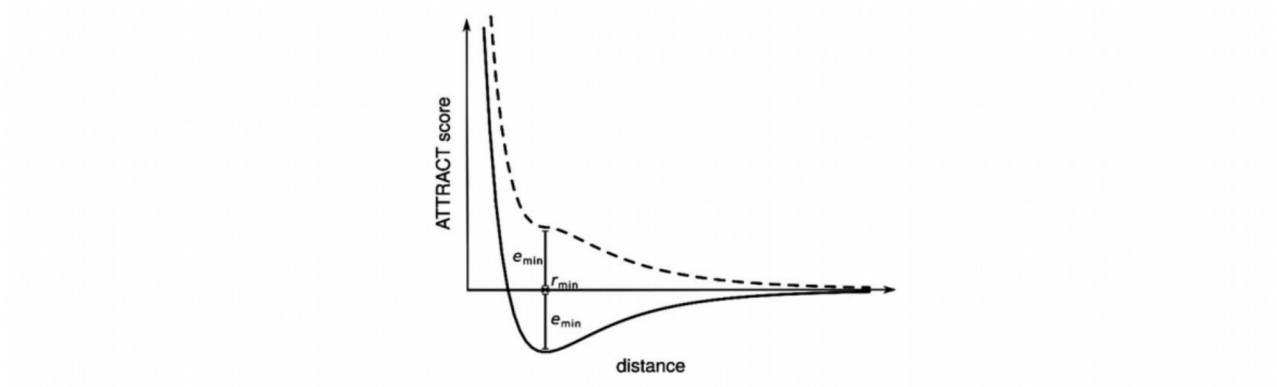


Figure 5.2. Représentation des deux formes de la fonction de score d'ATTRACT. La courbe continue représente le cas d'une paire attractive d'atomes avec un minimum d'énergie e_{min} à la distance r_{min}. La courbe en pointillés représente le cas d'une paire répulsive avec un point de selle e_{min} à la distance r_{min}. Figure extraite de Chéron et al (191).

5.2.3 Procédure de *docking* d'ATTRACT

Les calculs de docking réalisés avec ATTRACT suivent la procédure suivante :

- Des points de départ du ligand sont positionnés autour du récepteur par un algorithme modifié de Shrake et Rupley (69). Ces points de départ sont placés à la surface du récepteur à une distance légèrement supérieure au rayon de giration du ligand, afin d'éviter des recouvrements stériques entre les deux protéines au départ de la procédure de minimisation (Figure 5.3). Le nombre de points de départ est directement dépendant du rayon de giration du ligand, ainsi que de la surface totale du récepteur. Pour chacun de ces points de départ, 220 orientations différentes du ligand sont générées. Selon la taille du récepteur, plusieurs milliers à plusieurs dizaines de milliers de conformations de départ sont ainsi créées.

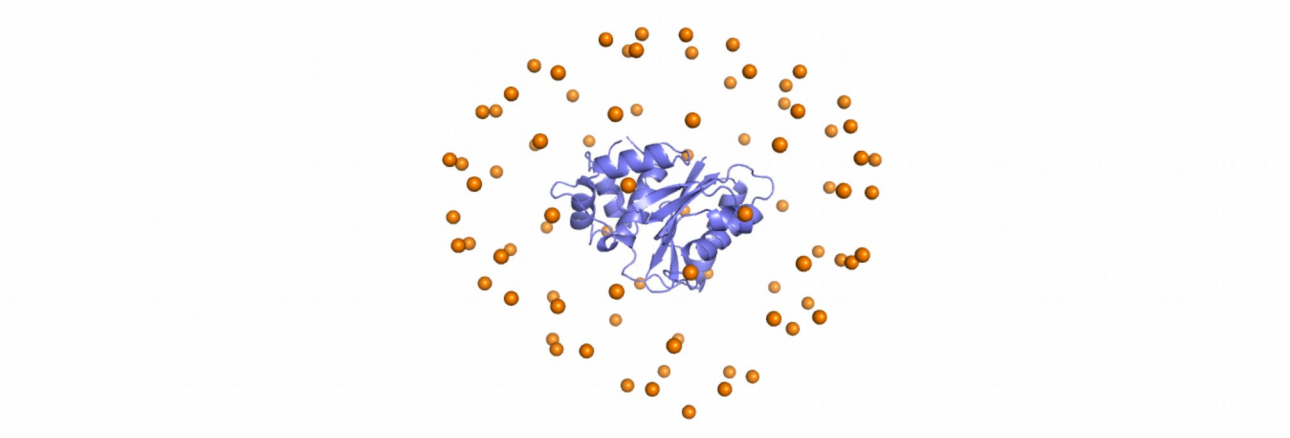


Figure 5.3. Exemple de positionnement des points de départ autour de la surface du récepteur. Le récepteur est représenté en bleu. Les positions de départ sont représentées par des sphères oranges. Elles sont placées de manière homogène autour de la surface du récepteur. Pour chaque position de départ 220 orientations différentes du ligand sont générées.

- Pour chacune des conformations de départ, plusieurs séries de minimisations d'énergie sont réalisées (généralement cinq) jusqu'à convergence dans un minimum local (voir section 5.2.3). Lors de la première étape de minimisation, une contrainte harmonique entre le centre de masse du récepteur et le pseudo-atome le plus proche du ligand est appliquée. Cette contrainte a pour but de rapprocher et de mettre en contact les deux protéines. Les étapes de minimisation suivantes permettent à la structure du ligand de converger dans un minimum local d'énergie. Il est à noter que, du fait de la manière dont ATTRACT est implémenté, plusieurs séries de minimisations sont nécessaires. En effet, pour évaluer les conformations obtenues, ATTRACT établit des listes de paires de pseudo-atomes n'appartenant pas à la même molécule (non liés). C'est à partir de ces listes de paires qu'est calculé le score de la conformation. Du fait de la modification des coordonnées du ligand par rapport au récepteur lors d'une série de minimisation, cette liste de paires de pseudo-atomes devrait être mise à jour à chaque pas de minimisation. Cependant cette procédure est coûteuse en temps de calcul. En conséquence, afin de réduire le coût computationnel, les listes de paires de pseudo-atomes ne sont générées qu'au début de chaque série de minimisation, puis remises à jour une dernière fois, une fois la convergence obtenue, pour l'évaluation finale de la conformation. De plus, durant ces étapes de minimisation, une limite sur les distances pour l'établissement des listes de paires de pseudo-atomes est fixée. Par exemple, avec une limite de 7Å, la liste créée n'inclura que des paires de pseudo-atomes distants de moins de 7Å. Cela permet d'éviter de réaliser un grand nombre de calculs sur des paires d'atomes très éloignées, et donc

n’influant pas sur le score final de la conformation. Cette limite permet de limiter les temps de calculs de manière très efficace. Les temps de calculs pour un couple de protéines sont du même ordre de grandeur que ceux de ZDOCK (129) ou FRODOCK (128).

5.2.4 Pourquoi ATTRACT ?

Au cours de cette thèse j’ai choisi d’utiliser le logiciel ATTRACT car nous avions besoin d’un outil (i) explorant de manière exhaustive la surface du récepteur et du ligand et (ii) pouvant réaliser cette exploration rapidement.

Beaucoup d’algorithmes, tels que HADDOCK (125), sont conçus pour être utilisés sous contraintes et ne permettent pas d’exploration exhaustive des surfaces protéiques à un coût computationnel acceptable. D’autres algorithmes comme ZDOCK (129) reposent sur l’utilisation de FFT pour réaliser un échantillonnage exhaustif des modes d’interactions entre les protéines mises en jeu. Plusieurs millions de conformations sur l’ensemble de la surface du récepteur peuvent ainsi être générées. Cependant, dans le cadre de notre étude, cette méthodologie présente des défauts: étant donné le très grand nombre de conformations analysées, seul un certain nombre d’entre elles (celles ayant les meilleurs scores) sont accessibles à l’utilisateur. Leur répartition n’est pas homogène sur la surface du récepteurs, les solutions étant généralement concentrées sur certaines régions du récepteur. Il serait bien entendu possible de prendre en compte un nombre supérieur de conformations jusqu’à obtenir un couverture totale de la surface du récepteur, mais cette méthodologie serait lourde du point de vue computationnel. La méthodologie employée par ATTRACT est donc mieux adaptée pour le problème qui nous intéresse. Nous avons estimé qu’ATTRACT permettait un échantillonnage et une évaluation rapide des surfaces protéiques et était donc un bon compromis.

5.3 Propriétés de la surface des protéines

5.3.1 Hydrophobicité

L’hydrophobicité est calculée selon l’échelle de Kyte et Doolittle (192). Les acides aminés ayant des valeurs négatives sur cette échelle sont hydrophiles (leur chaîne latérale affiche une préférence pour l’eau), tandis que les acides aminés ayant des valeurs positives sont dits hydrophobes.

5.3.2 *Stickiness*

La *stickiness* mesure la propension d'un type d'acide aminé à être présent aux interfaces protéiques par rapport à sa présence dans les surfaces protéique. Il s'agit d'une métrique purement statistique. Elle est corrélée avec certaine mesures d'hydrophobicité (106) telle que l'échelle d'hydrophobicité de Wimley-White (107).

Elle est définie de la manière suivante :

$$stk_i = \log\left(\frac{freqAAi_{interface}}{freqAAi_{surface}}\right)$$

avec stk_i la *stickiness* de l'acide aminé *i*, $freqAAi_{interface}$ la fréquence de l'acide aminé *i* aux interfaces protéiques et $freqAAi_{surface}$ la fréquence de l'acide aminé i aux surfaces protéiques. Un acide aminé "collant" possède une valeur de *stickiness* supérieure à 0, tandis qu'un acide aminé "non collant" possède une valeur de *stickiness* inférieure à 0.

Des échelles de *stickiness* ont été développées dans plusieurs études, sur des jeux de données différents (60,64,106). Leurs propriétés sont généralement similaires mais dépendent des spécificités du jeu de données utilisé, par exemple Ofran et Rost ont montré que la composition en acides aminés des interfaces homomériques est significativement différente de celle des interfaces hétéromériques. Par conséquent une échelle de *stickiness* développée sur des complexes homomériques n'aura pas les mêmes valeurs qu'une développée sur des complexes hétéromériques. Dans ce travail, nous référons systématiquement à l'échelle de propension à l'interaction développée par Levy et al (77,106) qui a été réalisée sur un jeu de données réunissant 397 structures de *E. coli.*

5.3.3 Variance circulaire

La variance circulaire mesure l'étendue de la distribution vectorielle d'un ensemble de points autour d'un point fixe dans l'espace. Cette mesure a été introduite pour l'étude de la topographie des protéines par Mezei (193). Dans le cas d'un atome appartenant à une protéine, la variance circulaire représente la densité en atomes autour de cet atome, ce qui permet de mesurer le degré d'enfouissement d'un atome dans une protéine. Concrètement, plus la variance circulaire d'un atome ou d'un acide aminé est faible, plus celui-ci est exposé, c'est-à-dire plus celui-ci appartient à une région protubérante. Inversement plus la variance circulaire d'un atome est élevée, plus celui-ci est enfouie dans la protéine (Figure 5.4). Par rapport à un calcul standard d'accessibilité au solvant,

la variance circulaire a pour avantage d'être moins sensible aux petits changements conformationnels (194). Elle s'écrit de la manière suivante :

$$CV_i = 1 - \frac{1}{n_i} \left| \sum_{j \neq i, r_{ij} \leqslant r_c} \frac{\vec{r}_{ij}}{\|\vec{r}_{ij}\|} \right|$$

avec CV_i la variance circulaire de l'atome *i*, n_i le nombre d'atomes distant de moins de r_c Å de l'atome *i*, r_{ij} la distance entre les atomes *i* et *j*, $\vec{r}_{ij}$ le vecteur de coordonnées *(i,j)* et $\|\vec{r}_{ij}\|$ la norme du vecteur $\vec{r}_{ij}$. r_c représente la distance maximale des atomes environnants. Au cours de ce travail nous avons utilisé r_c = 10 Å.

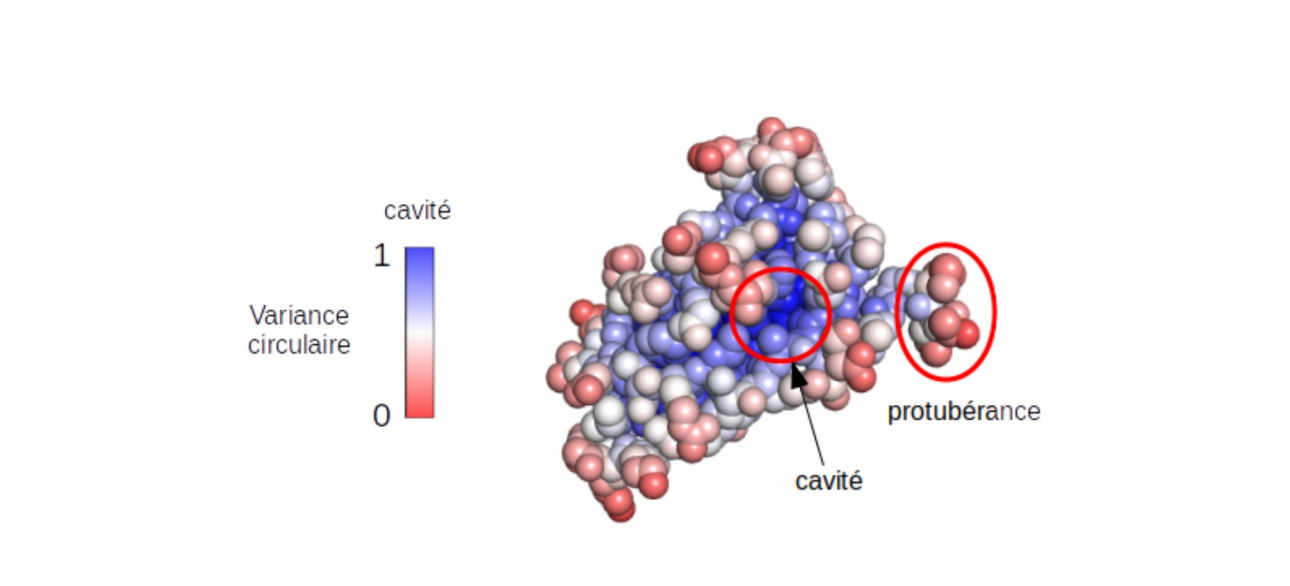


Figure 5.4. Variance circulaire des atomes d'une protéine. Les atomes sont colorés en fonction de leur variance circulaire. Les atomes enfouis ont une variance circulaire proche de 1 et sont colorés en bleu. Les atomes protubérants ont une variance circulaire proche de 0 et sont colorés en rouge.

5.4 Projection et cartographie

On distingue deux protocoles différents : le premier pour créer une carte de propriétés physico-chimiques ou évolutive de la surface d'une protéines (hydrophobicité, *stickiness*, variance circulaire, conservation de séquence), le deuxième pour créer les cartes d'énergie à partir des calculs de *docking* entre deux protéines.

5.4.1 Cartographie des propriétés de surface des protéines

La cartographie des propriétés de surface des protéines se fait en 5 étapes (figure 5.5):

1- Calcul des propriétés des résidus de surface de la protéine.

2- Génération d'un ensemble de particules autour de la surface de la protéine.

3- Assignation aux particules de la valeur de l'atome de la protéine le plus proche.

4- Projection sinusoïdale des coordonnées sphériques des CM des particules par rapport au CM de la protéine.

5- Division de la carte obtenue en grille et lissage.

1- Calcul des propriétés de surface de la protéine

La première étape consiste à calculer les valeurs de la propriétés d'intérêt des résidus de la protéine. L'hydrophobicité, la *stickiness* et la conservation de séquence sont calculées à l'échelle du résidu à partir des échelles de valeurs déterminées respectivement dans (192) et (106). La variance circulaire est calculée à l'échelle atomique.

2- Génération de particules autour de la protéine d'intérêt

La deuxième étape consiste à générer un ensemble de particules autour de la surface de la protéine d'intérêt à l'aide de l'algorithme d'échantillonnage de ATTRACT (126) (algorithme de Shrake et Rupley légèrement modifié (69). Chaque particule est générée à une distance de 5 Å de la surface de la protéine. La densité est de une particule par $Å^2$ (figure 5.5A).

3- Assignation de valeurs à chacune des particules

La valeur de l'atome le plus proche est assignée à chaque particule.

4- Projection des coordonnées des particules sur une carte en deux dimensions

Les coordonnées de chacune des particules sont exprimées en coordonnées sphériques par rapport au centre de masse de la protéine d'intérêt. Un angle θ est calculé par rapport à l'axe *z* (figure 5.5A**)** et un angle φ par rapport au plan formé par les axes *x* et *y*. Puis leurs coordonnées sont projetées sur un plan en deux dimensions par une projection sinusoïdale (figure 5.5B).

5- Division de la carte en une grille et lissage

La carte obtenue est ensuite divisée en une grille de dimension 72 x 36 cases. Les valeurs des particules contenues dans une même case sont moyennées. Un lissage est ensuite réalisé sur la carte en moyennant le score de chaque cellule avec les scores des cellules adjacentes (figure 5.5B et C).

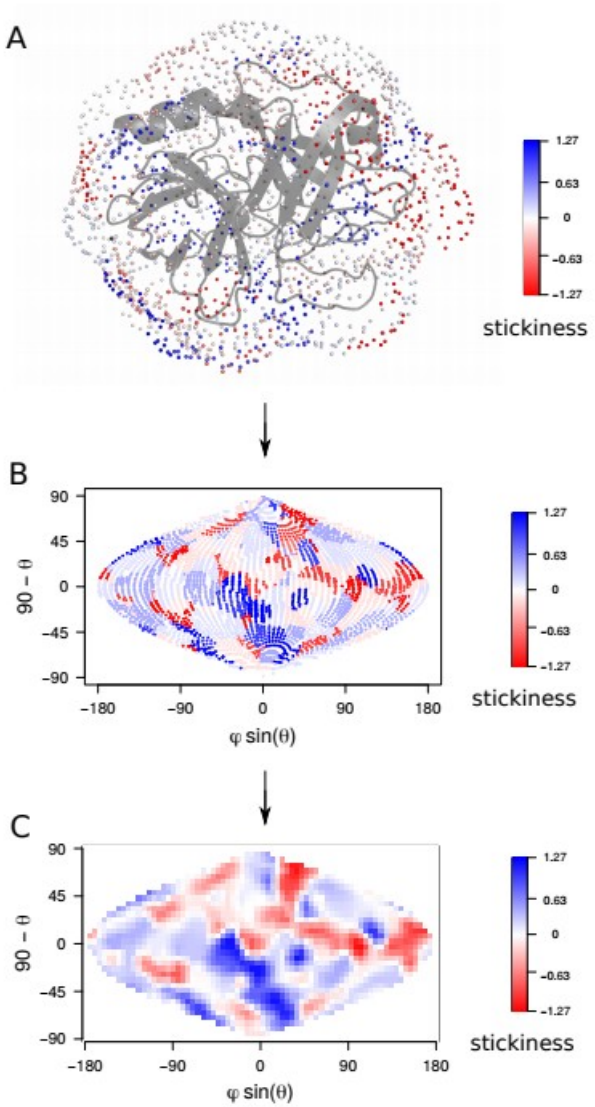


Figure 5.5. Procédure de création d'une carte de propriétés de surface. (A) Des particules sont générées de manière homogène autour de la surface de la protéine d'intérêt. Chaque particule est colorée en fonction de la valeur assignée à l'atome de la protéine le plus proche. (B) Les coordonnées sphériques des particules sont projetées sur une carte en deux dimensions par une projection sinusoïdale. (C) La carte obtenue est divisée en une grille de 72 x 36 cellules. Le score d'une cellule est la moyenne des valeurs de ses particules. La carte est lissée en moyennant la valeur de chaque cellule avec les valeurs des cellules adjacentes.

Par rapport à la méthodologie présentée, il est aussi possible de cartographier directement les positions des atomes des résidus de surface, ce qui rendrait les étapes 2 et 3 superflues. Cependant cette méthode peut poser problème en cas de cartographie de petites protéines du fait du faible nombre de résidus (et donc d'atomes) de surface, ce qui laisserait une grande partie des cellules de la carte non assignées. Notre procédure permet de générer un nombre suffisant de particules avec une distribution homogène autour de la surface de la protéine d'intérêt pour pouvoir cartographier les propriétés de surface de protéines avec un petit nombre de cellules non-assignées (et qui se voient assigner une valeur lors de l'étape lissage), y compris pour de petites protéines telles que l'ubiquitine (76 résidus). De plus la densité de particules générées est modulable est peut-être augmentée ou diminuée. A cela s'ajoute le fait que notre méthode a l'avantage de ne pas nécessiter de seuil d'accessibilité au solvant pour définir les résidus de surface, étant donné qu'ici un résidu est considéré comme résidu de surface si au moins un de ses atomes est assigné à une des particules (voir étape 3).

5.4.2 Cartographie des sites d'interaction protéiques

Dans cette étude, j'ai été souvent amené à vouloir visualiser et/ou projeter la localisation des sites d'interaction des protéines sur une carte en deux dimensions. J'ai donc mis en place une procédure pour projeter les sites d'interaction sur des cartes en deux dimensions, que nous appellerons cartes de sites d'interaction.

Cette fois nous assignons une valeur de 0 aux atomes des résidus n'appartenant à aucun site d'interaction, et un entier supérieur à 0 aux atomes des résidus appartenant à un site d'interaction. Les résidus appartenant à un même site d'interaction se voient attribuer le même entier, tandis que ceux appartenant à deux sites d'interaction différents (c'est-à-dire impliquant deux chaînes protéiques différentes) se voient attribuer des entiers différents (Figure 5.6).

Les trois premières étapes (voir section 5.4.1) sont identiques à celles utilisées pour la création des cartes de propriétés de surface. La quatrième étape (voir section 5.4.1 « Projection des coordonnées des particules sur une carte en deux dimensions ») est différente car au lieu de moyenner les valeurs des particules à l'intérieure d'une même cellule, c'est le score de l'entier majoritaire des particule présentes dans la cellule qui est assigné à celle-ci. En cas d'égalité, la plus haute valeur est assignée. S'il n'y a pas de particules dans une case, un score de 0 lui est assigné. De plus,le lissage final n'est pas réalisé.

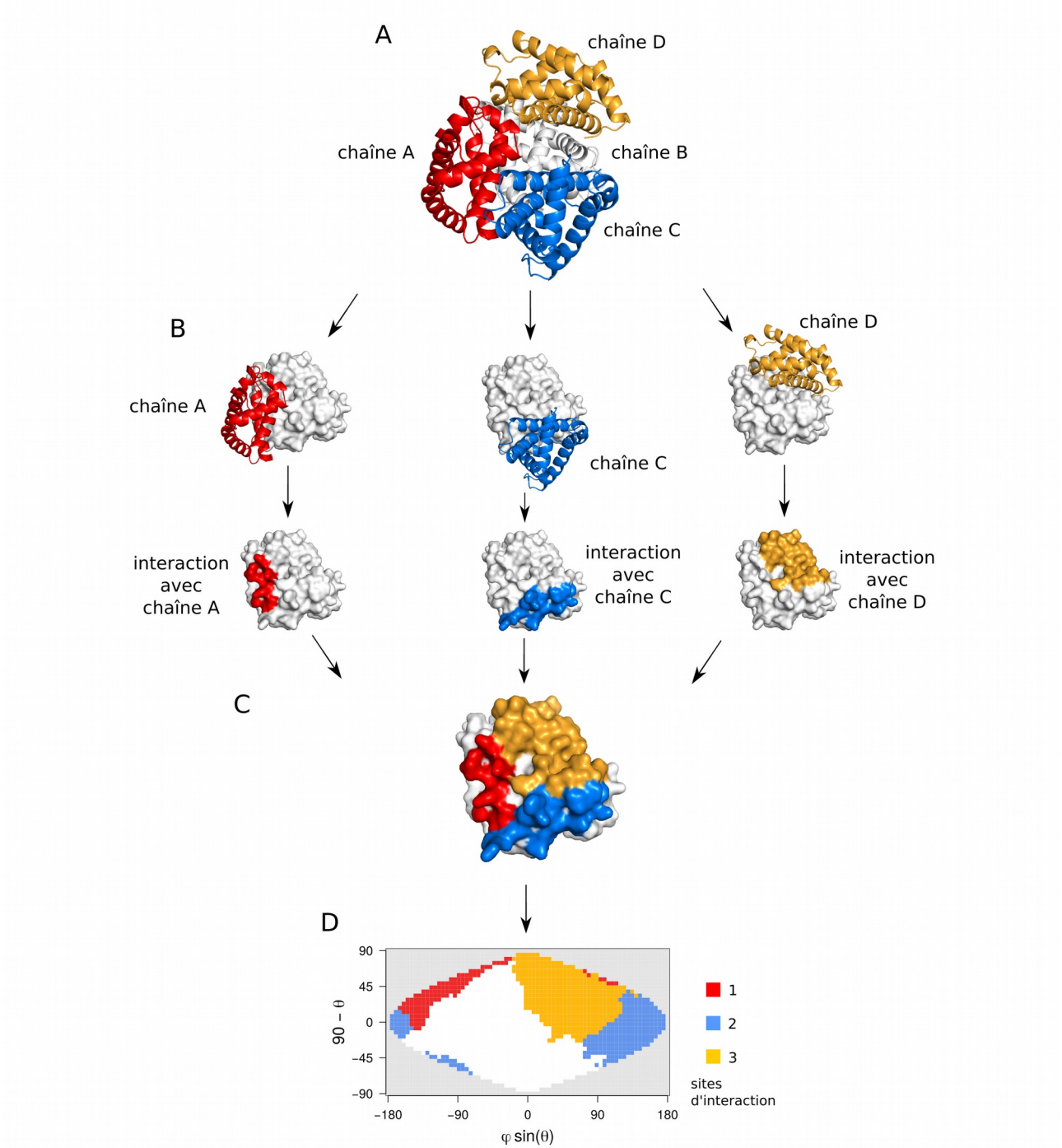


Figure 5.6. Cartographie des sites d'interaction d'une protéine sur une carte. (A) L'hémoglobine de *D. akajei* est composée de 4 sous-unités (pdb 1cg5). (B) La sous-unité B (en blanc, représentation surface) interagit avec les chaînes A, C et D colorées respectivement en rouge, bleu et orange et représentées en mode *cartoon*. (C) Les trois sites d'interaction de la chaîne B sont projetés sur la surface de la protéine puis (D) celle-ci est projetée sur une carte en deux dimensions suivant le protocole décrit dans la section 5.4. Les cellules appartenant à l'interface formée avec la chaîne A sont colorées en rouge, les cellules appartenant à l'interface formée avec la chaîne C sont colorées en bleu et celles avec la chaîne D en orange. Le reste des cellules est coloré en blanc.

5.4.3 Cartes d’énergie

La construction d’un carte d’énergie se fait en trois étapes auxquelles s’ajoute une étape de discrétisation optionnelle :

1- Calcul de *docking* entre les deux protéines d’intérêt.

2- Projection des centres de masse des solutions de *docking* sur une carte en deux dimensions par une projection sinusoïdale.

3- Création d’une carte continue à partir des coordonnées discrètes des solutions de *docking* suivi du lissage de la carte d’énergie.

4- Discrétisation optionnelle de la carte d’énergie en cinq classes d’énergie.

5- Binarisation de la carte de classes d’énergie en cinq cartes représentant chacune une classe d’énergie.

1- Calcul de *docking*

Un calcul de *docking* entre les deux protéines d’intérêt est réalisé avec le logiciel ATTRACT (136). Les coordonnées du récepteur sont maintenues fixes durant la procédure et les protéines sont considérées comme des objets rigides. Cinq étapes de minimisation d’énergie sont réalisées, chacune avec des paramètres spécifiques (voir tableau 5.1). Au terme de ces minimisations une réévaluation des conformations est effectuées en prenant en compte les interactions entre tous les pseudo-atomes des deux protéines. À la fin du calcul de *docking* les solutions redondantes sont éliminées. Toutes les solutions de *docking* ayant obtenu un score supérieur à 0 sont aussi éliminées. En effet, celles-ci présentent souvent des interpénétrations stériques entre pseudo-atomes. Étant donné que ce sont ces interpénétrations qui sont la cause des scores hautement défavorables, ces solutions ne sont pas considérées comme pertinentes.

À la fin de cette procédure plusieurs milliers à plusieurs dizaines de milliers de solutions de *docking* ont été générées. Ces solutions sont généralement réparties autour de la surface du récepteur (Figure 5.7A).

Tableau 5.1: paramètres utilisés lors de l'exécution d'un calcul de *docking* avec ATTRACT (voir section 5.2.3)

Étape de minimisation	seuil (Å²)	v_{max}
1	1500	50
2	500	60
3	150	60
4	80	100
5	80	100

- v_{max} correspond au nombre maximal de pas de minimisation autorisé pendant l'étape de minimisation d'énergie.
- seuil correspond au carré de la distance seuil maximale (en Å) autorisée entre deux pseudo-atomes pour l'établissement des listes de pseudo-atomes mises à jour au début de chaque étape de minimisation d'énergie.

2- Représentation du paysage énergétique en coordonnées sphériques

Le centre de masse du ligand de chaque solution de *docking* est représenté en coordonnées sphériques par rapport au centre de masse du récepteur. Un angle θ est calculé par rapport à l'axe *z* (Figure 5.7A**)** et un angle φ par rapport au plan formé par les axes *x* et *y*. À chaque centre de masse est associé le score de la solution de *docking* correspondante.

3- Projection cartographique des solutions de *docking*

Les coordonnées sphériques des solutions de *docking* retenues sont projetées sur un plan en deux dimensions par une projection sinusoïdale ($x = \varphi \sin(\theta)$; $y = 90 - \theta$). Cette projection a pour propriété de conserver les surfaces, ce qui signifie que deux zones de surface équivalente sur la carte sont de même aire sur la sphère ayant servie de base à la projection. Dans cette projection les parallèles forment des lignes droites et les méridiens ont une forme sinusoïdale (le méridien central étant droit) (Figure 5.7B).

La carte obtenue est divisée en par une grille de dimension 72 x 36 cellules. Pour chaque cellule, toutes les solutions de *docking* avec un score situé à moins de 2.7 $Kcal.mol^{-1}$ de la solution d'énergie la plus basse sont retenues, selon un protocole de filtrage des solutions de docking implémenté dans (166). Le score de chaque cellule est égal à la moyenne des scores des solutions de *docking* retenues. Si aucune solution de *docking* n'est présente dans la cellule, un score de 0 lui est

attribué. La carte obtenue est ensuite lissée Figure 5.7C : le score de chaque cellule est moyenné avec le score des huit cellules adjacentes. Les cellules ne possédant pas huit voisins dans la projection (cellules situées en bordure de la projection) sont éliminées lors de cette étape.

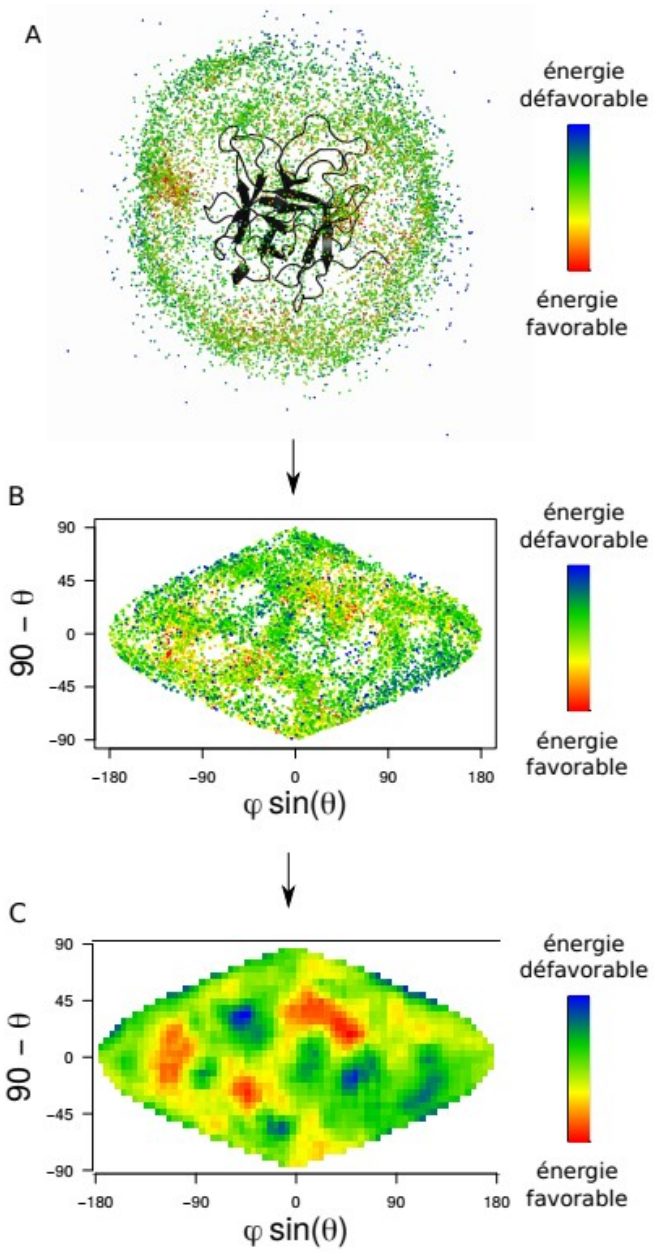


Figure 5.7. Création d'une carte d'énergie à partir d'un calcul de *docking*. (A) Les centres de masse du ligand (CM) de toutes les solutions de *docking* sont représentées en coordonnées sphériques par rapport au CM du récepteur. Les scores des solutions de *docking* sont assignés à leurs CM. Les CM des solutions de *docking* favorables sont donc colorés en rouge, ceux des solutions défavorables en bleu. (B) Les coordonnées sphériques des CM des solutions de *docking* sontt projetées sur une carte en deux dimensions par une projection sinusoïdale. (C) La projection obtenue est discrétisée puis lissée. La carte obtenue est appelée carte d'énergie.

4- Création des cartes de classes d'énergie

Les cartes d'énergie peuvent être discrétisées en carte de classes d'énergie (Figure 5.8). Celles-ci sont composées de cinq classes d'énergie représentant le "degré d'attractivité" du ligand pour la surface du récepteur. Les gammes de valeurs des cinq classes d'énergie sont calculées de la manière suivante :

1- Calcul de l'amplitude de l'intervalle des valeurs des scores d'énergie minimaux et maximaux de la carte

$$range_{map} = |max(E) - min(E)| \quad (1)$$

avec *max(E)* et *min(E)* respectivement le score d'énergie maximal et minimal de la carte d'énergie.

2- Calcul de l'amplitude des intervalles de valeur d'énergie de chaque classe

$$Cl_{range} = range_{map} / 5 \quad (2)$$

3- Calcul des gammes de valeurs de chaque classe d'énergie (i) de la carte

$$Icl_i = [maxE - range_{map} \times (i-1), maxE - range_{map} \times i] \quad (3)$$

Avec *i* allant de 1 à 5.

4- Création de la carte de classes d'énergie

La carte de classe d'énergie est créée en remplaçant chaque cellule de la carte d'énergie par la classe d'énergie à laquelle appartient sa valeur. Il est important de noter que cette discrétisation est dépendante des valeurs minimales et maximales de la carte d'énergie considérée. Les 5 classes d'énergie ordonnées des plus faibles (les plus favorables) aux plus fortes (les plus défavorables) seront représentées respectivement en rouge, jaune, vert clair, vert foncé et bleu (Figure 5.8B).

5- Binarisation des cartes d'énergie

Pour chaque classe d'énergies (pour chaque couleur), les cellules appartenant à la classe (à la zone de couleur) se voient attribuer un score de 1, les autres cellules un score de 0. Cinq cartes binaires sont donc ainsi générées pour une même protéine (Figure 5.8C).

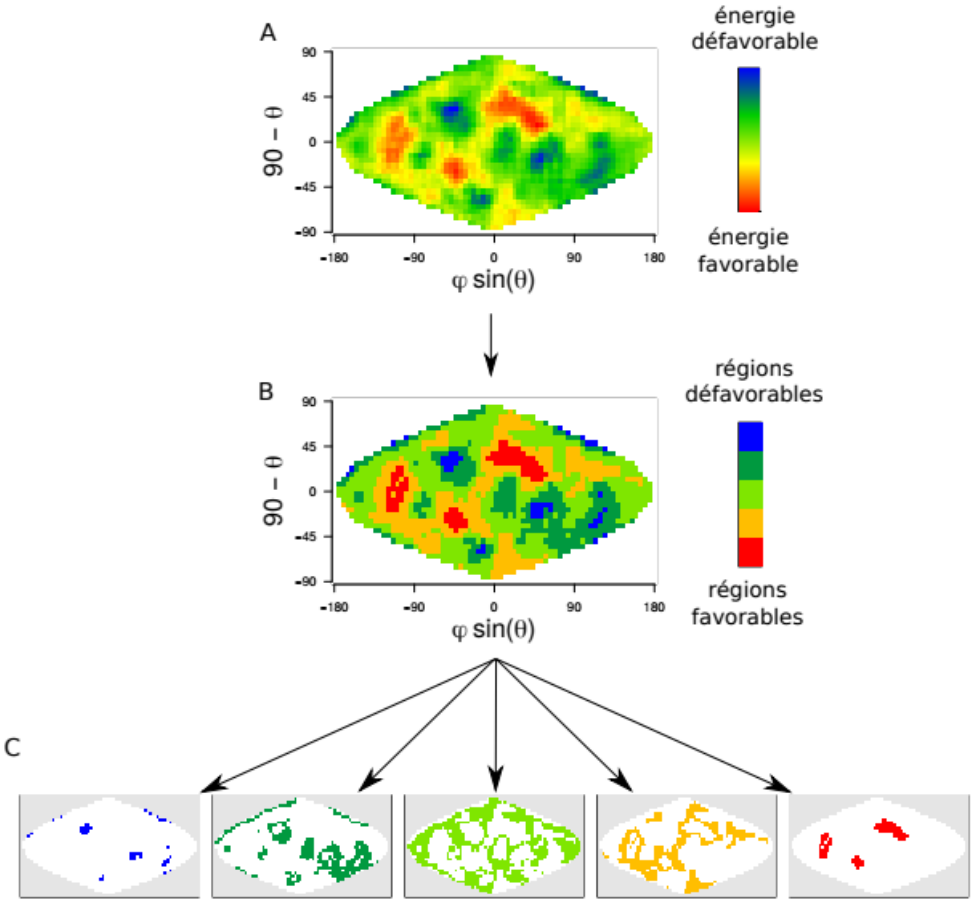


Figure 5.8. Discrétisation des cartes d'énergie. (A) Carte d'énergie. (B) La carte d'énergie est discrétisée en cinq classes d'énergie, chacune couvrant une gamme de scores d'énergie de même amplitude. (C) Les régions correspondant à chaque classe d'énergie sont extraites pour créer des cartes représentant chaque classe d'énergie séparément.

5.4.4 Dénombrement du nombre d'îlots par carte

Une fois que l'on a discrétisé une carte d'énergie, il est possible d'extraire les différents îlots d'une même classe d'énergie à partir de la carte discrète. Dans le cas d'une carte correspondant aux régions favorables à l'interaction (c'est-à-dire les cartes rouges) on pourra identifier les régions de la surface de la protéine qui sont favorables à l'interaction avec le ligand d'intérêt. Il est à noter que dans chaque carte binarisée, les cellules situées à l'extrême droite d'une carte sont considérées

voisins de celles situées à l'extrême gauche et inversement. En effet ces cellules appartiennent à un unique îlot et sont séparées artificiellement sur la carte à cause de la projection.

5.4.5 Création des cartes de potentiel d'interaction des surfaces protéiques (IPOPS)

Les cartes IPOPS sont des cartes de reflétant le potentiel de la surface d'une protéine à interagir avec un ensemble de ligands d'intérêt. Elles correspondent à la somme de cartes d'une même classe d'énergie d'un récepteur amarré avec un nombre *n* de ligands (Figure 5.9). Les cartes IPOPS sont réalisées en quatre étapes :

1- Somme des cartes d'énergie discrétisées binaires d'une classe d'énergie

Le score d'une cellule d'une carte IPOPS peut s'écrire :

$$IPS_{ij} = \sum_{k=1}^{l} SC_{ij}(k)$$

avec IPS_{ij} le score de la cellule de coordonnées (i,j), l le nombre de cartes binarisées, $SC_{ij}(k)$ le score de la cellule de coordonnées ij appartenant à la k^{ieme} carte binarisée.

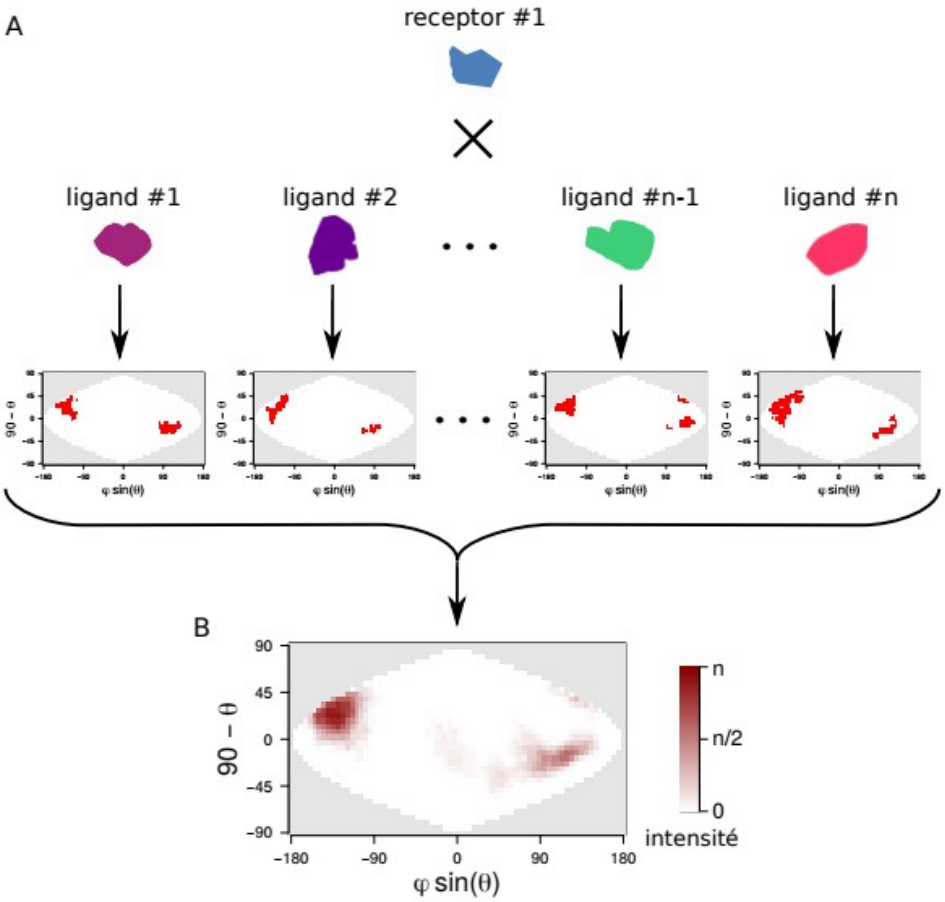


Figure 5.9. Création d'une carte IPOPS appliquée aux régions rouges. (A) Des calculs de *docking* sont réalisés entre une protéine récepteur et *n* ligands. De ces calculs de *docking* n cartes des régions rouges sont créées. (B) Les cartes sont sommées pour créer une carte IPOPS. La procédure est la même pour les autres classes d'énergie.

2. Élimination du bruit de fond

Afin d'éliminer les zones de faible intensité assimilables à du bruit de fond, j'ai utilisé une méthode de segmentation d'images implémentée dans le package *dpmixsim* de R. Ce package analyse la distribution des cellules d'une image et cherche à la modéliser par un mélange de plusieurs lois normales. Une fois le modèle estimé, il filtre l'image de manière à éliminer tous les cellules dont les intensités sont les plus faibles, c'est-à-dire qu'elles appartiennent à la loi normale de plus faible moyenne. Le « bruit de fond » de l'image est ainsi éliminé (Figure 5.10).

Le nombre de lois et les paramètres de ces lois étant inconnus, ils doivent être estimés. Le package *dpmixsim* implémente une méthode bayésienne non paramétrique (Processus de Dirichlet à mélange) dont les réalisations sont des mélanges de lois. Il permet de prendre en compte le caractère aléatoire du nombre de lois qui n'a donc pas besoin d'être défini à l'avance et sera estimé

à partir des données. Cette méthode repose sur des simulations MCMC (Monte Carlo Markov *Chain*) pour explorer des modèles de mélange gaussiens avec un nombre inconnu de composants.

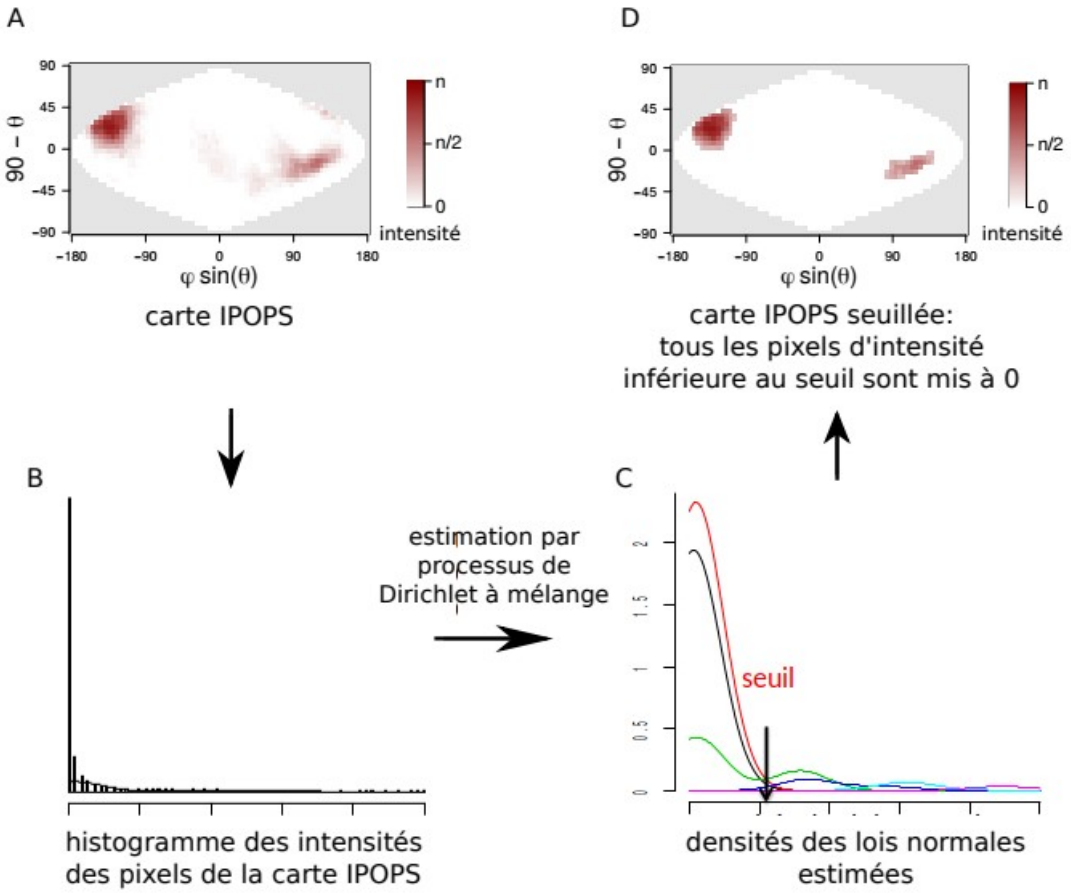


Figure 5.10. Elimination des cellules de faible intensité des cartes IPOPS. (A) Une carte IPOPS peut présenter des cellules de faible intensité assimilable à du bruit de fond. (B) La distribution de l'intensité des cellules est réalisée. (C) La distribution des intensités est approximée par plusieurs lois normales, permettant d'établir un seuil d'intensité des cellules. (D) Les intensités des cartes sont filtrées : toutes les intensités inférieures au seuil sont éliminées.

3. Élimination des îlots de trop petites taille

Après élimination du bruit de fond, il demeure des îlots de très petite taille que l'on ne peut considérer comme une zone d'intérêt (« favorable », ..., « défavorable » à l'interaction). On souhaite donc les éliminer. Pour cela encore j'ai utilisé une analyse de la distribution, mais cette fois-ci non plus des intensités mais de la taille des îlots. Ne pouvant utiliser comme précédemment le package *dpmixsim* qui prend en entrée une image, j'ai utilisé la fonction *normalmixEM* du package *mixtools* pour modéliser la densité de la taille des îlots par un mélange de lois normales. Là

aussi, c'est une méthode bayésienne qui est implémentée : l'estimation des paramètres des lois normales est réalisée par l'algorithme EM qui cherche à maximiser la vraisemblance du modèle. Contrairement au processus de Dirichlet, le nombre de lois normales doit être fixé. J'ai donc estimé les vraisemblances des modèles en faisant varier le nombre de lois normales de 2 à 10 et sélectionné le meilleur modèle. Comme précédemment, les îlots dont la taille appartenaient à la première distribution estimée ont été éliminées.

4- Assignation des îlots de forte et de faible intensité

Dans le cadre de l'analyse des cartes IPOPS des régions favorables (réalisées à partir des cartes d'énergie très favorables, c'est-à-dire correspondant aux régions rouges sur la carte d'énergie), j'ai séparé les îlots en deux types : les îlots favorables de forte intensité et les îlots favorables de faible intensité. Afin de réaliser cette dichotomie entre îlots de forte et de faible intensité, j'ai utilisé la même méthode que pour celle utilisée pour éliminer les îlots de petite taille (fonction *normalmixEM* du package *mixtools*). Cette méthode m'a permis de déterminer un seuil d'intensité moyenne d'un îlot. Si son intensité est inférieure à ce seuil, alors cet îlot est considéré comme étant de faible intensité. A l'inverse si son intensité moyenne est supérieure à ce seuil, celui-ci est considéré comme un îlot de forte intensité.

5.4.6 Comparaison des îlots des cartes IPOPS avec les sites d'interaction protéiques

Comparer les îlots des cartes IPOPS avec les sites d'interaction protéiques peut se faire (i) en comparant directement les îlots IPOPS avec les îlots des cartes de sites d'interaction ou (ii) en comparant les résidus prédits par les cartes IPOPS avec ceux appartenant respectivement aux sites d'interaction et aux surfaces protéiques.

La première approche peut se réaliser par un calcul de recouvrement entre les îlots IPOPS et les îlots des sites d'interaction. Cependant notre but ici est de montrer à quel point les îlots IPOPS sont recouverts par les sites d'interaction (voir section 6.3.4), et non pas à quels points ceux-ci sont similaires aux sites d'interaction. Nous avons donc privilégié une autre méthode, qui consiste à calculer la fraction des cellules des îlots IPOPS qui appartiennent effectivement aux îlots des sites d'interaction d'une carte de sites d'interaction, suivant la formule suivante :

$$Recouv_{cell}(i) = \frac{Card(cell_{IPOPS}(i) \cap cell_{SI}(i))}{Card(cell_{IPOPS}(i))}$$

avec $Recouv_{prot}(i)$ la fraction des cellules des îlots IPOPS de la protéine *i* appartenant à un site d'interaction, $cell_{SI}(i)$ les cellules de la carte de sites d'interaction de la protéine *i* appartenant à un site d'interaction, $cell_{IPOPS}(i))$ les cellules de la carte IPOPS rouge de la protéine *i* appartenant à un îlot IPOPS, *Card* le cardinal (nombre d'éléments) d'un ensemble. Notons que calculer $Recouv_{prot}(i)$ équivaut à calculer la PPV (voir section 5.5) des îlots IPOPS d'une protéine pour la prédiction de ses sites d'interaction. Le défaut de cette approche réside dans le fait qu'elle ne fournit aucune information au niveau moléculaire, notamment quels résidus sont prédits par les îlots IPOPS.

La deuxième approche est plus complexe car elle nécessite de déterminer quels sont les résidus appartenant aux îlots IPOPS (carte en deux dimensions) et de les comparer avec ceux appartenant aux sites d'interaction (structure en trois dimensions). Il faut donc comparer une structure 3D avec une carte 2D. Pour réaliser cette comparaison nous avons cartographié les coordonnées des résidus de surface en procédant d'une manière similaire à la création des cartes de propriétés physiques (voir section 5.4.1) :

1- génération d'un ensemble de particules autour de la protéine d'intérêt

2- à chaque particule est associé l'identifiant du résidu le plus proche

3- cartographie des positions de chaque particule sur la carte.

4- la carte est divisée en 72 x 36 cellules. A chaque cellule est associée la liste des particules dont les coordonnées sont contenues dans celle-ci. Cette liste de particules permet d'obtenir la liste des résidus associée à chaque cellule, cette liste contenant tout les résidus représentés par au moins une particule dans la cellule.

Avec cette procédure, à chaque cellule de la carte est associé un ensemble de résidus. Autrement dit avec cette méthode à chaque résidu de surface est associé un ensemble de cellules de la carte (ensemble pouvant être vide pour un résidu enfoui par exemple). Il convient de garder à l'esprit que les ensembles de cellules représentant les résidus sur la carte ne sont pas mutuellement exclusifs (des particules représentant plusieurs résidus peuvent être incluses dans une même cellule). Avec cette méthodologie il suffit qu'une particule associée à un résidu soit présente dans un îlot IPOPS pour que ce résidu soit prédit comme appartenant à un site d'interaction.

Suite à cette procédure, la fraction des résidus des îlots IPOPS appartenant effectivement aux sites d'interaction est calculée de la manière suivante :

$$Recouv_{res}(i) = \frac{Card\left(res_{IPOPS}(i) \cap res_{SI}(i)\right)}{Card\left(res_{IPOPS}(i)\right)}$$

avec $Recouv_{res}(i)$ la fraction des résidus détectés par les îlots IPOPS de la protéine *i* et appartenant à un site d'interaction, $res_{SI}(i)$ les résidus appartenant à un site d'interaction de la protéine *i*, $res_{IPOPS}(i)$ les résidus de la carte IPOPS rouge de la protéine *i* appartenant à un îlot IPOPS. *Card* désigne le cardinal d'un ensemble.

Une autre façon de procéder peut être de calculer et de projeter les coordonnées sphériques des atomes des résidus de surface sur une carte, puis de comparer ces cartes avec les cartes IPOPS. Cependant avec notre méthode nous n'avons pas à définir quels sont les résidus de surface. En effet ici les résidus de surface sont ceux qui sont associés à au moins une particule et donc tout ceux qui contribuent à la création des cartes de propriétés physiques (ou de sites d'interaction). De plus notre procédure nous permet d'être méthodologiquement cohérent avec la création des cartes d'énergie et de propriétés biophysiques que nous avons décrites plus haut (voir sections 5.4.1 et 5.4.3).

5.4.7 Calculs de la NIP

La NIP est un indice développé pour la prédiction de sites d'interaction protéiques. Elle a été développée dans (153) et a été utilisée (avec quelques modifications) dans d'autres études (156,159,160,165,166). Cet indice se base sur la fréquence d'un résidu à être présent dans les meilleures solutions de *docking*. Un résidu appartenant souvent aux meilleures solutions de *docking* étant prédit comme appartenant à un site d'interaction.

Pour calculer la NIP d'un récepteur P1 j'ai procédé de la manière suivante : J'ai réalisé un calcul de *docking* entre le récepteur et 100 ligands. Pour chaque calcul de *docking*, j'ai sélectionné la meilleure solution ainsi que toute celles dont le score est situé à moins de 2.7 kcal.mol^{-1} de cette solution, suivant le protocole établi dans (166). Puis j'ai calculé la IP (*Interaction Propensity*) de tous les résidus *i* de P1 pour chaque calcul de *docking*, de la manière suivante :

$$IP_{P1}(i) = \frac{N_{inter,P1}(i)}{N_{pos,P1}}$$

avec $IP_{P1}(i)$ la propension à l'interaction du résidu *i*, $N_{inter,P1}(i)$ le nombre de fois que le résidu i est vu comme appartenant à une solution de *docking et* $N_{pos,P1}$ le nombre de solutions de *docking* conservées pour la protéine P1.

La NIP (*Normalized Interaction Propensity*) est ensuite calculée de la façon suivante :

$$NIP_{P1}(i) = \frac{IP_{P1}(i) - <IP_{P1}(j)>_{j \in P_1}}{max(IP_{P1}(j))_{j \in P1} - <IP_{P1}(j)>_{j \in P_1}}$$

avec $IP_{P1}(i)$ la propension à l'interaction du résidu *i*, $<IP_{P1}(j)>_{j \in P_1}$ l'IP moyen des résidus de surface *j* de P1, et $max(IP_{P1}(j)_{j \in P_1})$ l'IP la plus élevée obtenue parmi les résidus de surface de P1. Les résidus ayant un score NIP supérieur à 0 sont prédits comme appartenant à un site d'interaction, ceux avec un score NIP inférieur ou égal à 0 sont prédits comme n'appartenant pas à un site d'interaction.

5.4.8 Calculs de distances entre cartes IPOPS

Les calculs de distance entre carte IPOPS réalisés dans la section 6.3.5 ont été réalisés avec la formule suivante : pour deux cartes A et B, la distance entre les deux cartes est égale au nombre de cellules ayant une valeur différente entre les deux cartes.

5.5 Critères d'évaluation des capacités prédictives des cartes

Plusieurs métriques ont été utilisées afin de mesurer la capacité des cartes d'énergie (discrétisées ou non en classes d'énergie) à identifier des protéines homologues (voir section 8) et la capacité des cartes IPOPS à prédire les sites d'interaction protéiques (section 6.3.6). Dans les formules qui suivent, les entités dites « positives » sont soit les paires de ligands homologues (travail de la section 8) soit les résidus appartenant à un site d'interaction (travaux des sections 6 et 7) et les entités dites « négatives » sont soit les paires de ligands non homologues soit les résidus appartenant à la surface d'une protéine mais pas à un site d'interaction. TP désigne le nombre de vrais positifs, c'est-à-dire le nombre d'entités positives prédites comme telles. TN désigne le nombre de vrais négatifs, c'est-à-dire le nombre d'entités négatives prédites comme telles. FP désigne le nombre de faux positifs, c'est-à-dire le nombre d'entités négatives prédites comme

positives et FN désigne le nombre de faux négatifs, c'est-à-dire le nombre d'entités positives prédites comme négatives.

Les mesures que j'ai employées sont les suivantes :

- La sensibilité : fraction de vrais positifs prédits comme positifs.

$$Sen = \frac{TP}{TP + FN}$$

- La spécificité : fraction de vrais négatifs prédits comme négatifs.

$$Spe = \frac{TN}{TN + FP}$$

- La PPV (*positive predictive value*) : fraction d'entités prédites comme positives et effectivement positives.

$$PPV = \frac{TP}{TP + FP}$$

- L'efficacité (*accuracy*) : évalue l'efficacité d'un prédicteur par la fraction d'entités correctement prédites :

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

- courbe ROC (*receiver operating characteristic*) : mesure de performance d'un classificateur binaire. Elle représente la sensibilité en fonction de la spécificité, ce qui se traduit graphiquement par une courbe affichant le taux de TP (vrais positifs) en fonction du taux de FP (faux positifs). Nous avons utilisé cette mesure dans la section 8 afin de mesurer notre capacité à identifier des paires de ligands homologues à partir de leurs distances de cartes d'énergie. Afin de quantifier l'efficacité des prédictions, nous avons calculé l'aire sous la courbe (AUC) des courbes ROC calculées. L'AUC est comprise entre 0 et 1. Une AUC de 1 signifie que le classificateur utilisé est parfaitement exact (aucun FP et aucun FN), une AUC de 0 signifie que le classificateur utilisé est parfaitement inexact (aucun TP et aucun TN) et une AUC de 0.5 signifie que le classificateur utilisé est équivalent a une prédiction aléatoire.

6 Développement d’un cadre théorique pour caractériser le potentiel d’interaction de la surface d’une protéine avec un jeu de partenaires d’intérêt

6.1 Introduction

La première partie de ma thèse est un travail méthodologique visant à mettre en place un cadre théorique pour caractériser le potentiel d’interaction de la surface d’une protéine avec un ensemble de partenaires dits « fonctionnels » ou non. Ceci impliquait donc d’être capable de caractériser d’un point de vue énergétique les interactions entre paires fonctionnelles de protéines et paires non-fonctionnelles ainsi que de caractériser le potentiel d’interaction de l’ensemble de la surface protéique. Le cadre théorique que j’ai mis en place repose donc sur des calculs de *docking* arbitraire et une représentation du paysage énergétique d’interaction d’une protéine avec un partenaire par des cartes d’énergie en deux dimensions reflétant le potentiel d’interaction de cette protéine avec ce partenaire. Afin de caractériser le potentiel d’interaction d’une protéine avec un jeu de partenaires d’intérêt, j’ai mis en place une nouvelle représentation en deux dimensions reposant sur la combinaison des cartes d’énergie de cette protéine avec les partenaires d’intérêt. Ces cartes, que nous appellerons cartes IPOPS (*Interaction Propensity Of Protein Surface*), permettent de caractériser la conservation du potentiel d’interaction de la surface d’une protéine pour un ensemble de partenaires. Autrement dit, ces cartes permettent d’identifier les régions de surface avec un haut potentiel d’interaction quelque soit la nature du partenaire, les régions favorables à l’interaction avec un sous-ensemble de partenaires ou les régions systématiquement défavorables à l’interaction. La section 6.2.2 présente le protocole pour générer les cartes IPOPS d’une protéine (voir aussi section 5.4.5).

6.2 Matériels et méthodes

6.2.1 Jeu de données

La stratégie a été éprouvée et évaluée sur un jeu de données composé lui même de deux jeux de données existants :

- un des jeux de données est composé de 348 structures protéiques réparties en familles d'homologues structuraux qui constituent le jeu de « récepteurs ».

- l'autre jeu de données est composé de 100 structures protéiques qui constituent le jeu de « ligands arbitraires ».

- Jeu de récepteurs

Ce jeu de données est composé de 348 structures protéiques réparties en 81 familles d'homologues structuraux. Les structures utilisées ont toutes été modélisées par homologie et sont extraites de la base de données PPI4DOCK (173). Cette base de données a été réalisée dans le but de fournir un large éventail de structures sous forme non liée (c'est-à-dire monomérique) à la communauté du *docking* protéique. En effet l'une des limites les plus importantes dans l'élaboration de méthodes de *docking* est le nombre limité de structures protéiques sur lesquelles calibrer les méthodes. Pour qu'un complexe protéique puisse être utilisé dans le jeu d'entraînement d'un algorithmes de *docking*, il est nécessaire d'être en possession (i) de la structure du complexe (qui représente la solution à obtenir) (ii) des structures sous forme monomérique (c'est-à-dire non liées) des sous-unités composant le complexe (structures à partir desquelles on va essayer de prédire la solution formée par le complexe). L'utilisation de modèles par homologie permet d'augmenter le nombre de structures disponibles pour l'évaluation des algorithmes de *docking*. Ces modèles sont donc une alternative viable aux structures déterminées expérimentalement. PPI4DOCK est à notre connaissance la base de données la plus exhaustive de structures protéiques sous forme non liée.

De plus PPI4DOCK présente l'avantage de contenir de nombreuses structures de complexes interologues (c'est-à-dire des paires des protéines qui ont des homologues qui interagissent dans un autre organisme). Or ceci est très précieux à terme pour pouvoir comparer les potentiels d'interaction de protéines homologues et investiguer la conservation du potentiel d'interaction dans une famille d'homologues. J'ai donc sélectionné toutes les familles composées d'au moins trois structures non redondantes à 70% d'identité de séquence, avec une couverture de l'alignement

structural d'au moins 75% entre les membres d'une même famille. Lorsqu'il y avait deux protéines redondantes (c'est-à-dire partageant plus de 70% d'identité de séquence), j'ai conservé celle ayant le TM-score (mesure de la similarité entre deux structures protéiques, (195)) le plus élevé avec son « patron » (structure expérimentale ayant servi à faire le modèle de la forme non-liée). L'hypothèse sous-jacente est que plus le TM-score entre le patron et le modèle est élevé, plus le modèle généré est fiable.

Le jeu de données ainsi obtenu contient 348 structures protéiques réparties en 81 familles d'homologues structuraux. Ces structures joueront le rôle de récepteurs dans la suite de l'expérience.

- Jeu de ligands arbitraires

Parallèlement à ce jeu de données, nous avons établi un jeu de structures protéiques issues du jeu de données réalisé par Wass et al (169). Ce jeu de données avait été défini lors d'une expérience de prédiction de partenaires protéiques par *docking* arbitraire. Le jeu était constitué de 922 structures protéiques sous forme non liée et non redondantes au niveau "superfamille" de la classification SCOP (196). Ce jeu de structures permet d'échantillonner une grande diversité de structures et de tailles de protéines. Dans notre procédure, chacune des 348 protéines du jeu de récepteurs est amarrée avec l'ensemble des protéines du jeu de ligands arbitraires. De manière à réduire le temps de calcul, je me suis limité à un échantillon de 100 structures choisies aléatoirement parmi le jeu de 922 ligands arbitraires. Nous verrons par la suite, qu'un sous-ensemble inférieur à 100 ligands est suffisant pour générer des cartes IPOPS robustes (voir section 6.3.7).

6.2.2 Procédure de *docking* et création des cartes IPOPS

J'ai réalisé des calculs de *docking* avec le logiciel ATTRACT (voir section 5.2) entre chacune des 348 structures du jeu de données "récepteurs" avec les 100 structures du jeu de données "ligands arbitraires" (Figure 6.1A-B) (soit un total de 34 800 calculs de *docking*). J'ai produit les 34 800 cartes d'énergie correspondantes (voir section 5.4.3). J'ai ensuite créé les cartes IPOPS correspondantes (voir section 5.4.5 et 6.1C-D) pour chacun des 348 récepteurs.

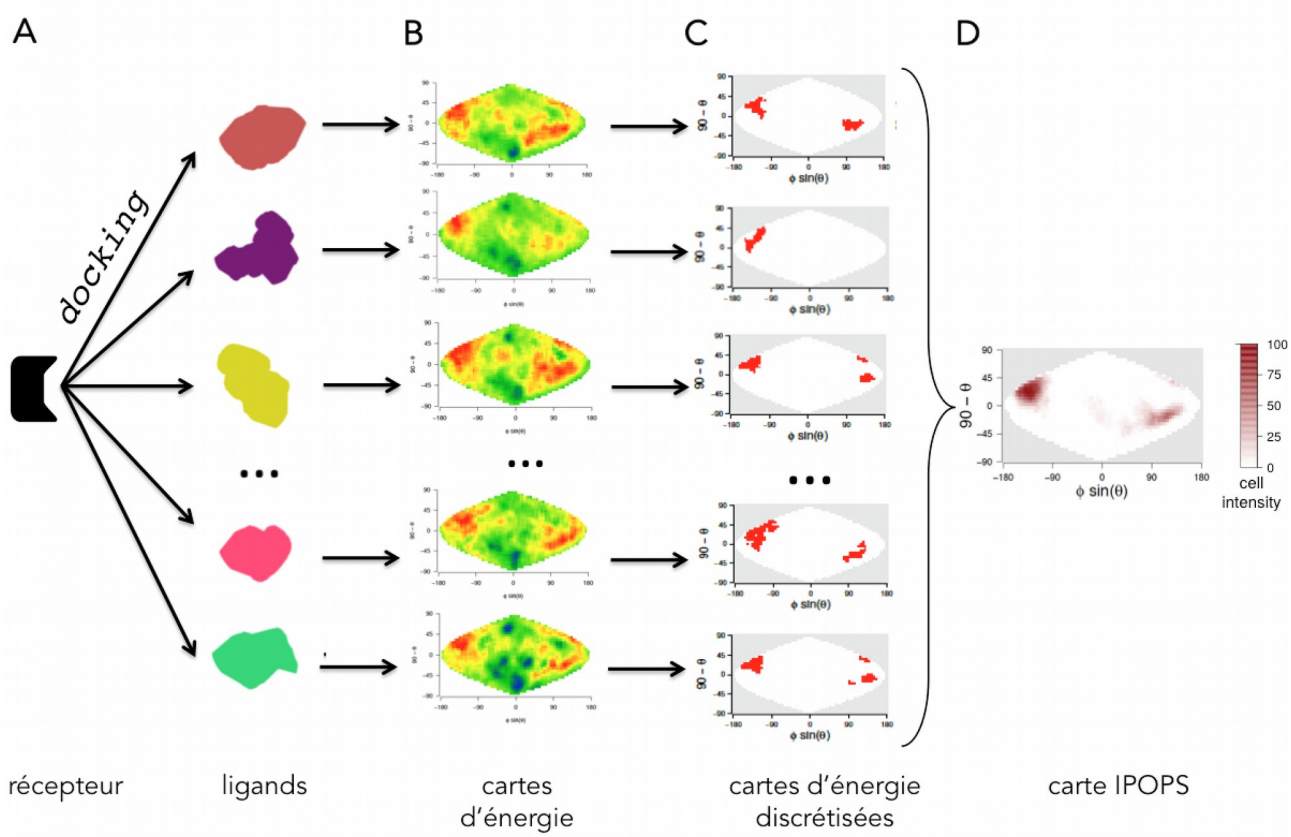


Figure 6.1. Protocole de création d'une carte IPOPS rouge pour un récepteur du jeu de données. (A) Un récepteur (jeu de récepteurs) est amarré avec 100 ligands (jeu de ligands arbitraires). (B) Pour chaque calcul de *docking* une carte d'énergie est créée. (C) Les régions favorables à l'interaction (en rouge) sont extraites des cartes d'énergie (voir section 5.4.3). (D) Une carte IPOPS rouge est calculée à partir cartes de classe d'énergie rouges (voir section 5.4.5). Cette procédure est répétée pour chacun des 348 récepteurs du jeu de données.

6.3 Résultats

6.3.1 Cartes d'énergie 2D

Un des objectifs de ma thèse était de caractériser l'ensemble du paysage énergétique d'interaction entre deux protéines. Le paysage énergétique d'interaction réfère à l'ensemble des solutions de *docking* et leurs scores d'énergie associés. J'ai choisi de représenter le paysage énergétique d'interaction entre deux protéines par des cartes d'énergie en deux dimensions. Comme le montrent les figures 5.7 et 6.1 la procédure est asymétrique. Nous garderons les mêmes conventions que celles utilisées couramment en *docking* protéine-protéine, à savoir, la protéine

maintenue fixe pendant la procédure sera appelée « récepteur » tandis que la protéine mobile sera appelée « ligand ». Ainsi une carte d'énergie représente le potentiel d'interaction de la surface d'un récepteur avec un ligand, projeté sur un plan en deux dimensions. Pour obtenir le potentiel d'interaction du ligand, il faut alors répéter la procédure en inversant les rôles entre le ligand et le récepteur. Pour rappel, ces cartes sont créées par la projection des centres de masses des ligands pour chaque solution de *docking* sur une carte en deux dimensions (figure 5.7). La carte est ensuite divisée en 72x36 cellules, puis le score de chaque case de la grille est lissé selon la procédure présentée en section 5.4.3. Ces cartes d'énergie présentent l'avantage d'être faciles à manipuler et à analyser. Les régions de plus basse énergie, soit les plus favorables à l'interaction sont représentées en rouge, tandis que les régions de plus haute énergie sont en bleu. Entre ces deux extrêmes se situent des régions présentant des niveaux d'énergie intermédiaires matérialisés par des couleurs allant du orange au vert foncé. Du fait des couleurs utilisées pour les représenter sur les cartes d'énergie, dans la suite de ce travail nous référerons aux régions de basse énergie (donc favorables à l'interaction) comme "régions rouges" et aux régions de haute énergie (donc défavorables à l'interaction) comme "régions bleues".

6.3.2 Cartes d'énergie : l'exemple de 2wo2_B

L'exemple de la protéine 2wo2_B est présenté dans la figure 6.2. Il s'agit d'une éphrine, une protéine connue pour interagir avec les récepteurs Eph, qui composent une sous-famille des protéines de type récepteur à activité tyrosine kinase (RTK). L'interaction entre l'éphrine et son récepteur est connue pour intervenir dans de nombreuses voies de signalisation, notamment pendant le développement embryonnaire (197). La figure 6.2B présente les cartes d'énergie calculées entre le récepteur 2wo2_B (éphrine-B2, (198)) et cinq ligands différents. Parmi ces ligands, 2wo2_A (récepteur à l'éphrine de type A) est un vrai partenaire de 2wo2_A (on parlera aussi de ligand natif) (198), tandis que 4l0p_A (199) est un homologue structural de 2wo2_B. Les trois ligands 1d4o_A (NADP(H) transhydrogénase), 1v30_A (protéine PH0828 hypothétique UPF0131) et 1ejg_A (crambine) ne sont pas connus pour interagir avec 2wo2_B et constituent trois ligands arbitraires.

La carte d'énergie résultant du *docking* du récepteur 2wo2_B avec son vrai partenaire présente une région rouge (région de basse énergie), et donc favorable à l'interaction dans la partie gauche de la carte. De façon intéressante, cette région correspond au site d'interaction expérimental de 2wo2_B avec 2wo2_A (dont le centre de masse est symbolisé par une croix sur la carte correspondante, première carte figure 6.2B). Par ailleurs, la carte d'énergie obtenue avec

l’homologue du ligand natif (4l0p_A) présente aussi une région rouge au niveau du site d’interaction du récepteur. De même, les cartes d’énergie obtenues avec les trois ligands arbitraires présentent une région rouge au niveau du site d’interaction du récepteur. Ces observations suggèrent que non seulement le partenaire natif, mais l’homologue de ce dernier ou les ligands arbitraires ont tendance à interagir favorablement sur la même région. En observant le reste de la surface, nous voyons que les cartes d’énergie obtenues avec les ligands arbitraires présentent d’autres régions rouges : la carte de 1v30_A possède trois régions rouges, une au niveau du site d’interaction natif de 2wo2_B, une sur la droite de la carte et une sur la partie inférieure au centre gauche de la carte. La carte de 1d4o_A possède deux régions rouges: une au niveau du site d’interaction de 2wo2_B et une dans la partie droite de la carte. La carte de 1ejg_A possède les mêmes régions rouges que celle de 1d4o_A. Les cartes produites par les deux ligands homologues sont très similaires.

Toutes les cartes possèdent une région bleue située au niveau centre-gauche de la carte. Une autre région bleue de taille variable selon les ligands est localisée dans la partie droite de la carte. Les régions jaune-orange sont fluctuantes selon les ligands tandis que les régions vertes recouvrent la plus grande partie de chacune des cartes d’énergie.

Cette analyse visuelle semble montrer (i) que la distribution des régions de différents niveaux d’énergie (régions rouges à bleues) n’est pas aléatoire, (ii) que les régions rouges sur la carte d’énergie peuvent indiquer des sites d’interaction à la surface du récepteur et (iii) que certaines régions rouges sur la carte sont conservées quelque soit la nature du ligand tandis que d’autres semblent spécifiques du ligand mis en jeu. Afin de généraliser et de quantifier ces observations pour un grand nombre de ligands, nous avons mis en place une nouvelle représentation permettant d’évaluer la conservation des régions de chaque niveau d’énergie en fonction de la nature des ligands. Ces cartes permettront de représenter le potentiel d’interaction d’un récepteur pour un jeu de ligands d’intérêt.

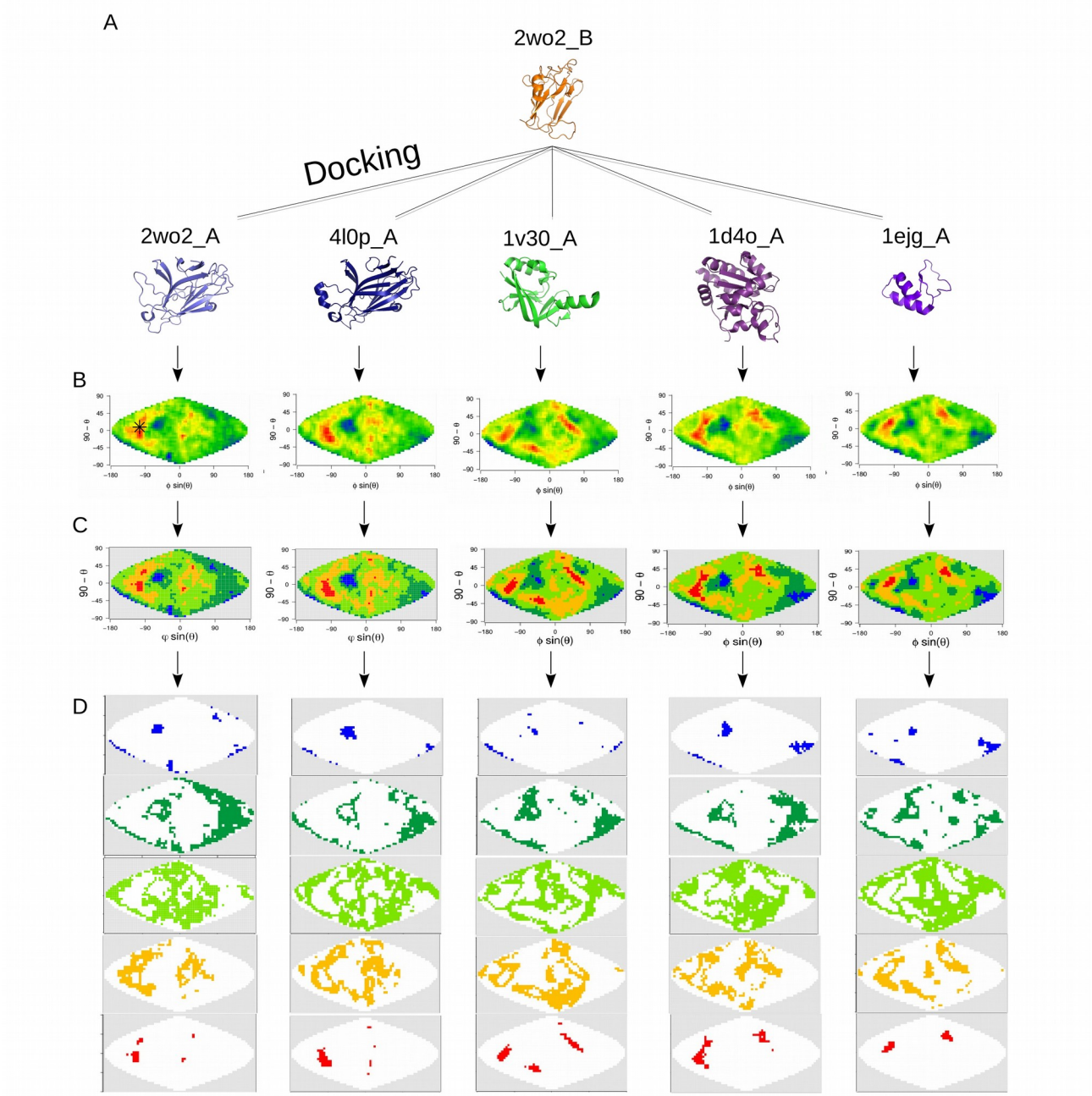


Figure 6.2. *Docking* de 2wo2_B avec des ligands partenaires et arbitraires. (A) La protéine 2wo2_B est amarrée avec 5 protéines. 2wo2_A est un partenaire fonctionnel, 4l0p_A un homologue structural de 2wo2_A et les trois autres protéines sont des partenaires arbitraires. (B) Cartes d'énergies de chaque calcul de *docking*. Le centre de masse du site d'interaction de 2wo2_A est symbolisé par une étoile sur la carte d'énergie de 2wo2_B amarré avec 2wo2_A. (C) cartes de classes d'énergie. (D) Cartes binarisées de classes d'énergie, chacune des cartes représentant une classe d'énergie. Voir section 5.4 pour plus de détails sur la méthodologie employée pour réaliser ces cartes.

6.3.2.1 Représentation de la conservation des régions de différentes classes d'énergie pour un jeu de ligands d'intérêt : création des cartes IPOPS

La procédure est décrite dans la partie 5.4.5. Pour rappel la première étape de cette procédure consiste à discrétiser les cartes d'énergie en cinq classes d'énergie, les régions bleues correspondant aux régions de haute énergie et donc à la classe « région défavorable » à l'interaction, les régions rouges correspondant aux régions de basse énergie et donc à la classe « région favorable » à l'interaction (voir figure 6.2A et B et section 5.4.3). Entre ces deux extrêmes, nous avons trois catégories d'énergie intermédiaires représentées par les couleurs vert, jaune et orange. A partir de ces cartes nous avons la possibilité d'extraire les régions de la classe d'énergie qui nous intéresse (figure 6.2C et D).

Ensuite, pour représenter la conservation de ces régions de différentes classes d'énergie pour un récepteur que l'on a amarré successivement avec un jeu de ligands d'intérêt. Nous avons défini de nouvelles cartes. Ces cartes, appelées cartes IPOPS (*Interaction Propensity Of Protein Surfaces*) sont calculées indépendamment pour chaque classe d'énergie et résultent de la somme de toutes les cartes d'une même classe d'énergie calculées pour un récepteur et un jeu de ligands (voir section 5.4.5 et Figure 6.1). L'intensité de chaque cellule varie de 0 à N avec N correspondant au nombre de ligands testés. L'intensité d'une cellule d'une carte d'une couleur donnée (à savoir, bleue, vert foncé, vert clair, jaune ou rouge) indique le nombre de fois où cette cellule a été associée à cette classe d'énergie. Ainsi une cellule ou un ensemble de cellules de forte intensité sur une carte IPOPS rouge révèle une région très favorable à l'interaction avec la plupart des ligands testés. Une cellule de faible intensité sur une carte IPOPS rouge révèle une région favorable à l'interaction pour un sous-ensemble de ligands et défavorable à la plupart des autres ligands. Plusieurs filtres successifs sont appliqués afin d'éliminer le « bruit de fond » (voir section 5.4.5). Les cellules d'une intensité inférieure à 22 (c'est à dire favorables à l'interaction pour moins de 22 ligands sur un total de 100) sont éliminées ainsi que les îlots de moins de quatre cellules. Ces cartes offrent une représentation synthétique de l'information portée par un grand nombre de cartes d'énergie en une seule carte, ce qui facilite d'autant l'analyse. Dans cette expérience, j'ai calculé les cartes IPOPS des 348 récepteurs à partir des calculs de *docking* réalisés entre chaque récepteur et les 100 ligands arbitraires. L'intensité des cellules des cartes IPOPS varie donc de 0 à 100. Au final, nous avons donc cinq cartes IPOPS (une par classe d'énergie) pour chacun des 348 récepteurs.

Les cartes IPOPS des différentes catégories de classe d'énergie du récepteur 2wo2_B sont représentées dans la figure 6.3. La carte IPOPS rouge montre que le site d'interaction expérimental de 2wo2_B présente une intensité forte (autour de 50). Cela montre que la majorité des ligands arbitraires amarrés avec 2wo2_B présentent des modes d'interactions favorables avec le récepteur au niveau du site d'interaction expérimental de ce dernier. Ce résultat est dans la lignée de ceux de Fernandez-Recio et al (153), Lopes et al (166) et Martin et Lavery (156), qui ont montré que lors de calculs de *docking* entre un récepteur et des ligands arbitraires, les solutions les plus favorables énergétiquement avaient tendance à s'accumuler autour des sites d'interaction expérimentaux. Une autre région de moins forte intensité, c'est-à-dire favorable à l'interaction pour un moins grand nombre de ligands (autour de 30), est aussi présente sur cette carte. Le site indiqué par cette région de faible intensité n'est pas connu et n'a pas été répertorié dans la PDB comme un site d'interaction connu.

En analysant la cartes IPOPS des régions bleues (figure 6.3E), nous observons que deux îlots bleus tendent à être conservés pour un grand nombre de ligands (environ 50). Les régions intermédiaires sont largement répandues sur la surface des protéines. Sur l'exemple du récepteur 2wo2_B, les 3 régions intermédiaires recouvrent à elles trois la majeure partie de la surface de la carte (figure 6.3C-D).

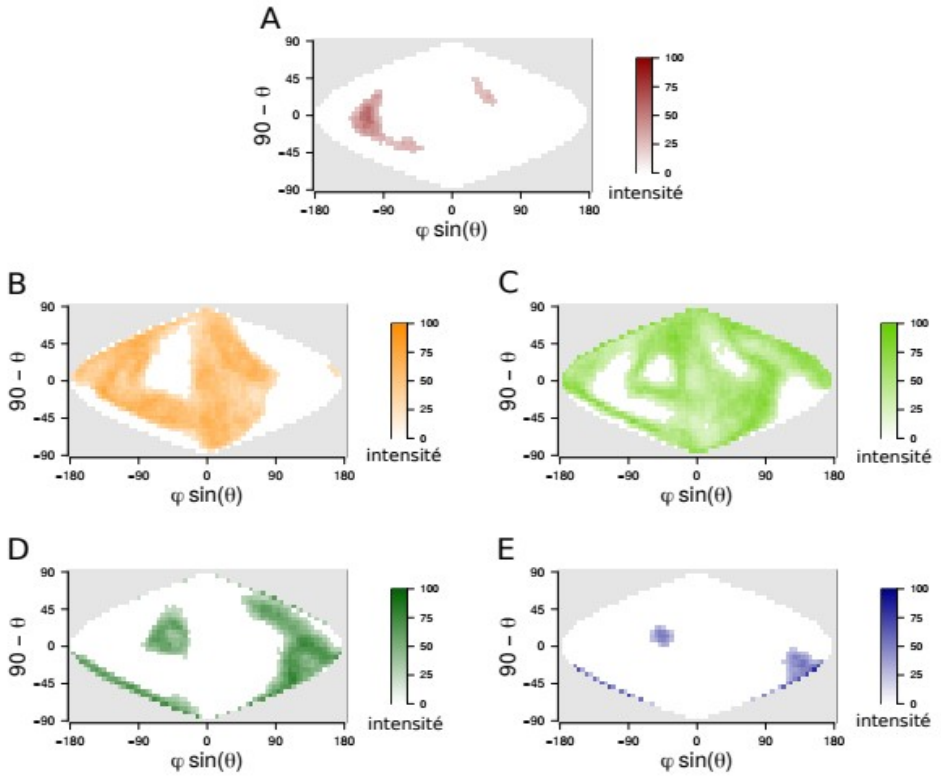


Figure 6.3.Cartes IPOPS du récepteur 2wo2_B. Ces carte a réalisée en sommant les cartes de classes d'énergie correspondantes extraites de 100 cartes d'énergie résultant du *docking* du récepteur 2wo2_B avec un ensemble de 100 ligands (voir section 5.4.5). (A) Carte IPOPS rouge. (B) Carte IPOPS jaune. (C) Carte IPOPS vert clair. (D) Carte IPOPS vert foncé. (E) Carte IPOPS bleue.

6.3.2.2 Taille des régions de différentes classes d'énergie

L'observation des cartes IPOPS de la protéine 2wo2_B, semble indiquer que les tailles des différentes classes d'énergie sont différentes. Afin de vérifier cette observation sur l'ensemble de notre base de données, j'ai calculé pour chaque classe d'énergie, le nombre de cellules présentant une intensité supérieure à 22 dans les 348 cartes IPOPS correspondantes (Figure 6.4).

Nous observons que les régions rouges et bleues sont plus petites et couvrent respectivement 108 et 111 cellules d'une carte IPOPS en moyenne, ce qui correspond à environ 7 % de la surface de la carte. Les régions intermédiaires sont beaucoup plus étalées: les régions vert foncé recouvrent 753 cellules (soit 49 % de la surface d'une carte), les régions vert clair recouvrent en moyenne 1208 cellules soit 78 % d'une carte. Les régions oranges recouvrent en moyenne 659 cellules, soit environ 43 % de la surface d'une carte. Il convient de garder à l'esprit que les régions identifiées

par les cartes IPOPS de différentes classes d'énergie peuvent être chevauchantes, puisqu'une région favorable pour un ligand ne l'est pas forcément pour un autre. Par conséquent une cellule peut posséder une intensité supérieure à 22 dans des cartes IPOPS de classes d'énergie différentes. Il semble que c'est d'autant plus le cas pour les régions intermédiaires.

Les régions favorables à l'interaction sont donc généralement restreintes à de petites portions de la surface des cartes. Les régions intermédiaires sont largement répandues répandues sur la surface de la carte tandis que les régions défavorables à l'interaction sont, de manière similaire aux régions favorables, limitées à de petites régions de la carte.

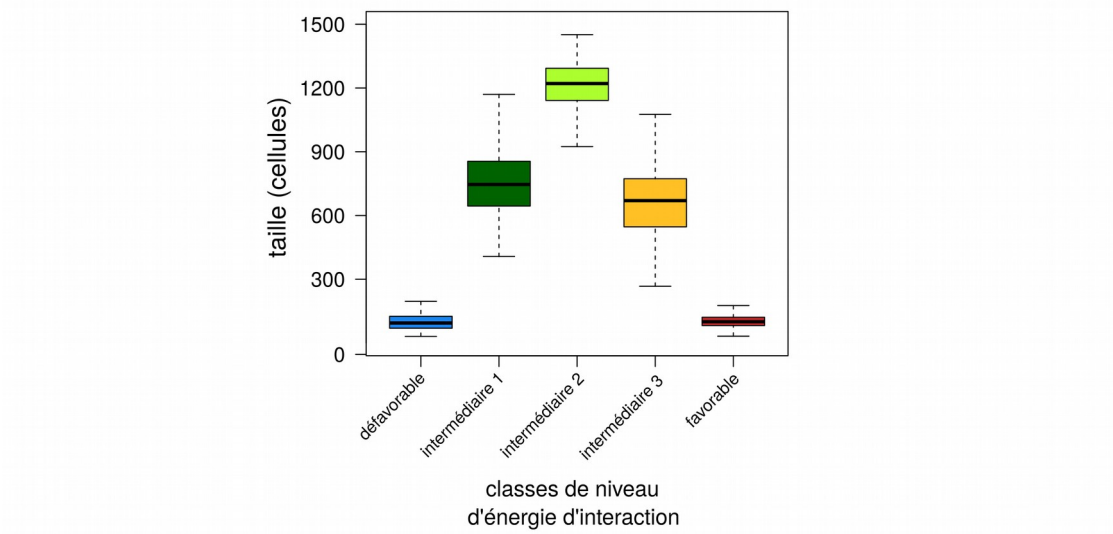


Figure 6.4. Boîtes à moustache du nombre de cellules d'intensité supérieure à 22 pour les 348 cartes IPOPS de chaque classe d'énergie.

6.3.2.3 Caractérisation des propriétés physico-chimiques et évolutives des régions de surface de différentes classes d'énergie

Afin de voir si les différentes régions de surface d'une protéine identifiées par les cinq cartes IPOPS qui lui correspondent, présentent des propriétés physico-chimiques et évolutives différentes, j'ai ensuite caractérisé ces propriétés pour les cinq types de régions de surface (à savoir les régions présentant un fort potentiel d'interaction identifiées par les cartes IPOPS rouges, celles qui au

contraire présentaient un faible potentiel d'interaction identifiées par les cartes IPOPS bleues et les régions intermédiaires). En particulier, j'ai analysé les propriétés suivantes : la *stickiness* (106), l'hydrophobicité selon l'échelle de Kyte-Doolittle (192), la variance circulaire (193) qui reflète le relief local de la surface protéique, la conservation évolutive des résidus matérialisée par le calcul de « trace évolutive » du logiciel JET (197) (voir section 5.3 pour plus de détails sur ces propriétés).

Pour réaliser cette comparaison, j'ai calculé les cartes 2D, c'est-à-dire projeté en deux dimensions ces différentes propriétés sur la surface de chacun des 348 récepteurs selon la méthodologie présentée dans la section 5.4.1.

La figure 6.5 représente les boîtes à moustache des valeurs de *stickiness*, d'hydrophobicité de Kyte-Doolittle, de variance circulaire et de la conservation évolutive calculée par le logiciel JET pour chaque classe d'énergie. Cette figure montre que la *stickiness*, l'hydrophobicité et la variance circulaire diminuent de manière significative avec les niveaux d'énergie de chaque classe (test de Tukey HSD (201), valeur p des différences observées entre chaque classe d'énergie < .001). En effet, les régions de plus basses énergie (en rouge) sont les plus "collantes", les plus hydrophobes et les plus planes tandis que les régions de haute énergie (en bleu) sont les moins « collantes », les moins hydrophobes et les plus protubérantes. De plus, plus les régions des surfaces protéiques sont enclines à l'interaction avec d'autres protéines, plus celles-ci sont conservées évolutivement.

Ce résultat est cohérent avec ce que l'on sait des propriétés des sites d'interaction connus pour être plus hydrophobes, plus conservés et généralement plus plans que le reste de la surface. Néanmoins, il est important de préciser que (i) les cinq classes ont été définies sur des critères énergétiques sans aucune connaissance des sites d'interaction expérimentaux et reflètent le potentiel d'interaction des 348 protéines à interagir avec un jeu de ligands arbitraires donc *a priori* avec des partenaires qui formeraient des interactions non-fonctionnelles. Autrement dit les régions privilégiées des partenaires non-fonctionnels présentent des propriétés en accord avec les sites d'interaction des partenaires fonctionnels. (ii) Le fait que les énergies de *docking* calculées entre partenaires natifs ou arbitraires soit anti-corrélées à la *stickiness* définie à partir d'un échantillon de complexes protéiques impliquant des couples fonctionnels (voir section 5.3.2) renforce fortement le concept « *stickiness* » comme la propension à interagir de façon non-fonctionnelle et fourni des éléments physiques solides pour proposer les régions « collantes » comme des vecteurs d'interactions non-fonctionnelles dans la cellule. Finalement, les îlots identifiés dans les cartes IPOPS rouges (c'est-à-dire ces régions de basse énergie quelque soit la nature du ligand) reflètent des régions de la surface des protéines avec un fort potentiel à engager des interactions non spécifiques avec un grand nombre de protéines. On peut émettre l'hypothèse que ces régions

présentent des propriétés universelles d'interaction qui conviennent à de nombreux ligands : elles sont plus hydrophobes et plus planes, et donc plus «collantes» que les autres régions.

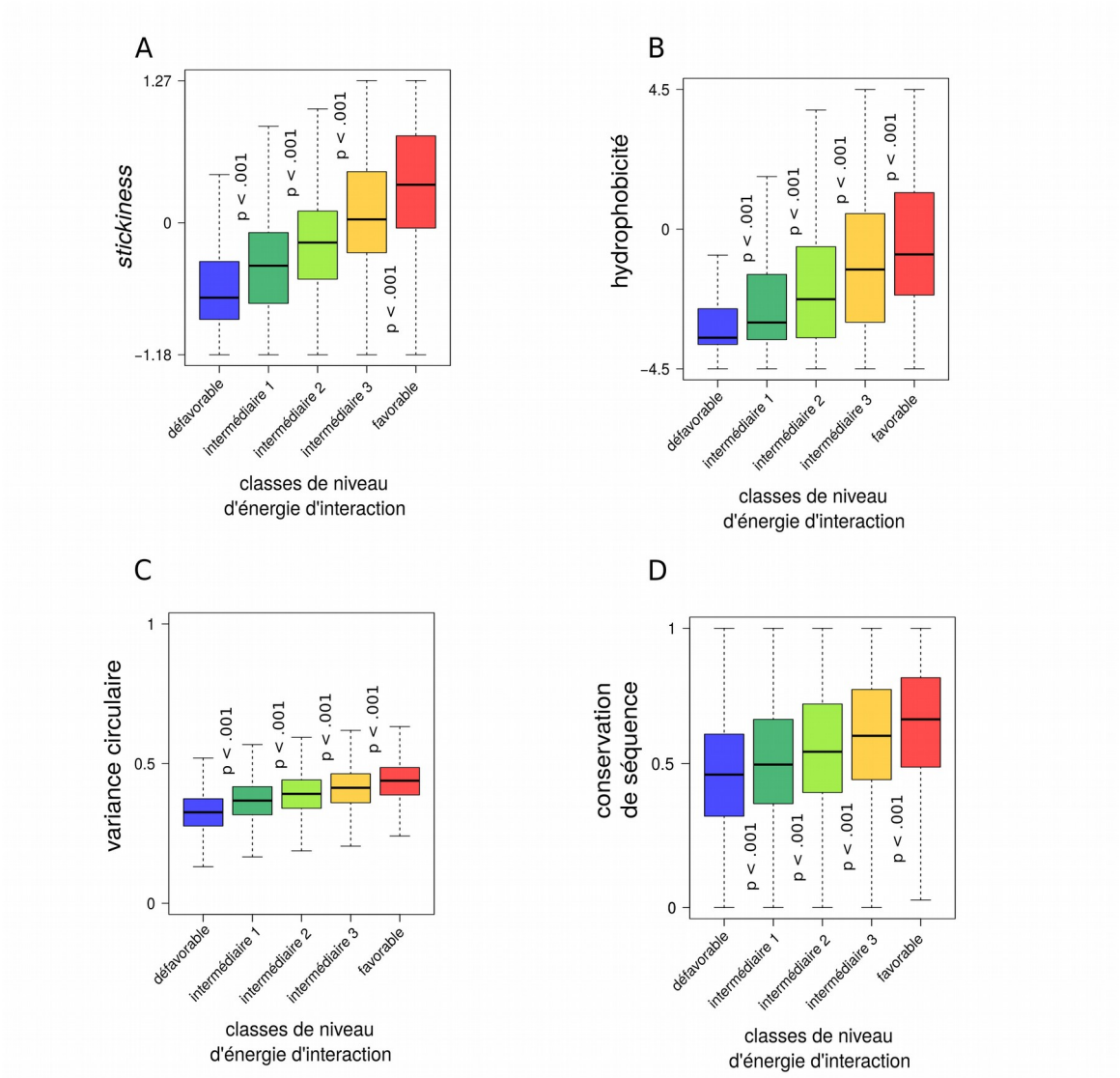


Figure 6.5. Boîtes à moustache des valeurs de *stickiness*, d'hydrophobicité, de variance circulaire, de conservation de séquence en fonction des classes d'énergie. Les valeurs des différents descripteurs utilisés sont calculées selon la méthodologie présentée dans la section 5.3. (A) *stickiness*. (B) hydrophobicité de Kyte-Doolittle. (C) variance circulaire. (D) conservation de séquence**.**

6.3.2.4 Exemples

Dans cette section je vais détailler trois exemples pour illustrer les relations entre le potentiel d'interaction des différentes régions des surfaces protéiques et leurs propriétés physico-chimiques et évolutives.

La protéine 1tgz_A est une protéase possédant une fonction essentielle dans les activités de sumoylation en clivant le propeptide SUMO (202), permettant l'activation de cette protéine. La figure 6.6A montre la cartes IPOPS rouge de la protéine 1tgz_A. Cette carte est composée d'une région de forte intensité localisée au centre de la carte et de deux régions d'intensités plus faibles localisées dans les parties gauche et droite de la carte. La figure 6.6 présente la carte des sites d'interaction expérimentaux de la protéine 1tgz_A. En comparant la carte IPOPS rouge avec celle des sites d'interaction expérimentaux, on observe que la région de forte intensité de la carte IPOPS correspond à la partie inférieure (sur la carte) du site d'interaction du récepteur. En particulier, cette région correspond à une région possédant une *stickiness* et une hydrophobicité élevée (figure 6.6C et D) qui est aussi plus bien plus conservée évolutivement que le reste de la surface (voir région bleu foncé sur la figure 6.6E). De façon intéressante, cette région se distingue aussi au niveau de son relief avec une valeur élevée de variance circulaire reflétant une région concave (figure 5F) qui correspond au site actif de la protéase et qui est situé approximativement au centre du site d'interaction. Il y a donc ici une très bonne complémentarité entre les cartes présentées : la région IPOPS rouge identifie clairement une sous-région du site d'interaction expérimental. En particulier, elle recouvre le site actif de ce dernier et correspond à une région à fort indice de *stickiness*, ce qui favorise les interactions avec d'autres protéines.

Bien que l'îlot de forte intensité identifié dans la carte IPOPS ne couvre qu'une partie du site d'interaction, il est important de noter que notre méthode repose sur la projection des centres de masse des solutions de *docking*. Par conséquent, une solution de *docking* favorable est représentée par un unique point sur une carte. Une région rouge sur une carte d'énergie correspond donc à un ensemble de centres de masses de solutions de *docking* énergétiquement favorables. Ainsi, la taille d'une région rouge sur une carte d'énergie ne peut être assimilée à la taille du site d'interaction mis en jeu dans les solutions de *docking* correspondantes. Une région de petite taille peut impliquer un site d'interaction de grande taille mais reflète une faible diversité de solutions favorables énergétiquement au niveau des positions du ligand. Cela explique pourquoi l'îlot IPOPS que nous mettons en évidence dans cette exemple (et sur nos cartes en général) possède une surface nettement plus petite que celle du site d'interaction de la protéine.

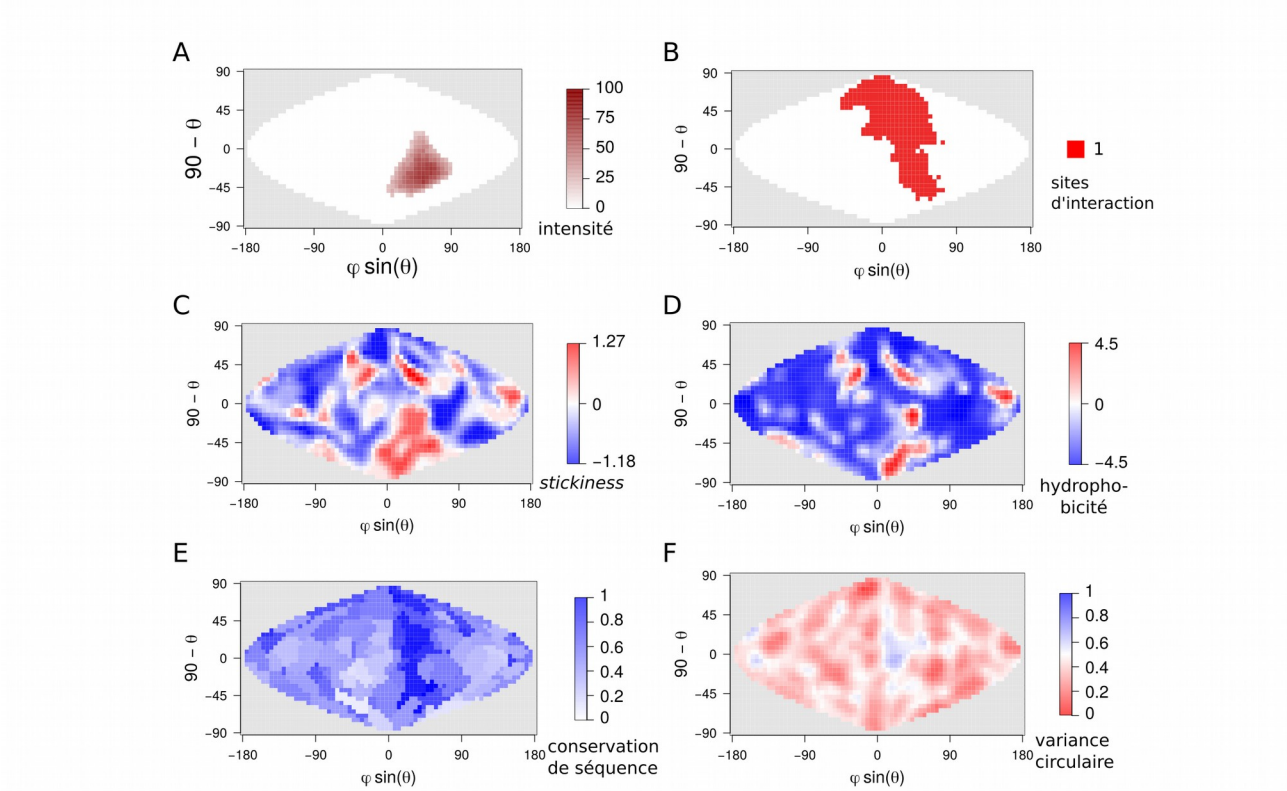


Figure 6.6. Cartes de propriétés de 1tgz_A (protéase spécifique de la sentrine 2). (A) carte IPOPS rouge. (B) carte de sites d'interaction projetés (voir section 5.4.2 pour la méthodologie). (C) carte de *stickiness*. (D) carte d'hydrophobicité. (E) carte de conservation évolutive (F) carte de variance circulaire. Dans les cartes de sites d'interaction, le blanc correspond à la surface non interagissante tandis que les autres couleurs correspondent chacune à un site d'interaction avec une chaîne dans l'unité biologique. Voir la section 5.4 pour plus de détails sur la procédure de création des cartes.

Un autre exemple intéressant est celui de 3rpf_B. Il s'agit d'une sous-unité catalytique de la molybdoptérine synthase. Comme son nom l'indique, elle permet la synthèse de molybdoptérines, une classe de précurseurs de cofacteurs enzymatiques (203).

La carte IPOPS rouge de cette protéine est composée de deux îlots de forte intensité localisés respectivement au centre et sur les bords droit et gauche de la carte (figure 6.7A). En comparant cette carte avec la carte des sites expérimentaux (figure 6.7B), on observe que les deux îlots de forte intensité identifiés par la carte IPOPS sont localisés au centre de chacun des deux sites d'interaction expérimentaux du récepteur. La partie inférieure du site d'interaction rouge (figure 6.7B) correspond à une région possédant une *stickiness* et une hydrophobicité élevée (figure 6.7C et D). Elle est aussi très conservée du point de vue évolutif (figure 6.76E) et, de manière similaire à l'exemple précédent, présente une région concave en son centre (figure 6.7F). Là encore cette région correspond au site d'interaction de la protéine avec son partenaire. De manière intéressante,

même si cette protéine est annotée comme un hétérotétramère d'après le logiciel PISA (204), les auteurs qui ont publié la structure sont en désaccord avec l'annotation PISA et annotent cette protéine comme un hétérodimère. Ils ne considèrent comme interface biologique que celle impliquant le site d'interaction coloré en rouge sur la carte et considèrent le site d'interaction coloré en jaune comme un contact cristallin (figure 6.7B). La carte IPOPS rouge montre une même intensité pour les deux îlots, qui semblent par ailleurs très conservés tous les deux (voir régions bleu foncé sur la carte de conservation évolutive, figure 6.7E). D'après ces résultats, on peut suggérer que la région identifiée par le second îlot (sur les bords gauche et droit de la carte IPOPS rouge) correspond à un site d'interaction fonctionnel, soit d'homo-tétramérisation soit d'interaction avec un autre partenaire encore non-identifié. Le fait que le second îlot soit conservé évolutivement suggère que cette région subit une pression évolutive et va dans le sens de cette hypothèse.

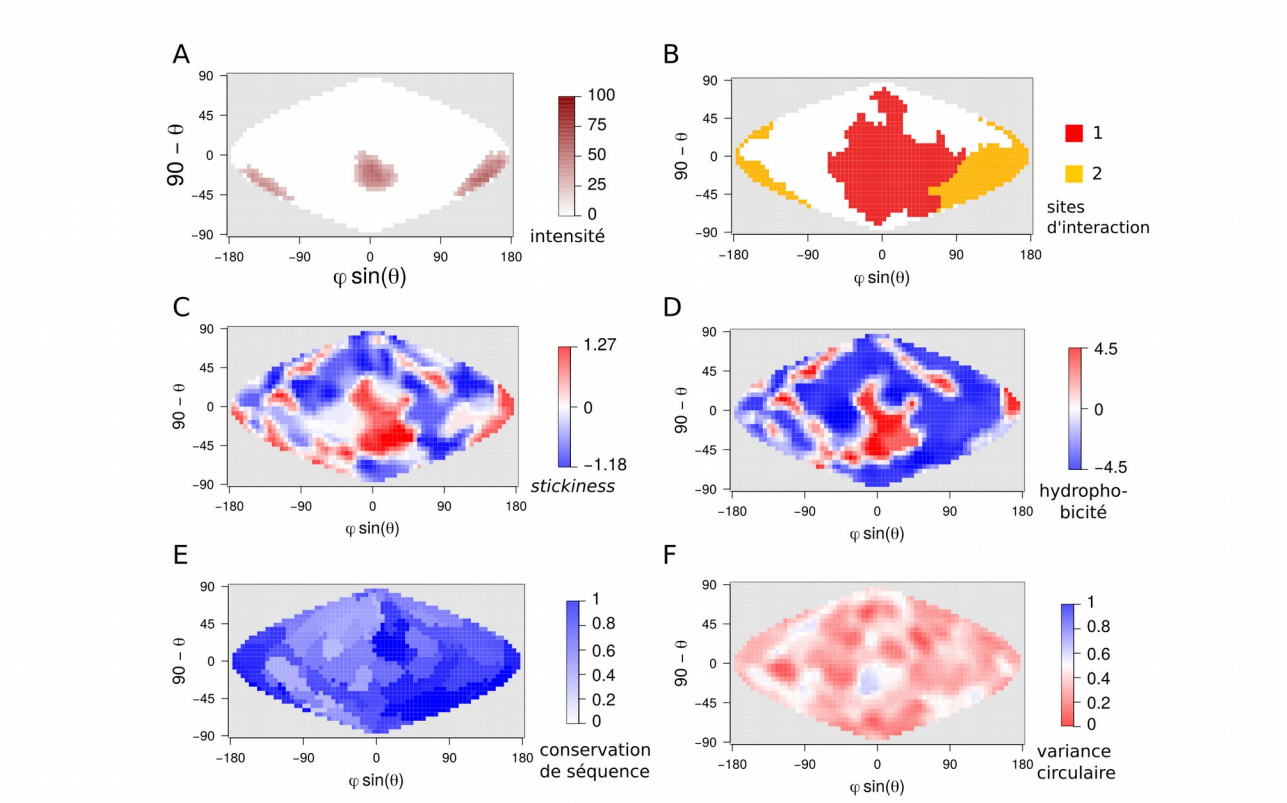


Figure 6.7. Cartes de propriétés de surface de 3rpf_B (sous-unité catalytique de la molybdoptérine synthase). (A) carte IPOPS rouge. (B) carte de sites d'interaction projetés. (C) carte de *stickiness*. (D) carte d'hydrophobicité. (E) carte de conservation évolutive. (F) carte de variance circulaire. Voir la section 5.4 pour plus de détails sur la procédure de création des cartes.

Enfin un troisième exemple est celui de la protéine 1rv6_W. Il s'agit d'un facteur de croissance placentaire (205). Il s'agit d'une facteur de croissance connu notamment pour son implication dans les évènements d'angiogenèse pathologique (205). La carte IPOPS de cette protéine est composée d'une région de forte intensité localisée sur la partie extrême gauche de la carte et qui se prolonge sur la partie droite. Nous observons une région d'intensité plus faible localisée vers le centre de cette carte (figure 6.8A). En comparant cette carte avec celle des sites d'interaction (figure 6.8B), on observe que la région de forte intensité de la carte IPOPS correspond aux sites d'interaction 1 et 2 (symbolisés par les couleurs rouge et bleue sur la carte des sites d'interaction expérimentaux). Ces deux sites d'interaction correspondent à des régions possédant une *stickiness* et une hydrophobicité élevée (figure 6.8C et D) et semblent aussi plus conservés évolutivement que le reste de la surface (figure 6.8F). Cependant, il est difficile de les distinguer d'après leur relief (figure 6.8E).

Le site d'interaction 3 (figure 6.8B, couleur jaune) correspond à une région de faible de intensité de la carte IPOPS rouge. Ce site d'interaction ne correspond pas à une région de forte hydrophobicité même s'il présente une *stickiness* relativement élevée. Il semble, comme les site d'interaction 1 et 2, un peu plus conservé évolutivement que le reste de la surface. Ce site d'interaction fonctionnel semble moins propice à l'interaction avec un grand nombre de partenaires (probablement car il est moins hydrophobe) que les autres régions indiquées par la région de forte intensité sur la carte IPOPS rouge.

Les observations réalisées sur ces trois exemples montrent qu'il existe à la surface des protéines des régions propices à l'interaction pour une large diversité de ligands, on parlera dans ce cas là d'îlots rouges ubiquitaires. Ces régions présentent généralement une *stickiness* et/ou une hydrophobicité élevée, comme nous l'avons vu quantitativement sur les 348 récepteurs. Le dernier exemple montre aussi l'existence de régions favorables à l'interaction de plus faible intensité. Ces régions plus sélectives semblent n'être favorables qu'à un nombre limité de ligands. Nous parlerons d'îlots rouges spécifiques. De façon intéressante, ces derniers ne présentaient pas les mêmes propriétés d'hydrophobicité et de *stickiness* que leurs îlots voisins de fort potentiel d'interaction avec un grand nombre de ligands. Plus généralement, on peut se demander si les îlots rouges ubiquitaires et spécifiques (présentant respectivement de fortes et de faibles intensités de potentiel d'interaction) présentent les mêmes propriétés. Pour cela, il a fallu mettre en place une procédure pour extraire automatiquement les îlots rouges ubiquitaires et spécifiques.

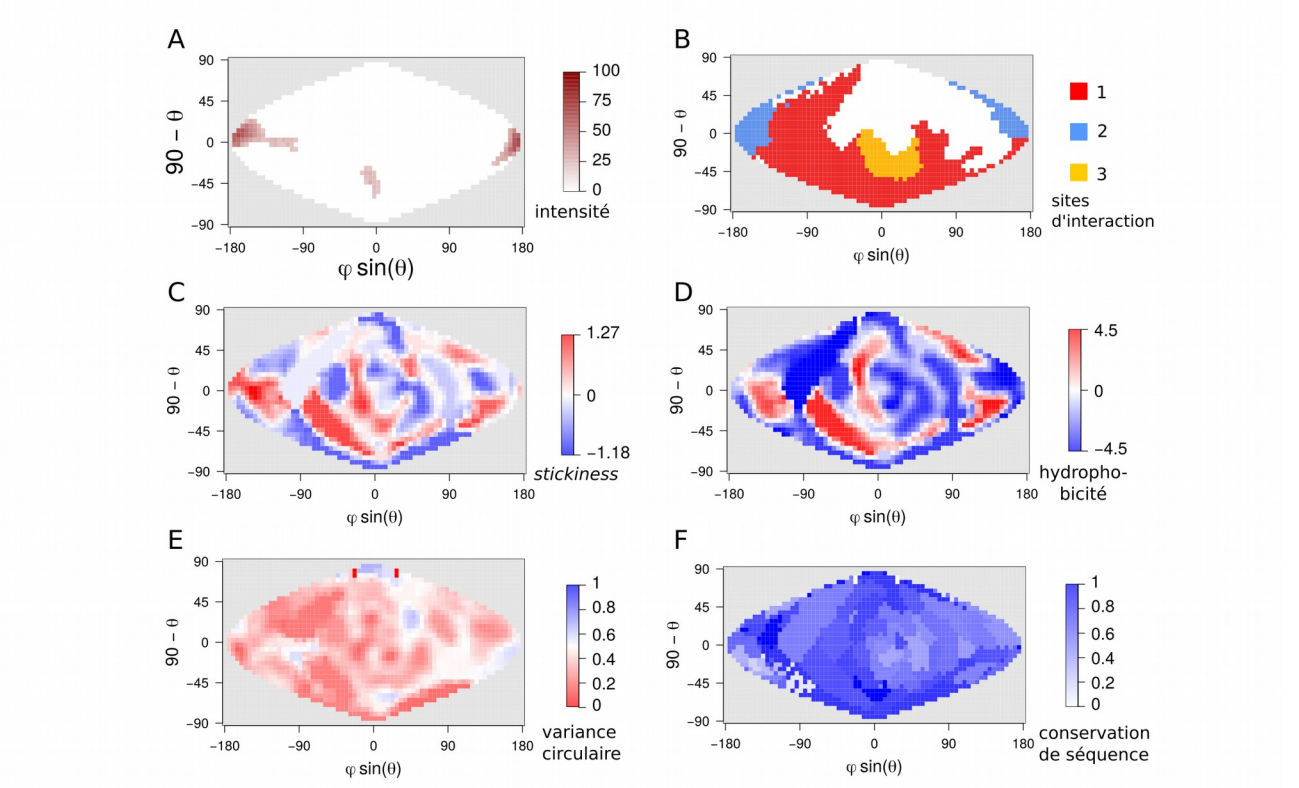


Figure 6.8. Cartes de propriétés de surface de 1rv6_W. (A) carte IPOPS rouge. (B) carte de sites d'interaction projetés (C) carte de *stickiness*. (D) carte d'hydrophobicité. (E) carte de variance circulaire. (F) carte de conservation évolutive. Voir la section 5.4 pour plus de détails sur la procédure de création des cartes.

6.3.3 Îlots rouges ubiquitaires et spécifiques : définitions et propriétés

6.3.3.1 Discrétisation des îlots rouges en deux catégories

La figure 6.9 montre la distribution de l'intensité moyenne des îlots rouges identifiés sur les 348 cartes IPOPS rouges. Cette distribution ne suit pas une loi normale, et semble présenter deux pics suggérant deux populations d'îlots. Nous avons donc entrepris de modéliser la distribution de l'intensité moyenne des îlots par plusieurs lois normales (figure 6.9). Nous avons testé des modèles à 1, 2, 3, 4 ou 5 lois normales (voir section 5.4.5). La meilleure vraisemblance est obtenue avec un modèle à deux lois normales. Cela montre que la distribution de l'intensité moyennes des îlots peut être approchée par un modèle avec deux populations distinctes d'îlots possédant deux distributions

différentes, chacune étant représentée par une loi normale. Ce modèle nous permet donc de faire la distinction entre deux catégories d'îlots : les îlots de forte intensité reflétant un fort potentiel d'interaction avec un grand nombre de ligands et les îlots de faible intensité reflétant un fort potentiel d'interaction avec un petit sous-ensemble de ligands. Le seuil permettant de discriminer les populations d'îlots de faible et de forte intensité correspond à une intensité de 28 (sur une échelle de 100), ce qui signifie que les îlots d'une intensité moyenne supérieure ou égale à 28 sont considérés comme des îlots de forte intensité et ceux d'intensité inférieure à 28 sont considérés comme des îlots de faible intensité. Nous référerons par la suite aux îlots de forte intensité comme des îlots rouges ubiquitaires et aux îlots de faible intensité comme des îlots rouges spécifiques.

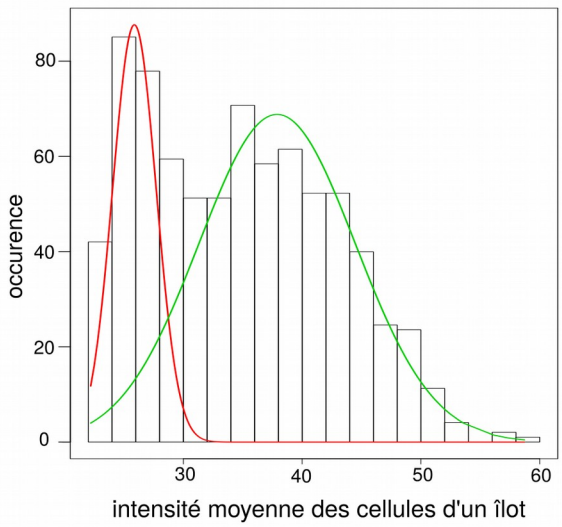


Figure 6.9. Distribution de l'intensité moyenne des îlots cartes IPOPS rouges du jeu de données. L'intensité moyenne d'un îlot est calculée comme étant l'intensité moyenne de l'ensemble des cellules le composant. La distribution est réalisée sur l'ensemble des îlots répertoriés sur les 348 cartes IPOPS rouges du jeu de données. La distribution en rouge représente la loi normale modélisée correspondant à la distribution des îlots rouges spécifiques, la distribution en vert celle correspondant aux îlots rouges ubiquitaires.

Les îlots rouges ubiquitaires représentent donc des régions de la surface d'une protéine fortement attractives pour une majorité de protéines, qu'elles soient des partenaires natifs ou des protéines arbitraires, tandis que les îlots rouges spécifiques représentent des régions favorables à l'interaction pour une minorité de ligands seulement. En moyenne les protéines présentent 1.59 îlots rouges ubiquitaires et 0.57 îlots rouges spécifiques. La figure 6.10 représente la distribution du

nombre d'îlots rouges ubiquitaires et rouges spécifiques identifiés à partir des 348 cartes IPOPS rouges. On observe que 208 cartes IPOPS ne présentent pas d'îlots rouges spécifiques, 96 en possèdent un, 31 en possèdent deux et 13 cartes possèdent plus de deux îlots rouges spécifiques. 190 cartes IPOPS possèdent un seul îlot rouge ubiquitaire, 114 en possèdent deux et 42 en possèdent plus de deux. Seulement deux protéines ne possèdent pas d'îlots rouges ubiquitaires. Le fait que la quasi-totalité des cartes IPOPS rouges possèdent au moins un îlot rouges ubiquitaires montre que dans la grande majorité des cas, les ligands ont tendances à avoir des interactions favorables avec le récepteur au niveau des mêmes régions de sa surface. Les récepteurs ne possédant pas d'îlots rouges ubiquitaires présentent des régions rouges de faible intensité sur une grande partie de la surface, dont la plupart des cellules de la carte sont en dessous du seuil d'intensité minimale.

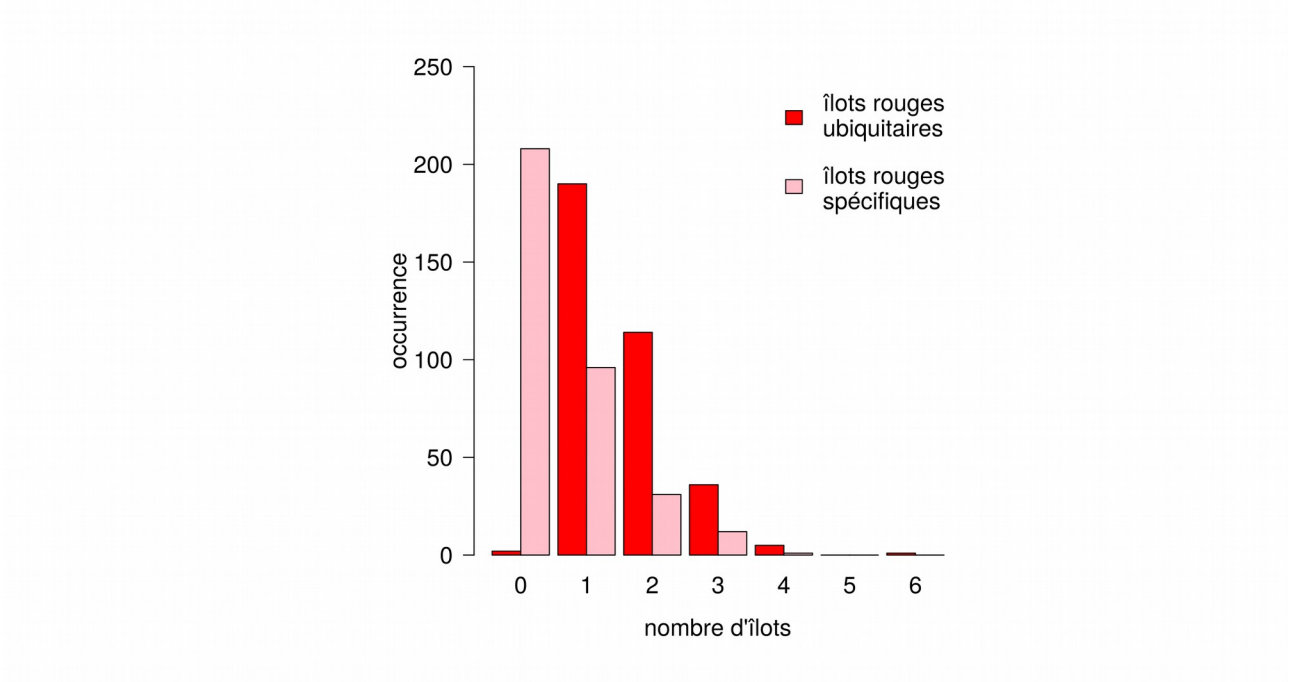


Figure 6.10. Distribution du nombre de îlots rouges spécifiques et rouges ubiquitaires par carte IPOPS rouge sur les 348 récepteurs. La distribution des îlots rouges spécifiques est représentée en rose, la distribution des îlots rouges ubiquitaires en rouge.

En plus d'être en moyenne moins nombreux que les îlots rouges ubiquitaires, les îlots rouges spécifiques sont aussi en moyenne significativement plus petits. En effet les îlots rouges ubiquitaires couvrent en moyenne 70 cellules, tandis que les îlots rouges spécifiques couvrent en

moyenne 15 cellules sur la carte (Figure 6.11). Cette différence de taille est attendue car les îlots rouges spécifiques, de part leur faible intensité, sont beaucoup plus “sensibles” au seuillage sur l’intensité des cellules effectué préalablement (voir section 5.4.5). Par ailleurs, comme je l’expliquais précédemment, ce sont les centres de masse qui sont projetés sur les cartes d’énergie, donc la taille d’un îlot reflète la diversité des modes d’interaction et/ou des formes des ligands mis en jeu. Elle est donc aussi directement influencée par le nombre de ligands qui se sont amarrés favorablement sur l’îlot correspondant et dépend donc de son intensité. Par la suite il faudra exploiter ce phénomène plus en détail pour éviter d’introduire d’éventuels biais.

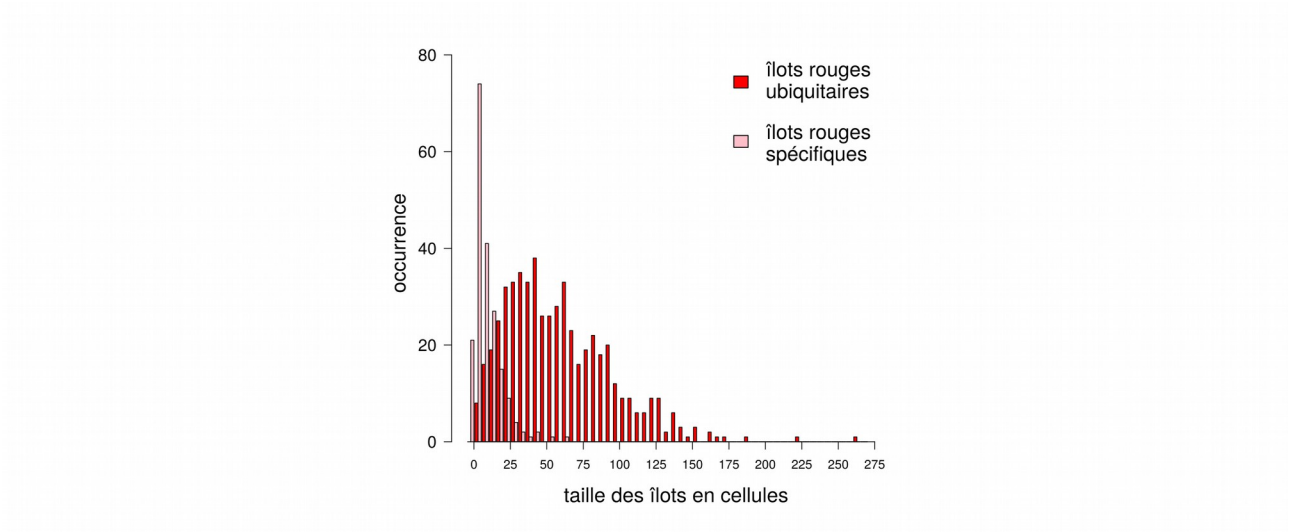


Figure 6.11. Distribution de la taille en nombre de cellules des îlots rouges spécifiques et ubiquitaires sur les 348 récepteurs. La distribution de la taille des îlots rouges spécifiques est représentée en rose, la distribution de la taille des îlots rouges ubiquitaires en rouge.

6.3.3.2 Propriétés physico-chimiques et évolutives des îlots rouges ubiquitaires et spécifiques

Afin de comprendre quelles sont les caractéristiques qui différencient ces deux catégories d’îlots, j’ai comparé leurs propriétés physico-chimiques et évolutives de la même manière que précédemment (voir section 6.3.2.3). La figure 6.12A montre que les îlots rouges spécifiques sont significativement moins collants que les îlots rouges ubiquitaires (test DSH de Tukey (201), valeur p <.001), de plus ils sont aussi moins hydrophobes (test DSH de Tukey, valeur p <.001) (Figure 6.12B). Ces deux résultats sont cohérents avec les indices de *stickiness* et d’hydrophobicité : les régions plus collantes et hydrophobes sont plus enclines à engager des interactions non-spécifiques

avec d'autres protéines. Les îlots rouges spécifiques se comportent plutôt comme les îlots des régions oranges du point de vue de la *stickiness* (test DSH de Tukey, valeur p = 0.58) mais se distinguent du point de vue de l'hydrophobicité (test DSH de Tukey, valeur p <.001) avec des valeurs intermédiaires entre les régions oranges et les îlots rouges ubiquitaires. Il semble donc difficile d'assimiler les îlots rouges spécifiques à ce type de régions. La conservation évolutive des îlots rouges sélectifs est aussi significativement plus faible que celle des îlots rouges ubiquitaires qui eux sont bien plus conservés que le reste de la surface (Figure 6.12D). Dans le cas de la variance circulaire, les propriétés des îlots rouges sélectifs sont proches de celles des îlots rouges ubiquitaires (test DSH de Tukey, valeur p = 0.96) (Figure 6.12C) : ces deux types d'îlots correspondent à des régions de surface plus planes que le reste de la surface des protéines, y compris que des régions oranges. Cela montre que, hormis le relief, les îlots rouges ubiquitaires et spécifiques présentent des propriétés de surface différentes. Les îlots rouges ubiquitaires semblent présenter les mêmes propriétés que les sites d'interaction décrits dans la littérature (généralement des sites d'interaction impliqués dans des interactions fonctionnelles), à savoir, ils sont plus hydrophobes, plus plans et plus conservés que le reste de la surface car soumis à la pression de sélection pour maintenir les assemblages auxquels ils participent. Les îlots rouges spécifiques semblent se comporter plutôt comme des régions de surface intermédiaires d'un point de vue énergétique (régions oranges). Ils sont moins hydrophobes, moins « collants » et moins conservés que les îlots rouges ubiquitaires. Nous pouvons émettre alors l'hypothèse que ces îlots reflètent généralement des régions enclines à l'interaction sans être pour autant des sites impliqués dans des interactions fonctionnelles (c'est-à-dire des sites fonctionnels). Ces îlots peuvent aussi refléter des sites d'interaction fonctionnels plus récents et qui n'ont pas encore acquis toutes les propriétés des îlots rouges ubiquitaires. En revanche, il est intéressant de voir que les îlots rouges spécifiques sont aussi plats que les îlots rouges ubiquitaires. Nous pouvons émettre l'hypothèse qu'une des premières conditions pour interagir favorablement avec plusieurs ligands est de présenter une surface plane, les autres propriétés pouvant être optimisées ensuite au cours de l'évolution pour donner naissance à des îlots rouges ubiquitaires. J'ai tenté d'aborder cette question complexe dans le chapitre suivant sur l'étude des complexes homomériques (section 7). Dans la suite de ce chapitre, afin de voir si les îlots rouges ubiquitaires correspondaient plus souvent à des sites d'interaction fonctionnels que les îlots rouges spécifiques, j'ai étudié la corrélation entre les différents types d'îlots rouges et les sites d'interaction identifiés expérimentalement chez les 348 récepteurs.

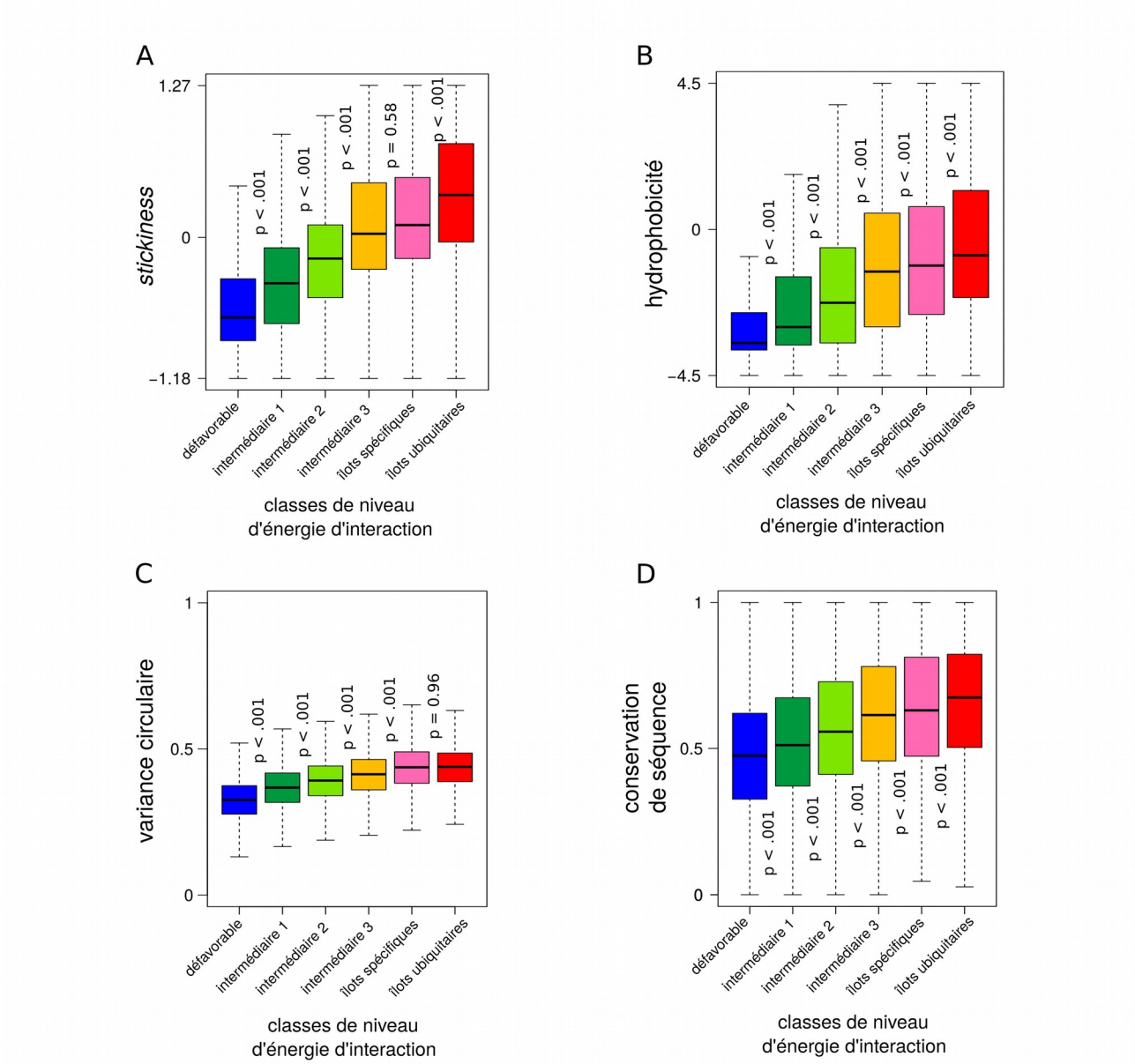


Figure 6.12. Boîtes à moustache des valeurs de variance circulaire, hydrophobicité, *stickiness*, conservation de séquence en fonction des classes d'énergie. Les valeurs des différents descripteurs utilisés sont calculées selon la méthodologie présentée dans la section 5.3. (A) *stickiness*. (B) hydrophobicité de Kyte-Doolittle. (C) variance circulaire. (D) conservation de séquence.

6.3.4 Les îlots rouges sont fortement corrélés avec les sites d'interaction

6.3.4.1 Recouvrement entre sites d'interaction et îlots rouges

Afin de quantifier à quel point les îlots rouges ubiquitaires et spécifiques correspondent effectivement à des sites d'interaction fonctionnels, j'ai calculé le recouvrement entre les deux types d'îlots rouges et les sites d'interaction connus pour chacun des 348 récepteurs.

Dans les résultats qui suivent il est important de noter que nous avons pris en compte la totalité des interfaces présentes dans l'unité biologique du fichier PDB de la protéine. En effet, de nombreuses structures utilisées dans notre jeu de données présentent plusieurs sites d'interaction connus. Ne pas tous les prendre en compte pourrait artificiellement réduire le recouvrement entre les îlots et les sites d'interaction expérimentaux.

Il faut aussi garder en tête que les sites d'interaction présents dans les unités biologiques des fichiers PDB de la base de donnée ne sont probablement pas exhaustifs, et que l'annotation des surfaces des 348 récepteurs en termes de sites d'interaction est sûrement incomplète. Nous sous-estimons donc le recouvrement entre îlots rouges et sites d'interaction.

J'ai utilisé deux méthodes pour comparer les îlots rouges avec les sites d'interaction. La première méthode consiste à procéder par recouvrement entre les sites d'interaction projetés sur des cartes en deux dimensions et les îlots rouges extraits des cartes IPOPS rouges. La seconde méthode consiste à comparer les résidus correspondant aux îlots rouges ubiquitaires et spécifiques avec les résidus appartenant aux sites d'interaction des 348 récepteurs (voir section 5.4.6). Le calcul de recouvrement que j'ai réalisé est asymétrique, il indique le nombre de cellules (méthode 1) ou de résidus (méthode 2) associés aux îlots rouges de la carte IPOPS rouge d'un récepteur qui correspondent effectivement à un site d'interaction identifié expérimentalement dans l'unité biologique du fichier PDB du récepteur correspondant.

Le recouvrement calculé sur l'ensemble des îlots rouges est de 0.58 avec la méthode de calcul de recouvrement entre cellules et de 0.53 avec la méthode de calcul de recouvrement entre résidus (tableau 6.1). Ce résultat signifie que 58 % des cellules constituant les îlots rouges ubiquitaires et spécifiques correspondent à des sites d'interaction dans les cartes de sites d'interaction projetés. En utilisant la méthode de calcul de recouvrement entre résidus, le

recouvrement est de 53 % des résidus appartenant aux îlots rouges (ubiquitaires et spécifiques confondus). Les îlots rouges des cartes IPOPS indiquent donc des sites d'interaction présents dans l'unité biologique des protéines d'intérêt. Lorsque nous distinguons les îlots rouges ubiquitaires et les îlots rouges spécifiques, nous obtenons alors un recouvrement de respectivement 0.59 pour les premiers et de 0.40 pour les seconds avec la méthode de calcul de recouvrement entre cellules, et de 0.55 et 0.38 avec la méthode de calcul de recouvrement entre résidus. Ce résultat montre que les îlots rouges ubiquitaires correspondent majoritairement à des sites d'interaction expérimentaux, puisque que plus de la moitié des résidus appartenant à un de ces îlots correspondent à un site d'interaction répertorié. Les îlots rouge spécifiques correspondent dans plus d'un tiers des cas à un site d'interaction expérimental. Cela semble suggérer qu'ils ne peuvent pas être assimilés à du bruit de fond et qu'ils indiquent des régions de surface présentant des propriétés physico-chimiques qui les rendent favorables à l'interaction. Par ailleurs, l'analyse des propriétés des ces îlots et la comparaison de leurs propriétés avec celles des autres régions semblent corroborer l'hypothèse des îlots rouges spécifiques comme régions enclines à interagir fonctionnellement ou non.

Tableau 6.1. Recouvrement des cellules ou résidus des îlots rouges ubiquitaires et spécifiques par les cellules ou résidus des sites d'interaction (voir section 5.4.6)

classe	$Recouv_{cellules}$	$Recouv_{résidus}$
Tout îlots	0.58	0.53
Îlots rouges ubiquitaires	0.59	0.55
Îlots rouges spécifiques	0.40	0.38

- $Recouv_{cellules}$ représente le recouvrement des cellules des îlots rouges par les cellules des sites d'interaction projetés
- $Recouv_{résidus}$ représente le recouvrement des résidus appartenant aux îlots rouges par ceux des sites d'interaction

6.3.4.2 Le recouvrement entre îlots rouges et sites d'interaction varie en fonction des protéines

J'ai montré que les îlots rouges ubiquitaires et spécifiques recouvrent largement les sites d'interaction protéiques. Je vais maintenant m'intéresser à étudier plus en détail ce résultat sur les

348 récepteurs. Pour cela, j'ai dans un premier temps calculé la distribution du recouvrement entre îlots rouges et les sites d'interaction obtenu pour chaque récepteur. La Figure 6.13 présente la distribution du recouvrement entre les résidus associés aux îlots rouges et ceux des sites d'interaction expérimentaux. J'ai distingué trois populations d'îlots (îlots rouges ubiquitaires, îlots rouges spécifiques et l'ensemble des îlots rouges) (Figure 6.13). Cette distribution est très hétérogène : en observant le recouvrement des îlots rouges ubiquitaires, 140 récepteurs présentent un recouvrement supérieur à 0.7, dont 70 présentent un recouvrement parfait (tous les résidus appartenant aux îlots rouges ubiquitaires appartiennent à un site d'interaction). En revanche, 95 récepteurs présentent un recouvrement inférieur à 0.3, parmi eux 38 présentent un recouvrement de 0 (signifiant qu'aucun résidu des îlots rouges de ces récepteurs ne correspond à un résidu de site d'interaction).

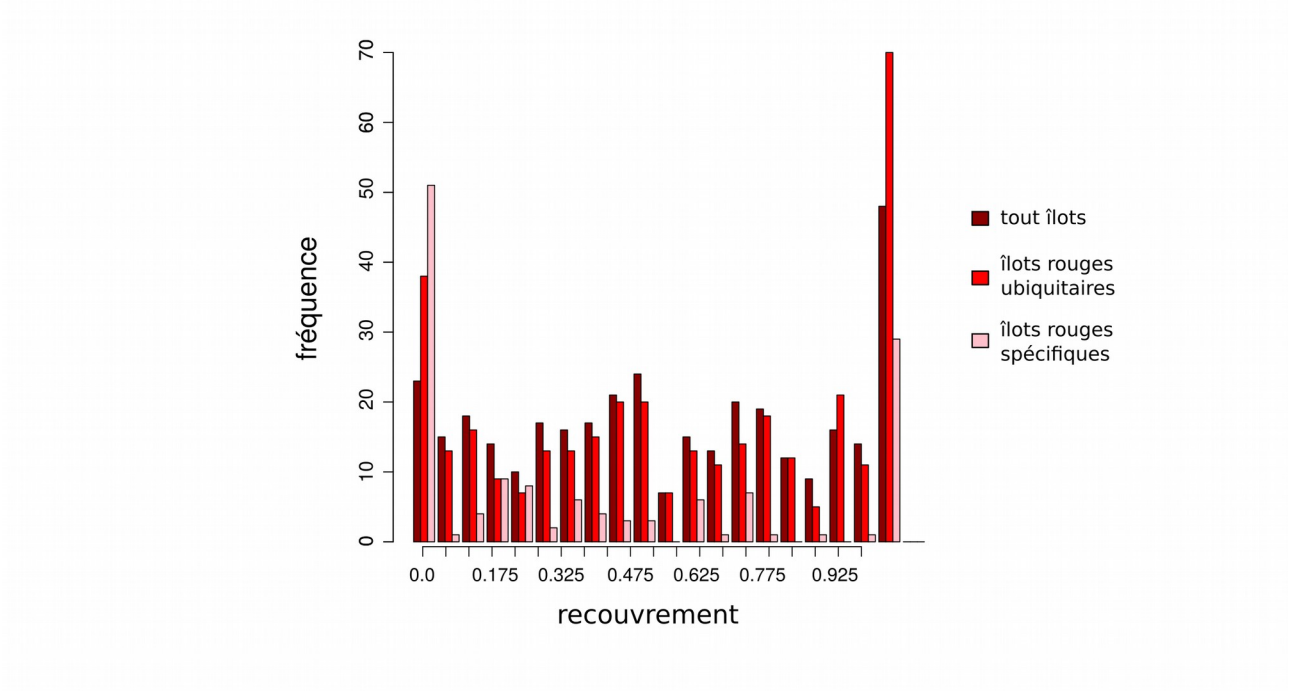


Figure 6.13. Distribution du recouvrement pour chacun des 348 récepteurs des résidus des sites d'interaction de chaque récepteur et de ceux correspondants aux îlots rouges.

Plusieurs exemples de protéines sont présentés dans la figure 6.14. Les figures 6.14A, C et E présentent les cartes IPOPS rouges des protéines 1cg5_B, 3rpf_B et 1tgz_A tandis que les figures 6.14B, D et F présentent leurs cartes de sites d'interaction projetés respectifs. 3rpf_B et 1tgz_A ont déjà été présentées plus haut (voir section 6.3.2.4), 1cg5_B est un monomère de l'hémoglobine de

D. akajei, et est impliquée dans le transport de l'oxygène. L'hémoglobine forme un hétérotétramère composé de deux homodimères (206). Les sous-unités des deux homodimères sont des homologues structuraux. Le cas de 1cg5_B est un cas d'école où les îlots rouges ubiquitaires et spécifiques correspondent parfaitement aux sites d'interaction. En effet le recouvrement calculé pour ce récepteur est de 1. Il est intéressant de voir que sa carte IPOPS rouge présente un îlot rouge ubiquitaire et un îlot rouge spécifique. L'îlot rouge ubiquitaire indique le site d'interaction numéro trois (en jaune sur la carte – figure 6.14B) tandis que l'îlot rouge spécifique correspond au premier site d'interaction (en rouge sur la carte des sites d'interaction projetés - figure 6.14B).

Dans l'exemple de 3rpf_B, les îlots de la carte IPOPS rouge correspondent encore parfaitement aux sites d'interaction (recouvrement de 0.86) (figure 6.14C et D).

La carte IPOPS de 1tgz_A possède un îlot rouge ubiquitaire situé en son milieu à droite (figure 6.14E). Cet îlot ne correspond que partiellement au site d'interaction de 1tgz_A. Le recouvrement est de 0.63, ce qui est au dessus du recouvrement moyen des îlots rouges avec les sites d'interaction (0.55) mais ce qui reflète aussi que l'îlot identifie une proportion importante de résidus en dehors du site d'interaction. Rappelons que les îlots reflètent la projection des centres de masse des solutions de *docking*. Les cellules en dehors du site d'interaction reflètent probablement des solutions au voisinage du site d'interaction (c'est-à-dire des poses de *docking* où le ligand interagit à proximité du site d'interaction). Il faut aussi noter que la projection du centre de masse d'un ligand dépend de son orientation (lorsqu'il n'est pas sphérique) et de sa forme. Ainsi, notre méthode n'identifie pas des sites d'interaction directement mais permet de refléter la diversité des modes d'assemblages d'une classe d'énergie donnée. Autrement dit, un grand îlot extrait d'une carte IPOPS rouge peut refléter une grande diversité de modes d'assemblages de basse énergie ou des assemblages de basse énergie impliquant la même région de surface du récepteur avec des ligands de tailles et de formes très différentes. Dans tous les cas, les régions identifiées par les îlots rouges correspondent à des régions avec un fort potentiel d'interaction qui peuvent impliquer les mêmes sites d'interaction ou non.

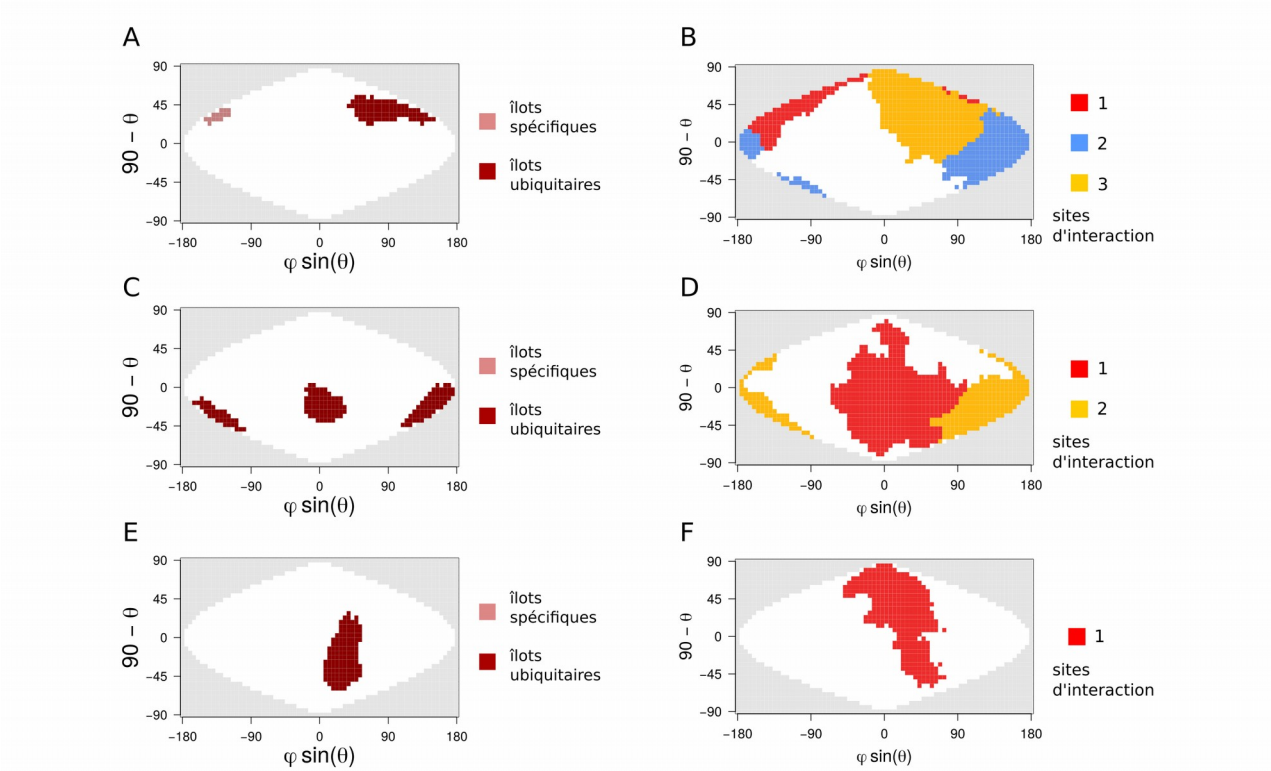


Figure 6.14. Cartes IPOPS rouges et de sites d'interaction des protéines 1cg5_B, 3rpf_B et 1tgz_A. (A) carte IPOPS rouge de 1cg5_B. (B) carte de sites d'interaction de 1cg5_B. (C) carte IPOPS rouge de 3rpf_B. (D) carte de sites d'interaction de 3rpf_B. (E) carte IPOPS rouge de 1tgz_A. (F) carte de sites d'interaction de 1tgz_A. Voir la section 5.4 pour la procédure de création des cartes.

Le cas de 1z7x_A est particulièrement intéressant (figure 6.15A-C). Il s'agit d'une ribonucléase 1 qui catalyse la dégradation d'ARN (207). Cette protéine présente un recouvrement de 0.32, ce qui signifie que seulement un tiers des résidus associés aux îlots rouges correspondent effectivement à un site d'interaction. La carte IPOPS rouge présente deux îlots spécifiques et un îlot ubiquitaire. Les régions indiquées par les îlots spécifiques correspondent bien à des sites d'interaction tandis que celle indiquée par l'îlot ubiquitaire localisé au centre de la carte, ne correspond à aucun des sites d'interaction de l'unité biologique de 1z7x_A. Nous pouvons nous demander dans quelle mesure cette région indique tout de même un sites d'interaction qui serait non identifié par ce fichier PDB.

Afin de répondre à cette question, j'ai utilisé le logiciel JET qui permet de prédire des sites d'interaction (200). Ce logiciel utilise des méthodes reposant sur l'analyse de séquences (conservation des résidus au cours de l'évolution) et de structures (propriétés physico-chimiques et

géométrie locale des résidus). Le fait que JET prédise la même région que celle identifiée par l'îlot ubiquitaire renforcerait l'hypothèse que cet îlot indique un site d'interaction fonctionnel. En effet, JET a été optimisé pour retrouver des site d'interaction fonctionnels en recherchant des régions de surface présentant des propriétés spécifiques des site d'interaction fonctionnels (à savoir, conservation des résidus, composition en acides-aminés et surfaces planes). La figure 6.15B représente la projection des scores JET, les zones blanches sur la carte correspondant aux régions non prédites comme des sites d'interaction et les zones bleues à des régions avec une forte probabilité de correspondre à des site d'interaction. De façon intéressante, nous observons que ce JET prédit effectivement un site d'interaction au niveau de la région indiquée par l'îlot ubiquitaire. Une recherche des partenaires de cette ribonucléase sur le site de Uniprot (208) ne permet pas de prouver que ce site correspond à un site d'interaction, cependant ce site forme une large interface cristalline dans un fichier pdb (pdb 1dza (209)). Ce résultat ne nous permet pas de conclure que ce site ne correspond pas à un site fonctionnel, mais suggère que l'îlot ubiquitaire pourrait peut-être être un site de forte propension à l'interaction.

Le cas de 1tnr_A est aussi très intéressant. Il s'agit du facteur beta de nécrose tumorale (210), une protéine impliquée dans de nombreux mécanismes tels que la réaction inflammatoire ou la prolifération cellulaire (211). Un îlot rouge ubiquitaire est localisé en dehors du site d'interaction présent dans le fichier PDB (figure 6.15A et C). Comme dans l'exemple précédent, nous observons que JET prédit effectivement un site d'interaction au niveau de la région indiquée par l'îlot rouge ubiquitaire (figure 6.15B). Une recherche des partenaires de cette protéine sur le site de Uniprot (208) montre que ce site correspond à un site d'homomérisation (pdb 4mxw (212)). Ce résultat confirme que l'îlot ubiquitaire indique effectivement un site d'interaction fonctionnel. Cela suggère que l'utilisation de JET qui repose sur la recherche de régions de surface conservées évolutivement et donc soumises à pression de sélection peut aider à discriminer les sites d'interaction fonctionnels des régions à forte propension l'interaction, c'est-à-dire enclines à interagir avec différents partenaires de façon non-fonctionnelle de par leur caractère « collant ».

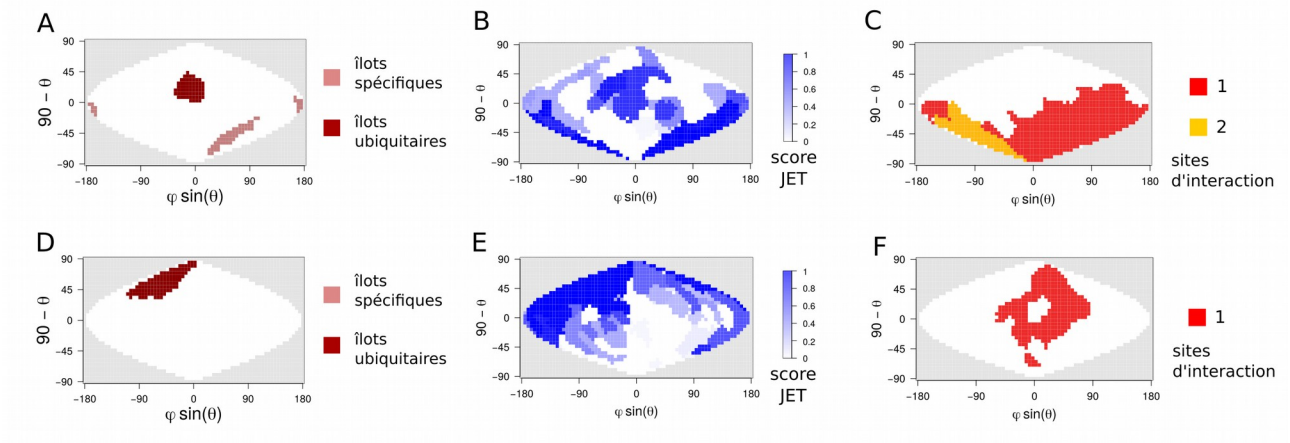


Figure 6.15. Cartes IPOPS rouge, de prédiction JET et sites d'interactions 1z7x_A et 1tnr_A. (A) carte IPOPS rouge de 1z7x_A. (B) carte de prédiction JET de 1z7x_A (C) carte des sites d'interactions projetés de 1z7x_A. (D) carte IPOPS rouge de 1tnr_A . (E) carte de prédiction JET de 1tnr_A. (F) carte des sites d'interactions projetés de 1tnr_A. Voir la section 5.4 pour la procédure de création des cartes.

Il existe aussi cependant des cas de protéines pour lesquelles nous ne pouvons expliquer les discordances entre les ilots rouges les sites d'interaction indiqués dans les unités biologiques des fichiers PDB de notre base de données. Par exemple la protéine 4ksk_A est une déubiquitinase spécifique de l'ubiquitine (213). Sa carte IPOPS rouge présente deux îlots ubiquitaires : un localisé sur la gauche de la carte, l'autre en son centre (figure 6.16A). Elle possède aussi un petit îlot spécifique dans la région supérieure gauche de la carte. L'îlot ubiquitaire localisé à gauche correspond au site d'interaction de 4ksk_A avec l'ubiquitine (figure 6.16B). L'autre îlot ubiquitaire et l'îlot spécifique ne correspondent pas à un site d'interaction connu. En comparant les îlots rouges avec la carte des scores JET projetés (figure 6.16C) nous voyons que les deux îlots ubiquitaires correspondent bien aux sites d'interaction prédits par JET ce qui n'est pas le cas de l'îlot spécifique. Ce dernier indique possiblement une région avec une forte propension à l'interaction qui ne correspond pas nécessairement à un site d'interaction fonctionnel. Une recherche sur le site d'Uniprot n'a pas permis de confirmer expérimentalement le caractère fonctionnel de l'îlot ubiquitaire au centre de la carte, cependant l'accord avec la prédiction de JET peut laisser supposer que ce site est fonctionnel mais n'a pas été caractérisé expérimentalement.

Un autre cas de figure est celui de la protéine 2nxx_B. Il s'agit d'un récepteur de l'ecdysone (une hormone stéroïde) (214). La carte IPOPS rouge de cette protéine (figure 6.16D) présente un îlot ubiquitaire au centre inférieur de la carte, ainsi que deux îlots spécifiques localisés respectivement dans les régions gauche inférieure et supérieure. Seul un îlot spécifique correspond à

un site d'interaction connu (figure 6.16E). En comparant les îlots rouges aux sites d'interaction prédits par JET nous voyons que la région indiquée par l'îlot spécifique en bas à gauche est aussi retrouvée par JET (figure 6.16F). Cependant ce n'est pas le cas des deux îlots restants. Soit ces deux îlots correspondent à du bruit soit ils indiquent des régions à forte propension à l'interaction qui ne correspondent pas à des sites fonctionnels ou du moins qui ne correspondent pas à des sites soumis à pression de sélection.

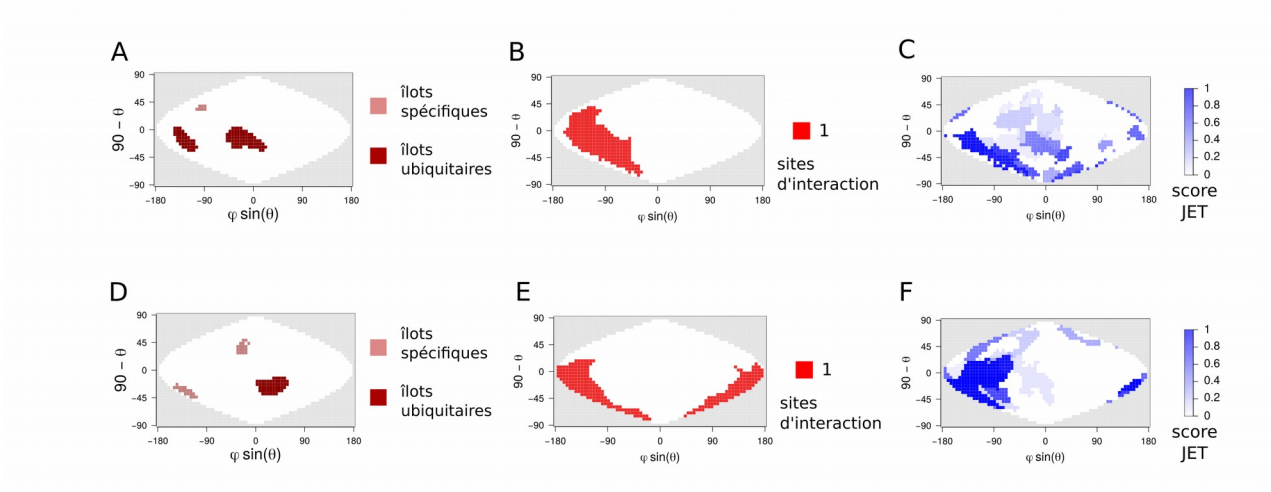


Figure 6.16. Cartes IPOPS rouges, d'interfaces et de prédiction JET des protéines 4ksk_A et 2nxx_B. (A) carte IPOPS rouge de 4ksk_A. (B) carte de sites d'interaction de 4ksk_A. (C) carte de prédiction JET de 4ksk_A. (D) carte IPOPS rouge de 2nxx_B. (E) carte de sites d'interaction de 2nxx_B. (F) carte de prédiction JET de 2nxx_B. Voir la section 5.4 pour la procédure de création des cartes.

En définitive, j'ai montré que les îlots ubiquitaires et spécifiques extraits des cartes IPOPS rouges ont un recouvrement important avec des sites d'interaction fonctionnels mais indiquent aussi des régions non annotées comme site d'interaction dans la PDB. Dans ce cas là plusieurs possibilités existent : il pourrait s'agir de sites d'interaction non répertoriés, de sites « collants » à fort potentiel d'interaction qui ne forment aucune interface fonctionnelle ou tout simplement de bruit lié à la méthode ou à des raisons techniques n'ayant pas de signification biologique. Par exemple, l'utilisation d'une structure d'une protéine tronquée, qui exposerait artificiellement une région hydrophobe pourrait présenter un fort potentiel d'interaction au niveau de cette région hydrophobe et l'îlot rouge qu'elle induirait n'aurait pas de sens biologique. L'utilisation conjointe de JET peut s'avérer très intéressante pour discriminer les sites soumis à pression de sélection et donc probablement fonctionnels, des régions à fort potentiel d'interaction avec un grand nombre de

ligands ou des sous-ensemble de ligands particuliers. Ce résultat est très intéressant pour le développement de méthodes de prédiction de sites à fort potentiel d'interaction fonctionnelles ou non-fonctionnelles.

6.3.4.3 Le recouvrement entre îlots rouges et sites d'interaction varie en fonction des familles protéiques

En plus de l'analyse des recouvrements entre îlots rouges et sites d'interaction par protéine, j'ai aussi investigué si le recouvrement entre les îlots rouges et les sites d'interaction était le même au sein d'une même famille d'homologues. La figure représente la distribution de la moyenne de recouvrement entre îlots rouges et sites d'interaction par famille (figure 6.17). Nous observons que 23 familles (27%) présentent un recouvrement moyen supérieure à 0.7. Ainsi, pour plus d'un quart des familles protéiques, les îlots rouges identifient des régions correspondant à des site d'interaction. A l'opposé 12 familles (15%) présentent un recouvrement moyen inférieure à 0.3. Les 46 autres familles présentent un recouvrement moyen situé entre ces deux extrêmes. Dans ces cas là, soit cela est dû à des îlots qui recouvrent partiellement des site d'interaction soit à des familles que je qualifierai de « hétérogènes », c'est-à-dire où les îlots identifiés et/ou les sites d'interaction annotés sont différents d'un récepteur à l'autre.

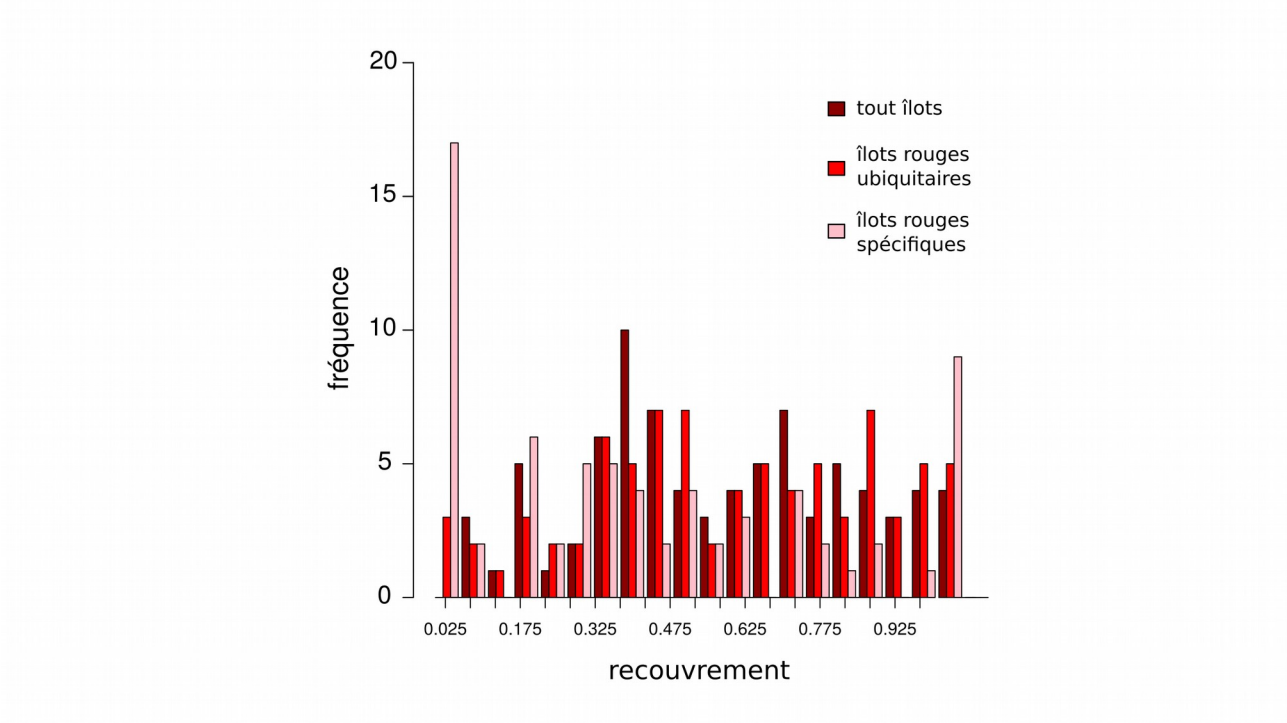


Figure 6.17. Distribution du recouvrement moyen pour chacune des 81 familles de récepteurs entre les résidus des sites d'interaction ceux correspondants aux îlots rouges.

Afin de mieux comprendre ces résultats, nous allons maintenant examiner deux cas de familles protéiques : un cas présentant un même taux de recouvrement entre tous les membres de la famille et une famille hétérogène. La figure 6.18 présente les cartes IPOPS rouges et les cartes des sites d'interaction projetés des membres de la famille des hémoglobines. Cette famille de protéine forme des hétérotétramères composés de deux dimères homologues. Les îlots identifiés pour tous les membres de cette famille présentent un recouvrement avec les site d'interaction de 1. En effet, pour chaque membre de la famille, nous observons un îlot rouge ubiquitaire situé en haut à droite de la carte qui correspond au site d'interaction numéro 3 (en jaune sur les cartes de site d'interaction projetés). Il s'agit d'un cas d'école de famille où les îlots sont homogènes d'un membre à l'autre et où ils indiquent des régions correspondant à des sites fonctionnels. Nous pouvons noter le cas de 1cg5_B qui présente un îlot spécifique supplémentaire en haut à gauche de sa carte IPOPS rouge (figure 6.18G). Cet îlot identifie une région qui correspond au site d'interaction 1 de 1cg5_A (en rouge sur la carte, figure 6.18H). De façon générale, il est étonnant que les autres site d'interaction (en rouge et en bleu sur la carte) ne soient pas du tout identifiés par des îlots rouges.

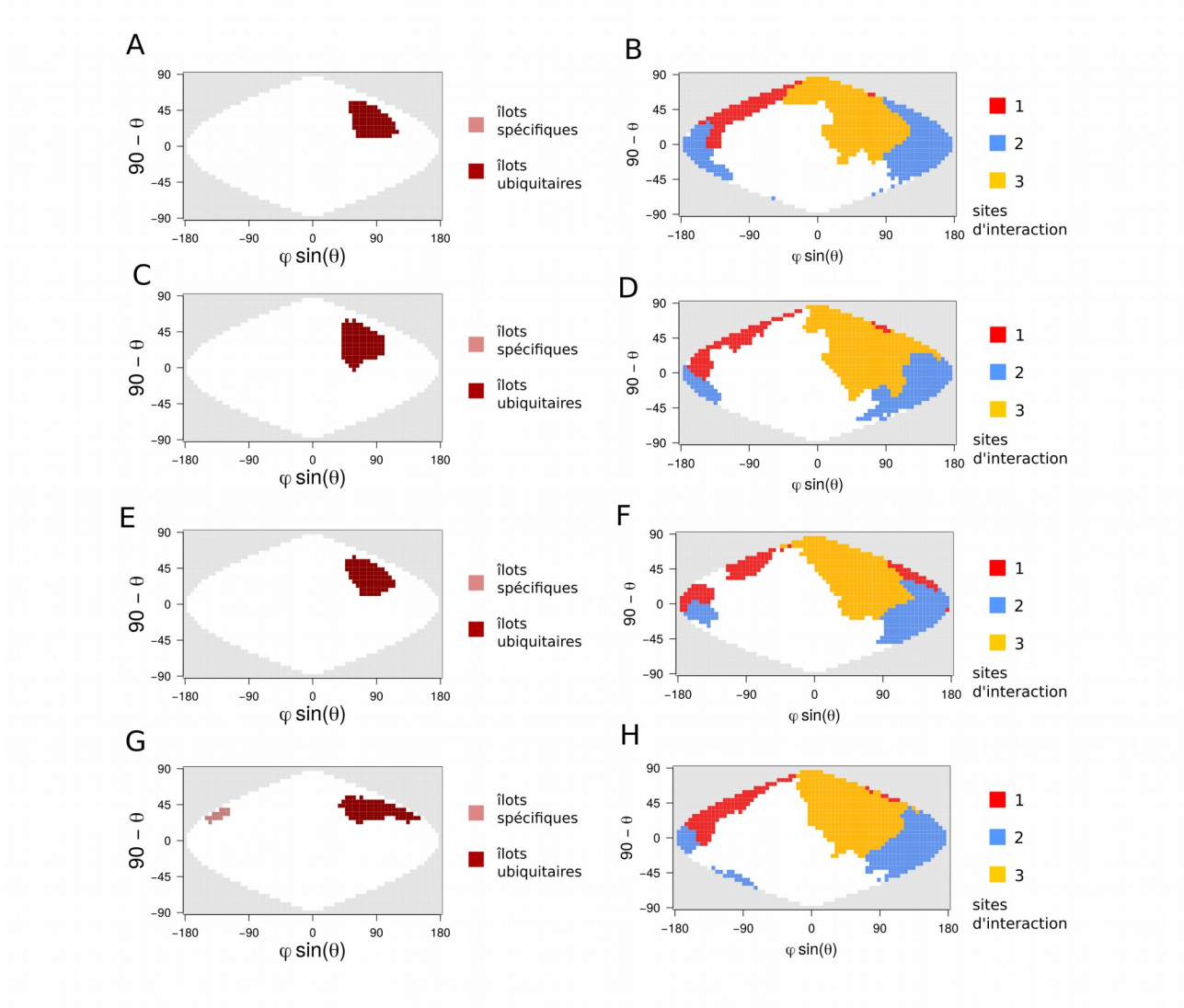


Figure 6.18. Cartes IPOPS rouges et de sites d'interaction de la famille des hémoglobines. (A) carte IPOPS rouge de 1wmu_B. (B) carte de sites d'interaction de 1wmu_B. (C) carte IPOPS rouge de 3bom_C. (D) carte de sites d'interaction de 3bom_C. (E) carte IPOPS rouge de 3d1k_B. (F) carte de sites d'interaction de 3d1k_B. (G) carte IPOPS rouge de 1cg5_B. (H) carte de sites d'interaction de 1cg5_B. Voir la section 5.4 pour la procédure de création des cartes.

A l'inverse la figure 6.19 présente les carte IPOPS rouges et les cartes des sites d'interaction projetés des trois membres de la famille des lectines. Nous pouvons tout de suite voir que les trois récepteurs présentent les mêmes sites d'interaction (voir leurs cartes de sites d'interaction projetés) tandis que leurs cartes IPOPS rouges présentent des différences notables. La carte IPOPS rouge de 1hwm_B (l'ébuline) présente un îlot rouge ubiquitaire sur le milieu supérieur de la carte (figure 6.19A). La région indiquée par cet îlot ne correspond pas au site d'interaction présent dans l'unité biologique ni à aucun autre site d'interaction connu. La carte IPOPS rouge de 1ggp_B (chaîne B de

la lectine 1) présente trois îlots ubiquitaires localisés au niveau du centre inférieur et supérieur de la carte, ainsi qu'un petit îlot spécifique dans la partie inférieure gauche de la carte. Il est intéressant d'observer qu'aucun des îlots spécifiques ne correspond à l'îlot spécifique de la carte IPOPS de 1hwm_B (figure 18A). Ils ne correspondent pas non plus au site d'interaction présent dans l'unité biologique de 1ggp_B ni à aucun site d'interaction répertorié pour cette protéine. La carte IPOPS de 1m2t_B (chaine B de la lectine 1 du gui) présente deux îlots ubiquitaires situés respectivement dans le milieu supérieur et inférieur de la carte (figure 6.19E). De même que pour les deux protéines précédentes, les deux îlots rouges ubiquitaires ne correspondent à aucun site d'interaction connu, bien que l'îlot de la partie inférieure de la carte présente un petit recouvrement avec le site d'interaction (figure 6.19F). Cet îlot est aussi retrouvé sur la carte IPOPS de 1ggp_B (figure 6.19C), tandis que l'îlot de la partie supérieure de la carte correspond à l'îlot de la carte IPOPS de 1hwm_B.

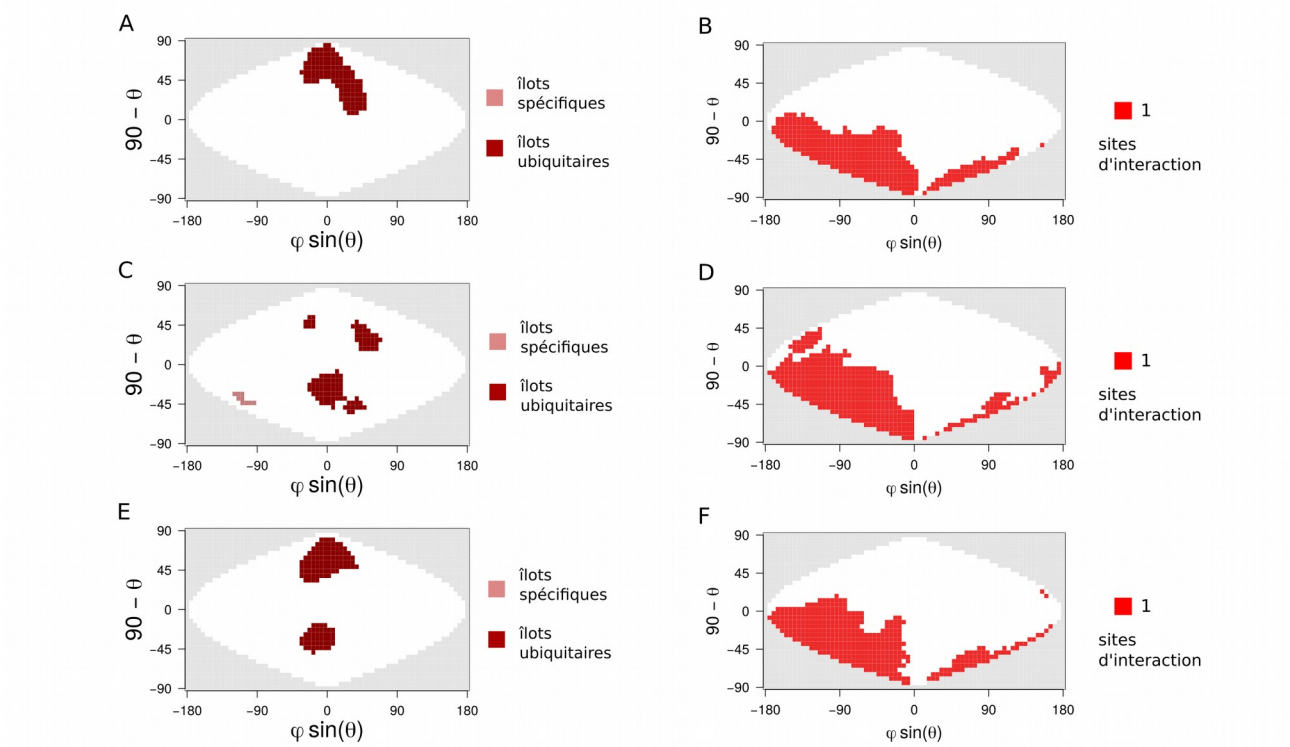


Figure 6.19. Cartes IPOPS et de sites d'interaction de la famille des lectines. (A) carte IPOPS rouge de 1hmw_B. (B) carte de sites d'interaction de 1hmw_B. (C) carte IPOPS rouge de 1ggp_B. (D) carte de sites d'interaction de 1ggp_B. (E) carte IPOPS rouge de 1m2t_B. (F) carte de sites d'interaction de 1m2t_B. Voir la section 5.4 pour la procédure de création des cartes.

6.3.5 Conservation des îlots rouges au sein des familles d'homologues

Je me suis intéressé plus en détail à la variabilité des cartes IPOPS rouges au sein des familles d'homologues afin d'étudier si les membres d'une même famille présentaient les mêmes cartes. Pour cela, j'ai calculé la distance entre toutes les paires de cartes IPOPS rouges des membres d'une même famille (voir section 5.4.8). Les cartes de distances de chaque famille sont présentées en annexes. Cependant par manque de temps, je n'ai pas trouvé de métrique adaptée pour analyser automatiquement la variabilité des cartes IPOPS au sein des familles, j'ai donc analysé les cartes de distances de chaque famille visuellement afin de bien appréhender les données et j'ai classé les 81 familles en quatre catégories :

- familles homogènes : les familles pour lesquelles les îlots rouges des membres de la famille sont localisés dans les mêmes régions.
- familles homogènes avec cas particulier(s) : familles homogènes dont un ou deux membres présentent des îlots rouges différents des autres membres
- familles séparées en deux sous-groupes homogènes.
- familles hétérogènes : familles où la plupart des membres présentent des îlots différents, autrement dit, familles pour lesquelles il est difficile d'établir un profil d'îlots.

J'ai dénombré 22 cas de familles homogènes, 22 cas de familles homogènes présentant un ou deux cas particuliers, deux familles subdivisées en deux sous-groupes homogènes et 35 cas de familles hétérogènes où aucune tendance ne se dégage.

Un exemple de famille au comportement homogène a déjà été présenté dans la figure 17 avec la famille des hémoglobines. Un exemple de famille homogène avec un cas particulier est celui de la famille des protéases eucaryotes (protéases apparentées aux trypsines, tel que la cathepsine ou l'élastase) (figure 6.20). Les trois protéines 1cgi_A, 1ldt_A et 1ppf_A présentent des cartes IPOPS similaires (figure 6.20A, C et E) avec un îlot rouge ubiquitaire localisé dans la partie inférieure de la carte, tandis que 1sgp_A possède un îlot rouge ubiquitaire localisé au milieu de la carte. La comparaison visuelle avec les cartes de sites d'interaction projetés de ces trois protéines (figure 6.20B, D, F et H) montre que les îlots rouges ubiquitaires indiquent des régions correspondant systématiquement au site d'interaction expérimental, alors que ce n'est pas le cas pour l'îlot rouge ubiquitaire de 1sgp_A. Soit cette protéine 1sgp_A possède des propriétés de surface différentes de ses homologues et/ou une fonction différente qui expliquent la différence observée sur sa carte IPOPS soit cet îlot n'est pas pertinent biologiquement et vient d'un biais de notre méthode.

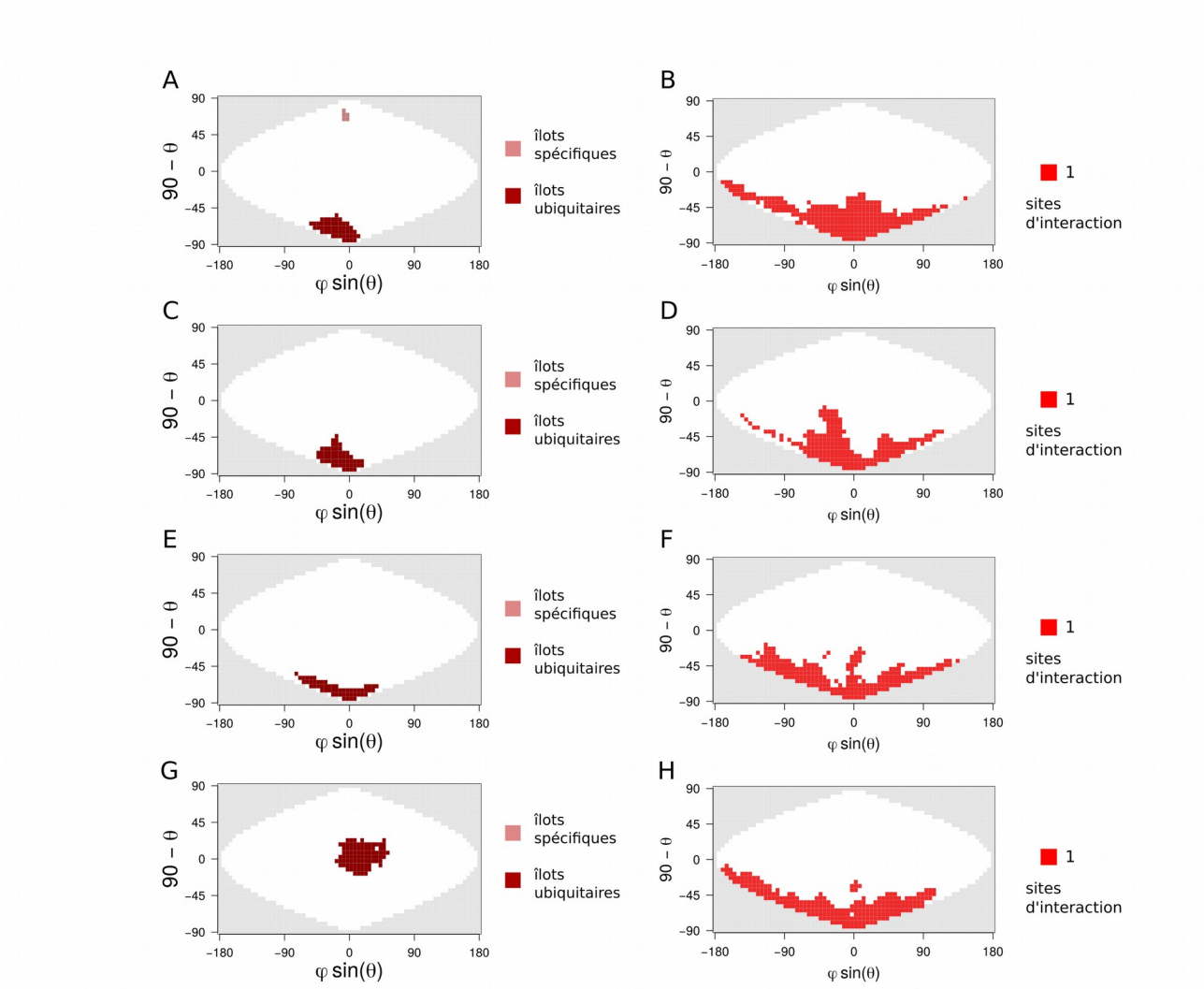


Figure 6.20. Cartes IPOPS rouges et de sites d'interaction de la famille des protéases eucaryotes. (A) carte IPOPS rouge de 1cgi_A. (B) carte des sites d'interaction de 1cgi_A. (C) carte IPOPS rouge de 1ldt_A. (D) carte des sites d'interaction de 1ldt_A. (E) carte IPOPS rouge de 1ppf_A. (F) carte des sites d'interaction de 1ppf_A. (G) carte IPOPS rouge de 1sgp_A. (H) carte des sites d'interaction de 1sgp_A. Voir la section 5.4 pour la procédure de création des cartes.

La famille d'inhibiteurs (type antithrombine) de protéinases est un exemple de famille hétérogène : les cartes IPOPS des 3 membres ne possèdent aucun îlot rouge ubiquitaire ou spécifique en commun (figure 6.21A,C et E), comme indiqué par notre mesure de distance (figure 6.21G). De plus une comparaison avec les cartes SI (figure 6.21B,D et F) montre que les îlots détectés avec notre méthode ne sont pas chevauchants avec les sites d'interactions présents dans l'unité biologique des protéines de la famille. Cela suggère que notre méthode n'a pas été capable de mettre en évidence des régions enclines à l'interaction. Cela peut-être dû au fait que ces

protéines soient un cas particulièrement difficile pour le *docking* protéique. La fonction de score d'ATTRACT pourrait peut-être être inadaptée dans le cas de cette protéine.

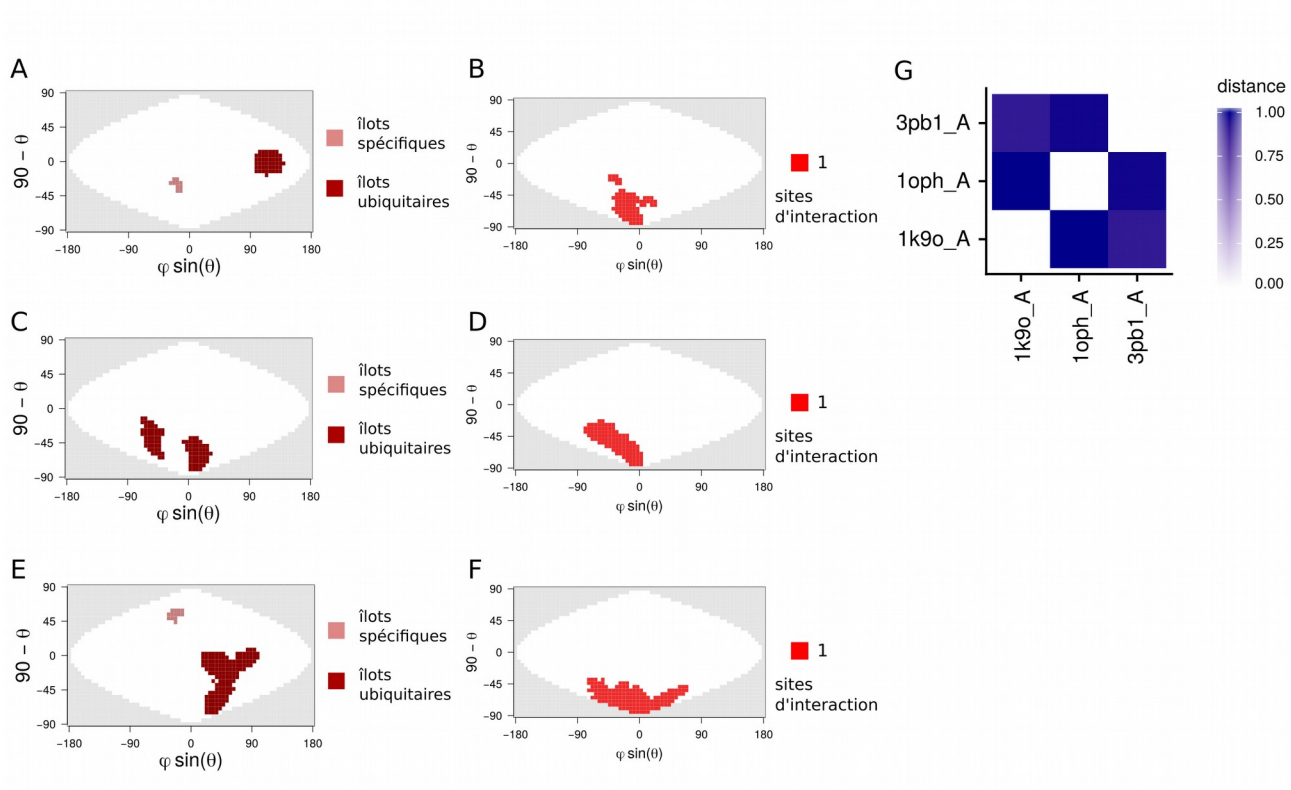


Figure 6.21. Cartes IPOPS rouges, cartes de sites d'interaction et matrice de distance de cartes d'une famille d'inhibiteur de protéinases. (A) carte IPOPS rouge de 1oph_A. (B) carte des sites d'interaction de 1oph_A. (C) carte IPOPS rouge de 1k9o_A. (D) carte des sites d'interaction de 1k9o_A. (E) carte IPOPS rouge de 3pb1_A. (F) carte des sites d'interaction de 3pb1_A. (G) Distances normalisées entre les cartes IPOPS rouges des trois membres de la famille (voir section 5.4.8 pour le calcul de distance). Voir la section 5.4 pour la procédure de création des cartes.

Le dernier cas de figure auquel nous avons été confronté est celui de familles étant séparées en 2 sous-groupes. La figure 6.22 présente l'exemple de la famille des facteurs de nécrose tumorale. Nous constatons que malgré les distances affichées dans la carte de distance (figure 6.22M), cette famille appartient en fait à la catégorie des familles homogènes : ses six membres possèdent un unique îlot rouge spécifique localisé systématiquement dans la partie supérieure gauche de la carte (figure 6.22A, C, E, G, I et L). Il est intéressant de voir que cette îlot persiste sur les six cartes IPOPS rouges des membres de la famille, malgré le fait que ces protéines fassent partie d'un assemblage de 6 ou 12 sous-unités et donc qu'une grande partie de leurs surfaces respectives (mis à part une exception, figure 6.22D) correspondent à des sites d'interaction (figure 6.22B, F, H, J et L).

La métrique utilisée classe la famille en deux sous-groupes du fait de la taille respective des îlots. Cependant nous ne pensons pas qu'un critère de taille soit pertinent dans le cas présent, le critère de localisation étant plus important.

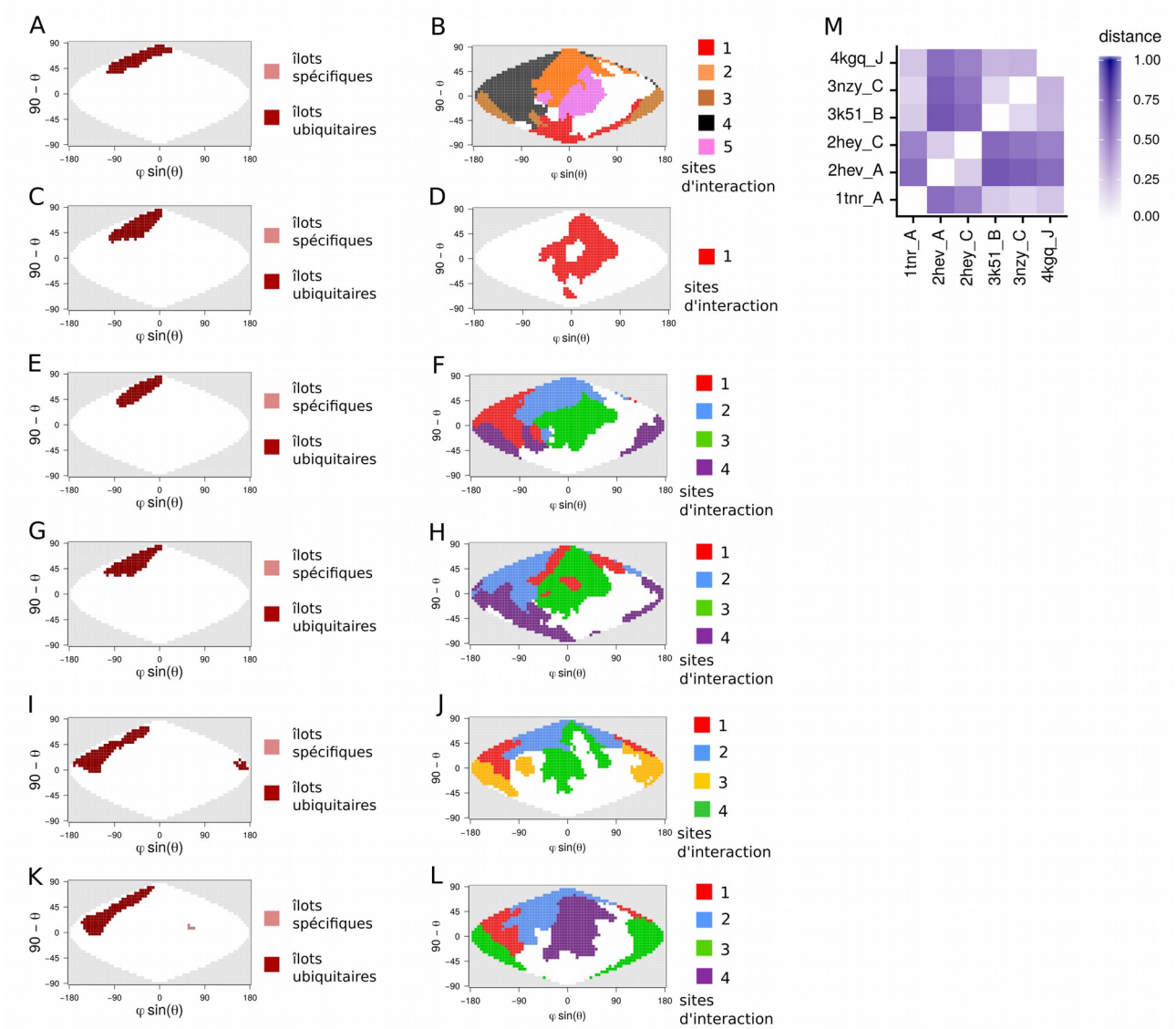


Figure 6.22. Cartes IPOPS rouges, cartes de sites d'interaction et matrice de distance de cartes d'une famille de facteurs de nécrose tumorale. (A) carte IPOPS rouge de 4kgq_J. (B) carte des sites d'interaction de 4kgq_J. (C) carte IPOPS rouge de 1tnr_A. (D) carte des sites d'interaction de 1tnr_A. (E) carte IPOPS rouge de 3k51_B. (F) carte des sites d'interaction de 3k51_B. (G) carte IPOPS rouge de 3nzy_A .(H) carte des sites d'interaction de 3nzy_A (I) carte IPOPS rouge de 2hev_A (J) carte des sites d'interaction de 2hev_A (K) carte IPOPS rouge de 2hey_A (L) carte des sites d'interaction de 2hey_A (M) Distances normalisées entre les cartes IPOPS rouges des trois membres de la famille. Voir la section 5.4 pour procédure de création des cartes et la section 5.4.8 pour le calcul de distance).

6.3.6 Application des cartes IPOPS rouges à la prédiction de sites d'interaction

Bien que nous n'ayons pas développé les cartes IPOPS pour la prédiction de sites d'interaction, le recouvrement que nous avons observé entre les îlots rouges extraits des cartes IPOPS et les sites d'interaction montrent que ces cartes peuvent être utilisées pour prédire la localisation des sites d'interaction d'une protéine. Afin de pouvoir évaluer les performances de ces cartes pour la prédiction de sites d'interaction, j'ai comparé les performances des cartes IPOPS en tant que prédicteur de localisation de sites d'interaction avec les performances du logiciel JET (200), ainsi que celles obtenues en utilisant la NIP (*Normalized Interaction Propensity* présentée en section 5.4.6) (166). Pour rappel, la NIP est une méthode utilisée pour prédire les résidus appartenant à un site d'interaction qui repose sur l'estimation de la propension d'un résidu à interagir à partir de calculs de *docking* arbitraires (voir section 5.4.6) (153,156,159,160,165,166). La philosophie sur laquelle reposent les cartes IPOPS et la NIP est très similaire. La différence principale réside dans le fait que la NIP est centrée sur les résidus de surface où l'on compte pour chaque résidu, le nombre de fois où il a été observé dans une interface d'une solution de *docking* de basse énergie tandis que les cartes IPOPS sont définies à partir de la projection des centres de masses des solutions de *docking*. Pour comparer les performances des trois approches, j'ai calculé différentes mesures de performance telles que la PPV (*positive predictive value*), la spécificité, la sensibilité et l'efficacité (voir section 5.5) pour leur capacité à prédire les résidus des sites d'interaction expérimentaux. Pour cela, j'ai utilisé comme précédemment le calcul du recouvrement entre les résidus des sites d'interaction expérimentaux et ceux prédits respectivement par JET, la NIP ou ceux associés aux îlots rouges des cartes IPOPS (voir section 5.5).

Le tableau 6.2 présente les résultats des trois approches pour la prédiction de sites d'interaction. Nous voyons que la prédiction reposant sur les cartes IPOPS présente une PPV nettement plus élevée que les deux autres approches (tableau 6.2) : 0.55 contre 0.45 et 0.41 pour l'approche fondée sur la NIP et JET respectivement. La spécificité de notre méthode est très élevée (0.92), là aussi clairement supérieure à celles utilisant le calcul de la NIP ou JET (0.73 et 0.63 respectivement). En revanche, la sensibilité de notre approche est bien plus faible avec une valeur de 0.24 contre 0.52 pour la prédiction fondée sur la NIP et 0.59 pour JET. Néanmoins, je rappelle que notre approche repose sur la projection des centres de masse des solutions de *docking*. Elle est plus adaptée à prédire la localisation d'un site d'interaction que l'ensemble des résidus qui le

composent. Enfin notre méthode possède la meilleure efficacité parmi les trois méthodes comparées : 0.69 contre 0.68 pour la prédiction fondée sur la NIP et 0.62 pour JET.

Ces résultats montrent que JET est la méthode présentant la sensibilité la plus élevée mais au prix d'une spécificité et d'une PPV plus faibles. JET prédit en effet un plus grand nombre de résidus (62 résidus en moyenne par protéine), au prix d'un plus fort taux de faux positifs. La prédiction fondée sur la NIP présente des performances similaires à celles de JET, bien que moins sensible et plus spécifique (49 résidus prédits en moyenne). Notre méthode prédit en moyenne 17 résidus par protéine, un nombre bien plus faible qu'avec les deux autres approches et la sensibilité s'en trouve évidemment affectée. Néanmoins, les points forts de notre méthode résident clairement dans la PPV et la spécificité. Le taux de faux positifs est plus faible avec notre approche.

La spécificité de notre méthode est excellente, cependant il convient de noter que c'est aussi le cas pour de nombreux prédicteurs de sites d'interaction. En effet, il existe un fort déséquilibre entre les classes positives (résidus appartenant à un site d'interaction) et négatives (résidus de surface n'appartenant pas à un site d'interaction) en faveur de la deuxième catégorie. Ce déséquilibre a pour résultat que les prédicteurs de sites d'interaction présentent souvent une spécificité élevée et une bonne sensibilité malgré une faible PPV. Nous pensons qu'avoir une bonne PPV est déterminant pour la prédiction de sites d'interaction dans la perspective d'expériences de mutagénèse dirigée, de *docking* sous contraintes ou de *design* de nouvelles interactions. En effet, avoir une fraction de faux positifs la plus faible possible permet d'augmenter les chances de réussite de l'expérience qui en découlera. Selon le type d'application envisagé, avoir une bonne PPV peut se révéler plus intéressante qu'avoir une très bonne sensibilité. Dans le cas du *docking* sous contraintes (les contraintes sont généralement utilisées pour réduire l'espace conformationnel de recherche, voir introduction section 2.2.3.1), avoir une liste de quelques résidus prédits avec confiance pour guider la procédure de *docking* est plus avantageux qu'une grande liste de résidus mais comprenant un fort taux de faux positifs.

Tableau 6.2: comparaison des performances de IPOPS, JET et la NIP (voir sections 5.4.6, 5.4.7 et 5.5 pour les méthodologies employées)

	PPV	sens	spe	ACC
Tout îlots rouges	0.53	0.24	0.9	0.69
Îlots rouges ubiquitaires	0.55	0.22	0.92	0.69
NIP	0.45	0.52	0.73	0.67
JET	0.41	0.59	0.63	0.62

- Tout îlots rouges : prédiction de sites d'interaction faisant usage des îlots rouges ubiquitaires et spécifiques
- Ilots rouges ubiquitaires : prédiction de sites d'interaction faisant usage des îlots rouges ubiquitaires
- NIP : *normalized interaction propensity*. Les scores ont été calculés avec la définition des résidus de surface et d'interface aboutissant à la plus haute efficacité ($\Delta rsa > 0$, $rsa = 5$), seuil NIP = 0.1
- JET : les scores sont calculés avec les paramètres suivants : seuil JET = 0

J'ai montré que notre approche peut être utilisée pour la prédiction de résidus de sites d'interaction protéiques. Pour aller plus loin, il conviendrait de comparer notre méthode avec plus de méthodes de prédiction de sites d'interaction, tel que dans (215). Il faut aussi tenir compte du temps de calcul de notre méthode. En effet, tout comme la méthode reposant sur le calcul de la NIP, notre méthode repose sur des calculs de *docking* arbitraire. Ici, nous avons défini les cartes IPOPS de chaque protéine à partir du *docking* de cette protéine et 100 protéines arbitraires. Ainsi, le temps de calcul nécessaire pour obtenir une carte IPOPS est conséquent (plusieurs dizaines d'heures avec un jeu de 100 ligands). Dans la perspective de proposer une utilisation en routine de notre méthode, j'ai analysé l'influence de la taille du jeu de ligands sur la variabilité des cartes IPOPS.

6.3.7 Influence de la taille du jeu de ligands sur la variabilité des cartes IPOPS

Ainsi, j'ai systématiquement recalculé les cartes IPOPS rouges à partir de tirages aléatoires de jeux de N ligands extraits du jeu de 100 ligands. J'ai choisi 10 tailles différentes d'échantillons, de respectivement 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 ligands. Pour chaque taille N d'échantillons, j'ai réalisé 100 tirages aléatoires de N ligands parmi les 100 ligands de départ. Dans tout ce qui suit,

nous réfèrerons aux 348 cartes IPOPS rouges réalisées avec les 100 ligands du jeu de départ comme les cartes de « référence » et aux cartes réalisées avec des sous-ensembles du jeu de 100 ligands comme les cartes de « sous-ensemble ».

Dans un premier temps, j'ai estimé pour chaque récepteur, la similarité de ses cartes IPOPS calculées sur les sous-ensembles de ligands avec sa carte IPOPS de « référence » en calculant le nombre de cellules de la carte assignées différemment (c'est-à-dire à un îlot rouge ubiquitaire, rouge spécifique ou non assignées à un îlot rouge) (voir section 5.4.8). La figure 6.23A présente pour chaque taille d'échantillon la moyenne des distances entre les cartes de « référence » et les cartes de « sous-ensemble » calculées sur les 348 récepteurs. De façon intéressante, nous observons qu'aucun plateau n'est atteint, même si la distance diminue rapidement (figure 6.23A) : les cartes réalisées avec cinq ligands uniquement présentent en moyenne une distance avec les cartes de références de 65 cellules sur les 1548 cellules de la carte (ce qui représente environ 4 % de la surface d'une carte), tandis que les cartes réalisées avec 20 ligands présentent une distance moyenne de 40 cellules (2.7 % de la surface d'une carte) et celles réalisées avec 50 ligands présente une distance moyenne de 23 (1 % de la surface d'une carte). Seulement, il est difficile de savoir si les distances calculées entre les cartes de « référence » et les cartes de « sous-ensembles » reflètent des apparitions/disparitions d'îlots ou si elles reflètent des îlots de tailles et formes différentes (c'est-à-dire dont les contours changent d'une carte à l'autre). Pour répondre à cette question, j'ai donc calculé la différence du nombre d'îlots ubiquitaires, spécifiques et des deux à la fois pour chaque carte de « sous-ensemble » par rapport à sa carte de « référence ». La figure 6.23B montre la variabilité en terme de nombre d'îlots des cartes de « sous-ensemble » par rapport aux cartes de « références » correspondantes. Les cartes de « sous-ensembles » cessent de varier significativement à partir d'un jeu de 30 ligands **(**risque de rejeter faussement l'égalité de deux variances (F-test), valeur p = 0.89**)**, ce qui montre qu'un jeu de cette taille est suffisant pour obtenir des cartes IPOPS stables et représentatives du potentiel d'interaction de la surface de la protéines évaluée. Néanmoins, nous pouvons remarquer que même au delà de 30 ligands, le nombre d'îlots total varie en moyenne de 0.32 et le nombre d'îlots ubiquitaires de 0.2. Nous pouvons donc nous demander dans quelle mesure les îlots qui apparaissent/disparaissent indiquent des régions qui correspondent à des sites d'interaction fonctionnels.

Pour cela, j'ai calculé le recouvrement entre les îlots rouges (ubiquitaires + spécifiques) extraits des cartes de « sous-ensembles » et les sites d'interaction (voir section 5.4.6). La figure 6.23C montre ces valeurs de recouvrement en fonction de la taille des échantillons de ligands. De manière étonnante, nous observons que ces valeurs varient très peu en fonction de la taille du jeu de

ligands. Le recouvrement moyen est de 0.52 pour un jeu de cinq ligands et cesse de varier significativement à partir d'un jeu de 20 ligands (risque de rejeter faussement l'égalité de deux variances (F-test), valeur p = 0.89). Si on se place dans une perspective de prédiction de résidus appartenant au site d'interaction, ce résultat est important car il montre qu'à partir de 20 ligands, l'identification des îlots correspondant à des sites d'interaction fonctionnels est robuste. Autrement dit, la prédiction de sous-ensembles de résidus de sites d'interaction est robuste à partir de 20 ligands. Cependant, si nous sommes intéressés par l'identification d'îlots spécifiques, il faudrait approfondir cette étude pour estimer l'effet de la taille et de la nature du jeu de ligands pour ces îlots qui par définition sont plus variables.

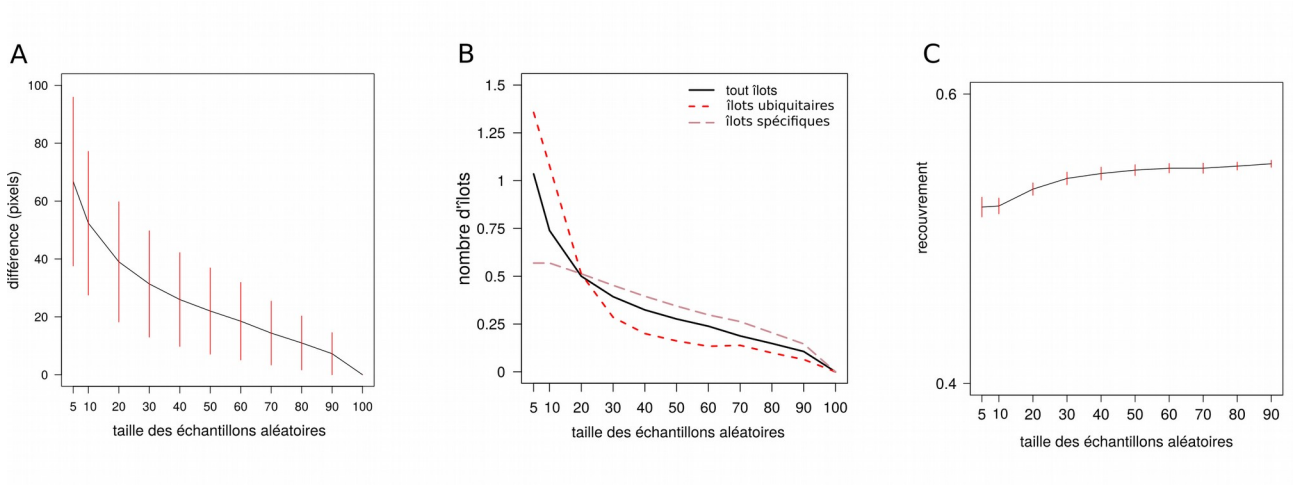


Figure 6.23. Variation entre les cartes IPOPS de référence et les cartes IPOPS réalisées par tirage aléatoire des ligands. La variation est calculée comme le nombre de cellules possédant des assignations différentes entre la carte IPOPS référence calculée avec 100 ligands et les cartes IPOPS réalisées avec un sous-échantillon des 100 ligands. Pour chaque taille d'échantillon 100 tirages aléatoires ont été effectués. Les barres verticales en rouge symbolisent l'écart type. (A) nombre de cellules présentant une assignation différente (B) différence de nombre d'îlots ubiquitaires, spécifiques et total. (C) Recouvrement entre résidus identifiés par les îlots rouges (totaux, ubiquitaires et spécifiques) et résidus des sites d'interaction.

7 Étude du potentiel d'interaction des sites d'homomérisation

Dans cette section, j'ai appliqué la méthodologie que j'avais développée (représentations et méthodes) dans la section précédente sur un autre système d'étude spécifique constitué de complexes homomériques.

Dans le travail précédent j'ai montré que les îlots rouges ubiquitaires (c'est-à-dire favorables à l'interaction pour un grand nombre de ligands) correspondent souvent à des sites d'interaction protéiques fonctionnels. Il sont en moyenne plus hydrophobes, forment des surfaces plus planes et sont plus conservés du point de vue évolutif. A l'opposé, les îlots rouges spécifiques (c'est-à-dire favorable à l'interaction pour un sous-ensemble de ligands) correspondent plus souvent à des régions non caractérisées comme sites d'interaction fonctionnels, et reflètent donc possiblement des (i) sites d'interaction non-fonctionnels (c'est-à-dire, qui peuvent interagir avec des protéines de l'environnement de façon non-fonctionnelle), ou peut-être (ii) des sites d'interaction fonctionnels ayant « émergés » récemment de la surface protéique et possédant ainsi des propriétés physico-chimiques proches de cette dernière. Cependant il est très difficile de vérifier cette dernière hypothèse.

Pour tenter d'apporter des éléments de réponse à cette question, je me suis intéressé à un jeu de données comprenant des assemblages protéiques homomériques homologues présentant des symétries différentes. Ce jeu de données avait été établi lors d'un travail précédent qui visait à caractériser l'évolution des assemblages homomériques (174). Comme je l'expliquerai plus loin, les assemblages homomériques homologues avec des symétries différentes présentent l'intérêt de fournir des protéines homologues possédant des sites d'interaction en commun ainsi que des sites d'interaction spécifiques de la symétrie de l'assemblage auquel elles participent.

Avant d'aborder mes résultats, je vais d'abord expliquer plus en détail quelles sont les caractéristiques et propriétés structurales des complexes homomériques.

7.1 Introduction

7.1.1 Homomères et symétries

L'homomérisation des protéines, c'est-à-dire l'assemblage de plusieurs sous-unités identiques pour former un complexe protéique, est un phénomène très répandu. En se basant sur les structures déterminées expérimentalement, on estime que 45 % des protéines d'organismes eucaryotes et 60 % des protéines d'organismes procaryotes forment des homomères dans des conditions physiologiques (50). La plupart des homomères arborent une symétrie (216), celle-ci pouvant être ouverte ou fermée. Les homomères à symétrie ouverte sont rares dans la cellule et constituent environ 3 % des structures connues (49,174). On peut citer les exemples classiques de l'actine et de la tubuline. A l'inverse les homomères à symétrie fermée sont très courants et constituent plus de 97 % des homomères dont la structure a été déterminée expérimentalement à ce jour.

Il existe plusieurs types de symétries fermées. On distingue par exemple les symétries de type cycliques, diédrales ou cubiques (tetraédrale, octaédrale et icosaédrale, figure 7.1) (49), les deux premières étant très largement majoritaires (49). Dans ce travail nous nous intéressons tout particulièrement à ce type d'homomères et nous référerons aux homomères cycliques composés de n sous-unités comme des Cn, et aux homomères de symétrie diédrale composé de $2n$ sous-unités comme des Dn.

Les homomères de symétrie cyclique sont caractérisés par un unique axe de rotation (figure 7.1), ils représentent plus de 70 % des structures d'homomères connues à ce jour, 62 % des protéines homomériques formant des C2 (homodimères).

Les homomères de symétrie diédrale sont caractérisés par trois axes de rotation (figure 7.1), ils représentent environ 25 % des structures d'homomères connues à ce jour. On peut aussi voir un homomère Dn comme l'assemblage de deux homomères Cn ou de n homomères C2.

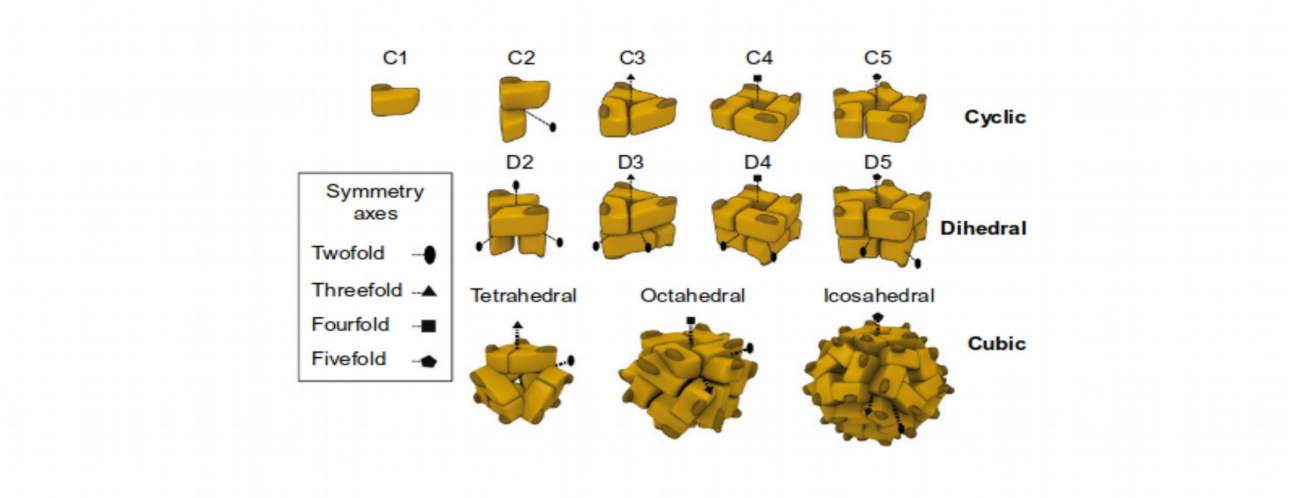


Figure 7.1. Les différents types de symétries fermées chez les homomères. Figure extraite de Levy et Teichmann (49).

7.1.2 Structure quaternaire et évolution des homomères

Les homomères homologues structuraux présentent généralement la même symétrie. En effet des homologues présentant une identité de séquence supérieure à 90 % partagent presque systématiquement la même structure quaternaire, tandis que pour des homologues présentant 30 à 40 % d'identité de séquence, la structure quaternaire est conservée dans environ 70 % des cas (figure 7.2) (174).

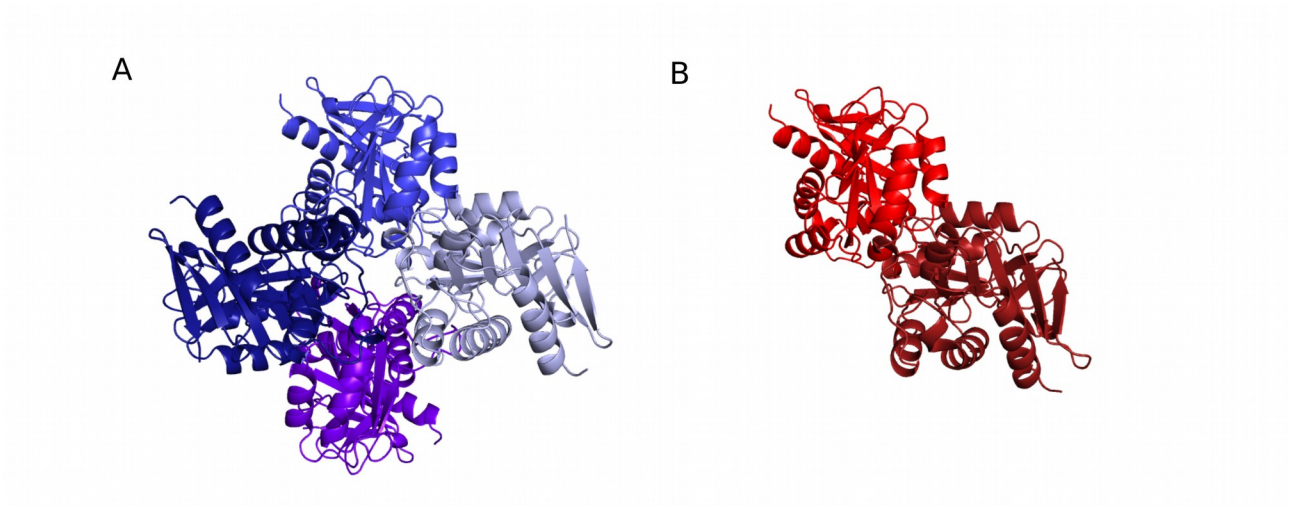


Figure 7.2. Exemple d'homomères homologues structuraux présentant un type de symétrie différent. (A) La D-Ribose-5-Phosphate Isomérase de *P. horikoshii* forme un tétramère (code pdb 1lk5). (B) La Ribose-5-Phosphate isomérase A de *H. influenza* forme un dimère (code pdb 1m0s).

7.1.3 Ordre d'assemblage des monomères

Levy et al (174) ont mis en évidence le fait que l'ordre d'assemblage des complexes homomériques lors du processus de formation du complexe reflète leur histoire évolutive, c'est-à-dire l'ordre des évènements qui ont eu lieu depuis l'existence de la protéine sous forme monomèrique jusqu'à l'existence de l'homomère sous sa forme actuelle – l'hypothèse étant que les homomères n'ont pas toujours existé sous cette forme et sont passés par des assemblages de plus en plus complexes au cours de l'évolution. Pour démontrer cela, les auteurs ont fait l'hypothèse que, lorsque la symétrie d'homomères homologues n'est pas conservée, les chemins reliant des géométries différentes sont les mêmes que les chemins évolutifs. Par exemple dans le cas d'un homomère C2 et de son homologue structural D2, le chemin géométrique est le suivant : deux monomères C1 vont s'assembler en un dimère C2 puis deux dimères C2 vont s'assembler un un tétramère D2. Dans le cas d'un homomère Dn (et donc composé de 2n sous-unités) il existe deux chemins possibles : celui-ci peut-être vu comme un assemblage de 2 complexes Cn ou de n complexes C2. Il existe donc plusieurs chemins évolutifs possibles pour parvenir à cette symétrie.

Pour vérifier cela, les auteurs ont commencé par quantifier le nombre de cas d'homomères homologues présentant des symétries différentes, et quels sont les types de symétrie de chacun des homologues. Ils ont créé un jeu de données constitués de 52 couples d'homomères homologues structuraux présentant des symétries différentes. Au sein de chacun de ces couples, un homologue arbore une symétrie cyclique et l'autre une symétrie diédrale (voir figure 7.2). Les auteurs ont fait l'hypothèse que, dans un couple d'homomères homologues, les interfaces en commun entre ces homologues auraient donc été présentes chez la protéine ancestrale à partir de laquelle les deux homologues ont évolué. J'appellerai ces interfaces, les interfaces « communes ». A l'opposé, les interfaces spécifiques d'une symétrie, que j'appellerai les interfaces « spécifiques », étaient donc absentes de la protéine ancestrale et seraient apparues après la divergence entre ces deux homologues. Nous pouvons donc raisonnablement émettre l'hypothèse que les interfaces « communes » étaient présentes chez l'ancêtre et sont donc plus anciennes que les interfaces « spécifiques » qui ne sont apparues que plus tardivement dans une symétrie donnée. A partir de ce jeu de données les auteurs ont montré que les interfaces « communes », c'est-à-dire les interfaces en commun entre les deux homomères homologues, sont plus grandes et donc probablement plus stables que les interfaces « spécifiques » dans 49 cas sur 52.

Les auteurs ont ensuite testé l'ordre d'assemblage des sous-unités de plusieurs complexes homomériques lors du processus de formation du complexe à l'aide de méthodes de spectrométrie

de masse. Ils ont pour cela étudié l'ordre de désassemblage de ces complexes (les auteurs ayant préalablement montré que l'ordre de désassemblage est l'inverse de l'ordre d'assemblage) en les soumettant à des conditions déstabilisantes. Ils ont montré que les interfaces les plus larges sont généralement les dernières à se désassembler, et donc par extension les premières à s'assembler.

Les interfaces les plus larges sont donc les premières à apparaître dans l'évolution de ces complexes et les premières à s'assembler dans l'ordre d'assemblage de ces complexes protéiques. En conséquence l'ordre d'assemblage et l'ordre d'apparition des complexes intermédiaires au cours de l'évolution sont les mêmes.

Le jeu de données qu'on établi les auteurs est donc particulièrement précieux car il fournit deux types de sites d'interactions avec des propriétés évolutives et physico-chimiques différentes : les sites d'interaction participant aux interfaces « communes » qui sont probablement plus anciens et dont les interfaces ont été montrées plus stables par spectrométrie de masse, et les sites d'interaction participant aux interfaces « spécifiques » qui sont probablement plus récents et dont les interfaces ont été montrées moins stables par spectrométrie de masse. Dans tous les cas, nous pouvons considérer que ces deux types de sites d'interaction constituent des sites d'interaction fonctionnels. Ce jeu de données constitue un jeu idéal pour exploiter le concept d'îlots rouges ubiquitaires et spécifiques et pour tenter de voir si les îlots rouges spécifiques correspondent à des cas de sites fonctionnels ayant émergé récemment.

J'ai donc appliqué la méthode que j'avais développée sur les 348 récepteurs extraits de la base de données de PPI4DOCK (173) afin de produire les cartes IPOPS pour chacune des protéines du jeu de données de complexes homomériques et extraire les îlots rouges « spécifiques » et « ubiquitaires ». J'ai ensuite comparé les recouvrements entre les régions identifiées par ces îlots avec les sites d'interaction « communs » et « spécifiques ». Puis, j'ai analysé les propriétés physico-chimiques et évolutives (conservation de séquence) des régions identifiées par les îlots rouges « ubiquitaires » et « spécifiques » ainsi que des sites d'interaction « communs » et « spécifiques » des protéines de la base de données.

7.2 Matériels et méthodes

7.2.1 Jeu de structures

Notre jeu de données est extrait du travail de Levy et al (174). Celui-ci est constitué de 33 paires de C2/D2, 13 paires C2/D3 et 6 paires C3/D3. J'ai supprimé les structures 1bj4, 1ixl, 1k9b et leurs homologues correspondants car celles-ci sont annotées comme de symétrie C1 dans la PDB, et ne possède donc pas d'interface dans leur unité biologique. J'ai aussi supprimé la paire 1uuy/1mkz car le premier complexe est annoté C2 et le second est annoté C3. Ils ne partagent pas d'interface en commun. J'ai enfin supprimé le complexe 1sru et son complexe homologue 1se8 car le premier présente un cas spécial de fusion de deux protéines, le dimère ancestral de 1sru étant structuralement superposable au monomère de 1se8. Par conséquent leurs sites respectifs de dimérisation ne sont pas comparables.

Notre jeu de données final est constitué de 32 paires de symétries C2/D2, 13 paires de symétries C2/D3 et 3 paires de symétries C3/D3, pour un total de 95 complexes protéiques (1ad3 étant à la fois homologue avec 1a4s et 1uzb et est donc présent dans deux couples d'homologues).

7.2.2 Définition des sites d'interaction communs et spécifiques

Au cours de ce travail nous avons défini comme sites d'interaction « communs » les sites d'interaction communs aux deux protéines homologues constituant une paire de symétrie (c'est-à-dire une paire de complexes homologues) et comme sites d'interaction « spécifiques », les sites d'interaction présents uniquement dans l'assemblage de plus haute complexité (à savoir, la plus grande stœchiométrie) pour une paire de complexes homologues, c'est-à-dire, l'assemblage diédral dans nos trois catégories C2/D2, C2/D3 et C3/D3.

7.2.3 Procédure de *docking*

J'ai réalisé un alignement structural avec le logiciel TM-align (188) pour chaque paire d'homomères afin de les placer dans le même référentiel, Ensuite, chaque monomère a été amarré avec ATTRACT contre lui même, contre son homologue et contre le jeu de 100 ligands arbitraires défini lors du travail précédent (section 6.2.1). J'ai ensuite produit les cinq types de cartes IPOPS

pour chaque monomère du jeu de données à partir des résultats de *docking* impliquant les 100 ligands arbitraires suivant le protocole décrit dans la section 5.4.5. Tout comme précédemment, j'ai ensuite extrait de chaque carte IPOPS rouge les îlots ubiquitaires et spécifiques. J'ai enfin produit pour chaque monomère, les cartes de sites d'interaction projetés en suivant la procédure décrite dans la section 5.4.2.

7.3 Résultats

7.3.1 Propriétés physico-chimique et évolutive des sites d'homomérisation

Dans un premier temps j'ai analysé les propriétés physico-chimiques des sites d'interaction ancestraux et secondaires. J'ai pour cela utilisé les même descripteurs des surfaces protéiques que dans la section 6.3.2.3, à savoir : *stickiness* (106), hydrophobicité de Kyte-Doolittle (192), variance circulaire (193) et conservation de séquence (200).

La figure 7.3 présente les boîtes à moustache de chacun de ces descripteurs en fonction des différentes catégories de résidus de la surface protéique : résidus appartenant aux sites d'interaction « communs », résidus appartenant aux sites d'interaction « spécifiques » et résidus n'appartenant à aucun des ces deux types de sites, et donc considérés comme résidus de surface. Bien entendu, il n'est pas exclu que ces résidus appartiennent à un autre site d'interaction non pris en considération dans le jeu de données. Ici, nous ne considérons que les sites d'interaction associés aux assemblages homomériques décrits dans le jeu de données. Nous observons (figure 7.3A) que les sites d'interaction « spécifiques » sont légèrement moins « collants » que les sites d'interaction « communs » (test DSH de Tukey (201), valeur p des différences observées entre chaque classe d'énergie < .001). De même la figure 7.3B montre que les sites d'interaction secondaires sont significativement moins hydrophobes que les sites d'interaction « communs » (valeur $p < .001$). De manière étonnante, la variance circulaire (figure 7.3C) n'est pas significativement différente entre la surface et les sites d'interaction « communs » (valeur $p = 0.96$). Enfin les sites d'interactions « communs » sont plus conservés évolutivement que les sites d'interaction « spécifiques », qui sont eux-même plus conservés que la surface (valeurs $p < .001$) (figure 7.3D).

Ces résultats montrent que les sites d'interaction « spécifiques » sont en moyennes moins hydrophobes et moins « collants » que les sites d'interaction « communs ». Ils doivent donc être en principe moins propices à l'interaction que ces derniers.

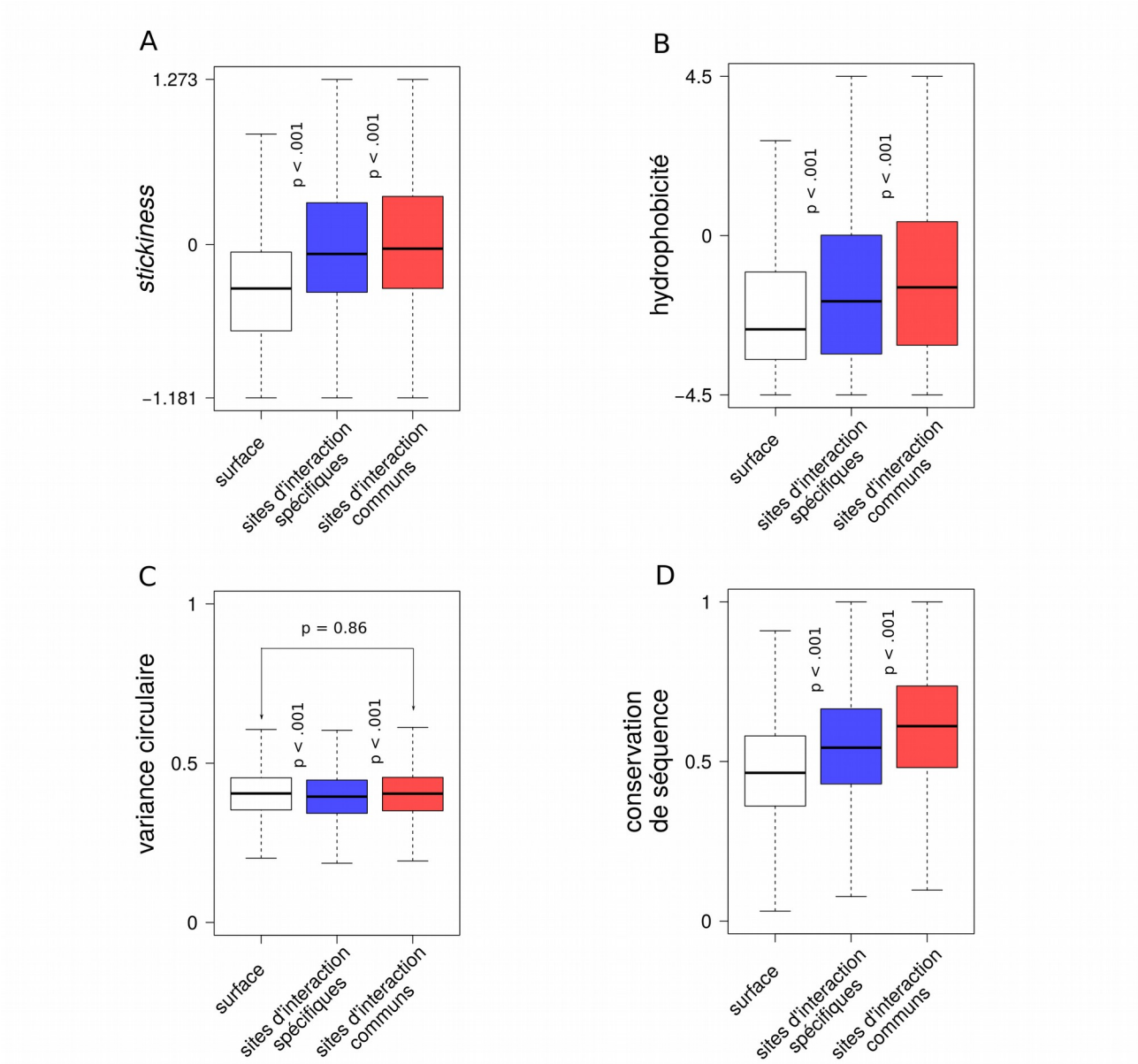


Figure 7.3. Propriétés physico-chimiques et évolutives des surfaces et des sites d'interaction communs et spécifiques. Les boîtes à moustache des propriétés des sites d'interaction communs aux homologues sont en bleu, Les boîtes à moustache des propriétés des sites d'interaction spécifiques à un homologue sont en rouge, Les boîtes à moustache des propriétés des surfaces sont en blanc. (A) *Stickiness*. (B) Hydrophobicité de Kyte-Doolittle. (C) Variance circulaire. (D) Conservation évolutive.

7.3.2 Tailles et nombres des îlots rouges ubiquitaires et spécifiques

Pour rappel, les îlots rouges ubiquitaires représentent des régions de la surface d'une protéine fortement attractives pour une majorité de protéines, que ces dernières soient des partenaires fonctionnels ou des protéines arbitraires. Les îlots rouges spécifiques ne représentent des régions favorables à l'interaction que pour une minorité de ligands.

La figure 7.4A représente la distribution du nombre d'îlots rouges ubiquitaires et rouges spécifiques identifiés sur les cartes IPOPS rouges de chaque monomère. Nous observons que tous les monomères excepté un, présente au moins un îlot ubiquitaire tandis qu'une majorité ne présente pas d'îlot spécifique. En moyenne les monomères présentent 1.3 îlots rouges ubiquitaires et 0.63 îlots rouges spécifiques. Nous pouvons nous demander dans quelle mesure les îlots rouges ubiquitaires correspondent justement aux sites d'interaction « communs » et les îlots rouges spécifiques aux sites d'interaction « spécifiques ».

En plus d'être en moyenne moins nombreux que les îlots rouges ubiquitaires, les îlots rouges spécifiques sont aussi en moyenne significativement plus petits. En effet les îlots rouges ubiquitaires couvrent en moyenne 70 cellules, tandis que les îlots rouges spécifiques couvrent en moyenne 15 cellules sur la carte figure 7.4B.

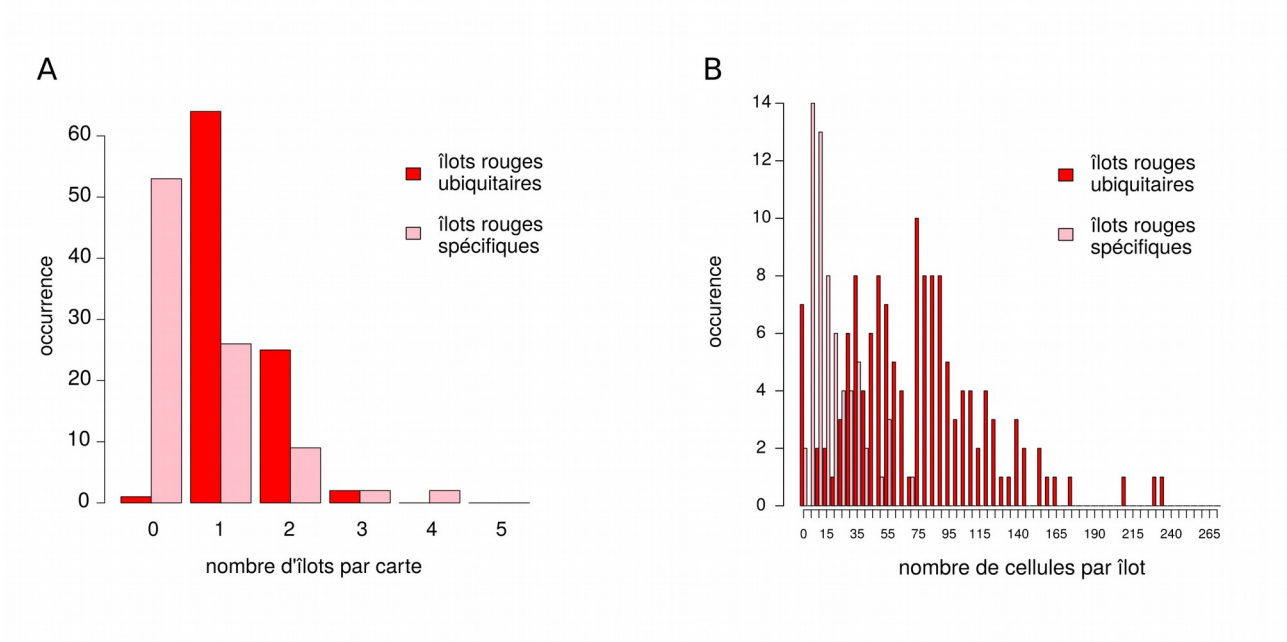


Figure 7.4. Distribution du nombre d'îlots par cartes et distribution de la taille des îlots.

7.3.3 Propriétés physico-chimiques et évolutive des îlots extraits des cartes IPOPS

Tout comme dans la section 6.3.2.3, j'ai analysé les propriétés physico-chimiques et évolutives des cinq types de régions de surface protéiques identifiées à partir des cartes IPOPS correspondantes (bleue, verte foncée, verte claire, jaune et rouge) afin de voir si la tendance observée chez les 348 récepteurs étudiés précédemment étaient toujours vraie dans le cas particulier des protéines formant des assemblages homomériques. J'ai calculé comme précédemment plusieurs descripteurs : *stickiness*, hydrophobicité de Kyte-Doolittle, variance circulaire et conservation de séquence.

La figure 7.5 représente les boîtes à moustache des valeurs de *stickiness*, d'hydrophobicité de Kyte-Doolittle, de variance circulaire et de la conservation évolutive calculée avec le logiciel JET (200) pour chaque classe d'énergie. De manière similaire aux observations réalisées dans les sections 6.3.2.3 et 6.3.3.2, la *stickiness*, l'hydrophobicité et la variance circulaire diminuent de manière significative avec les niveaux d'énergie de chaque classe (test DSH de Tukey (201), valeurs p des différences observées entre chaque classe d'énergie < .001) (figure 7.5A-C). En effet, les régions de plus basse énergie (en rouge) sont les plus "collantes", les plus hydrophobes et les plus planes tandis que les régions de haute énergie (en bleu) sont les moins « collantes », les moins hydrophobes et les plus protubérantes. Nous observons aussi la même différence de comportement entre les régions indiquées par les îlots rouges ubiquitaires et les îlots rouges spécifiques. En effet, les régions correspondant aux îlots spécifiques semblent avoir un comportement plus similaire aux régions jaunes qu'aux régions rouges ubiquitaires en étant moins hydrophobes (test DSH de Tukey, valeur p < .001), moins « collantes » (valeur p < .001) et moins conservées que ces dernières (test DSH de Tukey, valeur p < .001) (figure 7.5A,B et D). De plus leurs propriétés de *stickiness* ne sont pas significativement différentes de celles des régions jaunes (test DSH de Tukey, valeur p = 0.58, figure 7.5A). De manière étonnantes les îlots rouges spécifiques présentent une variance circulaire significativement plus élevée que les îlots rouges ubiquitaires (test DSH de Tukey, valeur p < .001, figure 7.5C). Ces derniers sont donc localisés dans des régions moins protubérantes de la surface des protéines.

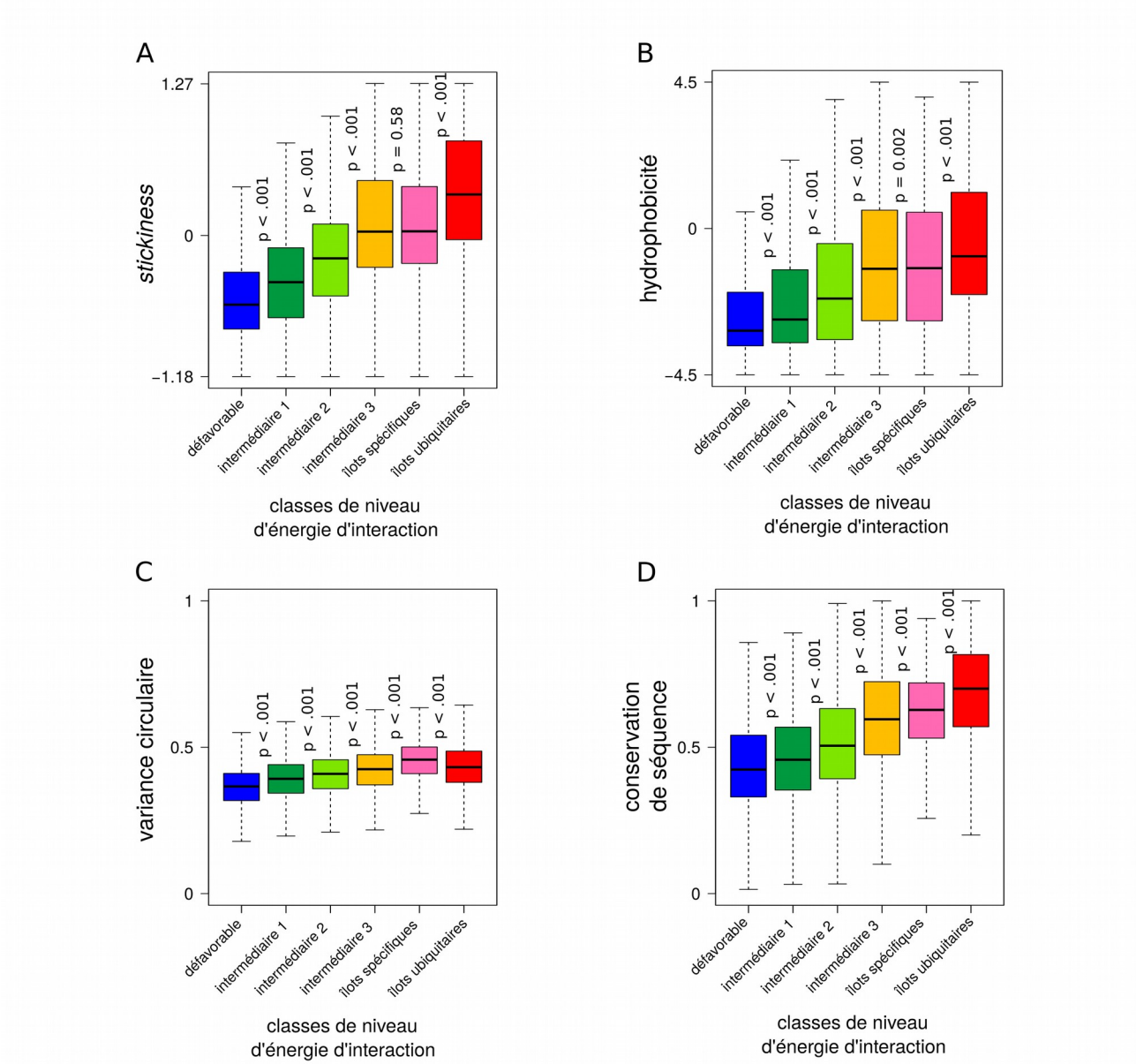


Figure 7.5. Boîtes à moustache des valeurs de variance circulaire, d'hydrophobicité, de *stickiness* et de conservation de séquence en fonction des classes d'énergie. Les valeurs des différents descripteurs utilisés sont calculées selon la méthodologie présentée dans la section 5.3. (A) *stickiness*. (B) hydrophobicité de Kyte-Doolittle. (C) variance circulaire. (D) conservation de séquence.

7.3.4 Les sites d’homomérisation présentent un fort recouvrement avec les îlots rouges ubiquitaires

7.3.4.1 Sites d’interaction chez les C2 et C3

J’ai ensuite comparé les résidus appartenant aux îlots des cartes IPOPS rouges avec ceux des sites d’interaction des homomères C2 et C3 (tableau 7.1). Le recouvrement obtenu est particulièrement élevé : 0.72 en considérant les résidus appartenant aux îlots rouges ubiquitaires et spécifiques, 0.75 en considérant ceux des îlots rouges ubiquitaires uniquement et 0.33 en considérant ceux des îlots rouges spécifiques uniquement. Tout d’abord, nous observons que les taux de recouvrement entre les îlots rouges et les sites d’interaction des homomères sont beaucoup plus élevés que ceux obtenus avec le jeu de données des 348 récepteurs extraits de PPI4DOCK (respectivement 0.53, 0.55 et 0.30). Cela semble montrer que les sites d’homomérisation de ces protéines présentent des propriétés très marquées qui les distinguent du reste de la surface, en particulier ces sites semblent avoir des propriétés les rendant très favorables à l’interaction avec un grand nombre de partenaires. Par ailleurs, il est très intéressant de voir que les régions correspondant aux îlots rouges ubiquitaires ont un très fort taux de recouvrement avec les sites d’homomérisation, ce qui est beaucoup moins le cas pour les îlots spécifiques.

Tableau 7.1. Recouvrement entre les îlots rouges et les sites d’homomérisation pour les homomères cycliques (voir section 5.4.6 pour la méthodologie)

	Sites d’interaction communs (homomères cycliques)
Tout îlots	0.72
Îlots rouges ubiquitaires	0.75
Îlots rouges spécifiques	0.33

7.3.4.2 Sites d'interaction chez les D2 et D3

J'ai ensuite comparé les résidus appartenant aux îlots des cartes IPOPS rouges avec les résidus appartenant aux sites d'interaction des homomères D2 et D3 (tableau 7.2). Pour rappel les complexes arborant une symétrie de type diédrale possèdent deux types d'interfaces : les interfaces « communes » qu'ils ont en commun avec leurs homologues de symétrie cyclique, et les interfaces « spécifiques » que leurs homologues respectifs ne possèdent pas.

En considérant les résidus des sites d'interaction « communs » et « spécifiques », le recouvrement calculé est très élevé : 0.83 pour les résidus identifiés par les îlots rouges ubiquitaires et spécifiques, 0.84 en ne considérant que les résidus des îlots rouges ubiquitaires et 0.71 en considérant uniquement les résidus des îlots rouges spécifiques. Ce résultat montre que la grande majorité des îlots rouges ubiquitaires et spécifiques identifient des résidus appartenant à des sites d'homomérisation.

J'ai ensuite étudié plus en détail si les sites d'interaction « communs » étaient mieux recouverts par les îlots rouges que les sites d'interaction « spécifiques » et si il y avait une correspondance entre les résidus identifiés par les îlots rouges ubiquitaires et ceux des sites d'interaction « communs » et les résidus des îlots spécifiques et les sites d'interactions « spécifiques ». Le tableau 7.2 présente ces recouvrements calculés sur les complexes D2 et D3. Nous observons un recouvrement de 0.58 entre les résidus appartenant aux îlots rouges ubiquitaires et les résidus appartenant aux sites d'interaction « communs », et un recouvrement de 0.27 entre les résidus des îlots rouges spécifiques et les résidus des sites d'interaction « communs ». Cela montre que les îlots rouges ubiquitaires correspondent en majorité aux sites d'interaction « communs », tandis que les résidus des îlots spécifiques y correspondent dans une moindre mesure. En revanche, nous observons un recouvrement de 0.25 entre les résidus appartenant aux îlots rouges ubiquitaires et les résidus appartenant aux sites d'interaction « communs » et un recouvrement de 0.44 entre les résidus des îlots rouges spécifiques et les résidus des sites d'interaction « spécifiques ».

Ces résultats sont particulièrement intéressants car ils montrent que (i) les îlots rouges spécifiques ont un recouvrement plus important avec les résidus des sites d'interaction « spécifiques » qu'avec les résidus des sites d'interaction « communs » et (ii) le recouvrement des îlots rouges ubiquitaires avec les résidus des sites d'interaction « communs » est deux fois plus important qu'avec ceux des sites d'interaction « spécifiques ». Cela montre que les premiers sont moins enclins à l'interaction avec un grand nombre de partenaires que les premiers.

Tableau 7.2. Recouvrement entre les îlots rouges et les sites d'homomérisation pour les homomères diédraux (voir section 5.4.6 pour la méthodologie)

	Sites d'interaction ancestraux (homomères diédraux)	Sites d'interaction secondaires (homomères diédraux)	Tous sites d'interaction (homomères diédraux)
Tout îlots	0.56	0.27	0.83
Îlots rouges ubiquitaires	0.58	0.25	0.84
Îlots rouges spécfiques	0.27	0.44	0.71

7.3.5 Exemple de paire d'homomères C2/D2: cas des superoxyde dismutases à manganèse

Les enzymes superoxyde dismutases à manganèse (MnSOD) sont des métalloprotéines connues pour leur activité d'élimination des dérivés réactifs de l'oxygène (DRO), toxiques pour la cellule (217). Nous avons dans notre jeu de données deux complexes homologues de cette famille de protéines (figure 7.6). Le premier (pdb 1gv3) correspond à un homodimère de type C2 provenant de la cyanobactérie *Nostoc sp.* (figure 7.6B). Le second (pdb 3mds) constitue un homotétramère de type D2 que l'on trouve chez la bactérie gram négative *Thermus thermophilus* (figure 7.6C). Les deux monomères partagent 58% d'identité de séquence et sont parfaitement superposables (RMSD entre les deux monomères de 1Å, figure 7.6A).

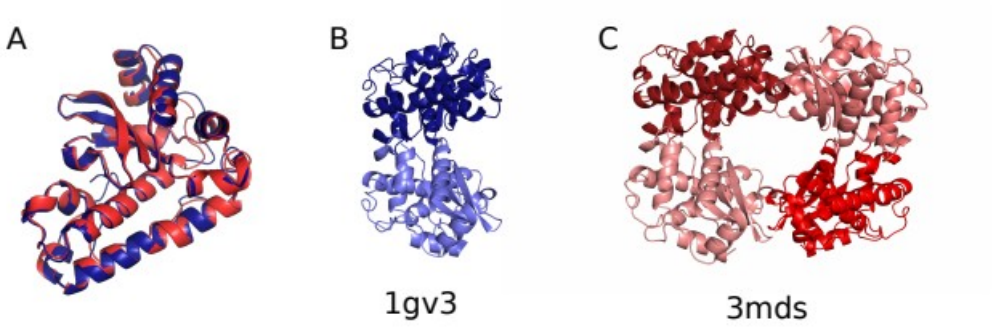


Figure 7.6. 1gv3_A et 3mds_A sont des homologues structuraux avec une structure quaternaire différente. (A) Les structures 1gv3_A (en bleu) et 3mds (en rouge) sont superposables et présentent un RMSD de 1Å. (B) 1gv3 forme un dimère. (C) 3mds forme un tétramère.

Il s'agit d'un cas intéressant qui m'a permis de réaliser l'expérience suivante (figure 7.7) : j'ai amarré chacune des protéines avec elle-même, puis j'ai amarré chacune des protéines avec son homologue. En réalisant cette expérience je voulais répondre aux questions suivantes :

1- Le *docking* de 1gv3 (C2) avec lui-même permet-il de retrouver son site d'interaction « commun » ?

2- Le *docking* de 3mds (D2) avec lui-même permet-il de retrouver son site d'interaction « commun », son site d'interaction « spécifique » ou les deux à la fois ?

3- Le *docking* de 3mds en tant que récepteur avec 1gv3 permet-il de retrouver son site d'interaction « commun », son site d'interaction « spécifique » ou les deux à la fois ?

4- Le *docking* de 1gv3 en tant que récepteur avec 3mds permet-il de retrouver son site d'interaction « commun », le site d'interaction « spécifique » de 3mds, ou les deux à la fois ?

La figure 7.7 présente les cartes de sites d'interaction des deux protéines ainsi que les cartes d'énergie issues des calculs de *docking* entre chaque monomère et lui-même et des calculs de *docking* croisés entre chaque monomère et son homologue. Tout d'abord, nous observons que le complexe 1gv3 étant de symétrie C2, ses sous-unités possèdent une unique site d'interaction, tandis que 3mds étant de symétrie D2, ses sous-unités présentent deux sites d'interaction (figure 7.7A-B). Nous observons que la carte d'énergie issue du *docking* de 1gv3_A avec lui-même présente une région rouge localisée au niveau du site de dimérisation (figure 7.7C), c'est-à-dire, du site d'interaction « commun ». La carte d'énergie issue du *docking* de 3mds_A présente quant à elle deux régions rouges : une localisée au niveau du site d'interaction « commun » de la protéine, et une localisée au niveau du site d'interaction « spécifique » de la protéine (figure 7.7C). De façon intéressante, nous pouvons voir que lorsque 3mds_A est amarré avec le récepteur 1gv3_A, la carte d'énergie ne présente qu'une seule région rouge correspondant au site d'interaction « commun » (figure 7.7D). Ainsi ce site est favorable à l'interaction avec l'homologue de 3mds_A mais la région correspondant au site d'interaction « spécifique » chez l'homologue 3mds_A ne présente pas de fort potentiel d'interaction (figure 7.7D). Autrement dit, 1gv3_A ne semble pas avoir la capacité d'interagir au niveau de cette région avec 3mds_A et ne semble pas capable de tétramériser. Au contraire, lorsqu'on amarre 1gv3_A sur 3mds_A, nous pouvons observer deux régions rouges sur la carte d'énergie correspondant respectivement aux sites d'interaction « communs » et « spécifiques ». Cela semble montrer que 3mds_A présente deux régions qui peuvent potentiellement interagir avec 1gv3_A.

J’ai ensuite calculé les cartes IPOPS rouges des deux monomères présentées (figure 7.7E). Sans surprise, la carte IPOPS rouge calculée pour 1gv3_A (symétrie C2) ne présente qu’un unique îlot rouge ubiquitaire localisé au niveau du site d’interaction « commun ». En revanche, la carte IPOPS rouge de 3mds (symétrie D2) présente un îlot rouge ubiquitaire localisé au niveau du site d’interaction « commun » et un îlot rouge spécifique au niveau du site d’interaction « spécifique ». Ce résultat montre que le site d’interaction « spécifique » est moins propice à interagir avec un grand nombre de ligands que le site « commun », autrement dit, qu’il présente un potentiel d’interaction plus faible que ce dernier. Ce résultat apporte par ailleurs un éclairage sur les propriétés des îlots rouges ubiquitaires et spécifiques. Ici, l’îlot rouge spécifique correspond au site d’interaction le plus récent (à savoir le site d’interaction « spécifique » du D2 et absent du C2) ce qui peut expliquer que ce dernier présente des propriétés physico-chimiques et évolutives différentes des îlots rouges ubiquitaires.

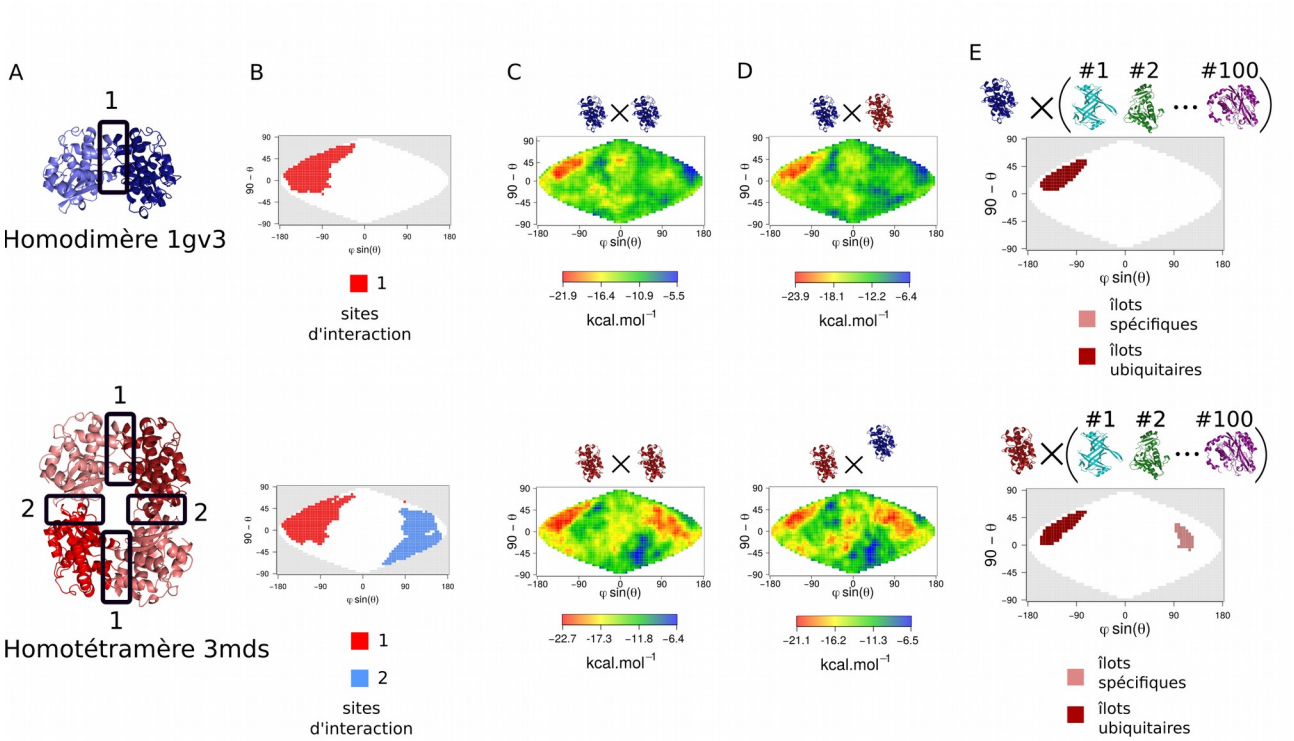


Figure 7.7. Expérience de *docking* sur la paire d’homologues 1gv3/3mds. (A) 1gv3 (en bleu) forme un homodimère tandis que 3mds (en rouge) forme un homotétramère à travers une interface supplémentaire. (B) Cartes des sites d’interactions projetés de 1gv3 et 3mds. Le site projeté commun au deux protéines est coloré en rouge, le site spécifique en bleu. (C) Cartes d’énergie résultant du *docking* de 1gv3 et de 3mds avec eux-mêmes. (D) Cartes d’énergie résultant du *docking* de 1gv3 avec 3mds et de 3mds avec 1gv3. (E) cartes IPOPS rouges de 1gv3 et 3mds.

7.4 Discussion

Les résultats obtenus sont très intéressants. Nous montrons que les sites d'interaction « communs » sont surreprésentés dans les îlots rouges ubiquitaires, et les sites d'interaction « spécifiques » sont surreprésentés dans les îlots rouges spécifiques. Ainsi, cette étude a permis d'apporter une explication sur les propriétés physico-chimiques et évolutives observées pour les îlots rouges ubiquitaires et spécifiques. En effet, nous avons montré que les îlots rouges spécifiques étaient moins hydrophobes, moins « collants » et moins conservés que les îlots ubiquitaires et correspondaient principalement aux sites d'interaction « spécifiques ». Ces derniers ont émergé plus récemment que les sites d'interaction « communs ». Cela va dans le sens de l'hypothèse émise précédemment que les régions identifiées par les îlots spécifiques correspondent à deux populations : une population de régions de surface avec un fort potentiel à interagir de façon non-fonctionnelle avec un sous-ensemble de protéines de l'environnement, et une population de régions impliquées dans des interactions fonctionnelles mais dont la capacité à interagir a émergé plus récemment (rappelons le cas de 1gv3_A qui n'est pas capable d'interagir via la région homologue au site d'interaction « spécifique » à 3mds_A) et qui présentent des propriétés plus proches des régions que l'on pourrait appeler « non-interagissantes » ou à potentiel d'interaction plus faible, telles que les régions intermédiaires « jaunes ». L'expérience de spectrométrie de masse réalisée dans (174) montre par ailleurs que ces régions semblent impliquées dans des assemblages moins stables que les régions identifiées par les îlots rouges ubiquitaires. Par ailleurs, cette étude a permis de caractériser du point de vue des propriétés physico-chimiques et évolutives mais surtout d'un point de vue énergétique les propriétés de ces deux types de sites d'interaction « communs » et « spécifiques ».

Cependant pour pouvoir finaliser notre travail, il faudrait refaire les calculs de *docking* en utilisant également les structures des dimères (ou trimères pour les paires d'homologues de symétrie C3/D3) pour évaluer leur capacité à tétramériser dans le cas des C2/D2 par exemple. En effet, dans certains cas, nous n'arrivons pas à identifier le site d'interaction « spécifique » au complexe diédral, probablement car ce site nécessite l'interaction avec les deux (voir trois pour les complexes de types D3) chaînes du complexe cyclique (le complexe diédral résulte de la dimérisation de deux dimères cycliques). Pour cela il faudrait réaliser les calculs de *docking* sur les formes dimérisées (ou trimérisée pour les cas C3/D3) afin de voir si nous arrivons à identifier les sites d'interaction « spécifiques » à partir des formes dimériques.

De plus je pense que le seuil utilisé pour filtrer le bruit des cartes IPOPS est trop strict. En effet le fait de considérer comme bruit de fond toutes les intensités de cellules inférieures à 22 (voir sections 5.4.5 et 6.3.3) a pour effet de supprimer la majorité des îlots rouges spécifiques. Ainsi, d'une part ces derniers sont représentés de façon marginale dans les cartes IPOPS, mais de plus quand ils sont présents, ceux-ci sont généralement de très petite taille. Il serait intéressant de recalculer ces cartes avec un seuil d'élimination du bruit de fond plus bas afin de voir si cela permet de rajouter de nouveaux îlots rouges spécifiques et/ou d'agrandir les contours des îlots rouges spécifiques existants.

Enfin nous pensons que notre méthode pourrait se révéler utile pour l'assignation de structure quaternaire de complexes homomériques. Récemment Yueh et al (218) ont réalisé une méthode permettant de discriminer entre les dimères biologiques et cristallographiques (dont l'interface est formée par un contact cristallin et pas une interface biologique) à partir de calculs de *docking*. Pour tester la capacité à se dimériser d'une protéine, ils réalisent le *docking* de celle-ci avec elle-même avec le logiciel de *docking* ClusPRO (131). Ils groupent ensuite les meilleures solutions. Si celles-ci sont localisées au niveau du site de dimérisation observé, alors la protéine est prédite comme un dimère biologique (en opposition à cristallographique), sinon les auteurs considèrent que la protéine ne forme pas de dimère en conditions natives. Je pense qu'il serait très intéressant de comparer notre approche avec celle de cette équipe. Nos cartes IPOPS reposent sur un concept différent, puisque que nous réalisons le *docking* de ligands arbitraires avec la protéine d'intérêt. Cependant nous avons montré que les îlots rouges de nos cartes correspondent en grande partie à des sites d'homomérisation.

En conclusion ce travail est très prometteur mais n'est pas encore achevé. Il reste encore de nombreuses pistes à explorer mais les résultats obtenus jusqu'à présent montrent que nos cartes IPOPS, et plus précisément les cartes IPOPS rouges présentent de nombreuses applications possibles. Nous allons maintenant voir dans le prochain chapitre comment l'exploitation des cartes IPOPS des autres couleurs, autrement dit, des cartes IPOPS correspondant aux régions dîtes non-interagissantes ou de plus faible potentiel d'interaction peuvent apporter une information importante sur les contraintes évolutives qui s'exercent sur l'ensemble de la surface des protéines.

8 Le paysage énergétique d'interaction des protéines est modelé par les partenaires fonctionnels ainsi que les partenaires non-fonctionnels

Il s'agit du dernier travail présenté dans cette thèse. Il faut garder à l'esprit en lisant la partie qui suit que, bien qu'il s'agisse du dernier travail présenté dans ce manuscrit, chronologiquement il s'agit en fait du premier. C'est pour un souci de clarté du manuscrit de thèse que cette partie a été placée en dernier. Mais c'est en réalisant cette étude que nous avons eu l'idée de réaliser les cartes IPOPS présentées dans les deux travaux précédents (voir sections 6 et 7).

Ce travail a consisté à comparer les paysages énergétiques d'interaction induits par des ligands homologues structuraux et des ligands non homologues, dans le but de déterminer si les cartes d'énergie induites par les premiers sont plus similaires entre elles que celles de ligands n'ayant aucune relation d'homologie. Pour ce faire avions alors été amenés à réaliser un calcul de *docking* croisé complet (calcul de *docking* entre toutes les combinaisons possibles de protéines du jeu de données). Du fait de notre méthodologie chaque récepteur était donc amarré successivement avec les 74 protéines du jeu de données. Nous avions alors constaté que les régions rouges (favorables à l'interaction) des cartes d'énergie étaient souvent localisées au même endroit quelques soient les ligands amarrés. C'est en voulant comparer la localisation de ces différentes régions à travers l'ensemble du jeu de ligands que nous avions eu l'idée de réaliser les cartes IPOPS présentées dans les deux chapitres précédents. Cependant nous n'avions pas encore fait tous les développements présentés dans les sections 6 et 7. C'est pour cela que ces cartes ne sont pas présentées ni appelées de la même manière dans le travail qui suit.

Title

Protein interaction energy landscapes are shaped by functional and also non-functional partners

Authors

H. Schweke[a], MH. Mucchielli[ab], S. Sacquin-Mora[c], W. Bei[a], A. Lopes[a]

Abstract

In the crowded cell, functional and non-functional interactions undergo severe competition. Understanding how a protein binds the right piece in the right way in this complex jigsaw puzzle is crucial and difficult to address experimentally. To interrogate how this competition constrains the behavior of proteins with respect to their partners or random encounters, we (i) performed thousands of docking simulations to systematically characterize the interaction energy landscapes of functional and non-functional protein pairs and (ii) developed an original theoretical framework based on two-dimensional energy maps that reflect the propensity of a protein surface to interact. Strikingly, we show that the interaction propensity of not only binding sites but also of the rest of protein surfaces is conserved for homologous partners, and this feature holds for both functional and non-functional partners. We highlight a new role for non-interacting regions in preventing non-functional interactions and guiding the interaction process toward functional interactions.

8.1 Introduction

Biomolecular interactions are central for many physiological processes and are of utmost importance for the functioning of the cell. Particularly protein-protein interactions have attracted a wealth of studies these last decades [1–5]. The concentration of proteins in a cell has been estimated to be approximately 2-4 million proteins per cubic micron [6]. In such a highly crowded environment, proteins constantly encounter each other and numerous non-specific interactions are likely to occur [7,8]. For example, in the cytosol of *S. cerevisiae* a protein can encounter no less than 2000 different proteins [9]. In this complex jigsaw puzzle, each protein has evolved to bind the right piece in the right way (positive design) and to prevent misassembly and non-functional interactions (negative design) [10–14]).

Consequently, positive design constrains the physico-chemical properties and the evolution of protein-protein interfaces. Indeed, a strong selection pressure operates on binding sites to maintain the functional assembly. For example, homologs sharing at least 30% sequence identity almost invariably interact in the same way [15]. Conversely, negative design prevents proteins to be trapped in the numerous competing non-functional interactions inherent to the crowded environment of the cell. Particularly, the misinteraction avoidance shapes the evolution and physico-chemical properties of abundant proteins, resulting in a slower evolution and less sticky surfaces than what is observed for less abundant ones [16–21]. The whole surface of abundant proteins is thus constrained, preventing them to engage deleterious non-specific interactions that could be of dramatic impact for the cell at high concentration [20]. Recently, it has been shown in *E. coli* that the net charge as well as the charge distribution on protein surfaces affect the diffusion coefficients of proteins in the cytoplasm [22]. Positively charged proteins move up to 100 times more slowly as they get caught in non-specific interactions with ribosomes which are negatively charged and therefore, shape the composition of the cytoplasmic proteome [22].

All these studies show that both positive and negative design effectively operate on the whole protein surface. Binding sites are constrained to maintain functional assemblies (i.e. functional binding modes and functional partners) while the rest of the surface is constrained to avoid non-functional assemblies. Consequently, these constraints should shape the energy landscapes of functional but also non-functional interactions so that non-functional interactions do not prevail over functional ones. This should have consequences (i) on the evolution of the propensity of a protein to interact with its environment (including functional and non-functional partners) and (ii)

on the evolution of the interaction propensity of the whole surface of proteins, non-interacting surfaces being in constant competition with functional binding sites. We can hypothesize that the interaction propensity of the whole surface of proteins is constrained during evolution in order to (i) ensure that proteins correctly bind functional partners**,** and (ii) limit non-functional assemblies as well as interactions with non-functional partners.

In this work, we focus on protein surfaces as a proxy for functional and non-functional protein-protein interactions. We investigate their interaction energy landscapes with native and non-native partners and ask whether their interaction propensity is conserved during evolution. With this aim in mind, we performed large-scale docking simulations to characterize interactions involving either native and/or native-related (i.e. partners of their homologs) partners or arbitrary partners. Docking simulations enable the characterization of all possible interactions involving either functional or arbitrary partners, and thus to simulate the interaction of arbitrary partners which is very difficult to address with experimental approaches. Docking algorithms are now fast enough for large-scale applications and allow for the characterization of interaction energy landscapes for thousand of protein couples. Typically, a docking simulation takes from a few minutes to a couple of hours on modern processors [23–25], opening the way for extensive cross-docking experiments [26–29]. Protein docking enables the exploration of the interaction propensity of the whole protein surface by simulating alternative binding modes. Here, we performed a cross-docking experiment involving 74 selected proteins docked with their native-related partners and their corresponding homologs, as well as arbitrary partners and their corresponding homologs. We represented the interaction energy landscape resulting from each docking calculation with a two dimensional (2D) energy map in order to (i) characterize the propensity of all surface regions of a protein to interact with a given partner (either native-related or not) and (ii) easily compare the energy maps resulting from the docking of a same protein with different homologous partners, thus addressing the evolution of the propensity of the whole protein surface to interact with homology-related partners either native or arbitrary.

8.2 Results

The interaction propensity of a protein to interact either with native-related or arbitrary partners is conserved during evolution

We ask whether the interaction propensity of a protein surface is conserved for homologous native-related partners**,** and whether this remains true for homologous arbitrary partners. For a protein A,

we refer as native-related partners its native partner (when its three dimensional (3D) structure is available) and native partners of proteins that are homologous to the protein A. Arbitrary pairs refer to pairs of proteins for which no interaction between them or their respective homologs has been experimentally characterized in the Protein Data Bank [30]. To test the aforementioned hypothesis, we assembled a dataset comprising 74 protein structures divided into 12 families of homologs (S1 Table and *Materials and Methods*). Each family displays different degrees of structural variability and sequence divergence in order to see the impact of these properties on the conservation of the interaction propensity inside a protein family. Each family has at least a native-related partner family (S1 Fig). Docking calculations were performed with the ATTRACT software [25]. ATTRACT enables a homogeneous and exhaustive conformational sampling and is well suited to investigate the propensity of the whole surface of a protein to interact with a given ligand. Our procedure is asymmetrical since we aim at characterizing the interaction propensity of a protein (namely the receptor) with a subset of proteins (namely the ligands). Therefore, a given receptor is docked with a subset of ligands (here the 74 proteins of the dataset) (Fig 1*A* and *Materials and Methods*). For each docking calculation, we produced a 2D energy map, which provides the distribution of interaction energies of all docking solutions over the whole receptor surface (Fig 1*B* and *Materials and Methods*, Fig 2*A*-*C*). The resulting energy map reflects the propensity of the whole surface of the receptor to interact with the docked ligand. One should notice that energy maps computed for two unrelated receptors are not comparable since their surfaces are not comparable. Therefore, the procedure is ligand-centered and allows only the comparison of energy maps produced by different ligands docked with the same receptor. The comparison of two energy maps enables the evaluation of the similarity of the interaction propensity of the receptor with the two corresponding ligands. In order to investigate the interaction propensity of all proteins of the dataset, each protein plays alternately the role of receptor and ligand. Consequently, the procedure presented in Fig 1 is repeated for the whole dataset where each protein plays the role of the receptor and is docked with the 74 proteins that play the role of ligands.

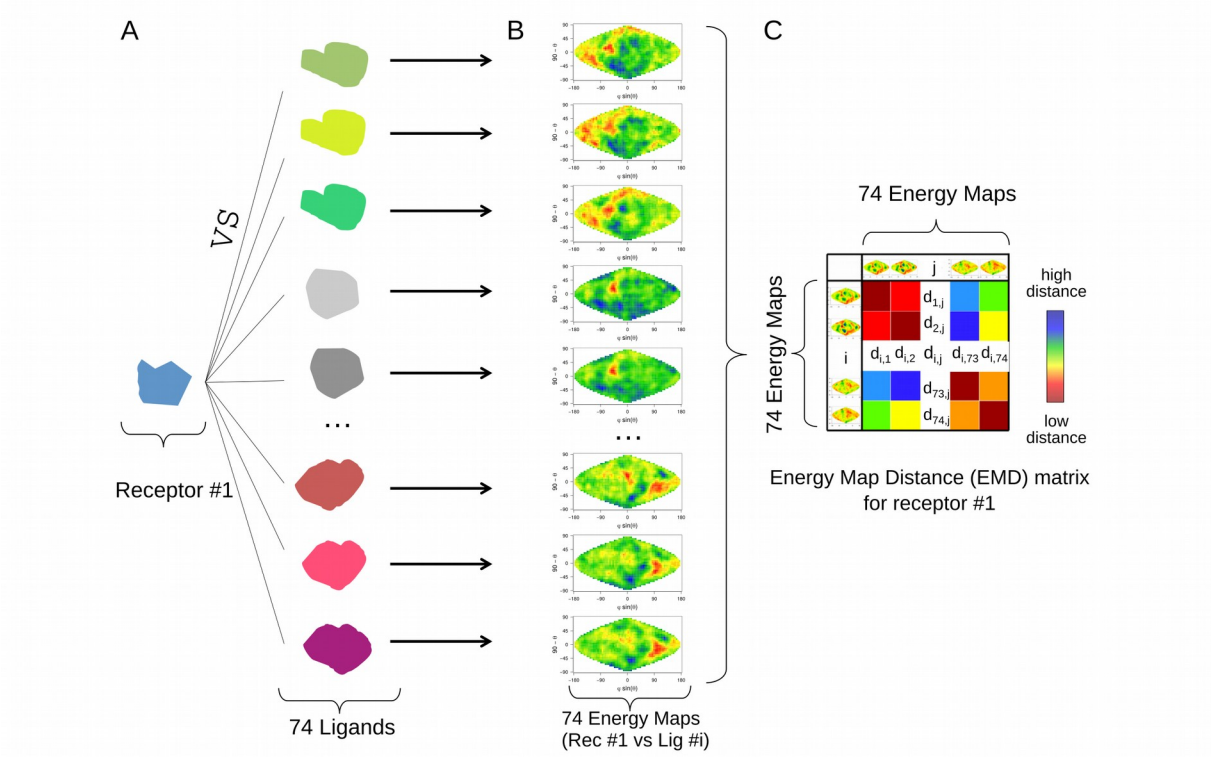


Fig 1. Experimental Protocol. (*A*) A receptor protein is docked with all proteins of the dataset (namely the ligands) resulting in 74 docking calculations. (*B*) For each docking calculation, an energy map is computed as well as its corresponding five-color and one-color energy maps, with the procedure described in Fig 2 and *Materials and Methods*. (*C*) An energy map distance (EMD) matrix is computed, representing the pairwise distances between the 74 energy maps resulting from the docking of all ligands with this receptor. Each cell (*i*,*j*) of the matrix represents the Manhattan distance between the two energy maps resulting from the docking of ligands *i* and *j* with the receptor. A small distance indicates that the ligands *i* and *j* produce similar energy maps when docked with this receptor. In other words, it reflects that the interaction propensity of this receptor is similar for these two ligands. To prevent any bias from the choice of the receptor, the whole procedure is repeated for each receptor of the database, leading to 74 EMD matrices.

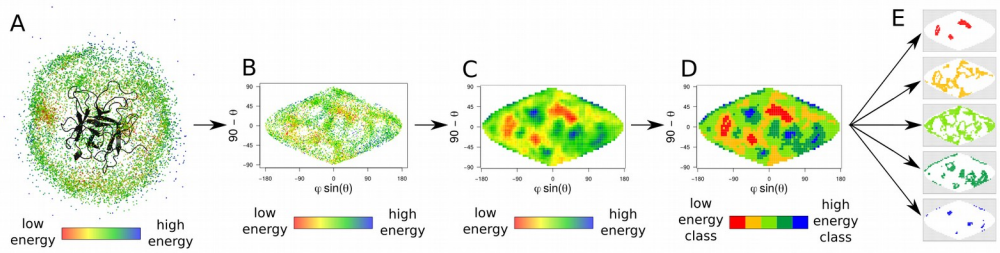


Fig 2. 2D asymmetrical representation of docking energy landscapes and resulting energy maps. (*A*) Three-dimensional (3D) representation of the ligand docking poses around the receptor. Each dot corresponds to the center of mass (CM) of a ligand docking pose. It is colored according to its docking energy score. (*B*) Representation of the CM of the ligand docking poses after an equal-area 2D sinusoidal projection. CMs are colored according to the same scale as in A. (*C*) Continuous energy map (see *Materials and Methods* for more details). (*D*) Five-color map. The energy map is discretized into five energy classes (*E*) One-color maps. Top to bottom: red, orange, green, dark green and blue maps highlight respectively hot, warm, lukewarm, cool and cold regions.

Fig 3*A* represents the energy maps computed for the receptor 2AYN_A, the human ubiquitin carboxyl-terminal hydrolase 14 (family UCH) docked with (i) its native partner (1XD3_B, ubiquitin-related family), a homolog of its partner (1NDD_B) and (ii) two arbitrary homologous ligands (1YVB_A and 1NQD_B from the papain-like family). For all four ligands, either native-related or arbitrary partners, docking calculations lead to an accumulation of low-energy solutions (hot regions in red) around the two experimentally known binding sites of the receptor. The first one corresponds to the interaction site with the native partner, ubiquitin (pdb id 2ayo). The second one corresponds to its homodimerisation site (pdb id 2ayn). This indicates that native-related but also arbitrary partners tend to bind onto the native binding sites of native partners as observed in earlier studies [29,31]. The same tendency is observed for all 74 ligands in the database (Fig 3*B*). Their 20 best docking poses systematically tend to accumulate in the vicinity of the two native interaction sites. Whereas the low-energy solutions for most ligands accumulate around the same interaction sites (i.e. the native binding sites), we observe that, globally, 2-D energy maps (i) seem to be more similar between ligands of a same family than between ligands belonging to different families (Fig 3*A*). The two energy maps obtained with the ligands of the native-related partners family both reveal two sharp hot regions around the native sites and a subset of well-defined cold regions (i.e. blue regions corresponding to high energy solutions) placed in the same area in the map's upper-right quadrant. In contrast, the energy maps obtained for the two ligands of the papain-like family display a large hot region around the two native binding sites of the receptor, extending to the upper-left and bottom-right regions of the map, suggesting a large promiscuous binding region for these ligands.

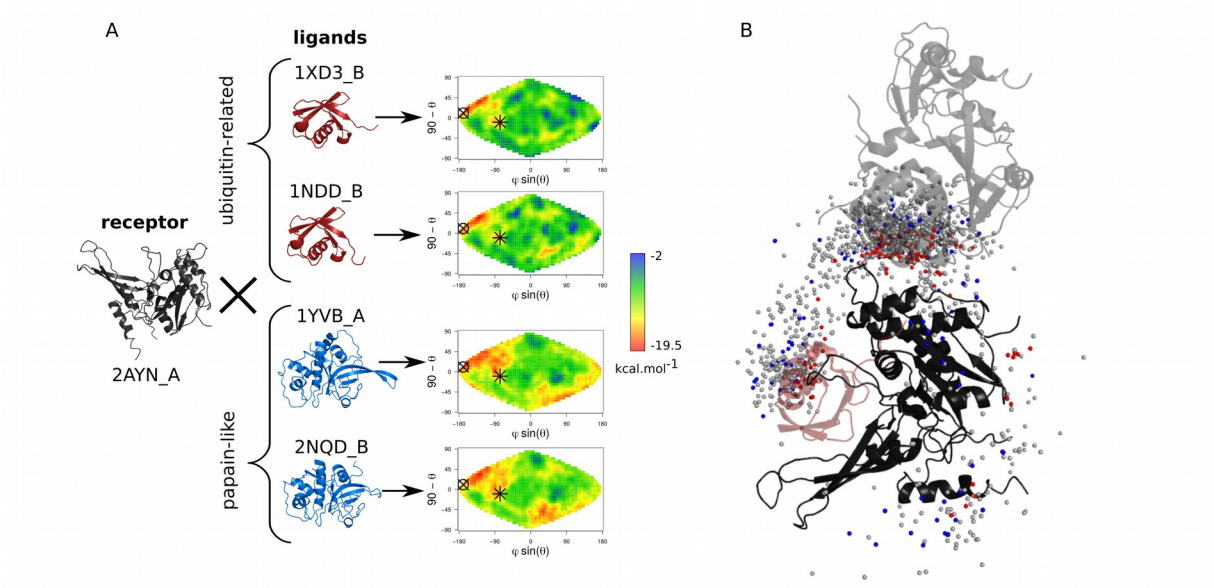


Fig 3. Subset of energy maps and of ligand docking poses for receptor 2AYN_A. (*A*) Examples of maps for the receptor 2AYN_A (ubiquitin carboxyl-terminal hydrolase (UCH) family) docked with the ligands 1XD3_B (native partner), 1NDD_B (homolog of the native partner), 1YVB_A and 2NQD_B (false partners). The star indicates the localization of the experimentally determined interaction site of the ubiquitin, the circle-cross indicates the homodimerization site of 2AYN_A. (*B*) Centers of mass (CM) of the 20 best docking poses obtained for each of the 74 ligands of the database docked with the receptors 2AYN_A. Receptor protein is represented in cartoon (black), its native ligand and its homodimere are represented in cartoon with transparency (red and black respectively). CMs of the ligands belonging to the ubiquitin-related family are colored in red, CMs of the proteins belonging to the papain-like family are colored in blue.

We ask whether the observation made for the receptor 2AYN_A, that energy maps produced with homologous ligands are more similar than those produced with unrelated ligands could be generalized to all proteins of the dataset. Therefore, we systematically compared the energy maps computed for a single receptor docked successively with the 74 ligands of the dataset by calculating of the Manhattan distance between each pair of maps (Fig 1*C* and *Materials and Methods*). The resulting distances are stored in an energy map distance (EMD) matrix, where each entry (*i*,*j*) corresponds to the distance $d_{i,j}$ between the energy maps of ligands *i* and *j* docked with the receptor of interest (Fig 1*C* and *Materials and Methods*). Consequently, a small distance $d_{i,j}$ between ligands *i* and *j* docked with the receptor *k*, reflects that their energy maps are similar. In other words, the interaction propensity of the surface of the receptor *k* is similar for both ligands *i* and *j*. The procedure is repeated for each receptor of the dataset resulting in 74 EMD matrices. In order to quantify the extent to which the interaction propensity of the receptor is conserved for homologous ligands, we investigate whether distances calculated between homologous ligand pairs (be they native-related to the receptor or not) are smaller than distances calculated between random pairs.

Fig 4 represents the boxplots of energy map distances calculated between random ligand pairs or between homologous ligand pairs docked with their native-related receptors or with the other receptors of the dataset. Homologous ligands docked either with their native-related or arbitrary receptors display significantly lower energy map distances than random ligand pairs (Wilcoxon test $p = 0$). This indicates that energy maps produced by homologous ligands docked with a given receptor are more similar than those produced with non-homologous ligands. Interestingly, this observation holds whether the receptor-ligand pair is a native pair or not. This suggests that the interaction propensity of a receptor is conserved for homologous ligands be they native-related or not.

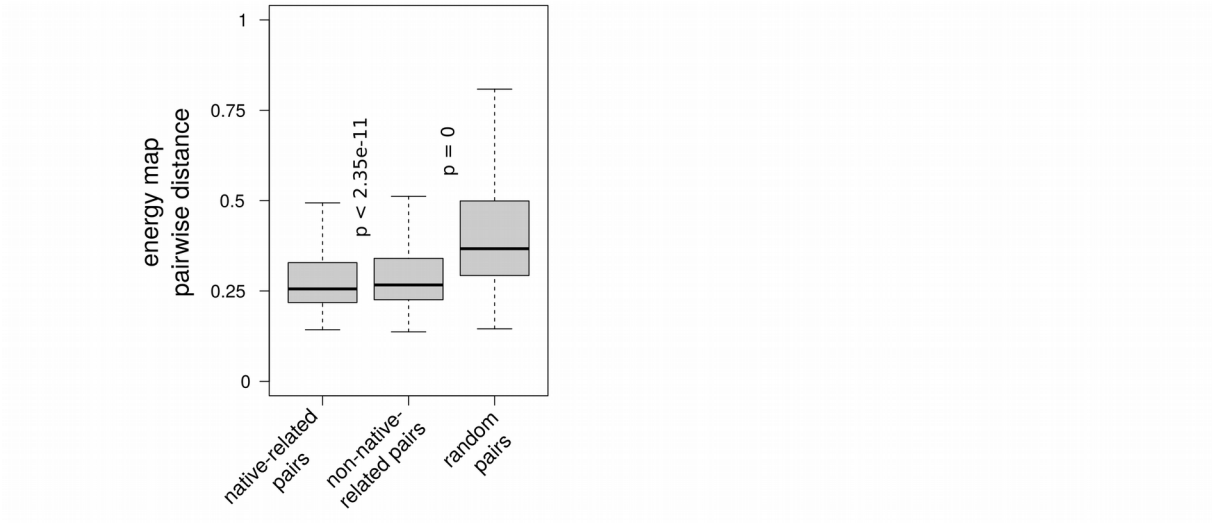


Fig 4. Boxplots of energy map pairwise distances between homologous ligand pairs from native-related partner families, homologous ligand pairs from arbitrary partner families and random ligand pairs. For each receptor, we computed (i) the average of energy map distances of pair of homologous ligands belonging to its native-related partner family(ies), (ii) the average of energy map distances of pair of homologous ligands belonging to its non-native-related partner families, and (iii) the average of energy map distances of random pairs. P-values are calculated with an unilateral Wilcoxon test.

Energy maps are specific to protein families

The results presented above prompt us to assess the extent to which the interaction propensity of a receptor is specific to the ligand families. In other words, we quantify the extent to which energy maps are specific to ligand families. If so, we should be able to retrieve ligand homology

relationships solely with the comparison of their corresponding 2D energy maps. Therefore, we tested our ability to predict the homologs of a given ligand based only on the comparison of its energy maps with those of the other ligands. In order to prevent any bias from the choice of the receptor, the 74 EMD matrices are averaged in an averaged distances matrix (ADM) (see *Materials and Methods*). Each entry *(i,j)* of the ADM corresponds to the averaged distance between two sets of 74 energy maps produced by two ligands *i* and *j*. A low distance indicates that the two ligands display similar energy maps whatever the receptor is. We computed a receiver operating characteristic (ROC) curve from the ADM (see *Materials and Methods*) which evaluates our capacity to discriminate the homologs of a given ligand from non-homologous ligands by comparing their respective energy maps computed with all 74 receptors of the dataset. The true positive set consists in the homologous protein pairs while the true negative set consists in any homology-unrelated protein pair. The resulting Area Under the Curve (AUC) is equal to 0.79 (Fig 5). We evaluated the robustness of the ligand's homologs prediction depending on the size of the receptor subset with a bootstrap procedure by randomly removing receptor subsets of different sizes (from 1 to 73 receptors). The resulting AUCs range from 0.769 to 0.79, and show that from a subset size of five receptors, the resulting prediction accuracy no longer significantly varies (risk of wrongly rejecting the equality of two variances (F-test) >5%)**,** and is thus robust to the nature of the receptor subset (S2 Fig). Finally, we evaluated the robustness of the predictions according to the number of grid cells composing the energy maps. Therefore, we repeated the procedure using energy maps with resolutions ranging from 144x72 to 48x24 cells. S2 Table presents the AUCs calculated with different grid resolutions. The resulting AUCs range from 0.78 to 0.8 showing that the grid resolution has a weak influence on the map comparison. All together, these results indicate that homology relationships between protein ligands can be detected solely on the basis of the comparison of their energy maps. In other words, the energy maps calculated for a given receptor docked with a set of ligands belonging to a same family are specific to these families. Interestingly, this observation holds for families displaying important sequence variations (S1 Table). For example, the AUC computed for the UCH and ubiquitin-related families are 0.98 and 0.88 respectively despite the fact that the average sequence identity of these families does not exceed 45% (S3 Fig and S1 Table). This indicates that energy maps are similar even for homologous ligands displaying large sequence variations.

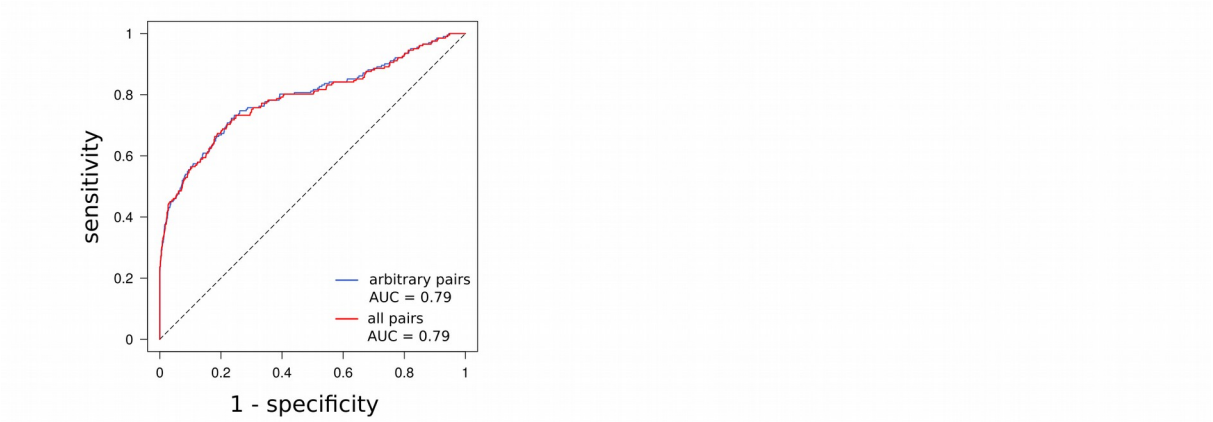


Fig 5. Receiver operating characteristic (ROC) curve and its Area Under the Curve (AUC). ROC are calculated on the averaged distance matrix (ADM) including either all pairs (blue) or only arbitrary pairs (red) (see *Materials and Methods* for more details).

We then specifically investigate the similarity of the energy maps produced by ligands belonging to a same family in order to see whether some ligands behave energetically differently from their family members. On the 74 ligands, only five (2L7R_A, 4BNR_A, 1BZX_A, 1QA9_A, 1YAL_B) display energy maps that are significantly different from those of their related homologs (Z-tests *p-values* for the comparison of the averaged distance of each ligand with their homologs versus the averaged distance of all ligands with their homologous ligands ≤ 5%). In order to identify the factors leading to differences between energy maps involving homologous ligands, we computed the pairwise sequence identity and the root mean square deviation (RMSD) between the members of each family. Interestingly, none of these criteria can explain the energy map differences observed within families (Fisher test *p* of the linear model estimated on all protein families >0.1) (see Fig 6*B-C* for the ubiquitin-related family, S4-14*B-C* Fig for the other families, and S3 Table for details). Fig 6*A* represents a subsection of the ADM for the ubiquitin-related family (i.e. the energy map distances computed between all the members of the ubiquitin-related family and averaged over the 74 receptors). Low distances reflect pairs of ligands with similar energy behaviors (i.e. producing similar energy maps when interacting with a same receptor) while high distances reveal pairs of ligands with distant energy behaviors. 2L7R_A distinguishes itself from the rest of the family, displaying high-energy map distances with all of its homologs. RMSD and sequence identity contribute modestly to the energy map distances observed in Fig 6*A* (Spearman correlation test p^{RMSD} = 0.01 and p^{seq} = 0.02 (S3 Table, Fig 6*B-C*)). Fig 6*D* shows a projection of the contribution from the electrostatic term in the energy function of ATTRACT on the surface of the seven ubiquitin-related family members (for more details, see S15 Fig and *Materials and Methods*). Fig

6*E* represents the electrostatic maps distances computed between all members of the family. 2L7R_A stands clearly out, displaying a negative electrostatic potential over the whole surface while its homologs harbor a remarkable fifty-fifty electrostatic distribution (Fig 6*D*). The negatively charged surface of 2L7R_A is explained by the absence of the numerous lysines that are present in the others members of the family (referred by black stars, Fig 6*D*). Lysines are known to be essential for ubiquitin function by enabling the formation of polyubiquitin chains on target proteins. Among the seven lysines of the ubiquitin, K63 polyubiquitin chains are known to act in non-proteolytic events while K48, K11, and the four other lysines polyubiquitin chains are presumed to be involved into addressing proteins to the proteasome [32]. 2L7R_A is a soluble UBL domain resulting from the cleavage of the fusion protein FAU [33]. Its function is unrelated to proteasomal degradation, which might explain the lack of lysines on its surface and the differences observed in its energy maps. Interestingly, the differences observed for the energy maps of 1YAL_B (Papain-like family) (S4 Fig) and 4BNR_A (eukaryotic proteases family) (S5 Fig) regarding their related homologs can be explained by the fact that they both display a highly charged surface. These two proteins are thermo-stable [34,35], which is not the case for their related homologs, and probably explains the differences observed in their relative energy maps. The V-set domain family is split into two major subgroups according to their averaged energy map distances (S6A Fig). The first group corresponds to CD2 proteins (1QA9_A and its unbound form 1HNF_A) and differs significantly from the second group (Z-test $p = 0.03$ and $p = 0.05$ respectively). The second group corresponds to CD58 (1QA9_B and its unbound form 1CCZ_A) and CD48 proteins (2PTT_A). Interestingly, CD2 is known to interact with its homologs (namely CD58 and CD48) through an interface with a striking electrostatic complementarity [36]. The two subgroups have thus evolved distinct and specific binding sites to interact together. We can hypothesize that they have different interaction propensities resulting in the differences observed between their corresponding energy maps. These five cases illustrate the capacity of our theoretical framework to reveal functional or biophysical specificities of homologous proteins that could not be revealed by classical descriptors such as RMSD or sequence identity.

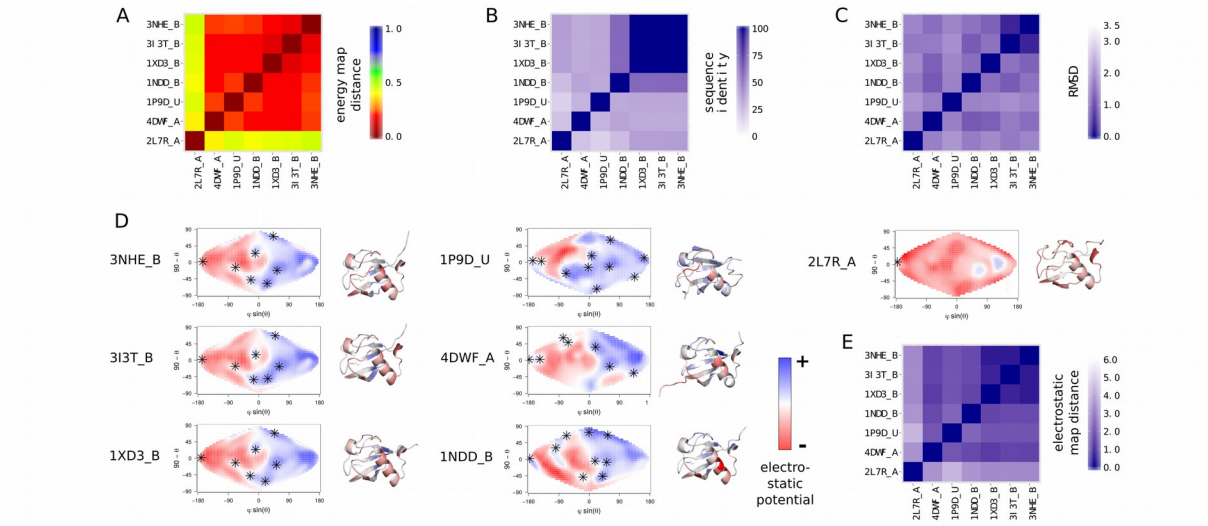


Fig 6. Ubiquitin-related family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the ubiquitin-related family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair (*i,j*) averaged over the 74 receptors of the dataset. (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see S15 Fig and *Materials and Methods*). On the electrostatic maps, lysines positions are indicated by stars. Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.

The AUC of 0.79 calculated previously with energy maps produced by the docking of either native-related or arbitrary pairs indicates that energy maps are specific to ligand families. To see whether this observation is not mainly due to the native-related pairs, we repeated the previous test while removing that time all energy maps computed with native-related pairs and calculated the resulting ADM. We then measured our ability to retrieve the homologs of each ligand by calculating the ROC curve as previously. The resulting AUC is still equal to 0.79, revealing that our ability to identify a ligand's homologs is independent from the fact that the corresponding energy maps were computed with native-related or arbitrary pairs (Fig 5). This shows that the energy maps are specific to protein families whether the docked pairs are native-related or not. Consequently, the propensity of the whole protein surface to interact with a given ligand is conserved and specific to the ligand family whether the ligand is native-related or not. This striking result may reflect both positive and negative design operating on protein surfaces to maintain functional interactions and to limit random interactions that are inherent to a crowded environment.

The interaction propensity of all surface regions of a receptor is evolutionary conserved for homologous ligands

To see whether some regions contribute more to the specificity of the maps produced by homologous ligands, we next dissected the effective contribution of the surface regions of the receptor defined according to their docking energy value, in the identification of ligand's homologs. We discretized the energy values of each energy map into five categories, leading to a palette of five energy classes (or colors) (see Fig 2*D* and *Materials and Methods*). These five-color maps highlight low-energy regions (i.e. hot regions in red), intermediate-energy regions (i.e. warm, lukewarm and cool regions in orange, light-green and dark-green respectively) and high-energy regions (i.e. cold regions in blue). We first checked that the discretization of the energy maps does not affect our ability to identify the homologs of each of the 74 ligands from the comparison of their five-colors maps. The resulting AUC is 0.77 (Table 1)**,** showing that the discretization step does not lead to an important loss of information.

Table 1. AUC obtained with different types of energy maps

type of map	continous energy maps	five-colors energy maps	red energy maps	orange energy maps	light green energy maps	dark green energy maps	blue energy maps
AUC	0.79	0.77	0.73	0.76	0.76	0.76	0.79

Table 1. The AUC are calculated from the ADM with the continuous energy maps (Fig 2*C*), the five-color energy maps (Fig 2*D*) and the one-color energy maps (Fig 2*E*) (see *Materials and Methods* for more details).

Then, we evaluated the contribution of each of the five energy classes separately in the ligand's homologs identification by testing our ability to retrieve the homologs of the 74 ligands from their one-color energy maps (either red, orange, yellow, green or blue) (see *Materials and Methods*). Table 1 shows the resulting AUCs. Interestingly, the information provided by each energy class taken separately is sufficient for discriminating the homologs of a given ligand from the rest of the dataset (Table 1). The resulting AUCs range from 0.76 to 0.79 for the orange, light green, dark green, and blue classes and are comparable to those obtained with all classes taken together (0.77). This shows that (i) warm, lukewarm, cool, and cold regions alone are sufficient to retrieve homology relationships between ligands and (ii) the localization on the receptor surface of a given energy class is specific to the ligand families. Hot regions are less discriminative and lead to an AUC of 0.73. In order to see how regions corresponding to a specific energy class are distributed over a receptor surface, we summed its 74 corresponding one-color maps into a stacked map (S16

Fig – see *Materials and Methods* for more details). For each color, the resulting stacked map reflects the tendency of a map cell to belong to the corresponding energy class. Fig 7 shows an example of the five stacked maps (i.e. for cold, cool, lukewarm, warm and hot regions) computed for the receptor 1P9D_U. Intermediates regions (i.e. warm, lukewarm and cool regions) are widespread on the stacked map while cold and hot regions are localized on few small spots (three and one respectively) no matter the nature of the ligand. S17 Fig shows for the receptor 1P9D_U the 12 blue and red stacked maps computed for each ligand family separately. We can see that some cold spots are specific to ligand families and that their area distribution is specific to families while all 12 ligand families display the same hot spot in the map's upper-right quadrant. These observations can be generalized to each receptor. On average, intermediate regions are widespread on the stacked maps and cover respectively 744, 1164 and 631 cells for cool, lukewarm and warm regions, while cold and hot regions cover no more than respectively 104 and 110 cells respectively (S18 Fig). Interestingly, hot regions are more colocalized than cold ones and are restricted to 2 distinct spots on average per stacked map**,** while cold regions are spread on 3.7 spots on average (t-Test p = 7.42e-13). These results show that ligands belonging to different families tend to dock preferentially on the same regions and thus lead to similar hot region distributions on the receptor surface. This observation recalls those made by *Fernandez-Recio et al*. [31], who showed that docking random proteins against a single receptor leads to an accumulation of low-energy solutions around the native interaction site and who suggested that different ligands will bind preferentially on the same localization.

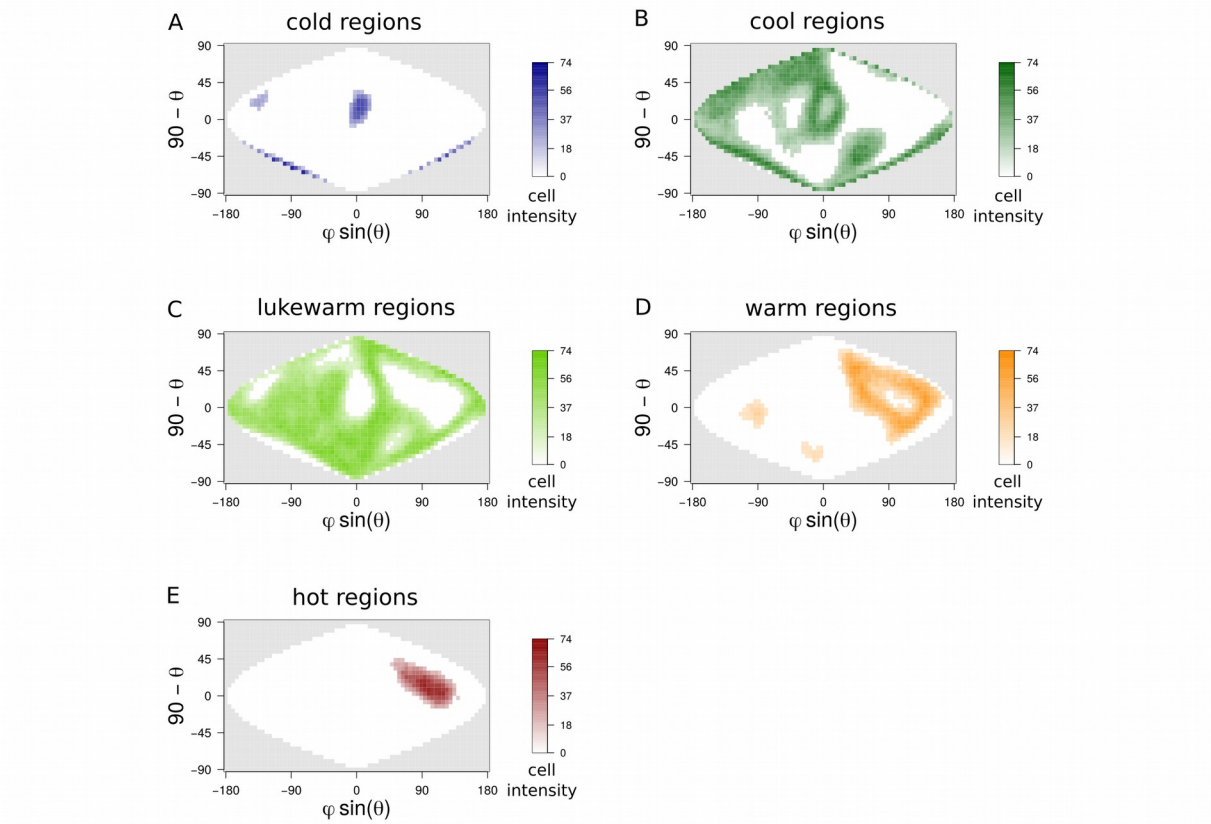


Fig 7. Stacked maps of 1P9D_U after the filtering of cells with too low intensity and areas of too small size. The protocol to generate stacked maps is presented in S16 Fig. (*A*) Blue stacked map (i.e. stacked cold regions). (*B*) Dark green stacked map (i.e. stacked cool regions). (*C*) Light green stacked map (i.e. stacked lukewarm regions). (*D*) Orange stacked map (i.e. stacked warm regions). (*E*) Red stacked map (i.e. stacked hot regions). One should notice that stacked maps of two different colors can overlap because a cell can be associated to different energy classes depending on the docked ligands. S17 Fig presents blue and red stacked maps of 1P9D_U computed for each ligand family.

We can hypothesize that hot regions present universal structural and biochemical features that make them more prone to interact with other proteins. To test this hypothesis, we computed for each protein of the dataset, the 2D projection of three protein surface descriptors (see *Materials and Methods* and S15 Fig): the Kyte-Doolittle (KD) hydrophobicity [37], the circular variance (CV) [38] and the stickiness [20]. The CV measures the density of protein around an atom and is a useful descriptor to reflect the local geometry of a surface region. CV values are comprised between 0 and 1. Low values reflect protruding residues and high values indicate residues located in cavities. Stickiness reflects the propensity of amino acids to be involved in protein-protein interfaces [20]. It has been calculated as the log ratio of the residues frequencies on protein surfaces versus their frequencies in protein-protein interfaces. For each receptor, we calculated the correlation between the docking energy and the stickiness, hydrophobicity or CV over all cells of the corresponding 2D

maps. We found a significant anti-correlation between the docking energy and these three descriptors (correlation test *p* between docking energies and respectively stickiness, hydrophobicity and CV < 2.2e-16, see S4 Table)). Fig 8 represents the boxplots of the stickiness, hydrophobicity and CV of each energy class (see S15 Fig and *Materials and Methods* section for more details). We observe a clear effect of these factors on the docking energy: cold regions (i.e. blue class) are the less sticky, the less hydrophobic and the most protruding while hot ones (i.e. red class) are the most sticky, the most hydrophobic and the most planar (Tukey HSD test [39], *p* of the differences observed between each energy classes < 2.2e-16). One should notice that stickiness has been defined from a statistical analysis performed on experimentally characterized protein interfaces and therefore between presumed native partners. The fact that docking energies (physics-based) calculated either between native-related or arbitrary partners is anti-correlated with stickiness (statistics-based) defined from native interfaces, strengthens the concept of stickiness as the propensity of interacting promiscuously and provides physics-based pieces of evidence for sticky regions as a proxy for promiscuous interactions.

We show that not only the area distribution on a receptor surface of hot regions but also those of intermediate and cold regions are similar for homologous ligands and are specific to ligand families (AUC ranging from 0.73 to 0.79) whether the ligands are native-related or not. This tendency is even stronger for intermediate and cold regions. Interestingly, the information contained in the cold regions that cover on average no more than 5.0% of the energy maps is sufficient to identify homology relationships between ligands.

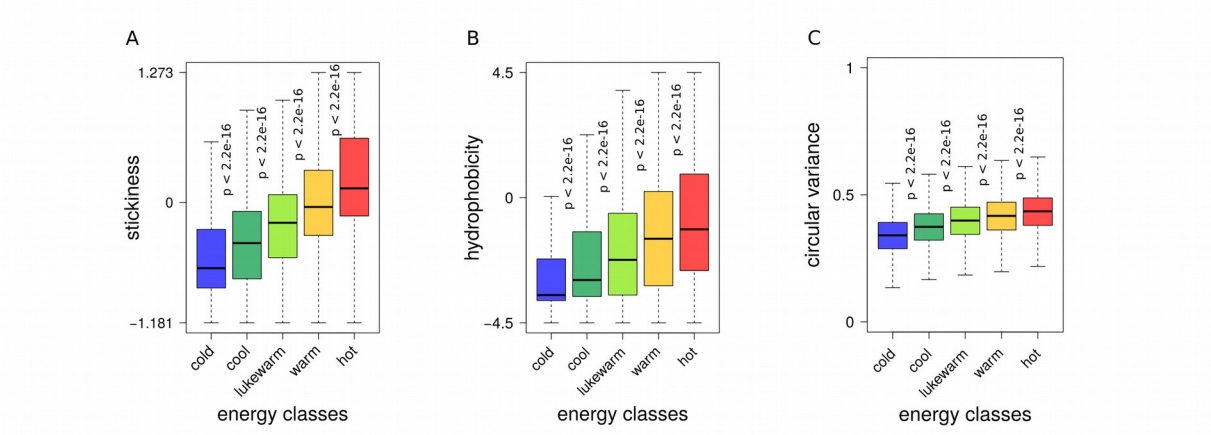


Fig 8. Boxplots of three descriptors of the protein surface. (*A*) the stickiness values, (*B*) the Kyte-Doolittle hydrophobicity and (*C*) the CV values, depending on the energy class. The stickiness, hydrophobicity and CV values are calculated for each protein following the protocol described in *Materials and Methods*. For each of these criteria, *p-values* between the median values of two "successive" energy classes were computed using the Tukey HSD statistical test [39].

8.3 Discussion

In this study, we address the impact of both positive and negative design on thousands of interaction energy landscapes by the mean of a synthetic and efficient representation of the docking energy landscapes: two-dimensional energy maps that reflect the interaction propensity of the whole surface of a protein (namely the receptor) with a given partner (namely the ligand). We show that all regions of the energy maps, including cold, intermediate and hot regions are similar for homologous ligands and are specific to ligand families whether the ligands are native-related or arbitrary. This reveals that the interaction propensity of the whole surface of proteins is constrained by functional and non-functional interactions, reflecting both positive and negative design operating on the whole surface of proteins, thus shaping the interaction energy landscapes of functional partners and random encounters. These observations were made on a dataset of 74 protein structures belonging to 12 families of structural homologs. 54 out of the 74 proteins of the dataset have at least one known partner in the dataset. For the 20 remaining proteins, we were not able to find evidences that they indeed interact with a protein of the dataset. However, we showed that the interaction propensity of a receptor is conserved for homologous ligands independently from the fact that these ligands correspond to native partners or not. Indeed, we showed that ligand homology relationships could be retrieved from their energy maps whether the maps were computed with native-related pairs or not (the corresponding AUCs calculated with and without native pairs both equal to 0.79).

While most studies that aim at depicting protein-protein interactions focus on native binding sites of proteins [12,40–44], we bring a new perspective on protein-protein interactions by providing a systematic and physical characterization of all regions of the surface of a protein in interaction with a given ligand (i.e. cold, intermediate and hot regions). Here, we address the energy behavior of not only known binding sites, but also of the rest of the protein surface, which plays an important role in protein interactions by constantly competing with the native binding site. We show that the interaction propensity of the rest of the surface is not homogeneous and displays regions with different binding energies that are specific to ligand families. This may reflect the negative design operating on these regions to limit non-functional interactions [12,14,45]. We can hypothesize that non-interacting regions participate to favor functional assemblies (i.e. functional assembly modes with functional partners) over non-functional ones and are thus evolutionary constrained by non-functional assemblies. The fact that cold regions seem to be more specific to ligand families than hot ones may be explained by the fact that they are on average more protuberant and more charged. They thus display more variability than hot ones. Indeed, there is more variability in being

positively or negatively charged and protuberant (with an important range of protuberant shapes) than in being neutral and flat. S19 Fig presents the electrostatic potential distribution of all energy classes. Cold regions display a larger variability of electrostatic potential (F-test, $p < 2.2e\text{-}16$) than hot regions that are mainly hydrophobic thus displaying neutral charge distributions in average. Consequently, a same hot region may be attractive for a large set of ligands while a cold region may be unfavorable to specific set of ligands, depending on their charges, shapes and other biophysical properties.

On the other hand, we show that hot regions are very localized (4.9% of the cells of an energy map) and tend to be similar no matter the ligand. Similarly to protein interfaces that have been extensively characterized in previous studies [2,40,41], hot regions are likely to display universal properties of binding, i.e. they are more hydrophobic and more planar, and thus more “sticky” than the other regions. They may provide a non-specific binding patch that is suitable for many ligands. However, we can hypothesize that native partners have evolved to optimize their interfaces (positive design) so that native interactions prevail over non-native competing ones. Indeed, we have previously shown that the docking of native partners lead to more favorable binding energies than the docking of non-native partners when the ligand is constrained to dock around the receptor’s native binding site [28,46]. All these results suggest a new physical model of protein surfaces where protein surface regions, in the crowded cellular environment, serve as a proxy for regulating the competition between functional and non-functional interactions. In this model, intermediate and cold regions play an important role by preventing non-functional assemblies and by guiding the interaction process towards functional ones and hot regions may select the functional assembly among the competing ones through optimized interfaces with the native partner.

In this work, we used and extended the application of the 2D energy map representation developed in [31] to develop an original theoretical framework that enables the efficient, automated and integrative analysis of different protein surface features. 2D maps provide the area distribution of a given feature on the whole protein surface and their discretization enables the study of a given surface property (e.g. protuberance, planarity, stickiness, positively charged regions, or cold and hot regions for example). They are easy to manipulate and their straightforward comparison enables (i) the study of relationships between different surface properties through the comparison of their area distributions on a protein surface and (ii) the highlight of the evolutionary constraints exerted on a given feature by comparing its area distribution on the surfaces of homologous proteins. Particularly, this enables the identification and characterization of hot regions on a protein surface which can be either specific or conserved for all ligands and opens up new possibilities for the

development of novel methods for protein binding sites prediction and their classification as functional or promiscuous in the continuity of previous developments based on arbitrary docking [28,29,31,46].

Our framework provides a proxy for further protein functional characterization as shown with the five proteins discussed in the *Results* section *Energy maps are specific to protein families.* The comparison of their respective energy maps enables us to reveal biophysical and functional properties that could not be revealed with classical monomeric descriptors such as RMSD or sequence identity. Indeed, our framework can reflect the energy behavior of a protein interacting with a subset of selected partners either functional or arbitrary, thus revealing functional and systemic properties of proteins. This work goes beyond the classical use of binary docking to provide a systemic point of view of protein interactions**,** for example by exploring the propensity of a protein to interact with hundreds of selected ligands, and thus addressing the behavior of a protein in a specific cellular environment. Particularly, exploring the dark interactome (i.e. non-functional assemblies and interactions with non-functional partners) can provide a wealth of valuable information to understand mechanisms driving and regulating protein-protein interactions. Precisely, our 2D energy maps based strategy enables its exploration in an efficient and automated way.

8.4 Materials and Methods

Protein dataset

The dataset comprises 74 protein structures divided into 12 families of structural homologs (see S1 Table for a detailed list of each family). Each family is related to at least one other family (its native-related partners family) through a pair of interacting proteins for which the 3D structure of the complex is characterized experimentally (except the V set domain family: the two native partners are homologous and belong to the same family) (S1 Fig). Each family is composed of a monomer selected from the protein-protein docking benchmark 5.0 [47] in its bound and unbound forms, which is called the master protein. Each master protein has a native partner (for which the 3D structure of the corresponding complex has been characterized experimentally) in the database, which is the master protein for another family, except the V set domain family, which is a self-interacting family. When available, we completed families with interologs (i.e. pairs of proteins which have interacting homologs in an other organism) selected in the INTEREVOL database [48] according to the following criteria: (i) experimental structure resolution better than 3.25 Å, (ii) minimum alignment coverage of 75% with the rest of the family members and (iii) minimum sequence identity of 30% with at least one member of the family. Since we were limited by the number of available interologs, we completed families with unbound monomers homologous to the master following the same criteria and by searching for their partners in the following protein-protein interactions databases [49–54]. We consider that all members of a family correspond to native-related partners of all members of their native-related partner family. To address the impact of conformational changes of a protein on its interaction energy maps, we added different NMR conformers. We show that energy maps involving pairs of conformers are significantly more similar than those obtained for other pairs of homologous ligands (unilateral Wilcoxon test, $p < 2.2e\text{-}16$) showing that the conformational changes in a protein (lower than 3Å) have a low impact on the resulting energy maps (S20 Fig).

Docking experiment and construction of energy maps

A complete cross-docking experiment was realized with the ATTRACT software [25] on the 74 proteins of the dataset, leading to 5476 (74 x 74) docking calculations (Fig 1*A*). ATTRACT uses a coarse-grain reduced protein representation and a simplified energy function comprising a pseudo

Lennard-Jones term and an electrostatic term. The calculations took approximately 20000 hours on a 2.7GHz processor. Prior to docking calculations, all PDB structures were prepared with the DOCKPREP software [55].

During a docking calculation, the ligand L_i explores exhaustively the surface of the receptor R_k (whose position is fixed during the procedure), sampling and scoring thousands of different ligand docking poses (between 10000 and 50000 depending on the sizes of the proteins) (Fig 2*A*). For each protein couple R_k-L_i, a 2D energy map is computed which shows the distribution of the energies of all docking solutions over the receptor surface. To compute these maps, for all docking poses, the spherical coordinates (φ, θ) (with respect to the receptor center of mass (CM)) of the ligand CM are represented onto a 2D map in an equal-area 2D sinusoidal projection (Fig 2*B*) (see [31] for more details). Each couple of coordinates (φ, θ) is associated with the energy of the corresponding docking conformation (Fig 2*B*). A continuous energy map is then derived from the discrete one, where the map is divided into a grid of 36 x 72 cells. Each cell represents the same surface and, depending on the size of the receptor, can span from 2.5 $Å^2$ to 13$Å^2$. For each cell, all solutions with an energy score below 2.7 $kcal/mol^{-1}$ from the lowest solution of the cell are retained, according to the conformations filtering protocol implemented in [28]. The average of the retained energy scores is then assigned to the cell. If there is no docking solution in a cell, a score of 0 is assigned to it. Finally, the energies of the cells are smoothed, by averaging the energy values of each cell and of the eight surrounding neighbors (Fig 2*C*).

For each map, the energy values are discretized into five energy classes of same range leading to a discrete five-colors energy map (Fig 2*D*). The range is calculated for each energy map and spans from the minimum to the maximum scores of the map cells. The range of the energy classes of the map R_k-L_i is equal to (maxE – minE)/5, where maxE and minE correspond to the maximal and minimal energy values in the R_k-L_i map. Each five-colors energy map is then split into five one-color maps, each one representing an energy class of the map (Fig 2*E*). The continuous, five-colors and one-color energy maps are calculated for the 5476 energy maps.

Comparison of energy maps and identification of ligand's homologs

Since, we cannot compare energy maps computed for two unrelated receptors, the procedure is ligand-centered and only compares energy maps produced with different ligands docked with the same receptor. The referential (i.e. the receptor) is thus the same (in other words all grid cells are comparable) for all the energy maps that are compared. For each receptor R_k, we computed a 74x74

energy map distance (EMD) matrix where each entry (*i,j*) corresponds to the pairwise distance between the energy maps R_k-L_i and R_k-L_j resulting from the docking of the ligands L_i and L_j on the receptor R_k (Fig 1). The pairwise distance $d_{Man}(R_k\text{-}L_i, R_k\text{-}L_j)$ between the energy maps is calculated with a Manhattan distance according to

$$d_{Man}\left(R_k L_i, R_k L_j\right) = \sum_{n=1}^{36} \sum_{m=1}^{72} \left| a_{nm} - b_{nm} \right|$$

where a_{nm} and b_{nm} are the cells of row index *n* and column index *m* of the energy maps R_k-L_i and R_k-L_j respectively. Low distances reflect pairs of ligands that induce similar energy maps when they are docked on the same receptor. The procedure presented in Fig 1 is repeated for each receptor of the database resulting in 74 EMD matrices. The 74 EMD matrices are averaged into an averaged distances matrix (ADM). Each entry (*i,j*) of the ADM reflects the similarity of the R_k-L_i and R_k-L_j energy maps averaged over all the receptors R_k in the dataset. In order to estimate the extent to which family members display similar energy maps when they are docked with the same receptor, we tested our ability to correctly identify the homologs of the 74 ligands from the only comparison of its energy maps with those of the other ligands. Because, energy maps are receptor-centered, we cannot compare the energy maps computed for two unrelated receptors. The procedure consists in the comparison of energy maps produced with different ligands docked with a same receptor. Two ligands (*i,j*) are predicted as homologs according to their corresponding distance (*i,j*) in the ADM. Values close to zero should reflect homologous ligand pairs, while values close to one should reflect unrelated ligand pairs. A Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC) are computed from the ADM. True positives (TP) are all the homologous ligand pairs and predicted as such, true negatives (TN) are all the unrelated ligand pairs and predicted as such. False positives (FP) are unrelated ligand pairs but incorrectly predicted as homologous pairs. False negatives (FN) are homologous ligand pairs but incorrectly predicted as unrelated pairs. ROC curves and AUC values were calculated with the R package pROC [56]. The ligand's homologs identification was also realized using the five-color energy maps or the one-color energy maps taken separately. The five energy class regions display very different sizes, with median ranging from 63 and 66 cells for the blue and red regions to 633 cells for the yellow one. To prevent any bias due to the size of the different classes, we normalized the Manhattan distance by the size of the regions compared in the map. The rest of the procedure is the same than those used for continuous energy maps (Fig 1).

To visualize the area distribution of the regions of a given energy class for all ligands on the receptor surface, the 74 corresponding one-color maps are summed into a stacked map where each cell's intensity varies from 0 to 74 (S16 Fig). To remove background-image from these maps, i.e. cells with low intensity (intensity < 17) and the areas of small size (< 4 cells), we used a Dirichlet process mixture model simulation for image segmentation (R package *dpmixsim*) [57].

2D projection of monomeric descriptors of protein surfaces

We computed KD hydrophobicity [37], stickiness [20], CV [38] maps of each protein of the dataset, in order to compare their topology with the energy maps. Prior to all, proteins belonging to the same families were structurally aligned with TM-align [58] in order to place them in the same reference frame, making their maps comparable. Particles were generated around the protein surface with a slightly modified Shrake-Rupley algorithm [59]. The density of spheres is fixed at $1Å^2$, representing several thousands particles per protein. Each particle is located at 5Å from the surface of the protein. The CV, stickiness and KD hydrophobicity values of the closest atom of the protein are attributed to each particle. We also generated electrostatic maps reflecting the distribution of the contribution of the coulombic term as encoded in the ATTRACT force field on a protein surface. The procedure is slightly different: each particle *i* has a +1 positive charge, and receives the coulombic value:

$$Q_i = \sum_{j=1}^{n} q_i q_j / E r_{ij}$$

with *n* the number of pseudo-atom in the protein, q_i the charge of the particle, q_j the charge of the pseudo-atom *j*, r_{ij} the distance between the particle *i* and the pseudo-atom *j*, and ε a distant-dependent dielectric constant ($\varepsilon = 15 r_{ij}$). CV was calculated following the protocol described in [38] on the all-atom structures. Stickiness, electrostatics and hydrophobicity were calculated on ATTRACT coarse-grain models. Pseudo-atom charges are defined according to the ATTRACT force field [25]. After attributing a value to each particle, the position of their spherical coordinates is represented in a 2-D sinusoidal projection, following the same protocol as described in Fig 2 and *Materials and Methods* section *Docking experiment and construction of energy maps*. The map is then smoothed following the protocol in Fig 2.

Acknowledgment

We thank F. Fraternali, R. Guerois, E. Laine, M. Montes for their constructive comments on the manuscript. SSM work is supported by the “Initiative d’Excellence” program from the French State (Grant “DYNAMO”, ANR-11-LABX-0011-01). HS work is supported by a French government fellowship. This work was supported by GENCI-CINES (grant No. x2016077460).

8.5 References

[1] Garzón JI, Deng L, Murray D, Shapira S, Petrey D, Honig B. A computational interactome and functional annotation for the human proteome. ELife Sciences 2016;5:e18715. doi:10.7554/eLife.18715.

[2] Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. Q Rev Biophys 2008;41:133–80. doi:10.1017/S0033583508004708.

[3] Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nature Biotechnology 2009;27:157–67. doi:10.1038/nbt1519.

[4] Nooren IMA, Thornton JM. Diversity of protein–protein interactions. The EMBO Journal 2003;22:3486–92. doi:10.1093/emboj/cdg359.

[5] Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007;450:973–82. doi:10.1038/nature06523.

[6] Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays 2013;35:1050–5. doi:10.1002/bies.201300066.

[7] McGuffee SR, Elcock AH. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. PLoS Comput Biol 2010;6:e1000694. doi:10.1371/journal.pcbi.1000694.

[8] Yu I, Mori T, Ando T, Harada R, Jung J, Sugita Y, et al. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. Elife 2016;5. doi:10.7554/eLife.19274.

[9] Levy ED, Kowarzyk J, Michnick SW. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. Cell Rep 2014;7:1333–40. doi:10.1016/j.celrep.2014.04.009.

[10] Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 2002;99:2754–9. doi:10.1073/pnas.052706099.

[11] Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI. Robust protein protein interactions in crowded cellular environments. Proc Natl Acad Sci USA 2007;104:14952–7. doi:10.1073/pnas.0702766104.

[12] Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci USA 2009;106:10159–64. doi:10.1073/pnas.0812414106.

[13] Karanicolas J, Corn JE, Chen I, Joachimiak LA, Dym O, Peck SH, et al. A de novo protein binding pair by computational design and directed evolution. Mol Cell 2011;42:250–60. doi:10.1016/j.molcel.2011.03.010.

[14] Garcia-Seisdedos H, Empereur-Mot C, Elad N, Levy ED. Proteins evolve on the edge of supramolecular self-assembly. Nature 2017;548:244–7. doi:10.1038/nature23320.

[15] Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–98.

[16] Pál C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. Genetics 2001;158:927–31.

[17] Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 2008;134:341–52. doi:10.1016/j.cell.2008.05.042.

[18] Zhang J, Maslov S, Shakhnovich EI. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. Molecular Systems Biology 2008;4:210.

[19] Heo M, Maslov S, Shakhnovich E. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci USA 2011;108:4258–63. doi:10.1073/pnas.1009392108.

[20] Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci USA 2012;109:20461–6. doi:10.1073/pnas.1209312109.

[21] Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci USA 2012;109:E831-840. doi:10.1073/pnas.1117408109.

[22] Schavemaker PE, Śmigiel WM, Poolman B. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. Elife 2017;6. doi:10.7554/eLife.30084.

[23] Ritchie DW, Venkatraman V. Ultra-fast FFT protein docking on graphics processors. Bioinformatics 2010;26:2398–405. doi:10.1093/bioinformatics/btq444.

[24] Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS ONE 2011;6:e24657. doi:10.1371/journal.pone.0024657.

[25] de Vries S, Zacharias M. Flexible docking and refinement with a coarse-grained protein model using ATTRACT. Proteins 2013;81:2167–74. doi:10.1002/prot.24400.

[26] Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. Towards the prediction of protein interaction partners using physical docking. Mol Syst Biol 2011;7:469. doi:10.1038/msb.2011.3.

[27] Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y. Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods. BMC proceedings, vol. 7, BioMed Central; 2013, p. S6.

[28] Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A. Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. PLoS Comput Biol 2013;9:e1003369. doi:10.1371/journal.pcbi.1003369.

[29] Vamparys L, Laurent B, Carbone A, Sacquin-Mora S. Great interactions: How binding incorrect partners can teach us about protein recognition and function. Proteins 2016;84:1408–21. doi:10.1002/prot.25086.

[30] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.

[31] Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein–protein interaction sites from docking energy landscapes. Journal of Molecular Biology 2004;335:843–865.

[32] Xu P, Duong DM, Seyfried NT, Cheng D, Xie Y, Robert J, et al. Quantitative Proteomics Reveals the Function of Unconventional Ubiquitin Chains in Proteasomal Degradation. Cell 2009;137:133–45. doi:10.1016/j.cell.2009.01.041.

[33] Welchman RL, Gordon C, Mayer RJ. Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol 2005;6:599–609. doi:10.1038/nrm1700.

[34] Molnár T, Vörös J, Szeder B, Takáts K, Kardos J, Katona G, et al. Comparison of complexes formed by a crustacean and a vertebrate trypsin with bovine pancreatic trypsin inhibitor - the key to achieving extreme stability? FEBS J 2013;280:5750–63. doi:10.1111/febs.12491.

[35] Sumner IG, Harris GW, Taylor MA, Pickersgill RW, Owen AJ, Goodenough PW. Factors effecting the thermostability of cysteine proteinases from Carica papaya. Eur J Biochem 1993;214:129–34.

[36] Wang JH, Smolyar A, Tan K, Liu JH, Kim M, Sun ZY, et al. Structure of a heterophilic adhesion complex between the human CD2 and CD58 (LFA-3) counterreceptors. Cell 1999;97:791–803.

[37] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–32.

[38] Mezei M. A new method for mapping macromolecular topography. Journal of Molecular Graphics and Modelling 2003;21:463–72. doi:10.1016/S1093-3263(02)00203-6.

[39] Tukey JW. Comparing Individual Means in the Analysis of Variance. Biometrics 1949;5:99–114. doi:10.2307/3001913.

[40] Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–98.

[41] Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins: Structure, Function, and Bioinformatics 2002;47:334–343.

[42] Li X, Keskin O, Ma B, Nussinov R, Liang J. Protein–protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. Journal of Molecular Biology 2004;344:781–795.

[43] Keskin O, Ma B, Nussinov R. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 2005;345:1281–94. doi:10.1016/j.jmb.2004.10.077.

[44] Andreani J, Faure G, Guerois R. Versatility and invariance in the evolution of homologous heteromeric interfaces. PLoS Comput Biol 2012;8:e1002677. doi:10.1371/journal.pcbi.1002677.

[45] Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. J Mol Biol 2014;426:2632–52. doi:10.1016/j.jmb.2014.04.017.

[46] Sacquin-Mora S, Carbone A, Lavery R. Identification of protein interaction partners and protein–protein interaction sites. Journal of Molecular Biology 2008;382:1276–1289.

[47] Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. Journal of Molecular Biology 2015;427:3031–3041.

[48] Faure G, Andreani J, Guerois R. InterEvol database: exploring the structure and evolution of protein complex interfaces. Nucleic Acids Res 2012;40:D847-856. doi:10.1093/nar/gkr845.

[49] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 2004;32:D449–51. doi:10.1093/nar/gkh086.

[50] Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 2005;33:D418–24. doi:10.1093/nar/gki051.

[51] Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes H-W, et al. MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 2006;34:D436–41. doi:10.1093/nar/gkj003.

[52] Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9. doi:10.1093/nar/gkj109.

[53] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct—open source resource for molecular interaction data. Nucleic Acids Res 2007;35:D561–5. doi:10.1093/nar/gkl958.

[54] Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res 2016;44:W529–35. doi:10.1093/nar/gkw363.

[55] Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: combining techniques to model RNA-small molecule complexes. RNA 2009;15:1219–30. doi:10.1261/rna.1563609.

[56] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77. doi:10.1186/1471-2105-12-77.

[57] Ferreira da Silva AR. A Dirichlet process mixture model for brain MRI tissue classification. Med Image Anal 2007;11:169–82. doi:10.1016/j.media.2006.12.002.

[58] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–9. doi:10.1093/nar/gki524.

[59] Saladin A, Fiorucci S, Poulain P, Prévost C, Zacharias M. PTools: an opensource molecular docking library. BMC Struct Biol 2009;9:27. doi:10.1186/1472-6807-9-27.

[60] Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 2014;42:D304-309. doi:10.1093/nar/gkt1240.

[61] Snedecor GW, Cochran WC. Statistical methods. lowa state university press, Ames. Iowa N Vadivukkarasi et Al 1989.

8.6 Supporting information

Protein identifier	sequence identity (%)	RMSD (Å²)	Structural family
1YVB_A (b) 2GHU_A (u) 2NQD_B (b) 3F75_A (b) 3BWK_C (b) 3RVV_A (b) 1YAL_B (u) 3IMA_A (b)	43	1.57	Papain-like
1QA9_A (b) 1CCZ_A (u) 1QA9_B (b) 1HNF_A (u) 2PTT_A (b) 2PTT_B (b)	33.6	1.74	V set domains (antibody variable domain-like)
1XD3_A (b) 1UCH_A (u) 1CMX_A (b) 2WDT_A (b) 3IFW_A (b)	45.8	1.71	Ubiquitin carboxyl-terminal hydrolases (UCH-L)
2AYO_A (b) 2AYN_A (u) 3I3T_A (b) 3NHE_A (b) 2Y6E_A (u)	41.1	2.35	Ubiquitin carboxyl-terminal hydrolases (UCH)
3FN1_A (b) 2LQ7_A_2 (u) 2LQ7_A_13 (u) 2LQ7_A_15 (u) 2LQ7_A_19 (u)	100	2.13	Ubiquitin activating enzymes (UBA)
3FN1_B (b) 2EDI_A_1 (u) 2EDI_A_7 (u) 2EDI_A_10 (u) 1YH2_A (u) 1FXT_A (b) 2GMI_A (b) 4P5O_G (b)	48	1.79	UBC-related
1AVW_B (b) 1BA7_A (u) 3I2A_B (u) 3BX1_B (b) 4AN6_A (u) 4J2Y_A (b) 4IHZ_A (u)	37.1	2.22	Kunitz (STI) inhibitors
1M9X_B (b) 4J93_A (u) 2WLV_B (u) 2XGU_A (u)	52.7	1.93	Retrovirus capsid proteins, N-terminal core domain
1XD3_B (b) 3NHE_B (b) 3I3T_B (b) 1NDD_B (u) 2L7R_A (u) 1P9D_U (u) 4DWF_A (u)	44.2	1.64	Ubiquitin-related
1YVB_B (b) 1CEW_I (u) 3IMA_B (b) 4N6O_B (b) 2CH9_B (u)	40	1.99	Cystatins
1M9X_A (b) 2CPL_A (u) 3K2C_A (u) 1MZW_A (u) 2RMC_G (u) 2NUL_B (u)	52.2	1.53	Cyclophilins (peptidylprolyl isomerase)
1AVW_A (b) 1QQU_A (u) 1FXY_B (u) 1HNE_C (u) 1ZJD_A (b) 1BZX_A (b) 4BNR_A (b) 3I29_A (b)	49.6	1.45	Eukaryotic proteases

S1 Table. List of proteins of the dataset and their structural families. Proteins are referred by their PDB identifiers, followed by their chain identifier. The NMR conformers are referred with their conformation identifier. The conformational state of the structures are indicated in brackets ((b) for bound conformation, (u) for unbound conformation). Structural families are named according to the SCOPe database [60] at the family level. Averaged sequence identity and RMSD are given for each family.

grid resolution	144x72	120x60	100x50	80x40	72x36	60x30	48x24
AUC	0.8	0.8	0.8	0.79	0.79	0.78	0.78

S2 Table. AUC according to the grid resolution used for the energy maps. A linear model was constructed from the dataset constituted of all the intra-family ligand pairs (202 protein pairs). This model allows the estimation of the linear correlation between the three descriptors and the pairwise ADM distance. The model takes into account the individual contribution of each descriptor as well as their crossed contributions with each other. The p-value of each individual contribution calculated over the 202 pairs is estimated with a Fisher test and are given in the table line “all proteins”. We then individually looked each family to see whether the contribution of the descriptors is dependent from the family. Inside each family, the number of protein pairs is too small to estimate a linear model. Consequently, we used a Spearman correlation coefficient test to estimate the p-value of each contribution.

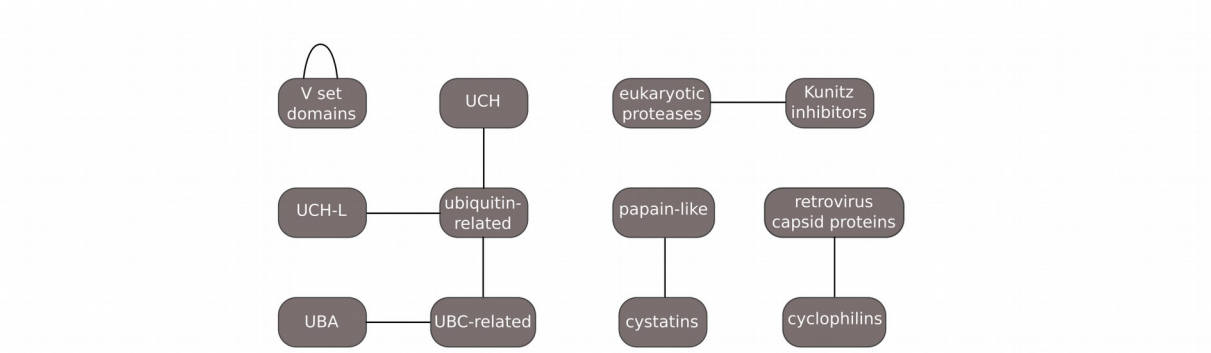
protein family	sequence identity	RMSD	electrostatics potential
ubiquitin-related	0.02396	0.01164	4.228e-5
V set domains	0.00651	0.5359	0.1773
UCH-L	0.07388	0.1076	0.3466
UCH	0.5109	0.00651	0.3104
UBA	NA	0.00117	0.04791
UBC-related	0.00566	0.3972	0.01296
Kunitz inhibitors	0.00527	0.00358	2.989e-5
retrovirus capsid proteins	0.00481	0.04156	0.00481
papain-like	0.8075	0.4613	9.09e-7
cystatins	0.0706	0.9867	0.08972
cyclophilins	0.03993	0.2539	0.00028
eukaryotic proteases	0.3861	0.2273	5.634e-7
all proteins	0.64989	0.91385	2.59e-7

S3 Table. Estimation of the effective contribution of sequence identity, RMSD and electrostatic distance in the pairwise ADM distances for each ligand pair belonging to a same family. The correlation is computed between each cell of the 74 energy maps of each of

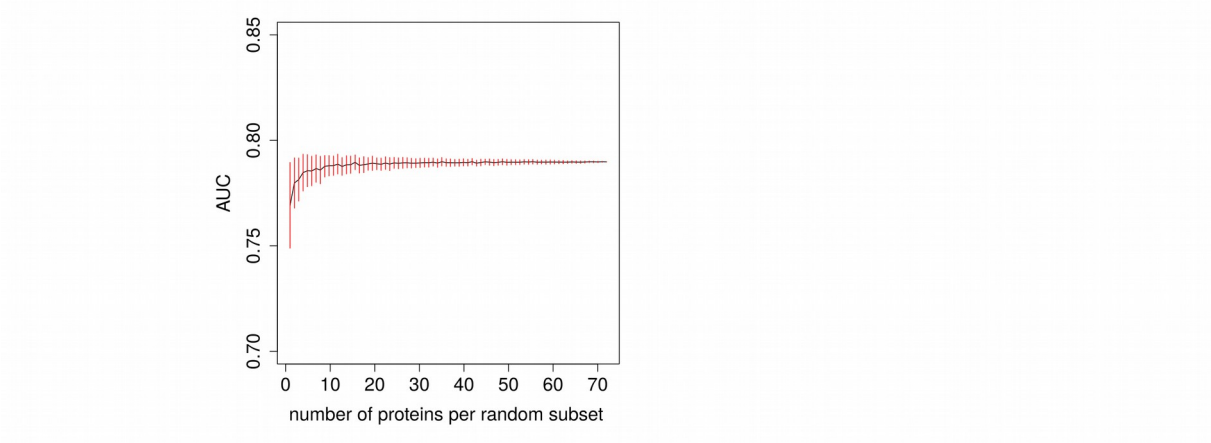
the 74 receptors and the corresponding cell in receptor's maps of stickiness, hydrophobicity and CV.

	Spearman correlation	p-value
energy vs stickiness	-0.36	< 2.2e-16
energy vs hydrophobicity	-0.24	< 2.2e-16
energy vs CV	-0.26	< 2.2e-16

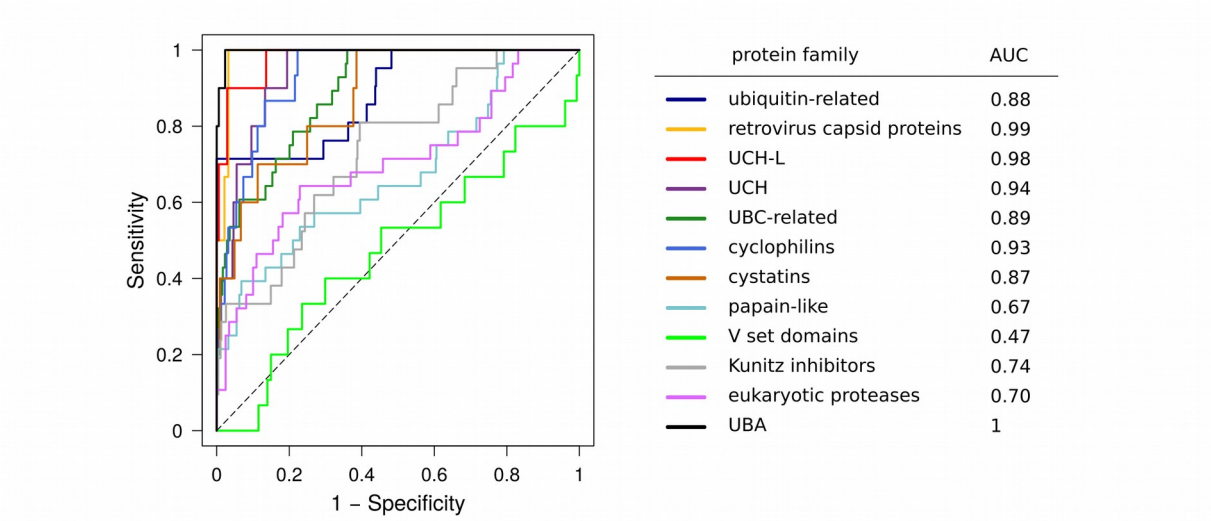
S4 Table. Correlation between energy scores and stickiness, hydrophobicity and circular variance (CV). The grid resolution corresponds to the number of cells composing the energy maps. The AUC is calculated following the same protocol used in the main text (see *Materials and Methods*).



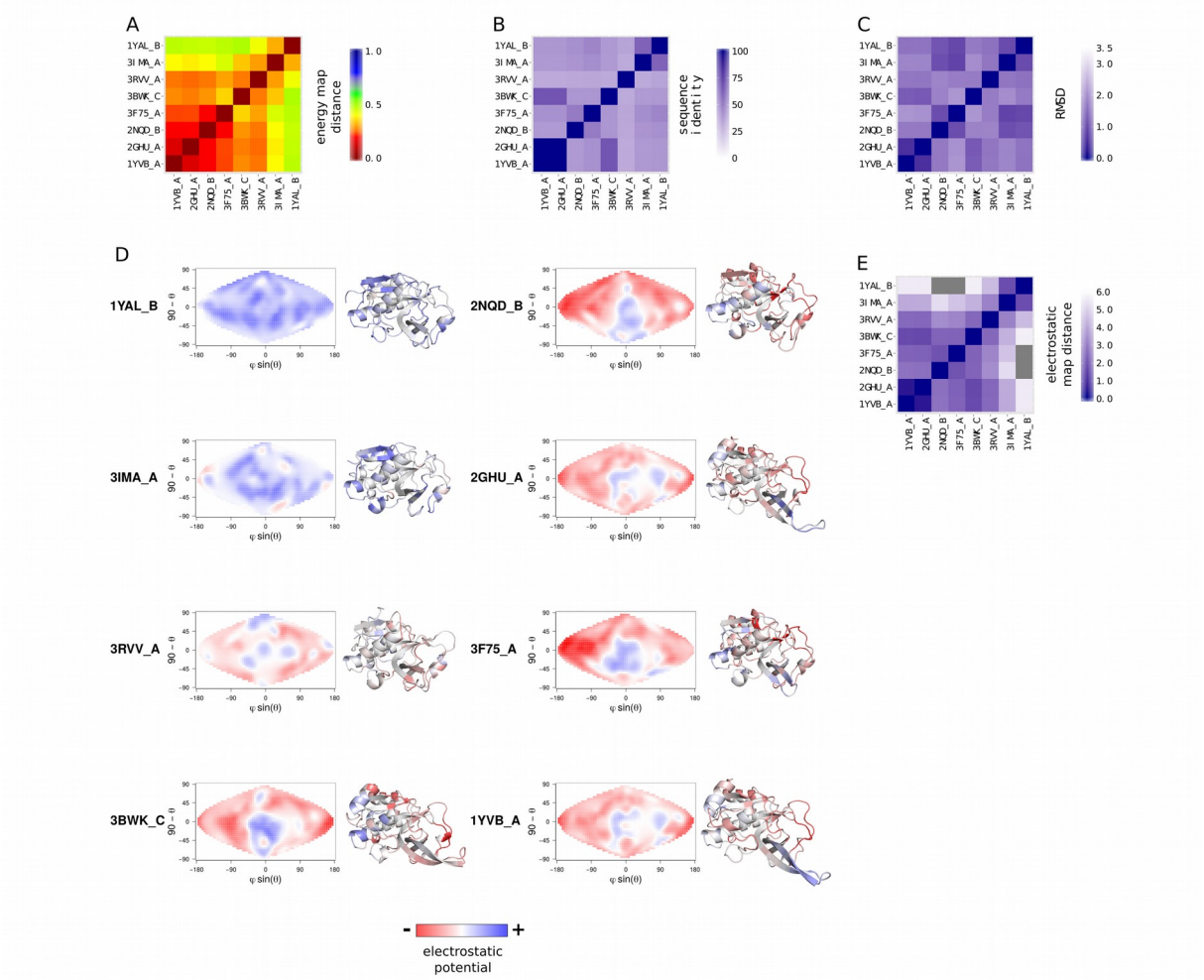
S1 Fig. Interactions between structural families of the dataset. Interactions are symbolized by links between families. An interaction is established between two families when, there is at least one PDB reporting a structure of complex involving members of the two families [30]. Consequently, all members of a family do not necessarily have its native partner in its native-related partner family. The V set domains family is a special case of self-interacting family, where members form dimers of structural homologs.



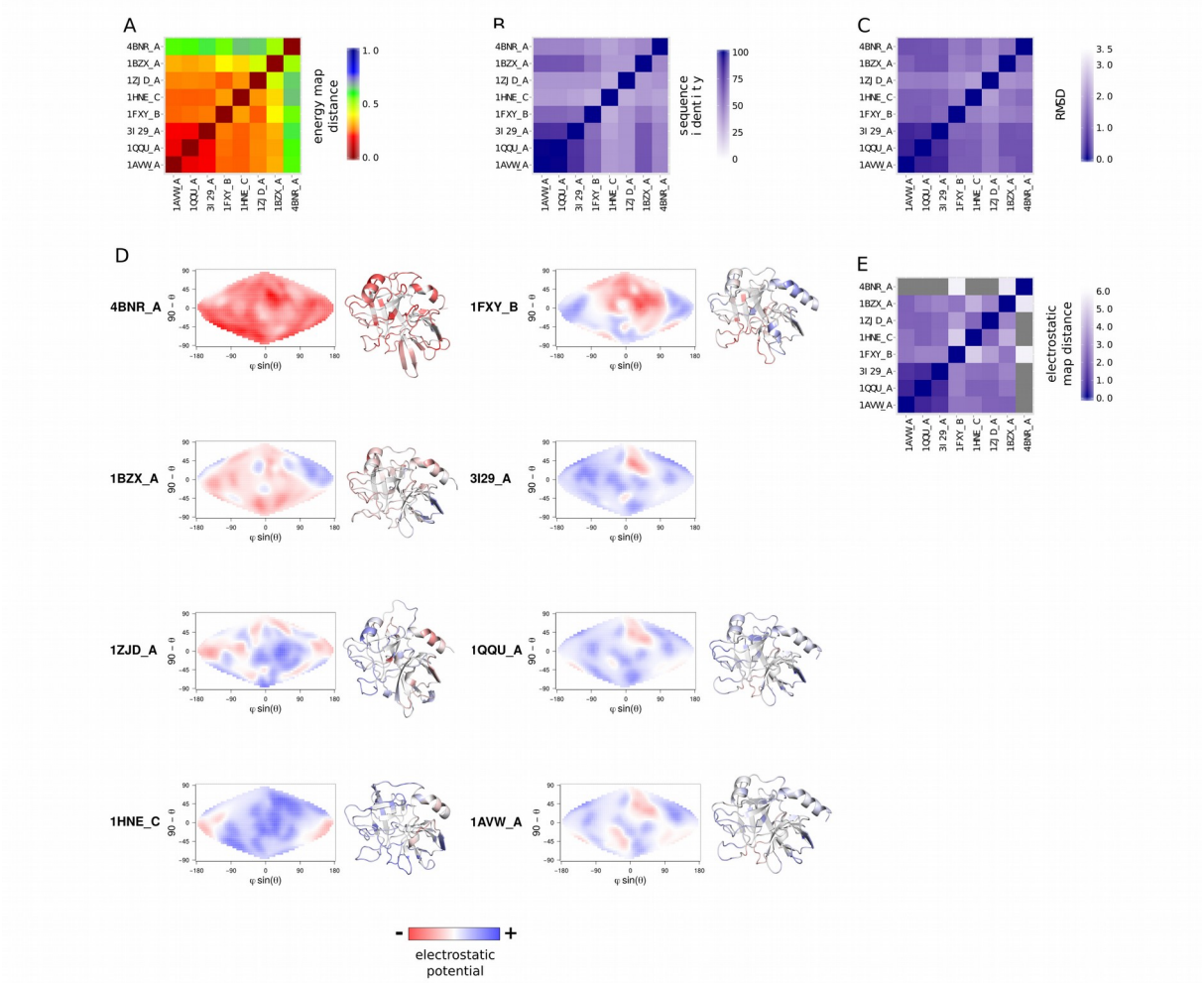
S2 Fig. AUC values calculated on random subsets of receptor of different sizes. The AUC is computed following the protocol described in Fig. 1 with random subsets composed from 1 to 73 receptors. Receptors of each subset are randomly chosen among the 74 receptors of the dataset. For each subset size, the procedure is repeated 100 times. Red vertical lines indicate the standard deviation of the AUC for each subset size. Above a subset size of five receptors, the AUC does not significantly fluctuate (risk of wrongly rejecting the equality of two variances (F-test) >5% [61]).



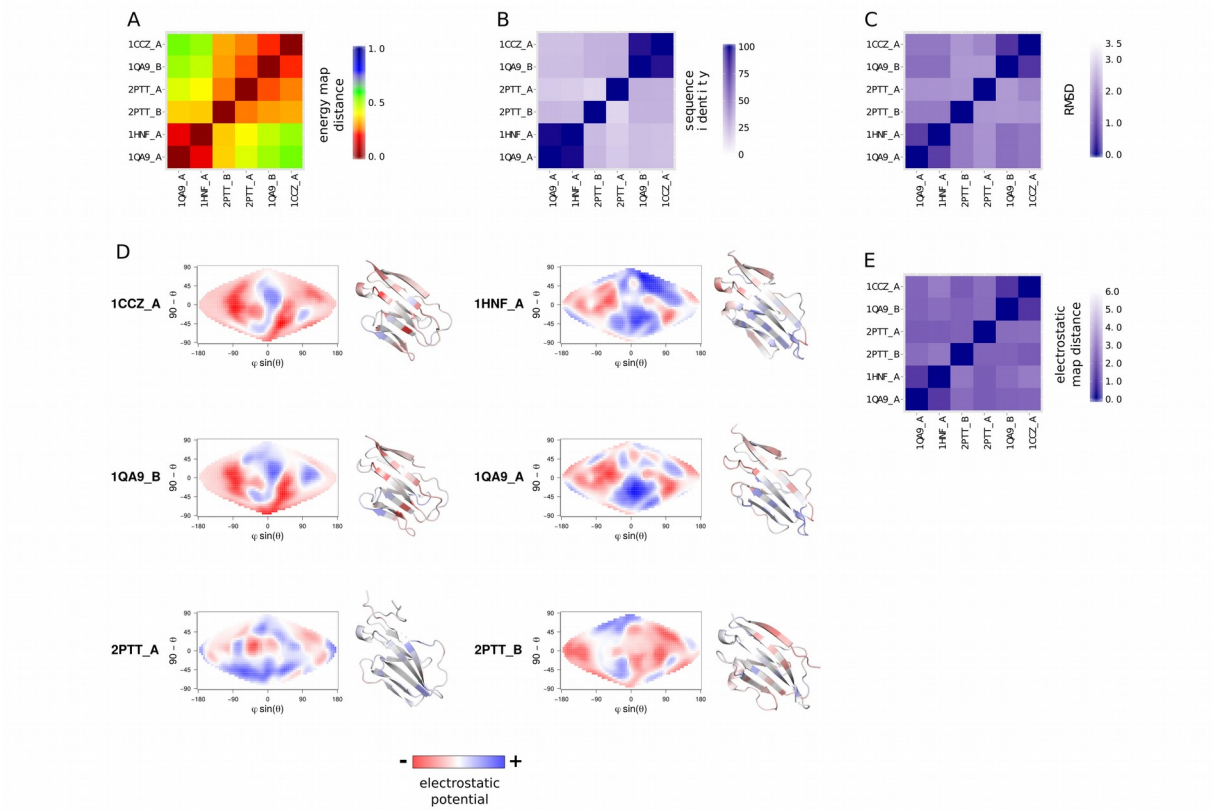
S3 Fig. Receiver operating characteristic (ROC) curve and Area Under this Curve (AUC) calculated for each family.



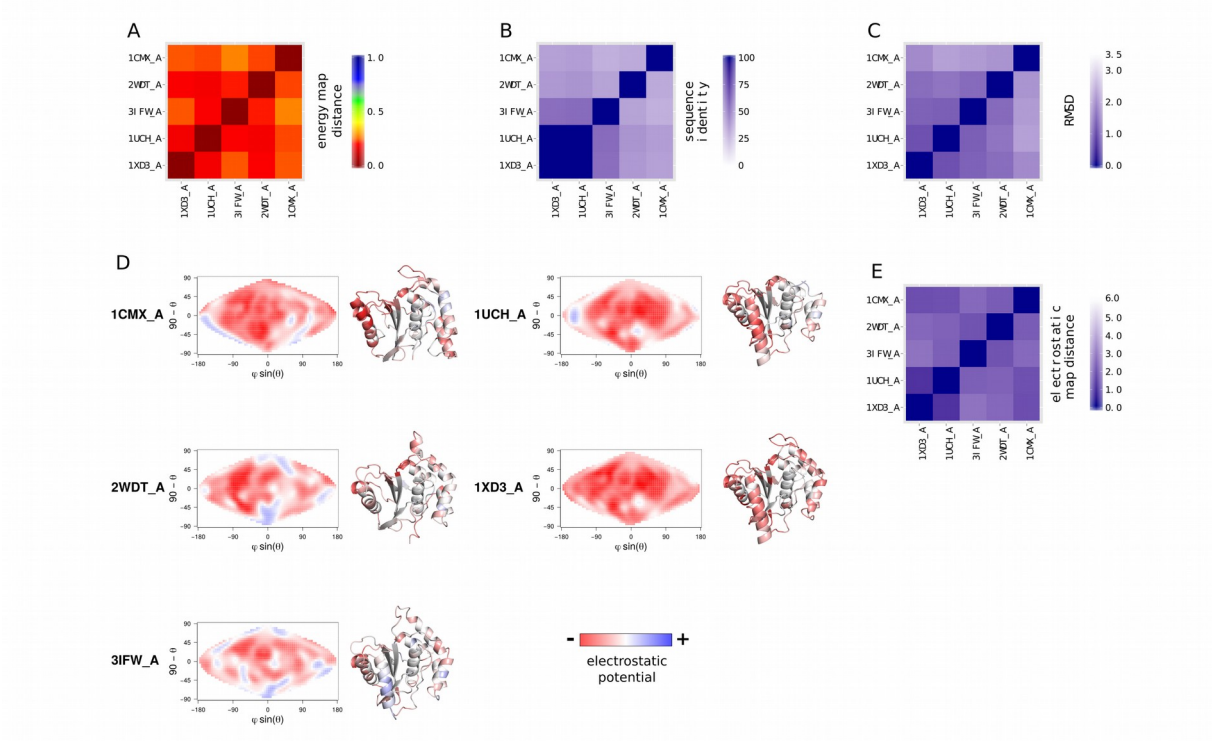
S4 Fig. Papain-like family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the papain-like family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



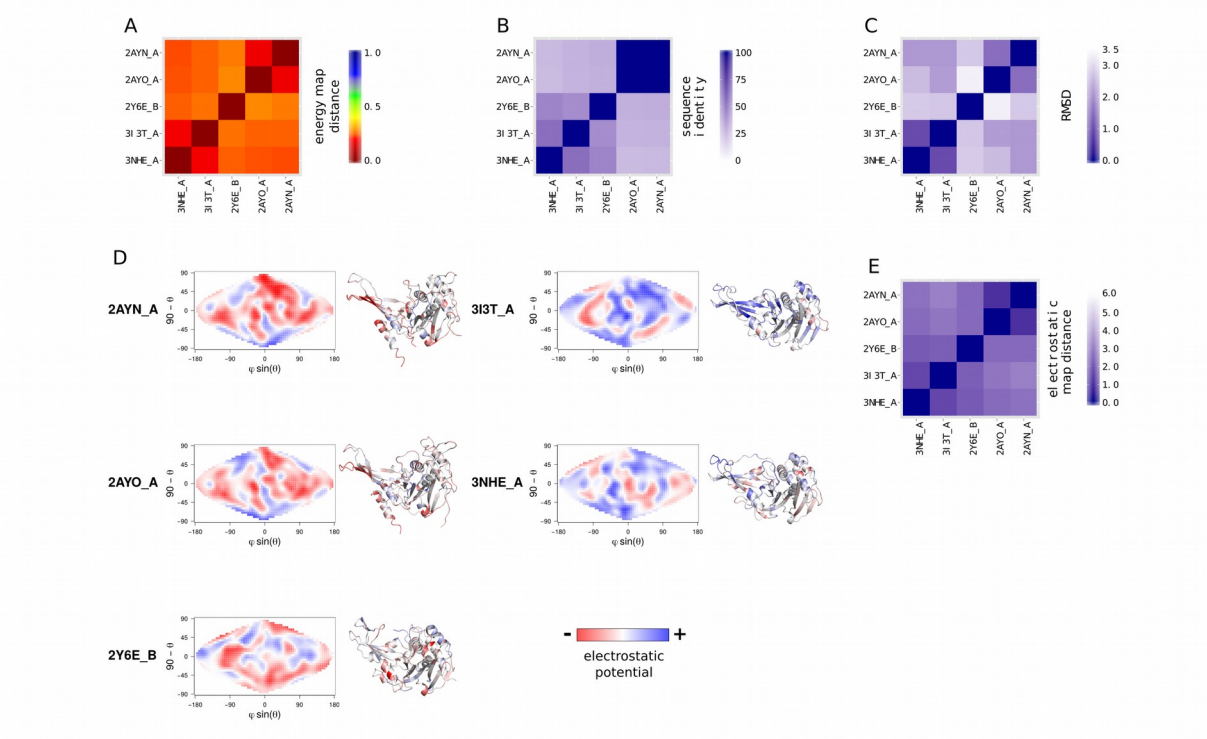
S5 Fig. Eukaryotic-proteases family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the Eukaryotic proteases family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



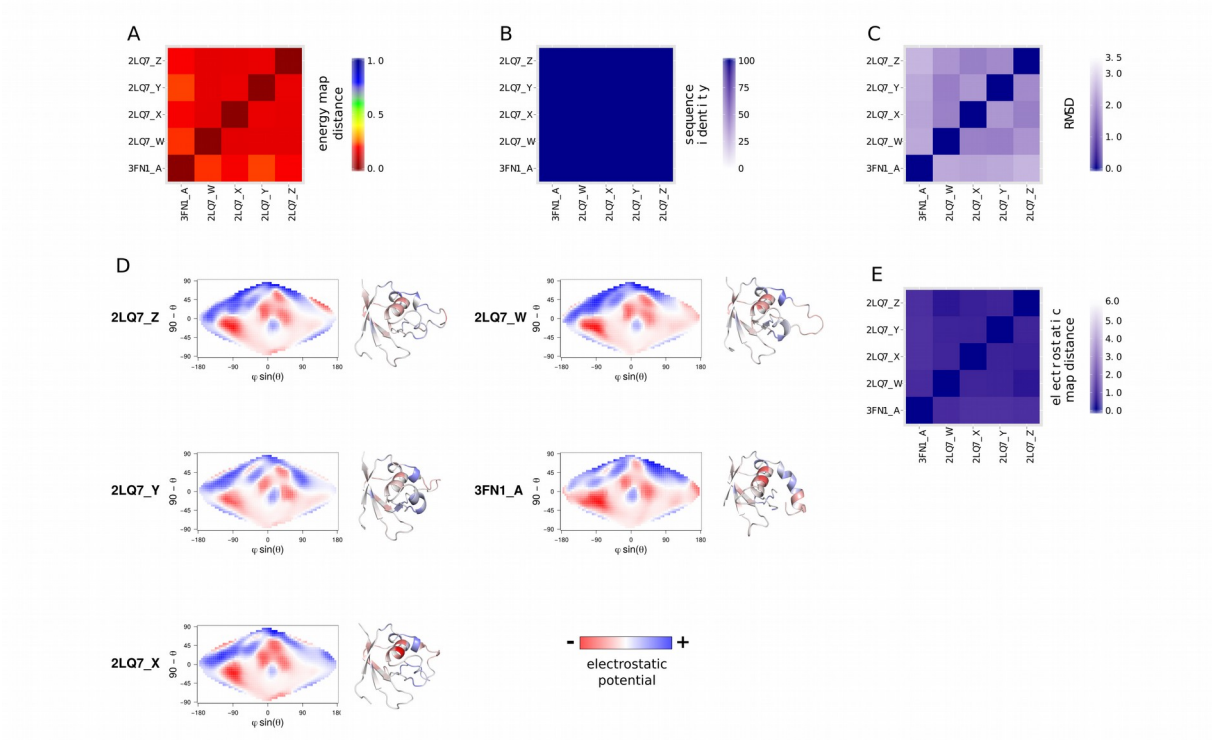
S6 Fig. V set domains family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the V set domain family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the six members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



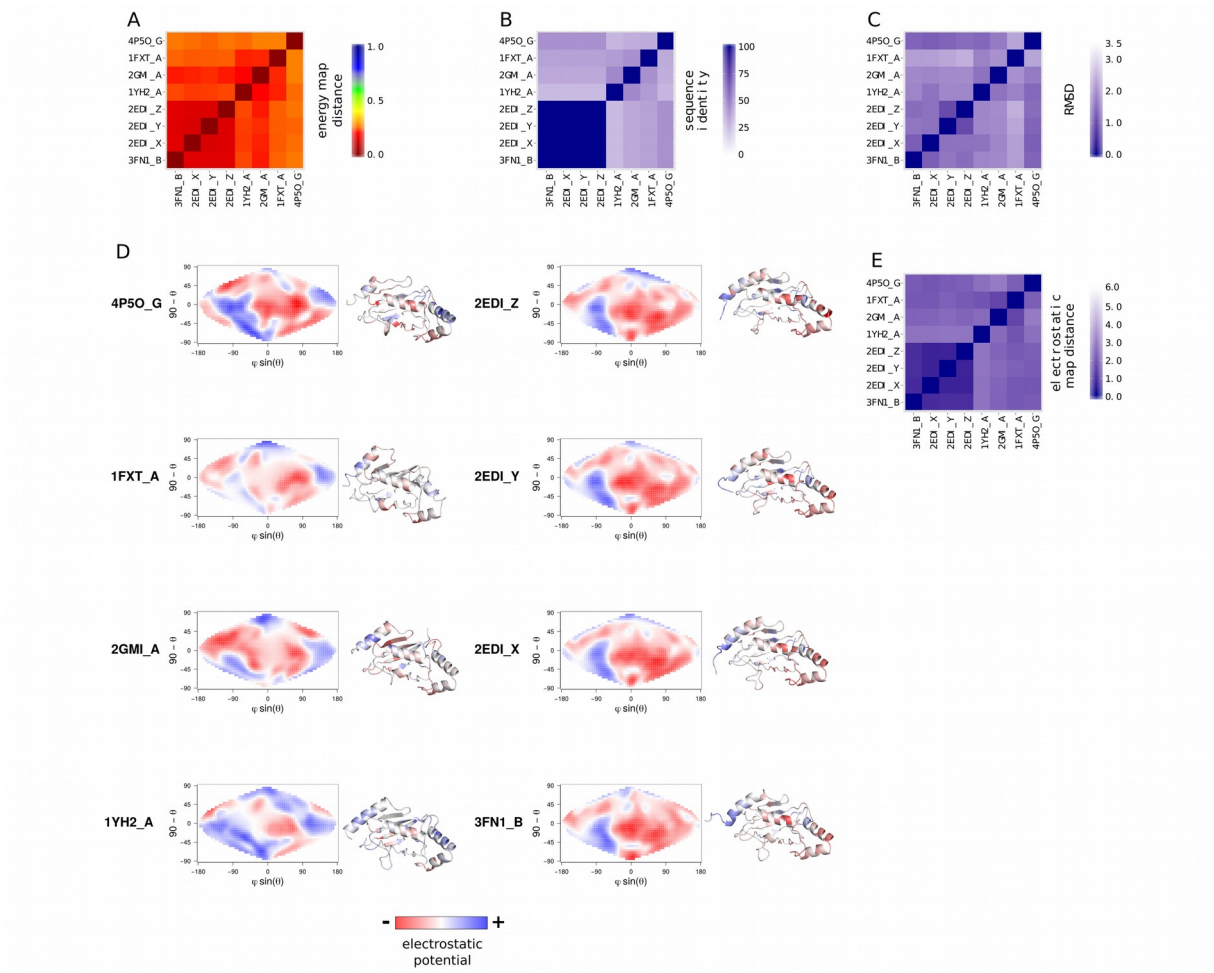
S7 Fig. UCH-L family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the UCH-L family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



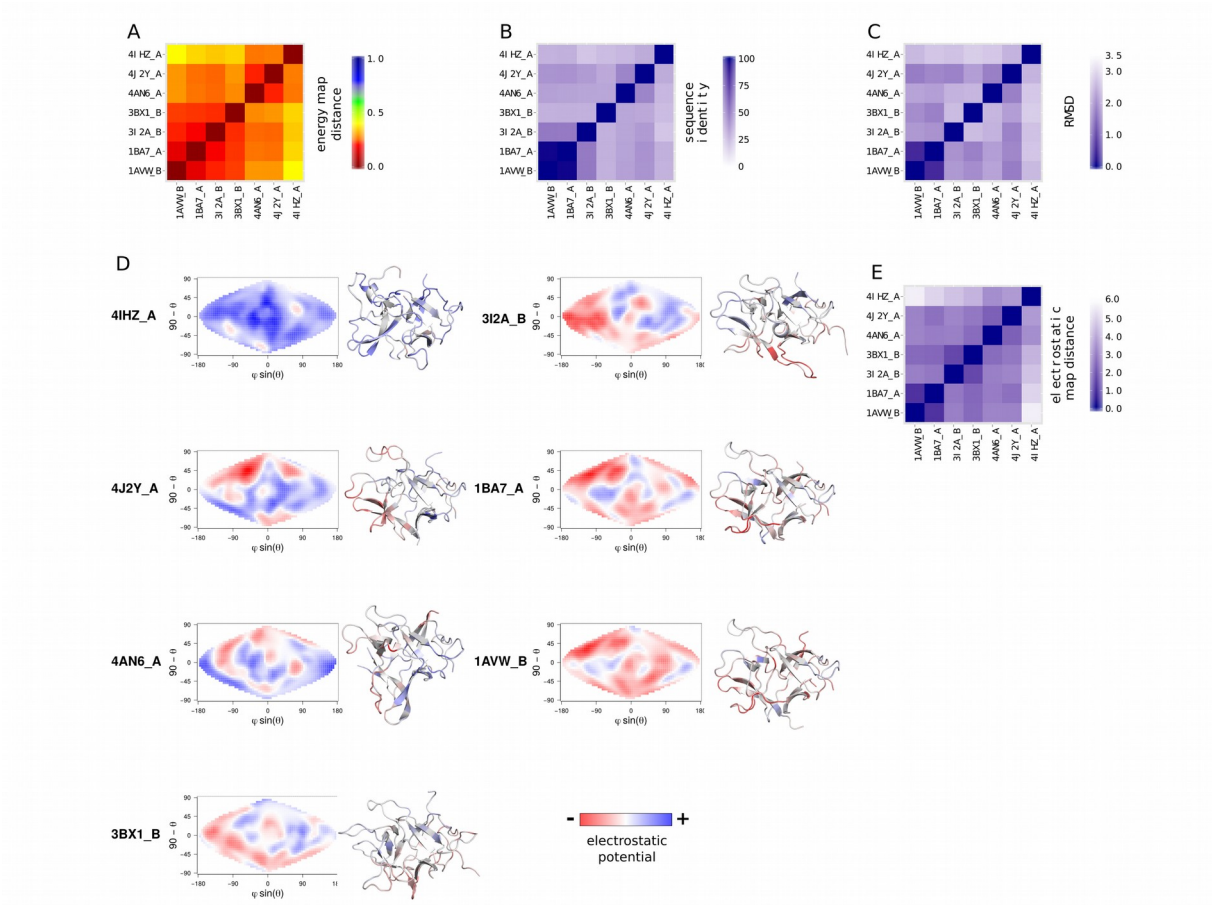
S8 Fig. UCH family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the UCH family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



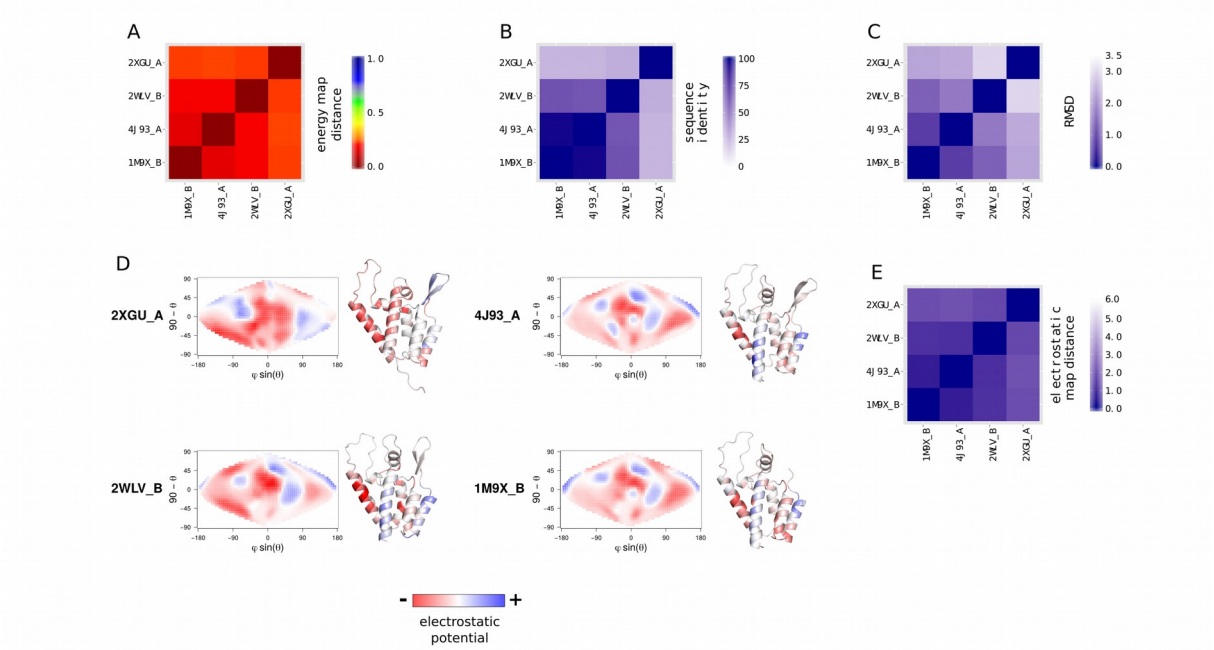
S9 Fig. Ubiquitin activating enzymes family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the Ubiquitin activating enzymes family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



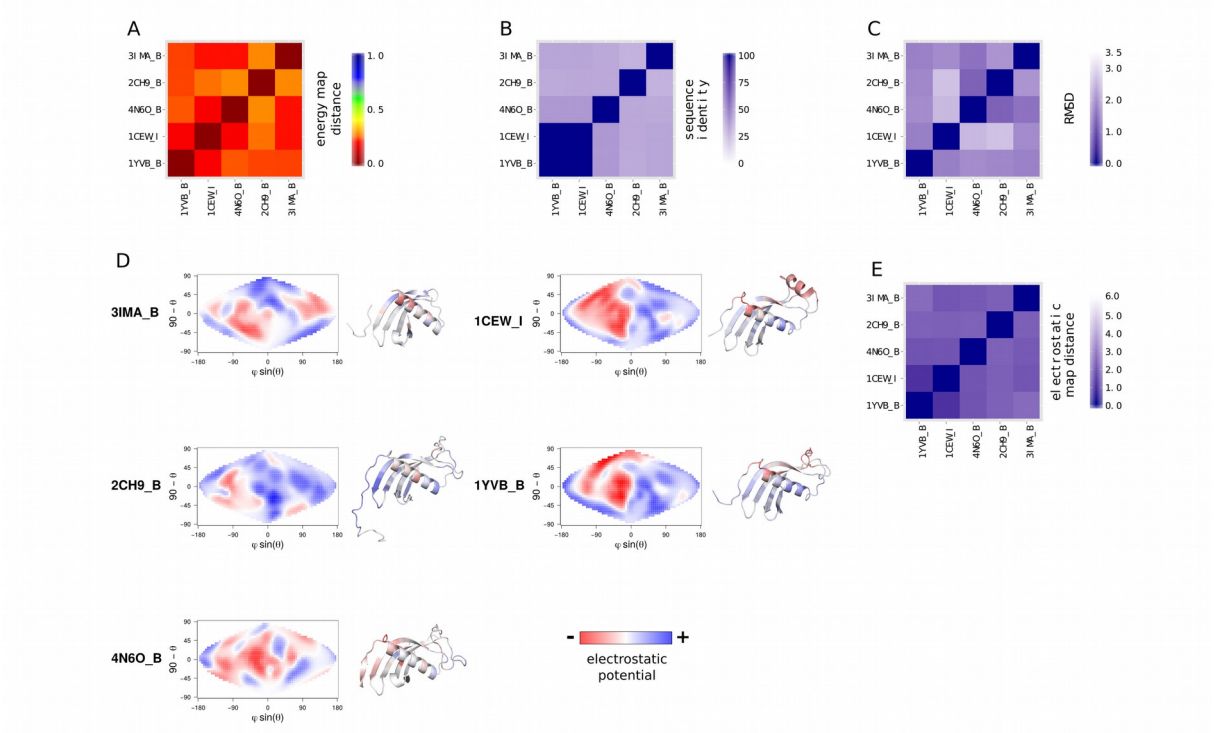
S10 Fig. UBC-related family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the UBC-related family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



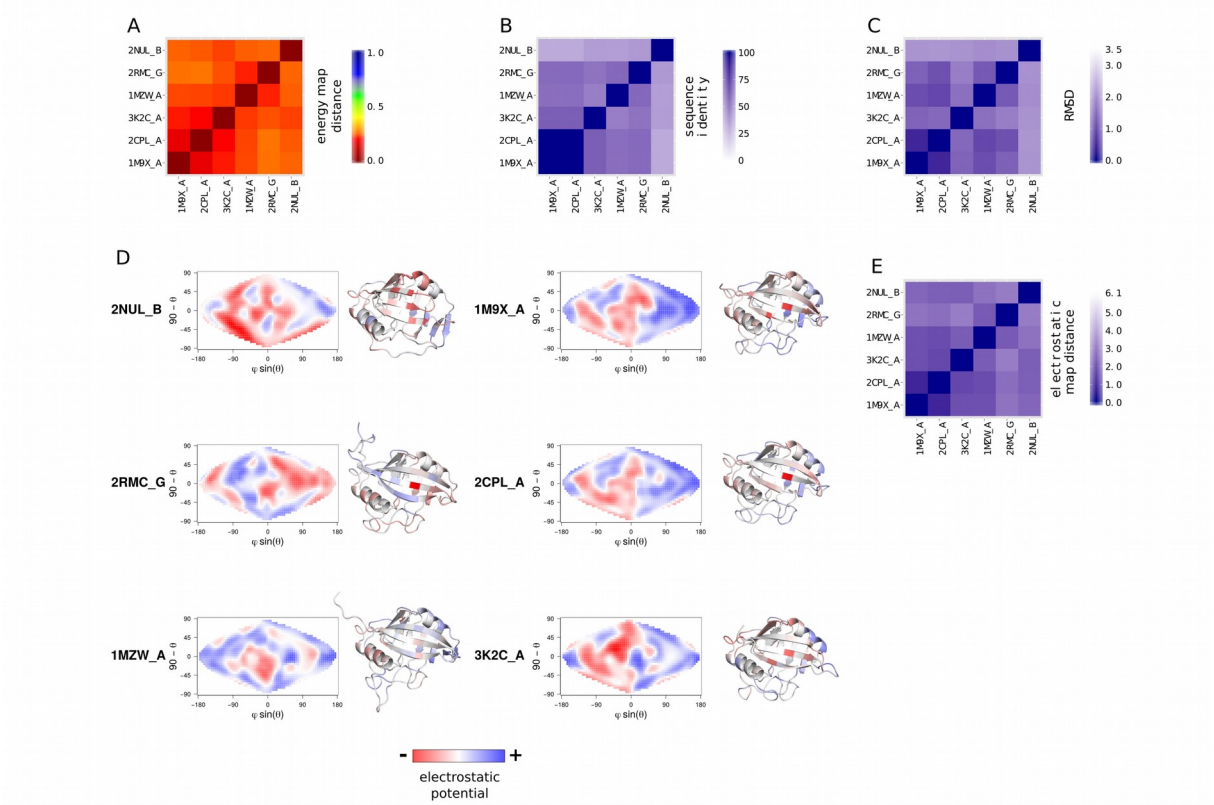
S11 Fig. Kunitz (STI) inhibitors family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the Kunitz (STI) inhibitors family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



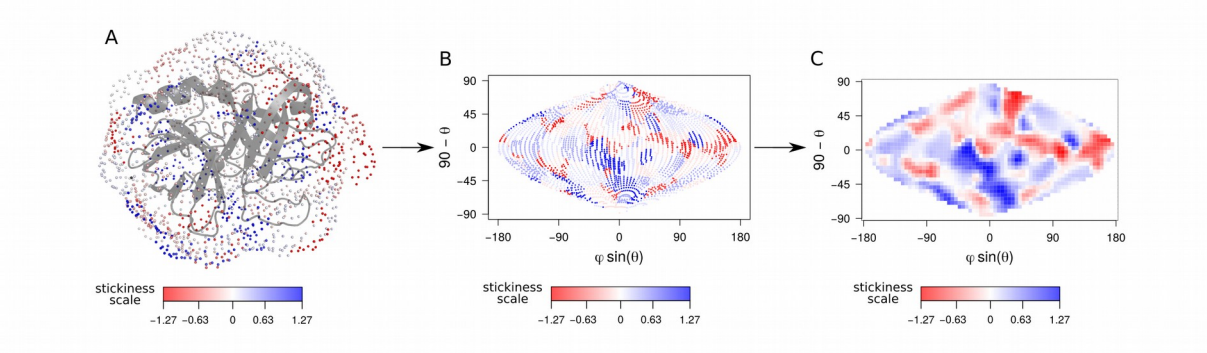
S12 Fig. Retrovirus capsid proteins family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the retrovirus capsid proteins family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



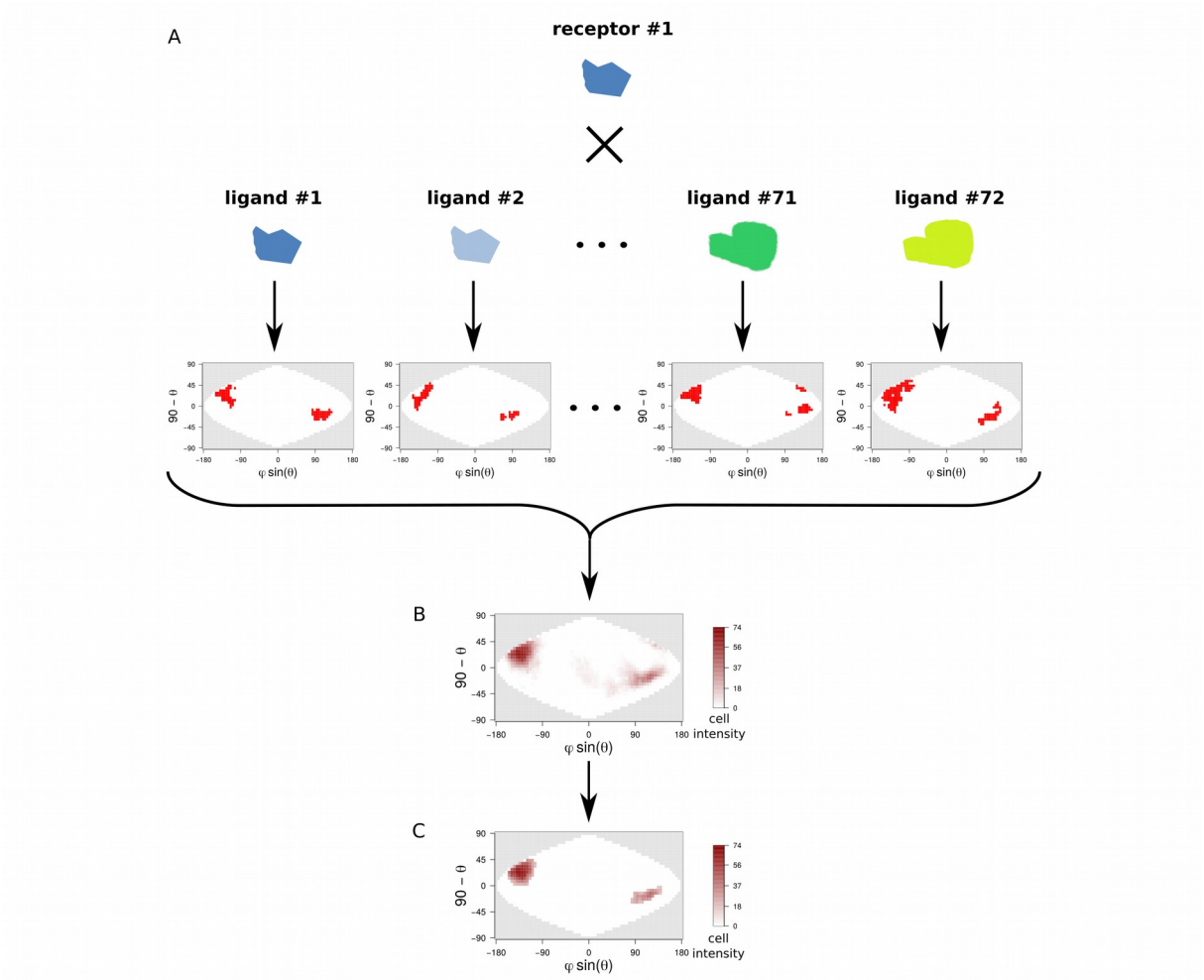
S13 Fig. Cystatins family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the cystatins family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



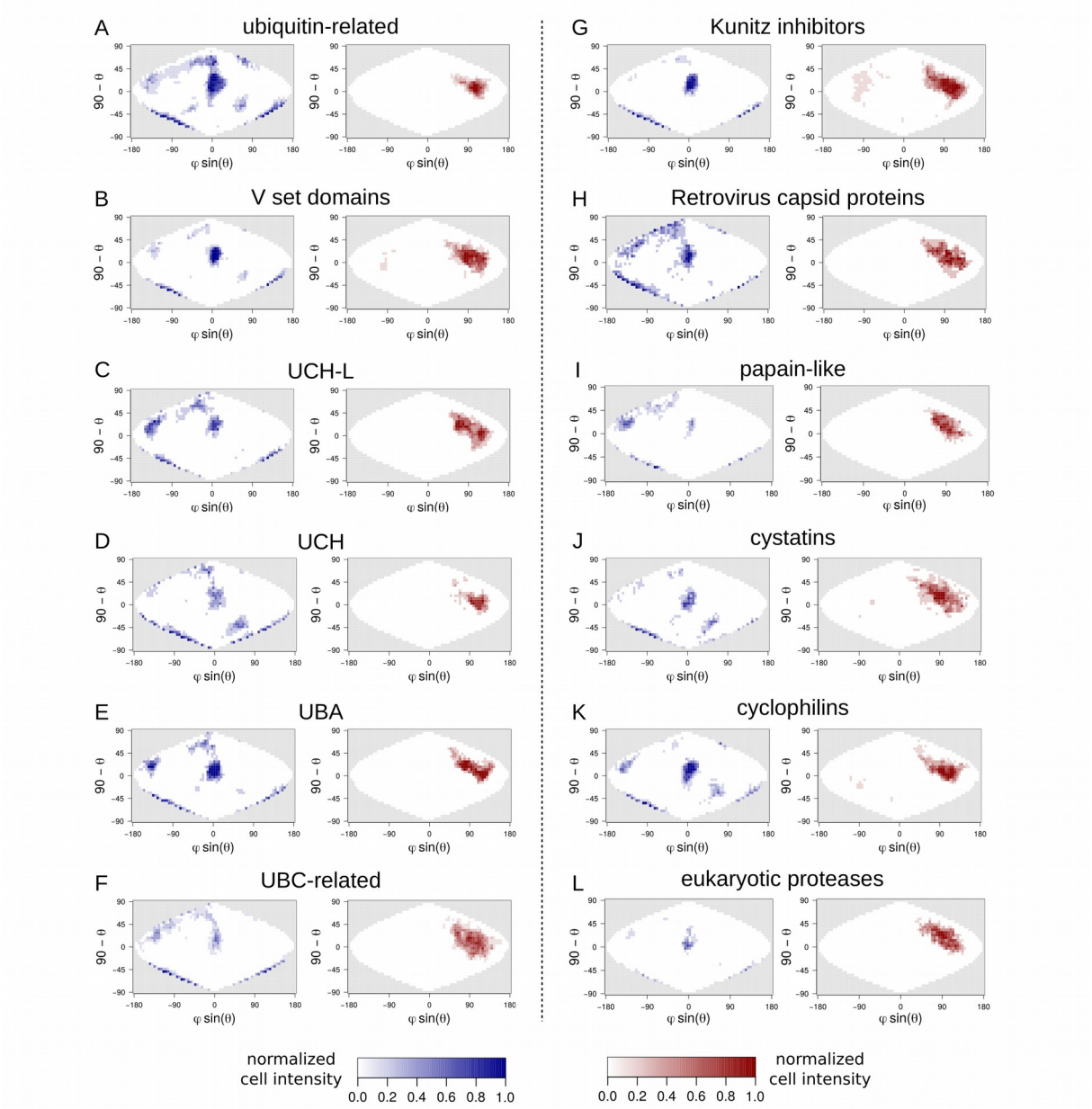
S14 Fig. Cyclophilins family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the cyclophilins family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.



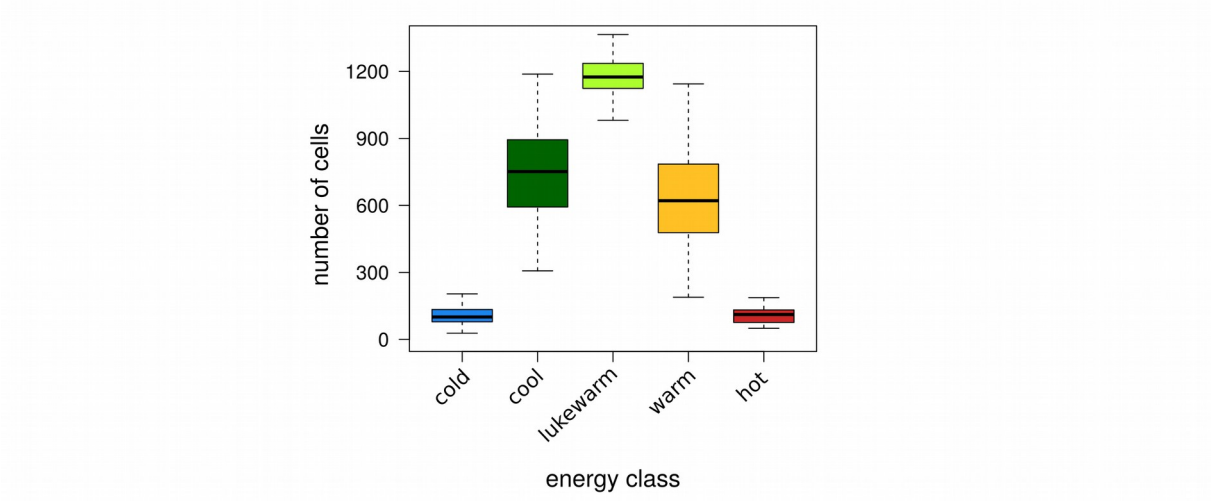
S15 Fig. Generation of electrostatics, stickiness, hydrophobicity and circular variance (CV) maps. Here is presented an example of generation of the stickiness map for the structure 1AVW_A. (*A*) Generation of particles with a slightly modified Shrake-Rupley algorithm [59] around the protein surface, leads to a homogenous shell of particles with a $1Å^2$ density. Each sphere is located at 5Å from the surface of the protein. The stickiness value of the closest atom of the protein is attributed to each particle. In this example, spheres are colored according to the stickiness of the protein surface. The procedure is similar for hydrophobicity and CV. (*B*) The spherical coordinates of each sphere is represented on a 2-D map with an equal-area sinusoidal projection, following the same protocol as described in Fig. 2 and *Materials and Methods*. Each resulting dot is colored according to the same scale of (*A*). (*C*) The map is smoothed following the protocol in Fig. 2 and *Materials and Methods*. The scale is the same as in (*A*).



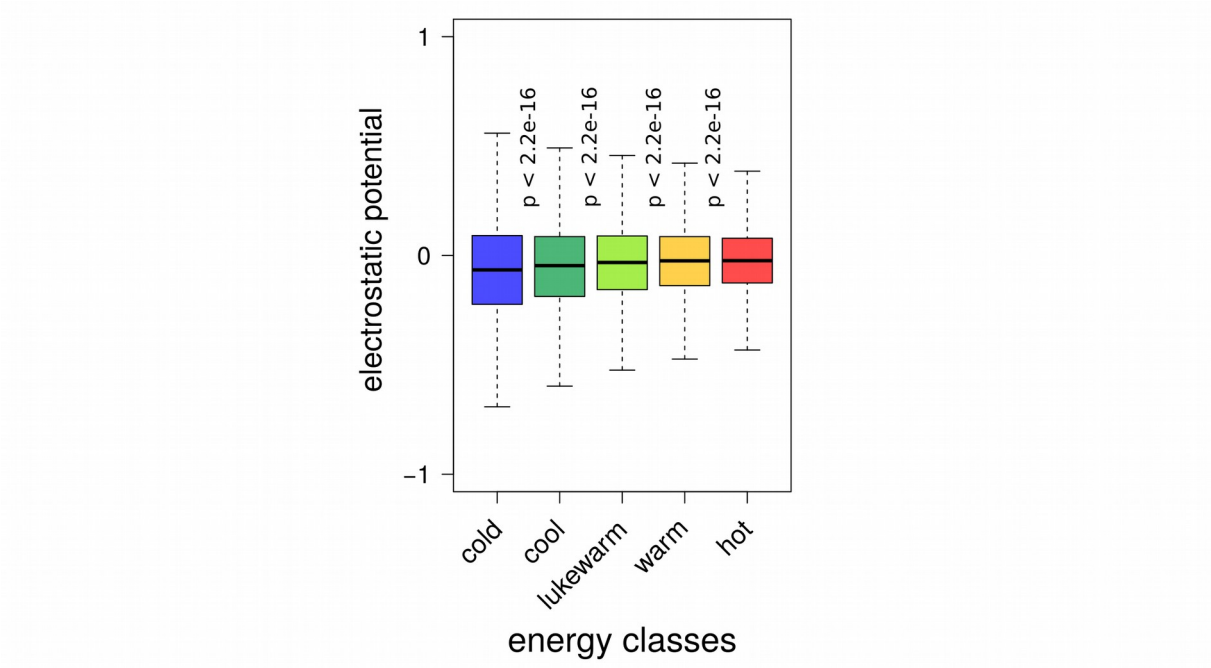
S16 Fig. Generation of stacked maps of a receptor. (*A*) Calculation of the 74 one-color maps (red ones in the example) of receptor #1. A value of one is associated to colored cells while zero is assigned to white cells. (*B*) Sum of the 74 one-color maps into a stacked map. Cell's intensity varies from 0 to 74 and corresponds to the number of time the cell is colored over the 74 ligands. (*C*) Filtering of the cells of low cell intensity (intensity < 17) and areas of too small size (< 4 cells) with a Dirichlet process mixture model simulation for image segmentation [57]. The procedure is repeated for each color stacked map.



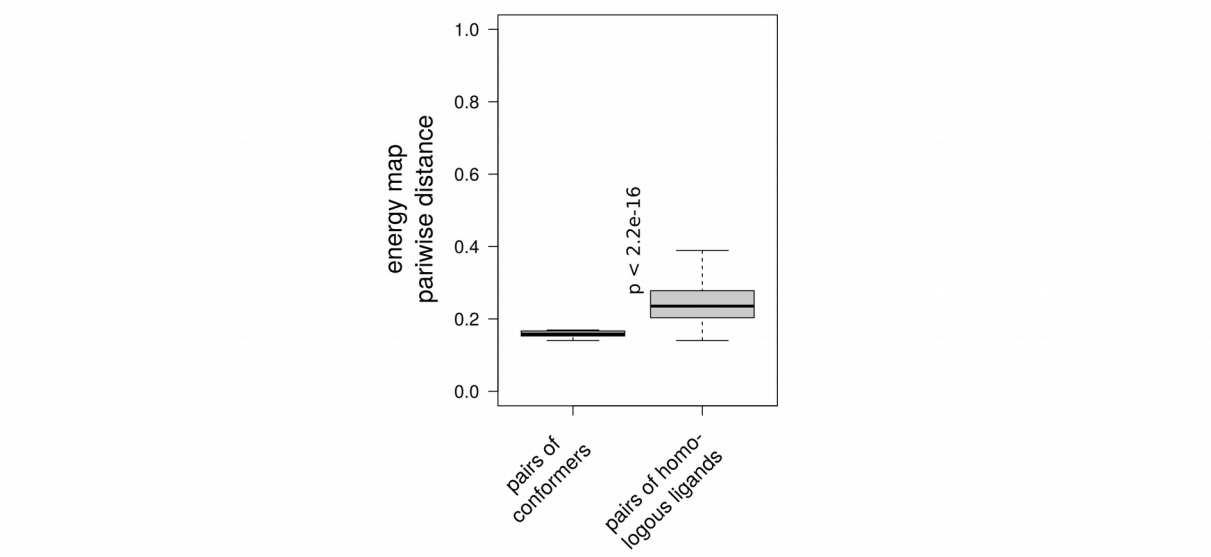
S17 Fig. Blue and red stacked maps of 1P9D_U computed for each ligand family. (*A-L*) We compute the one-color stacked map of each family as the sum of the one-color maps resulting from the docking of each ligand of a same family with 1P9D_U.



S18 Fig. Boxplots of the size (in number of cells) of each energy class for all stacked maps. One should notice that the sum of the sizes of the 5 energy classes is superior to 1548 cells, which is the total size of a map, because a same cell of a stacked map can be assigned to several energy classes (Fig 8).



S19 Fig. Boxplots of the electrostatic potential of the protein surfaces depending on the energy class. The electrostatic potential is calculated for each protein following the protocol described in *Materials and Methods*. *p-values* between the variances of two "successive" energy classes were computed using the F-test.



S20 Fig. Boxplots of energy map pairwise distances between ligand pairs of conformers and pairs of homologous ligands (i.e. non-conformers pairs). For each receptor, we computed (i) the average of energy map distances of pairs of conformers, (ii) the average of energy map distances of pairs of homologous ligands. P-values are calculated with an unilateral Wilcoxon test.

8.7 Remarque sur la métrique employée pour la comparaison des cartes

Après avoir présenté ce travail, nous avons eu des remarques très intéressantes sur le fait qu'il existe d'autres métriques plus adaptées au problème que nous voulions résoudre (comparaison de cartes d'énergie) que la distance de Manhattan que nous employions à l'époque. Suite à ces discussions nous avons testé une nouvelle métrique, la distance du cantonnier (plus connue sous le nom de EMD (*Earth mover's distance*)), pour comparer les cartes d'énergie des protéines.

L'utilisation de cette métrique à l'aide de la *Structural Bioinformatics Library* (219) a permis d'obtenir de meilleurs résultats de prédiction des paires de protéines homologues et non homologues qu'avec la distance de Manhattan. L'AUC obtenue est de 0.81 contre 0.79 pour la distance de Manhattan sur les cartes d'énergie non discrétisées. Comme tous les résultats fournis dans mon article sont ceux obtenus avec la distance de Manhattan, je les ai conservés dans ma thèse.

Si la distance EMD produit de meilleurs résultats c'est que, contrairement à la distance de Manhattan, cette métrique ne compare pas des cartes cellule à cellule : elle recherche tout d'abord dans chaque carte une signature (dans nos cartes, des « zones » neutres ou de faible ou forte énergies) puis compare les signatures des deux cartes. Elle est donc plus adaptée à notre problématique. Sa formule est la suivante :

$$EMD(P,Q)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\,d_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}}$$

où P et Q sont les signatures respectives des deux cartes à comparer. Ces signatures contiennent respectivement n et m zones : $P=\{(p_1,w_{p_1}),\ldots,(p_n,w_{p_n})\}$ et $Q=\{(q_1,w_{q_1}),\ldots,(q_m,w_{q_m})\}$.

p_i et q_j sont respectivement la ième et la jème zone des signatures P et Q et w_{p_i} et w_{q_j} les poids respectifs de ces zones.

d_{ij} est la distance « au sol » entre les zones p_i et q_j, c'est à dire l'écart entre les distributions des niveaux d'énergies des zones p_i et q_j.

f_{ij} est le flot lié au transport de la zone p_i vers la zone q_j. Il est lié à la distance sur la carte entre les zones p_i et q_j et est calculé de façon à minimiser le coût global.

9 Conclusion

L'objectif de ma thèse était de mettre en place un cadre théorique pour pouvoir étudier les interactions protéiques avec un regard systémique, à savoir dans leur environnement cellulaire. Autrement dit, l'objectif était de développer de nouvelles méthodes et systèmes de représentations pour tenter d'aborder la question de la compétition entre interactions fonctionnelles et non-fonctionnelles, en particulier d'étudier comment cette compétition contraint les propriétés de surface des protéines. Cette question était très ambitieuse mais j'ai pu apporter des éléments de réponse quant aux contraintes de l'environnement qui s'exercent sur la surface des protéines et quant à l'évolution du potentiel d'interaction de ces dernières. Par ailleurs, j'ai pu montrer que ce cadre théorique ouvrait la voie à de nombreuses applications.

Comment la compétition entre interactions fonctionnelles et non-fonctionnelles se répercute sur les propriétés des surfaces protéiques

Le cadre théorique que j'ai développé au cours de ma thèse, reposant principalement sur une représentation synthétique du potentiel d'interaction de l'ensemble de la surface d'une protéine avec un jeu de protéines d'intérêt permet d'étudier les interactions protéiques avec une résolution atomique (ou « résidus ») tout en gardant un point de vue systémique. En particulier, j'ai montré que cela permettait d'étudier à la fois l'effet du *positive design* (optimisation des sites d'interaction pour maintenir les assemblages fonctionnels) et *negative design* (contrainte exercée sur le reste de la surface pour que la protéine ne soit pas « piégée » dans des interactions non-fonctionnelles) sur la surface des protéines. J'ai étudié les propriétés physico-chimiques et évolutives non pas des sites d'interaction fonctionnels uniquement mais aussi du reste de la surface qui joue un rôle essentiel dans la régulation des interactions en étant en constante compétition avec les sites d'interaction. J'ai pu montrer que le reste de la surface n'est pas homogène et est constitué de régions présentant différents potentiels d'interaction avec « l'environnement », bien qu'ici cet environnement ait été simulé par un jeu de 100 ligands arbitraires - j'ai montré qu'à partir d'une vingtaine de ligands, les cartes IPOPS étaient robustes. J'ai pu montrer que les différentes régions des surfaces protéiques présentaient des propriétés physico-chimiques et évolutives spécifiques et que leur distribution sur la surface des protéines était spécifique des familles de ligands testées, reflétant le *negative design*

opérant sur le reste de la surface pour limiter les interactions non-fonctionnelles. D'autre part, j'ai montré que les régions favorables à l'interaction étaient très localisées (environ 5% des cellules d'une carte d'énergie) et ont tendance à être similaires quel que soit le ligand. De manière similaire aux sites d'interaction fonctionnels qui ont été largement caractérisés dans des études précédentes (64,77), les régions de la classe d'énergie favorable (rouge) semblent présenter des propriétés universelles de liaison, c'est-à-dire qu'elles sont plus hydrophobes et plus planes, et donc plus "collantes" que les autres régions. Elles fournissent ainsi une surface « d'accueil » à un grand nombre de ligands. Cependant, nous pouvons émettre l'hypothèse que les couples fonctionnels ont coévolué pour optimiser leurs interfaces (*positive design*), de sorte que les interactions fonctionnelles prévalent sur les non-fonctionnelles. En effet, il a été précédemment montré que le *docking* entre protéines formant des couples fonctionnels conduisait à des énergies d'interaction plus favorables que des partenaires arbitraires lorsque le ligand était contraint de s'amarrer sur le site d'interaction expérimental du récepteur (165,166). Autrement dit, bien que le site d'interaction fonctionnel d'un récepteur soit favorable à de nombreux ligands, le partenaire participant à l'assemblage fonctionnel gagne généralement la compétition avec les autres ligands grâce à une interface optimisée pour maintenir l'assemblage fonctionnel. Tous ces résultats suggèrent un nouveau modèle physique des surfaces protéiques où les différentes régions de surface, dans l'environnement cellulaire encombré, participent à la régulation de la compétition entre interactions fonctionnelles et non-fonctionnelles. Dans ce modèle, les régions intermédiaires et de haute classe d'énergie (allant de orange à bleu) jouent un rôle essentiel en empêchant les assemblages non-fonctionnels et en orientant le processus d'assemblage vers les modes d'interaction fonctionnels. Les régions favorables (rouges) sélectionnent alors l'assemblage fonctionnel parmi les concurrentes au travers d'interfaces optimisées avec le partenaire natif.

J'ai aussi pu montrer par l'intermédiaire des cartes IPOPS qui fournissent un regard systémique sur les interactions que les régions favorables à l'interaction se distinguaient en deux types de régions (les îlots rouges ubiquitaires et les îlots rouges spécifiques). Les îlots rouges ubiquitaires correspondent à des régions favorables à un grand nombre de ligands. Elles sont plus hydrophobes, plus collantes et leurs résidus sont plus conservés d'un point de vue évolutif que les régions identifiées par les îlots spécifiques. Cependant, elles sont aussi planes que les régions des îlots rouges spécifiques. J'ai pu montrer que les îlots rouges ubiquitaires et spécifiques correspondent souvent à des sites d'interaction annotés, mais peuvent aussi correspondre à des régions non annotées comme sites d'interaction. Dans ces cas là, nous pouvons alors nous demander dans quelle mesure ces régions correspondent à des sites d'interaction fonctionnels, mais non

annotés dans les unités biologiques des structures de notre base de données, ou encore à des régions avec un fort potentiel d'interaction avec un nombre important de ligands (c'est-à-dire « collantes »). Pour cela il est possible de faire une recherche de l'ensemble des sites d'interaction répertoriés dans la PDB. Cependant, il faut garder à l'esprit que le nombre de sites fonctionnels est largement sous-estimé dans la PDB (13). Dans le cas ou les îlots détectés ne correspondent pas à des interfaces répertoriées, il pourrait être judicieux de corroborer la localisation de ces îlots avec un logiciel tel que JET (200). L'avantage d'un tel outil est qu'il repose sur la détection de régions de surface dont la séquence est conservée du point de évolutif. Nous pouvons raisonnablement émettre l'hypothèse que les sites d'interaction prédits par JET sont soumis à la pression de sélection et donc jouent un rôle fonctionnel. Cependant il faut garder à l'esprit que toutes les interfaces fonctionnelles ne présentent pas nécessairement de trace de conservation évolutive (c'est le cas par exemple des interactions anticorps/antigènes). Quand elles ne correspondent pas à des sites d'interaction fonctionnels, les régions identifiées par les îlots rouges peuvent correspondre à des régions dites « collantes », seulement cette hypothèse n'est pas simple à vérifier. Il serait intéressant de valider expérimentalement ces cas là. Pour ce faire, nous pourrions par exemple regarder s'ils correspondent à des contacts cristallins (souvent non-fonctionnels mais représentant des modes alternatifs d'interaction), ou encore à des régions vues en interaction par des expériences de *cross-link*. De façon intéressante, j'ai pu montrer que les îlots rouges ubiquitaires correspondaient plus souvent à des sites d'interaction fonctionnels que les îlots rouges spécifiques. L'analyse sur les homomères (section 7) a montré que les îlots rouges spécifiques pouvaient correspondre à des sites fonctionnels plus récents (sites d'interaction « spécifiques » des assemblages diédraux).

Il reste encore beaucoup de travail à réaliser pour mieux décortiquer les propriétés physico-chimiques, évolutives et fonctionnelles de ces deux types de régions mais l'hypothèse que l'on peut émettre est que les îlots rouges ubiquitaires correspondent majoritairement à des sites d'interaction fonctionnels, plus hydrophobes, plus plans et plus conservés (au niveau de la séquence des résidus qui les composent) que le reste de la surface. Ils présentent des propriétés qui les rendent attractifs pour un grand nombre de ligands, mais l'interface qu'ils forment avec leurs partenaires fonctionnels est optimisée de façon à ce que les assemblages fonctionnels prennent le dessus sur les assemblages non-fonctionnels. Les îlots rouges spécifiques présentent des propriétés de surface différentes des îlots ubiquitaires (hormis pour le relief de la surface) et plus proches des régions de la classe d'énergie intermédiaire (régions orange). Ces derniers peuvent soit correspondre à des sites d'interaction fonctionnels, peut-être récents ou nécessitant des propriétés différentes des îlots ubiquitaires, ou à des régions de fort potentiel d'interaction avec un sous-ensemble de protéines de

l’environnement qui n’ont pas été optimisées et qui ne présentent donc pas toutes les propriétés des sites d’interaction canoniques.

Le cadre théorique que j’ai développé ouvre la voie à de nombreuses applications

J’ai montré comment les cartes IPOPS rouges ouvrent la voie au développement de nouvelles méthodes de prédiction de sites d’interaction et à leur classification en site d’interaction fonctionnels ou « collants ». Les îlots rouges pourraient aussi être utilisés comme contraintes dans les algorithmes de *docking* pour restreindre l’espace conformationnel de recherche. La prochaine étape consisterait à étudier le recouvrement de ces îlots (ubiquitaires et spécifiques) avec les différentes régions des sites d’interaction présentées dans (77), à savoir le *rim*, le *core* et le *support*. De même, il serait intéressant d’étudier si les résidus appartenant à des îlots rouges mais pas à des sites d’interaction sont localisés dans les bordures d’îlots rouges correspondant à des sites d’interaction ou au contraire d’îlots rouges n’ayant aucun recouvrement avec un site d’interaction.

Par ailleurs, je n’ai pas eu le temps de l’exploiter pleinement (mis à part les quels cas documentés dans la section 8) mais la comparaison des cartes IPOPS calculées pour les membres d’une même famille d’homologue fournit un point d'entrée pour une caractérisation fonctionnelle supplémentaire de ces homologues car elle peut révéler des propriétés biophysiques et fonctionnelles des protéines qui n'auraient pas pu être révélées avec des descripteurs classiques tels que le RMSD ou l'identité de séquence (voir section 8). Le prochain travail à faire dans ce sens est d’étudier les cas de familles hétérogènes (cas où certains membres présentent des cartes IPOPS différentes) parmi les 81 familles du jeu des 348 récepteurs. Nous pouvons émettre l’hypothèse que ces protéines présentant des potentiels d’interaction différents interagissent avec des partenaires différents.

L'ensemble de données utilisées dans ce travail devait répondre à des contraintes liées à la structure des protéines et aux familles structurales et/ou d’interologues disponibles et ne représentait donc pas un véritable environnement cellulaire encombré. J’ai tout de même pu montrer que notre stratégie permet d’explorer le potentiel d’une protéine à interagir avec des centaines de partenaires sélectionnés, et donc d’aborder le comportement d’une protéine dans un environnement cellulaire spécifique. J’ai donc en parallèle de ces travaux réalisé des calculs de *docking* croisé sur l’ensemble des protéines du cytosol de *S. cerevisiae* dont la structure (expérimentale ou modélisée) était disponible (600 protéines). Les structures ont été extraites de la base de données Interactome3D

(220). L’objectif à moyen terme de ce projet est de caractériser le comportement d’une protéine avec des protéines d’un même environnement cellulaire. Pour l’instant la quasi totalité des calculs de *docking* a été réalisée. La quantité de calculs réalisés est d’environ 2 000 000 d’heures, que nous avons pu réaliser grâce à trois allocations DARI nous donnant accès aux grilles nationales de calculs Occigen et TGCC. Même si je n’ai pas eu le temps d’analyser les données issues de ces calculs, je pense que la méthodologie que j’ai mise au point est un premier pas nécessaire pour pouvoir exploiter ce jeu de données.

Ce cadre théorique va au-delà de l'utilisation classique du *docking* « binaire » protéine-protéine pour fournir un point de vue systémique des interactions entre protéines avec une résolution à l’échelle du résidu et ouvre ainsi la voie à de nouveaux développements pour la caractérisation et la compréhension du fonctionnement des protéines dans un environnement encombré.

Références

1. Vickery HB. Origin of the word protein. Nature. 1951;(168):244.

2. Vauquelin L-N, Robiquet, Pierre Jean. The discovery of a new plant principle in Asparagus sativus. Ann Chim. 1806;57:88–93.

3. Meyer CE, Rose WC. the spatial configuration of a-amino-b-hydroxy-n-butyric acid. J Biol Chem. 1936 Jul 20;115(3):721–129.

4. Sanger F. The terminal peptides of insulin. Biochem J. 1949;45(5):563–74.

5. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature. 1958 Mar;181(4610):662–6.

6. Perutz MF, Rossmann MG, Cullis AF, Muirhead H, Will G, North ACT. Structure of Hæmoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. Nature. 1960 Feb;185(4711):416–22.

7. Blake CC, Koenig DF, Mair GA, North AC, Phillips DC, Sarma VR. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. Nature. 1965 May 22;206(4986):757–61.

8. Kartha G, Bello J, Harker D. Tertiary Structure of Ribonuclease. Nature. 1967 Mar;213(5079):862–5.

9. Wyckoff HW, Hardman KD, Allewell NM, Inagami T, Johnson LN, Richards FM. The Structure of Ribonuclease-S at 3.5 A Resolution. J Biol Chem. 1967 Sep 10;242(17):3984–8.

10. Drenth J, Jansonius JN, Koekoek R, Swen HM, Wolthers BG. Structure of Papain. Nature. 1968 Jun;218(5145):929–32.

11. Perutz MF, Muirhead H, Cox JM, Goaman LCG. Three-dimensional Fourier Synthesis of Horse Oxyhaemoglobin at 2.8 Å Resolution: The Atomic Model. Nature. 1968 Jul;219(5150):131–9.

12. Blundell TL, Cutfield JF, Cutfield SM, Dodson EJ, Dodson GG, Hodgkin DC, et al. Atomic positions in rhombohedral 2-zinc insulin crystals. Nature. 1971 Jun 25;231(5304):506–11.

13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

14. Braun P, Gingras A-C. History of protein-protein interactions: from egg-white to complex networks. Proteomics. 2012 May;12(10):1478–98.

15. Hedin SG. Trypsin and Antitrypsin. Biochem J. 1906;1(10):474–83.

16. Svedberg T. Mass and Size of Protein Molecules. Nature. 1929 Jun;123(3110):871.

17. Svedberg T, Chirnoaga E. THE MOLECULAR WEIGHT OF HEMOCYANIN. J Am Chem Soc. 1928 May 1;50(5):1399–411.

18. van Holde KE. Reflections on a century of protein chemistry. Biophys Chem. 2002 Dec;100(1–3):71–9.

19. Sanadi DR, Littlefield JW, Bock RM. STUDIES ON α-KETOGLUTARIC OXIDASE II. PURIFICATION AND PROPERTIES. J Biol Chem. 1952 Aug 1;197(2):851–62.

20. Koike M, Reed LJ, Carroll WR. α-Keto Acid Dehydrogenation Complexes I. PURIFICATION AND PROPERTIES OF PYRUVATE AND α-KETOGLUTARATE DEHYDROGENATION COMPLEXES OF ESCHERICHIA COLI. J Biol Chem. 1960 Jul 1;235(7):1924–30.

21. Laemmli UK. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. Nature. 1970 Aug;227(5259):680–5.

22. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988 Jan 29;239(4839):487–91.

23. Fischer EH. Phosphorylase and the origin of reversible protein phosphorylation. Biol Chem. 2010 Mar;391(2–3):131–7.

24. Pawson T, Scott JD. Protein phosphorylation in signaling – 50 years and counting. Trends Biochem Sci. 2005 Jun 1;30(6):286–90.

25. Alberts B, Miake-Lye R. Unscrambling the puzzle of biological machines: The importance of the details. Cell. 1992 Feb;68(3):415–20.

26. Fields S, Song O. A novel genetic system to detect protein-protein interactions. Nature. 1989 Jul 20;340(6230):245–6.

27. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, et al. Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci. 2000 Feb 1;97(3):1143–7.

28. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci. 2001 Apr 10;98(8):4569–74.

29. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, et al. A comprehensive analysis of protein–protein interactions inSaccharomyces cerevisiae. Nature. 2000 Feb;403(6770):623–7.

30. Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002 Jan;415(6868):141–7.

31. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature. 2002 Jan 10;415(6868):180–3.

32. Goll J, Uetz P. The elusive yeast interactome. Genome Biol. 2006;7(6):223.

33. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature. 2006 Mar;440(7084):637–43.

34. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, et al. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. Science. 2008 Oct 3;322(5898):104–10.

35. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006 Mar 30;440(7084):631–6.

36. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, et al. Towards a proteome-scale map of the human protein–protein interaction network. Nature. 2005 Oct;437(7062):1173–8.

37. Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, et al. A Proteome-Scale Map of the Human Interactome Network. Cell. 2014 Nov;159(5):1212–26.

38. Snider J, Kotlyar M, Saraon P, Yao Z, Jurisica I, Stagljar I. Fundamentals of protein interaction network mapping. Mol Syst Biol. 2015 Dec 17;11(12):848–848.

39. Agapito G, Guzzi PH, Cannataro M. Visualization of protein interaction networks: problems and solutions. BMC Bioinformatics. 2013;14 Suppl 1:S1.

40. Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. Nat Biotechnol. 2009 Oct;27(10):921–4.

41. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan;43(Database issue):D447-452.

42. Aloy P, Russell RB. Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol. 2006 Mar;7(3):188–97.

43. Nooren IMA, Thornton JM. NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. EMBO J. 2003 Jul 15;22(14):3486–92.

44. Acuner Ozbabacan SE, Engin HB, Gursoy A, Keskin O. Transient protein-protein interactions. Protein Eng Des Sel. 2011 Sep 1;24(9):635–48.

45. Kastritis PL, Bonvin AM. Molecular origins of binding affinity: seeking the Archimedean point. Curr Opin Struct Biol. 2013 Dec;23(6):868–77.

46. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. Structure. 2010 Oct;18(10):1233–43.

47. Grubb GJJ. Hitting the Hot Spots of Cell Signaling Cascades. Science. 2006 Apr 21;312(5772):377–8.

48. Rudolph MJ, Wuebbens MM, Rajagopalan KV, Schindelin H. Crystal structure of molybdopterin synthase and its evolutionary relationship to ubiquitin activation. Nat Struct Biol. 2001 Jan;8(1):42–6.

49. Levy ED, Teichmann SA. Chapter Two - Structural, Evolutionary, and Assembly Principles of Protein Oligomerization. In: Progress in Molecular Biology and Translational Science. Academic Press; 2013. p. 25–51. (Oligomerization in Health and Disease; vol. 117).

50. Marsh JA, Rees HA, Ahnert SE, Teichmann SA. Structural and evolutionary versatility in protein complexes with uneven stoichiometry. Nat Commun. 2015 Dec;6(1).

51. Sanghani PC, Robinson H, Bosron WF, Hurley TD. Human glutathione-dependent formaldehyde dehydrogenase. Structures of apo, binary, and inhibitory ternary complexes. Biochemistry. 2002 Sep 3;41(35):10778–86.

52. Chothia C, Janin J. Principles of protein–protein recognition. Nature. 1975 Aug;256(5520):705–8.

53. Janin J, Chothia C. Stability and specificity of protein-protein interactions: The case of the trypsin-trypsin inhibitor complexes. J Mol Biol. 1976 Jan;100(2):197–211.

54. Argos P. An investigation of protein subunit and domain interfaces. Protein Eng Des Sel. 1988;2(2):101–13.

55. Janin J, Chothia C. The structure of protein-protein recognition sites. J Biol Chem. 1990 Sep 25;265(27):16027–30.

56. Nooren IM., Thornton JM. Structural Characterisation and Functional Significance of Transient Protein–Protein Interactions. J Mol Biol. 2003 Jan;325(5):991–1018.

57. Jones S, Thornton JM. Analysis of Protein-Protein Interaction Sites using Surface Patches. J Mol Biol. 1997 Sep 12;272(1):121–32.

58. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins Struct Funct Genet. 2002 May 15;47(3):334–43.

59. Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol. 1999 Feb 5;285(5):2177–98.

60. Ofran Y, Rost B. Analysing Six Types of Protein–Protein Interfaces. J Mol Biol. 2003 Jan;325(2):377–87.

61. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci. 2005 Aug 2;102(31):10930–5.

62. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci U S A. 1996 Jan 9;93(1):13–20.

63. Janin J. Principles of protein-protein recognition from structure to thermodynamics. :9.

64. Janin J, Bahadur RP, Chakrabarti P. Protein–protein interaction and quaternary structure. Q Rev Biophys. 2008 May;41(02).

65. Reichmann D, Rahat O, Cohen M, Neuvirth H, Schreiber G. The molecular architecture of protein–protein binding sites. Curr Opin Struct Biol. 2007 Feb;17(1):67–76.

66. Tsai C-J, Lin SL, Wolfson HJ, Nussinov R. A Dataset of Protein–Protein Interfaces Generated with a Sequence-order-independent Comparison Technique. J Mol Biol. 1996 Jul;260(4):604–20.

67. Tsai C-J, Lin SL, Wolfson HJ, Nussinov R. Protein-Protein Interfaces: Architectures and Interactions in Protein-Protein Interfaces and in Protein Cores. Their Similarities and Differences. Crit Rev Biochem Mol Biol. 1996 Jan;31(2):127–52.

68. Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. J Mol Biol. 1971 Feb 14;55(3):379-IN4.

69. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol. 1973 Sep 15;79(2):351–71.

70. Prasad Bahadur R, Chakrabarti P, Rodier F, Janin J. A Dissection of Specific and Non-specific Protein–Protein Interfaces. J Mol Biol. 2004 Feb;336(4):943–55.

71. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci. 2008 Dec 31;3(5):717–29.

72. Tsai C-J, Lin SL, Wolfson HJ, Nussinov R. Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. Protein Sci. 2008 Dec 31;6(1):53–64.

73. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998 Jul;280(1):1–9.

74. Keskin O, Ma B, Nussinov R. Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues. J Mol Biol. 2005 Feb;345(5):1281–94.

75. Li X, Keskin O, Ma B, Nussinov R, Liang J. Protein–Protein Interactions: Hot Spots and Structurally Conserved Residues often Locate in Complemented Pockets that Pre-organized in the Unbound States: Implications for Docking. J Mol Biol. 2004 Nov;344(3):781–95.

76. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci. 2003 May 13;100(10):5772–7.

77. Levy ED. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. J Mol Biol. 2010 Nov;403(4):660–70.

78. Chothial C, Lesk AM. The relation between the divergence of sequence and structure in proteins. 1986;5(4):823–6.

79. Aloy P, Ceulemans H, Stark A, Russell RB. The Relationship Between Sequence and Interaction Divergence in Proteins. J Mol Biol. 2003 Oct;332(5):989–98.

80. Faure G, Andreani J, Guerois R. InterEvol database: exploring the structure and evolution of protein complex interfaces. Nucleic Acids Res. 2012 Jan;40(D1):D847–56.

81. Andreani J, Faure G, Guerois R. Versatility and Invariance in the Evolution of Homologous Heteromeric Interfaces. PLoS Comput Biol. 2012 Aug 30;8(8):e1002677.

82. Pazos F, Valencia A. Protein co-evolution, co-adaptation and interactions. EMBO J. 2008 Oct 22;27(20):2648–55.

83. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol. 1997 Aug 29;271(4):511–23.

84. Andreani J, Guerois R. Evolution of protein interactions: From interactomes to interfaces. Arch Biochem Biophys. 2014 Jul;554:65–75.

85. Tonddast-Navaei S, Skolnick J. Are protein-protein interfaces special regions on a protein's surface? J Chem Phys. 2015 Dec 28;143(24):243149.

86. Illmer P, Erlebach C, Schinner F. A practicable and accurate method to differentiate between intra- and extracellular water of microbial cells. FEMS Microbiol Lett. 1999 Sep 1;178(1):135–9.

87. Gombert AK, Moreira dos Santos M, Christensen B, Nielsen J. Network Identification and Flux Quantification in the Central Metabolism of Saccharomyces cerevisiae under Different Conditions of Glucose Repression. J Bacteriol. 2001 Feb 15;183(4):1441–51.

88. Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values: Insights & Perspectives. BioEssays. 2013 Dec;35(12):1050–5.

89. Jorgensen P. Systematic Identification of Pathways That Couple Cell Growth and Division in Yeast. Science. 2002 Jul 19;297(5580):395–400.

90. Ho B, Baryshnikova A, Brown GW. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. Cell Syst. 2018 Feb 28;6(2):192-205.e3.

91. Levy ED, Kowarzyk J, Michnick SW. High-Resolution Mapping of Protein Concentration Reveals Principles of Proteome Architecture and Adaptation. Cell Rep. 2014 May;7(4):1333–40.

92. McGuffee SR, Elcock AH. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. PLoS Comput Biol. 2010 Mar 5;6(3):e1000694.

93. Yu I, Mori T, Ando T, Harada R, Jung J, Sugita Y, et al. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. eLife. 2016 Nov 1;5.

94. Harada R, Tochio N, Kigawa T, Sugita Y, Feig M. Reduced Native State Stability in Crowded Cellular Environment Due to Protein–Protein Interactions. J Am Chem Soc. 2013 Mar 6;135(9):3696–701.

95. Monteith WB, Cohen RD, Smith AE, Guzman-Cisneros E, Pielak GJ. Quinary structure modulates protein stability in cells. Proc Natl Acad Sci. 2015 Feb 10;112(6):1739–42.

96. McConkey EH. Molecular evolution, intracellular organization, and the quinary structure of proteins. Proc Natl Acad Sci. 1982 May 1;79(10):3236–40.

97. Zhang J, Maslov S, Shakhnovich EI. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. Mol Syst Biol. 2008 Aug 5;4.

98. Johnson ME, Hummer G. Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. Proc Natl Acad Sci. 2011 Jan 11;108(2):603–8.

99. Pal C, Papp B, Hurst LD. Highly Expressed Genes in Yeast Evolve Slowly. 2001;158:927–31.

100. Popescu CE, Borza T, Bielawski JP, Lee RW. Evolutionary Rates and Expression Level in Chlamydomonas. Genetics. 2006 Mar;172(3):1567–76.

101. Rocha EPC, Danchin A. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. Mol Biol Evol. 2004 Jan;21(1):108–16.

102. Drummond DA, Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008 Jul;134(2):341–52.

103. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci. 2005 Oct 4;102(40):14338–43.

104. Drummond DA, Raval A, Wilke CO. A Single Determinant Dominates the Rate of Yeast Protein Evolution. Mol Biol Evol. 2006 Feb 1;23(2):327–37.

105. Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci. 2012 Apr 3;109(14):E831–40.

106. Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci. 2012 Dec 11;109(50):20461–6.

107. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Mol Biol. 1996 Oct;3(10):842–8.

108. Schavemaker PE, Śmigiel WM, Poolman B. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. eLife. 2017 Nov 20;6:e30084.

109. Johansson H, Jensen MR, Gesmar H, Meier S, Vinther JM, Keeler C, et al. Specific and Nonspecific Interactions in Ultraweak Protein–Protein Associations Revealed by Solvent Paramagnetic Relaxation Enhancements. J Am Chem Soc. 2014 Jul 23;136(29):10277–86.

110. Levitt M. The birth of computational structural biology. Nat Struct Mol Biol. 2001 May;8(5):392–3.

111. Schlick T, Collepardo-Guevara R, Halvorsen LA, Jung S, Xiao X. Biomolecular modeling and simulation: a field coming of age. Q Rev Biophys. 2011 May;44(02):191–228.

112. Vakser IA. Protein-Protein Docking: From Interaction to Interactome. Biophys J. 2014 Oct;107(8):1785–93.

113. Greer J, Bush BL. Macromolecular shape and surface maps by solvent exclusion. Proc Natl Acad Sci. 1978 Jan 1;75(1):303–7.

114. Wodak SJ, Janin J. Computer analysis of protein-protein interaction. J Mol Biol. 1978 Sep 15;124(2):323–42.

115. Janin J. Assessing predictions of protein-protein interaction: The CAPRI experiment. Protein Sci. 2005 Feb 1;14(2):278–83.

116. Janin J. Protein–protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst. 2010;6(12):2351.

117. Szilagyi A, Zhang Y. Template-based structure modeling of protein–protein interactions. Curr Opin Struct Biol. 2014 Feb;24:10–23.

118. Kundrotas PJ, Lensink MF, Alexov E. Homology-based modeling of 3D structures of protein–protein complexes using alignments of modified sequence profiles. Int J Biol Macromol. 2008 Aug 15;43(2):198–208.

119. Tuncbag N, Keskin O, Nussinov R, Gursoy A. Fast and accurate modeling of protein–protein interactions by combining template-interface-based docking with flexible refinement. Proteins Struct Funct Bioinforma. 2012 Apr 1;80(4):1239–49.

120. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci. 2012 Jun 12;109(24):9438–41.

121. Sinha R, Kundrotas PJ, Vakser IA. Docking by structural similarity at protein-protein interfaces. Proteins. 2010 Nov 15;78(15):3235–41.

122. Vakser IA. Low-resolution structural modeling of protein interactome. Curr Opin Struct Biol. 2013 Apr;23(2):198–205.

123. Anishchenko I, Kundrotas PJ, Vakser IA. Modeling complexes of modeled proteins. Proteins Struct Funct Bioinforma. 2017;85(3):470–478.

124. Huang S-Y. Search strategies and evaluation in protein–protein docking: principles, advances and challenges. Drug Discov Today. 2014 Aug;19(8):1081–96.

125. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. J Am Chem Soc. 2003 Feb;125(7):1731–7.

126. Zacharias M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci. 2003 Jun;12(6):1271–82.

127. Ritchie DW, Venkatraman V. Ultra-fast FFT protein docking on graphics processors. Bioinformatics. 2010 Oct 1;26(19):2398–405.

128. Garzon JI, Lopéz-Blanco JR, Pons C, Kovacs J, Abagyan R, Fernandez-Recio J, et al. FRODOCK: a new approach for fast rotational protein–protein docking. Bioinformatics. 2009 Oct 1;25(19):2544–51.

129. Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z. Integrating statistical pair potentials into protein complex prediction. Proteins Struct Funct Bioinforma. 2007 Nov 15;69(3):511–20.

130. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci. 1992 Mar 15;89(6):2195–9.

131. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res. 2004 Jul 1;32(Web Server):W96–9.

132. Schneidman-Duhovny D, Inbar Y, Polak V, Shatsky M, Halperin I, Benyamini H, et al. Taking geometry to its edge: Fast unbound rigid (and hinge-bent) docking. Proteins Struct Funct Genet. 2003 Jul 1;52(1):107–12.

133. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science. 1983 Aug 19;221(4612):709–13.

134. Schneider SS and M, Zacharias M. Flexible Protein-Protein Docking. In: Selected Works in Bioinformatics. 2011.

135. de Vries SJ, Zacharias M. Fast and accurate grid representations for atom-based docking with partner flexibility. J Comput Chem. 2017;38(17):1538–1546.

136. de Vries S, Zacharias M. Flexible docking and refinement with a coarse-grained protein model using ATTRACT: Flexible Protein-Protein Docking and Refinement. Proteins Struct Funct Bioinforma. 2013 Dec;81(12):2167–74.

137. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, et al. Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations. J Mol Biol. 2003 Aug;331(1):281–99.

138. Moal IH, Bates PA. SwarmDock and the Use of Normal Modes in Protein-Protein Docking. Int J Mol Sci. 2010 Sep 28;11(10):3623–48.

139. Bonvin AM. Flexible protein–protein docking. Curr Opin Struct Biol. 2006 Apr;16(2):194–200.

140. May A, Zacharias M. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein-protein docking. Proteins Struct Funct Bioinforma. 2007 Aug 29;70(3):794–809.

141. Cheng TM-K, Blundell TL, Fernandez Recio J. pyDock: Electrostatics and desolvation for effective scoring of rigid body protein–protein docking. Proteins Struct Funct Bioinforma. 2007 Aug 1;68(2):503–15.

142. Fiorucci S, Zacharias M. Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. Proteins. 2010 Nov 15;78(15):3131–9.

143. Andreani J, Faure G, Guerois R. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. Bioinformatics. 2013 Jul 15;29(14):1742–9.

144. Yu J, Vavrusa M, Andreani J, Rey J, Tufféry P, Guerois R. InterEvDock: a docking server to predict the structure of protein–protein interactions using evolutionary information. Nucleic Acids Res. 2016 Jul 8;44(W1):W542–9.

145. Quignot C, Rey J, Yu J, Tufféry P, Guerois R, Andreani J. InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs. Nucleic Acids Res. 2018 Jul 2;46(W1):W408–16.

146. Bordner AJ, Gorin AA. Protein docking using surface matching and supervised machine learning. Proteins Struct Funct Bioinforma. 2007 Apr 19;68(2):488–502.

147. Bernauer J, Aze J, Janin J, Poupon A. A new protein protein docking scoring function based on interface residue properties. Bioinformatics. 2007 Mar 1;23(5):555–62.

148. Palma PN, Krippahl L, Wampler JE, Moura JJG. BiGGER: A new (soft) docking algorithm for predicting protein interactions. Proteins Struct Funct Genet. 2000 Jun 1;39(4):372–84.

149. Chae M-H, Krull F, Lorenzen S, Knapp E-W. Predicting protein complex geometries with a neural network. Proteins Struct Funct Bioinforma. 2010 Mar;78(4):1026–39.

150. Moal IH, Torchala M, Bates PA, Fernández-Recio J. The scoring of poses in protein-protein docking: current capabilities and future directions. BMC Bioinformatics. 2013;14(1):286.

151. Pierce B, Weng Z. ZRANK: Reranking protein docking predictions with an optimized energy function. Proteins Struct Funct Bioinforma. 2007 Mar 20;67(4):1078–86.

152. Mosca R, Pons C, Fernández-Recio J, Aloy P. Pushing Structural Information into the Yeast Interactome by High-Throughput Protein Docking Experiments. PLoS Comput Biol. 2009 Aug 28;5(8):e1000490.

153. Fernández-Recio J, Totrov M, Abagyan R. Identification of Protein–Protein Interaction Sites from Docking Energy Landscapes. J Mol Biol. 2004 Jan;335(3):843–65.

154. Grosdidier S, Fernández-Recio J. Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics. 2008;9(1):447.

155. Fernández-Recio J. Prediction of protein binding sites and hot spots: Prediction of protein binding sites. Wiley Interdiscip Rev Comput Mol Sci. 2011 Sep;1(5):680–98.

156. Martin J, Lavery R. Arbitrary protein−protein docking targets biologically relevant interfaces. BMC Biophys. 2012;5(1):7.

157. Engelen S, Trojan LA, Sacquin-Mora S, Lavery R, Carbone A. Joint Evolutionary Trees: A Large-Scale Method To Predict Protein Interfaces Based on Sequence Sampling. PLoS Comput Biol. 2009 Jan 23;5(1):e1000267.

158. Segura J, Jones PF, Fernandez-Fuentes N. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. BMC Bioinformatics. 2011;12(1):352.

159. Vamparys L, Laurent B, Carbone A, Sacquin-Mora S. Great interactions: How binding incorrect partners can teach us about protein recognition and function: Predicting Binding Sites From Cross-Docking. Proteins Struct Funct Bioinforma. 2016 Oct;84(10):1408–21.

160. Lagarde N, Carbone A, Sacquin-Mora S. Hidden partners: Using cross-docking calculations to predict binding sites for proteins with multiple interactions. Proteins Struct Funct Bioinforma. 2018 Jul;86(7):723–37.

161. Joung JK, Ramm EI, Pabo CO. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. Proc Natl Acad Sci. 2000 Jun 20;97(13):7382–7.

162. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Séraphin B. A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol. 1999 Oct;17(10):1030–2.

163. Pazos F, Valencia A. Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng Des Sel. 2001 Sep 1;14(9):609–14.

164. Sun J, Xu J, Liu Z, Liu Q, Zhao A, Shi T, et al. Refined phylogenetic profiles method for predicting protein-protein interactions. Bioinforma Oxf Engl. 2005 Aug 15;21(16):3409–15.

165. Sacquin-Mora S, Carbone A, Lavery R. Identification of Protein Interaction Partners and Protein–Protein Interaction Sites. J Mol Biol. 2008 Oct 24;382(5):1276–89.

166. Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A. Protein-Protein Interactions in a Crowded Environment: An Analysis via Cross-Docking Simulations and Evolutionary Information. PLoS Comput Biol. 2013 Dec 5;9(12):e1003369.

167. Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, et al. Protein–protein docking benchmark 2.0: An update. Proteins Struct Funct Bioinforma. 2005 Aug 1;60(2):214–6.

168. Maheshwari S, Brylinski M. Across-proteome modeling of dimer structures for the bottom-up assembly of protein-protein interaction networks. BMC Bioinformatics. 2017 May 12;18(1).

169. Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. Towards the prediction of protein interaction partners using physical docking. Mol Syst Biol. 2011;7(1):469–469.

170. Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y. Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods. 2013;10.

171. Ohue M, Matsuzaki Y, Uchikoga N, Ishida T, Akiyama Y. MEGADOCK: An All-to-All Protein-Protein Interaction Prediction System Using Tertiary Structure Data. Protein Pept Lett. 2014 Jun;21(8):766–78.

172. Zhang C, Tang B, Wang Q, Lai L. Discovery of binding proteins for a protein target using protein–protein docking-based virtual screening. Proteins Struct Funct Bioinforma. 2014 Oct 1;82(10):2472–82.

173. Yu J, Guerois R. PPI4DOCK: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets. Bioinformatics. 2016 Dec 15;32(24):3760–7.

174. Levy ED, Erba EB, Robinson CV, Teichmann SA. Assembly reflects evolution of protein complexes. Nature. 2008 Jun;453(7199):1262–5.

175. Bugayevskiy LM, Snyder J. Map Projections: A Reference Manual. CRC Press; 1995. 356 p.

176. Kobe B, Deisenhofer J. Mechanism of ribonuclease inhibition by ribonuclease inhibitor protein based on the crystal structure of its complex with ribonuclease A. J Mol Biol. 1996 Dec 20;264(5):1028–43.

177. Pellegrino S, Radzimanowski J, de Sanctis D, Erba EB, McSweeney S, Timmins J. Structural and Functional Characterization of an SMC-like Protein RecN: New Insights into Double-Strand Break Repair. Structure. 2012 Dec;20(12):2076–89.

178. Fanning DW, Smith JA, Rose GD. Molecular cartography of globular proteins with application to antigenic sites. Biopolymers. 1986 May;25(5):863–83.

179. J Barlow D, M Thornton J. Interactive map projection algorithm for illustrating protein surfaces. J Mol Graph. 1986 Jun;4(2):97–100.

180. Pawłowski K, Godzik A. Surface Map Comparison: Studying Function Diversity of Homologous Proteins. J Mol Biol. 2001 Jun;309(3):793–806.

181. Sasin JM, Godzik A, Bujnicki JM. SURF's UP! — Protein classification by surface comparisons. J Biosci. 2007 Jan;32(1):97–100.

182. Kontopoulos DG, Vlachakis D, Tsiliki G, Kossida S. Structuprint: a scalable and extensible tool for two-dimensional representation of protein surfaces. BMC Struct Biol. 2016 Dec;16(1).

183. Koromyslova AD, Chugunov AO, Efremov RG. Deciphering Fine Molecular Details of Proteins' Structure and Function with a *Protein Surface Topography (PST)* Method. J Chem Inf Model. 2014 Apr 28;54(4):1189–99.

184. Ravikumar KM, Huang W, Yang S. Coarse-Grained Simulations of Protein-Protein Association: An Energy Landscape Perspective. Biophys J. 2012 Aug;103(4):837–45.

185. Kim YC, Hummer G. Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. J Mol Biol. 2008 Feb;375(5):1416–33.

186. Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Research. 2016 Feb 18;5:189.

187. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: Combining techniques to model RNA–small molecule complexes. RNA. 2009 Jun 1;15(6):1219–30.

188. Zhang Y. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005 Apr 11;33(7):2302–9.

189. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 8;32(5):1792–7.

190. Schrödinger LLC. The PYMOL Molecular Graphics System, Ver. 1.7r0. Schrodinger, New York; 2014.

191. Chéron J-B, Zacharias M, Antonczak S, Fiorucci S. Update of the ATTRACT force field for the prediction of protein-protein binding affinity. J Comput Chem. 2017 Jun 5;38(21):1887–90.

192. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982 May;157(1):105–32.

193. Mezei M. A new method for mapping macromolecular topography. J Mol Graph Model. 2003 Mar;21(5):463–72.

194. Ceres N, Pasi M, Lavery R. A Protein Solvation Model Based on Residue Burial. J Chem Theory Comput. 2012 Jun 12;8(6):2141–4.

195. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins. 2004 Dec 1;57(4):702–10.

196. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995 Apr 7;247(4):536–40.

197. Rohani N, Canty L, Luu O, Fagotto F, Winklbauer R. EphrinB/EphB Signaling Controls Embryonic Germ Layer Separation by Contact-Induced Cell Detachment. PLOS Biol. 2011 Mar 1;9(3):e1000597.

198. Bowden TA, Aricescu AR, Nettleship JE, Siebold C, Rahman-Huq N, Owens RJ, et al. Structural plasticity of eph receptor A4 facilitates cross-class ephrin signaling. Struct Lond Engl 1993. 2009 Oct 14;17(10):1386–97.

199. Forse GJ, Uson ML, Nasertorabi F, Kolatkar A, Lamberto I, Pasquale EB, et al. Distinctive Structure of the EphA3/Ephrin-A5 Complex Reveals a Dual Mode of Eph Receptor Interaction for Ephrin-A5. PLOS ONE. 2015 May 20;10(5):e0127081.

200. Laine E, Carbone A. Local Geometry and Evolutionary Conservation of Protein Surfaces Reveal the Multiple Recognition Patches in Protein-Protein Interactions. PLOS Comput Biol. 2015 Dec 21;11(12):e1004580.

201. Tukey JW. Comparing Individual Means in the Analysis of Variance. Biometrics. 1949;5(2):99–114.

202. Reverter D, Lima CD. A Basis for SUMO Protease Specificity Provided by Analysis of Human Senp2 and a Senp2-SUMO Complex. Structure. 2004 Aug;12(8):1519–31.

203. Wuebbens MM, Rajagopalan KV. Mechanistic and Mutational Studies of *Escherichia coli* Molybdopterin Synthase Clarify the Final Step of Molybdopterin Biosynthesis. J Biol Chem. 2003 Apr 18;278(16):14523–32.

204. Krissinel E, Henrick K. Inference of Macromolecular Assemblies from Crystalline State. J Mol Biol. 2007 Sep;372(3):774–97.

206.

205. Christinger HW, Fuh G, de Vos AM, Wiesmann C. The Crystal Structure of Placental Growth Factor in Complex with Domain 2 of Vascular Endothelial Growth Factor Receptor-1. J Biol Chem. 2004 Mar 12;279(11):10382–8.

206. Chong KT, Miyazaki G, Morimoto H, Oda Y, Park SY. Structures of the deoxy and CO forms of haemoglobin from Dasyatis akajei, a cartilaginous fish. Acta Crystallogr D Biol Crystallogr. 1999 Jul;55(Pt 7):1291–300.

207. Johnson RJ, McCoy JG, Bingman CA, Phillips GN, Raines RT. Inhibition of Human Pancreatic Ribonuclease by the Human Ribonuclease Inhibitor Protein. J Mol Biol. 2007 Apr;368(2):434–49.

208. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017 Jan 4;45(D1):D158–69.

209. Pous J, Canals A, Terzyan SS, Guasch A, Benito A, Ribó M, et al. Three-dimensional structure of a human pancreatic ribonuclease variant, a step forward in the design of cytotoxic ribonucleases. J Mol Biol. 2000 Oct 13;303(1):49–60.

210. Banner DW, D'Arcy A, Janes W, Gentz R, Schoenfeld HJ, Broger C, et al. Crystal structure of the soluble human 55 kd TNF receptor-human TNF beta complex: implications for TNF receptor activation. Cell. 1993 May 7;73(3):431–45.

211. Bauer J, Namineni S, Reisinger F, Zöller J, Yuan D, Heikenwälder M. Lymphotoxin, NF-κB, and cancer: the dark side of cytokines. Dig Dis Basel Switz. 2012;30(5):453–68.

212. Sudhamsu J, Yin J, Chiang EY, Starovasnik MA, Grogan JL, Hymowitz SG. Dimerization of LTβR by LTα1β2 is necessary and sufficient for signal transduction. Proc Natl Acad Sci U S A. 2013 Dec 3;110(49):19896–901.

213. Rivkin E, Almeida SM, Ceccarelli DF, Juang Y-C, MacLean TA, Srikumar T, et al. The linear ubiquitin-specific deubiquitinase gumby regulates angiogenesis. Nature. 2013 Jun 20;498(7454):318–24.

214. Iwema T, Billas IM, Beck Y, Bonneton F, Nierengarten H, Chaumot A, et al. Structural and functional characterization of a novel type of ligand-independent RXR-USP receptor. EMBO J. 2007 Aug 22;26(16):3770–82.

215. Maheshwari S, Brylinski M. Predicting protein interface residues using easily accessible on-line resources. Brief Bioinform. 2015 Nov 1;16(6):1025–34.

216. Schulz GE. The Dominance of Symmetry in the Evolution of Homo-oligomeric Proteins. J Mol Biol. 2010 Jan;395(4):834–43.

217. Holley AK, Bakthavatchalu V, Velez-Roman JM, St. Clair DK. Manganese Superoxide Dismutase: Guardian of the Powerhouse. Int J Mol Sci. 2011 Oct 21;12(10):7114–62.

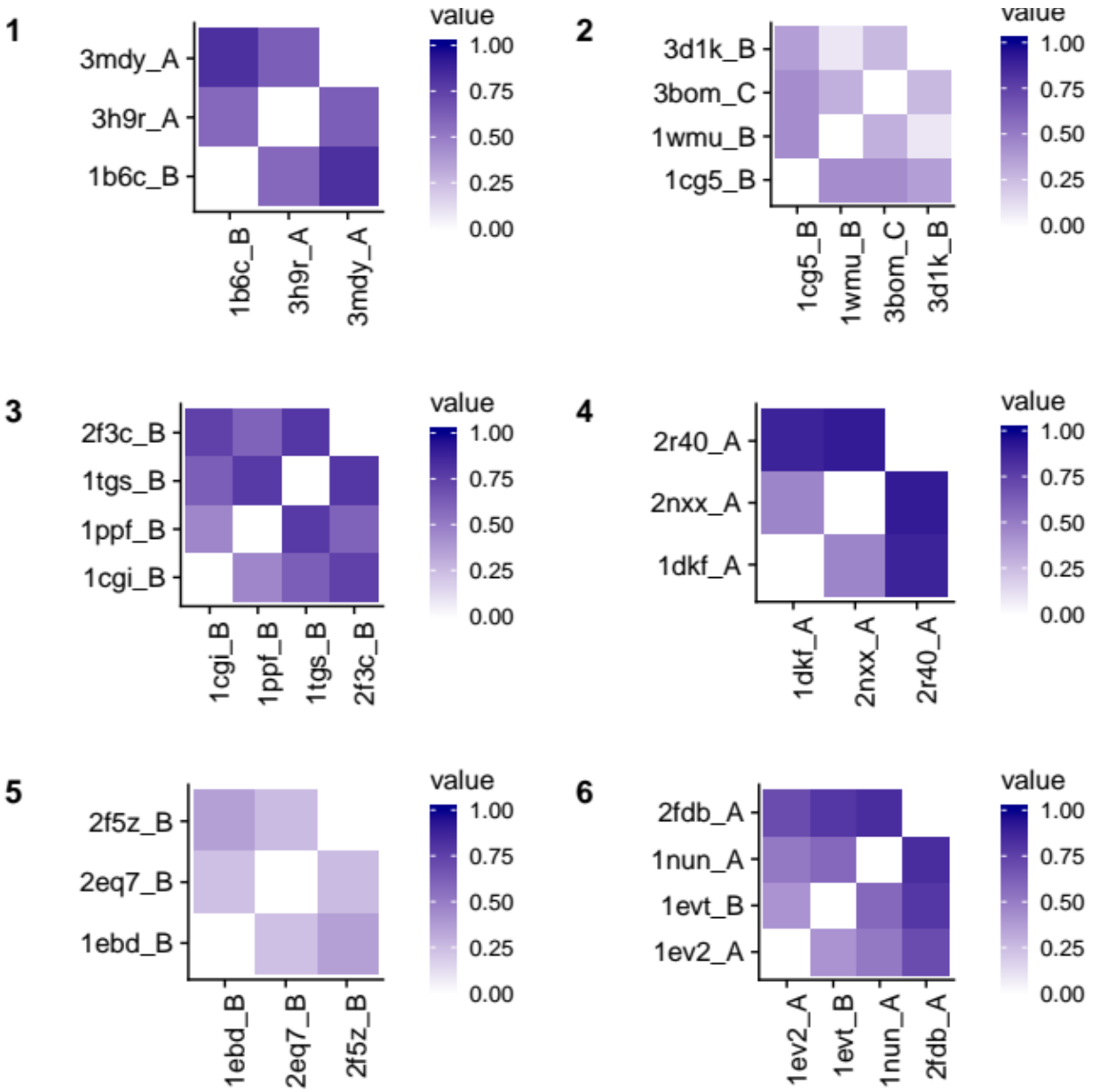
218. Yueh C, Hall DR, Xia B, Padhorny D, Kozakov D, Vajda S. ClusPro-DC: Dimer Classification by the Cluspro Server for Protein–Protein Docking. J Mol Biol. 2017 Feb;429(3):372–81.

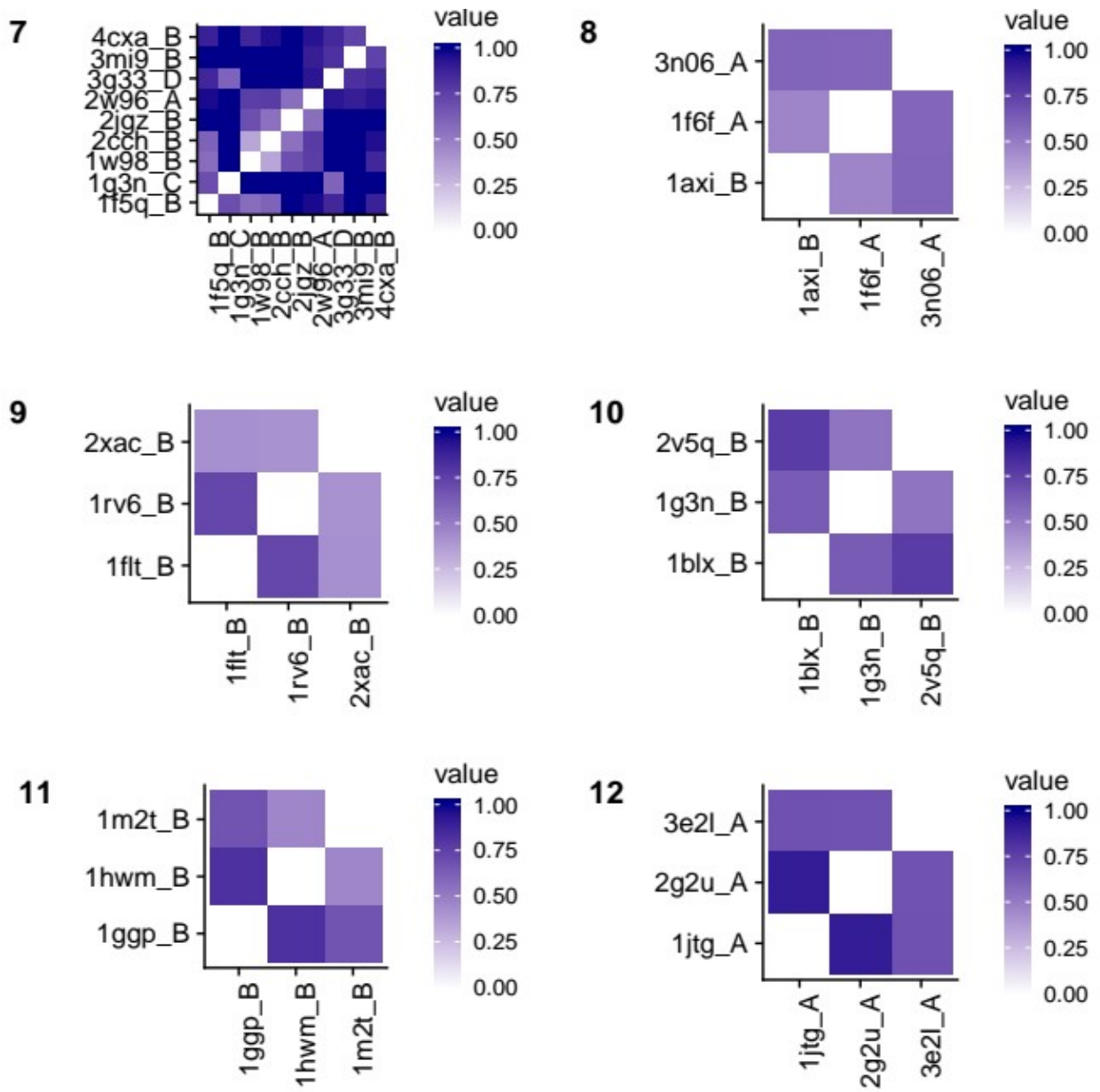
219. Cazals F, Dreyfus T. The structural bioinformatics library: modeling in biomolecular science and beyond. Bioinformatics. 2017 Jan 5;33(7):997–1004.

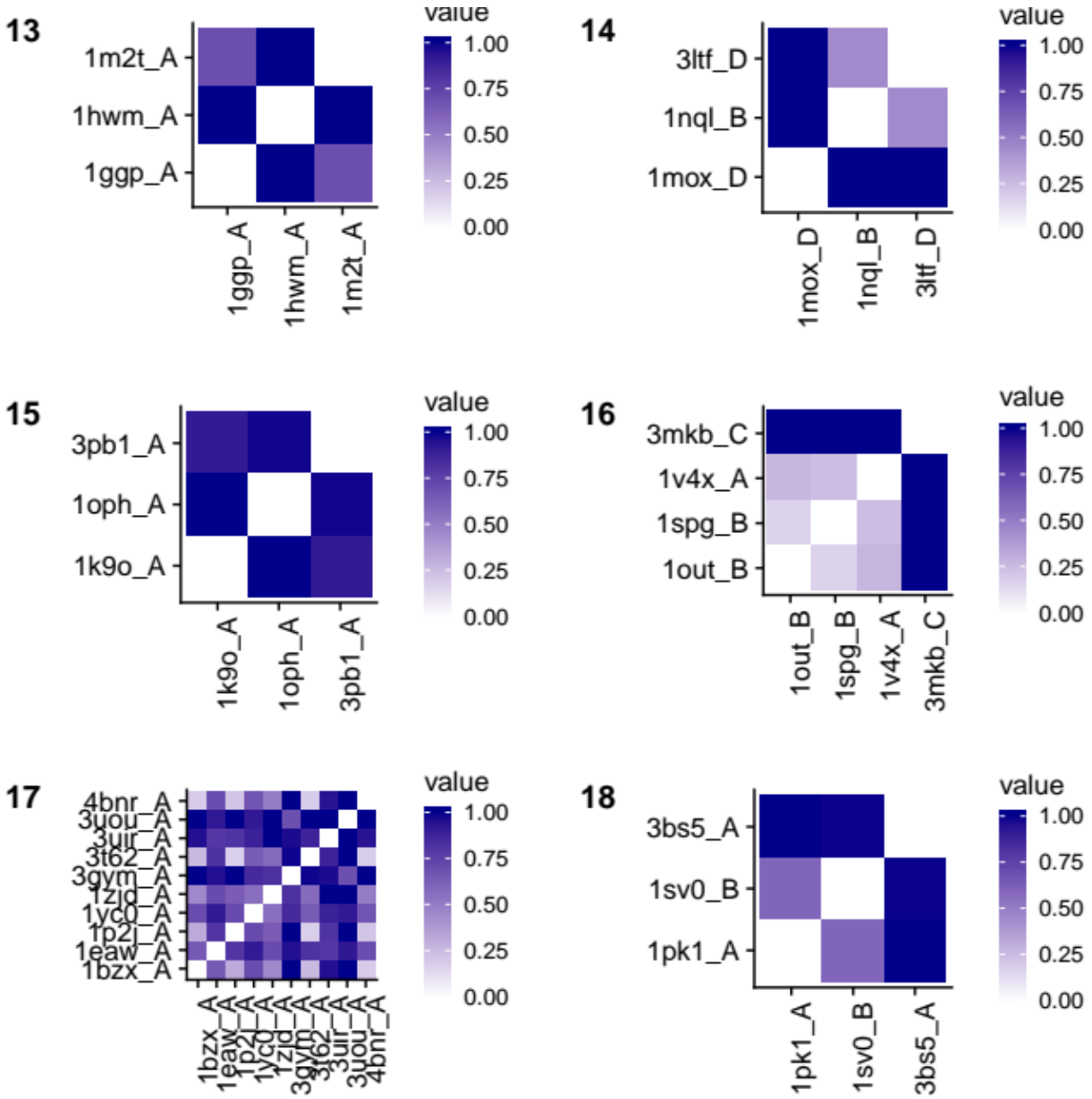
220. Mosca R, Céol A, Aloy P. Interactome3D: adding structural details to protein networks. Nat Methods. 2013 Jan;10(1):47–53.

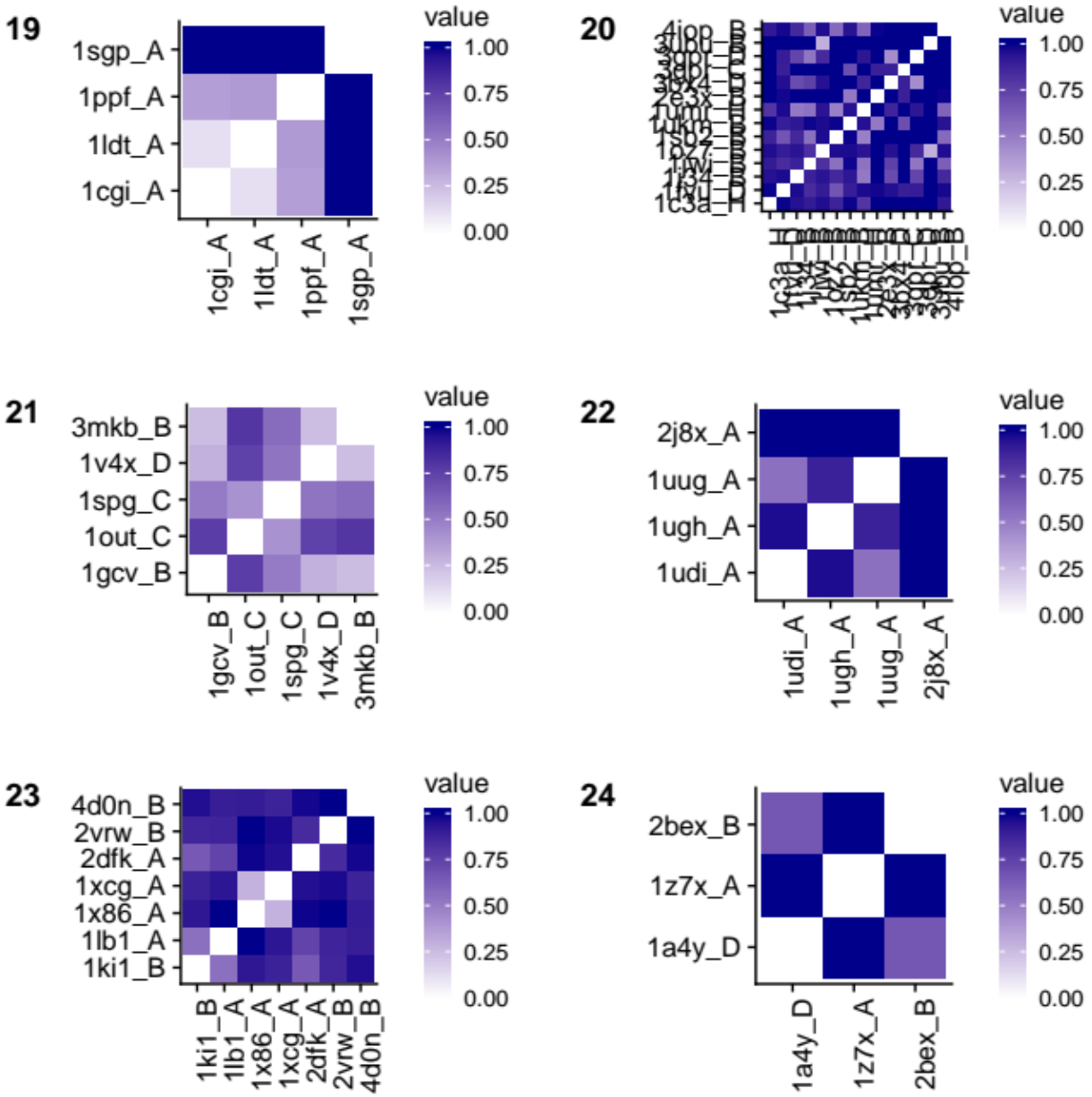
Annexes

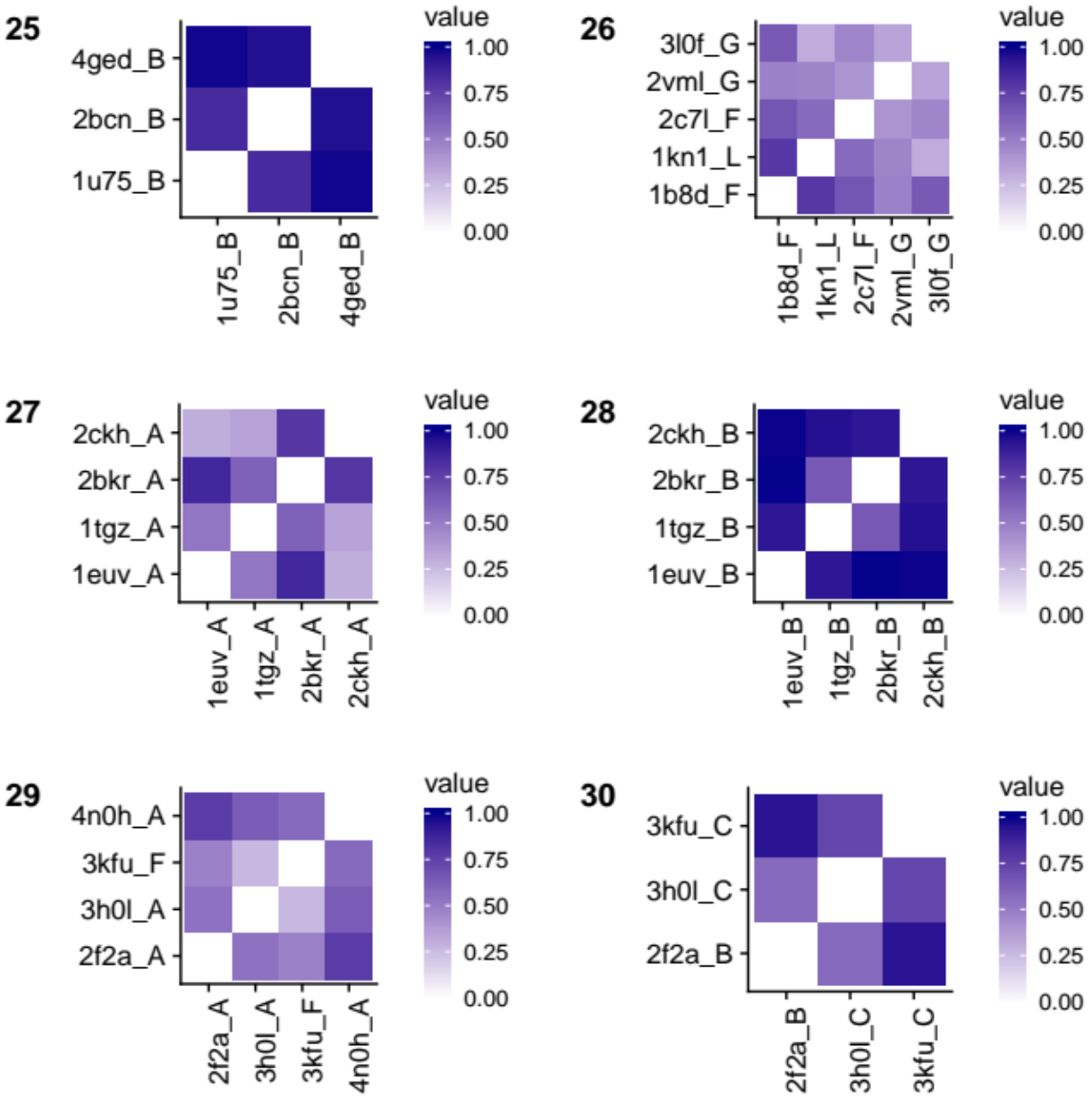
En annexe sont présentées les matrices de distance entre les cartes IPOPS rouges des 81 familles d'homologues structuraux du jeu de données utilisé dans la section 6. Pour plus de détails sur la métrique utilisée, voir la section 5.4.8.

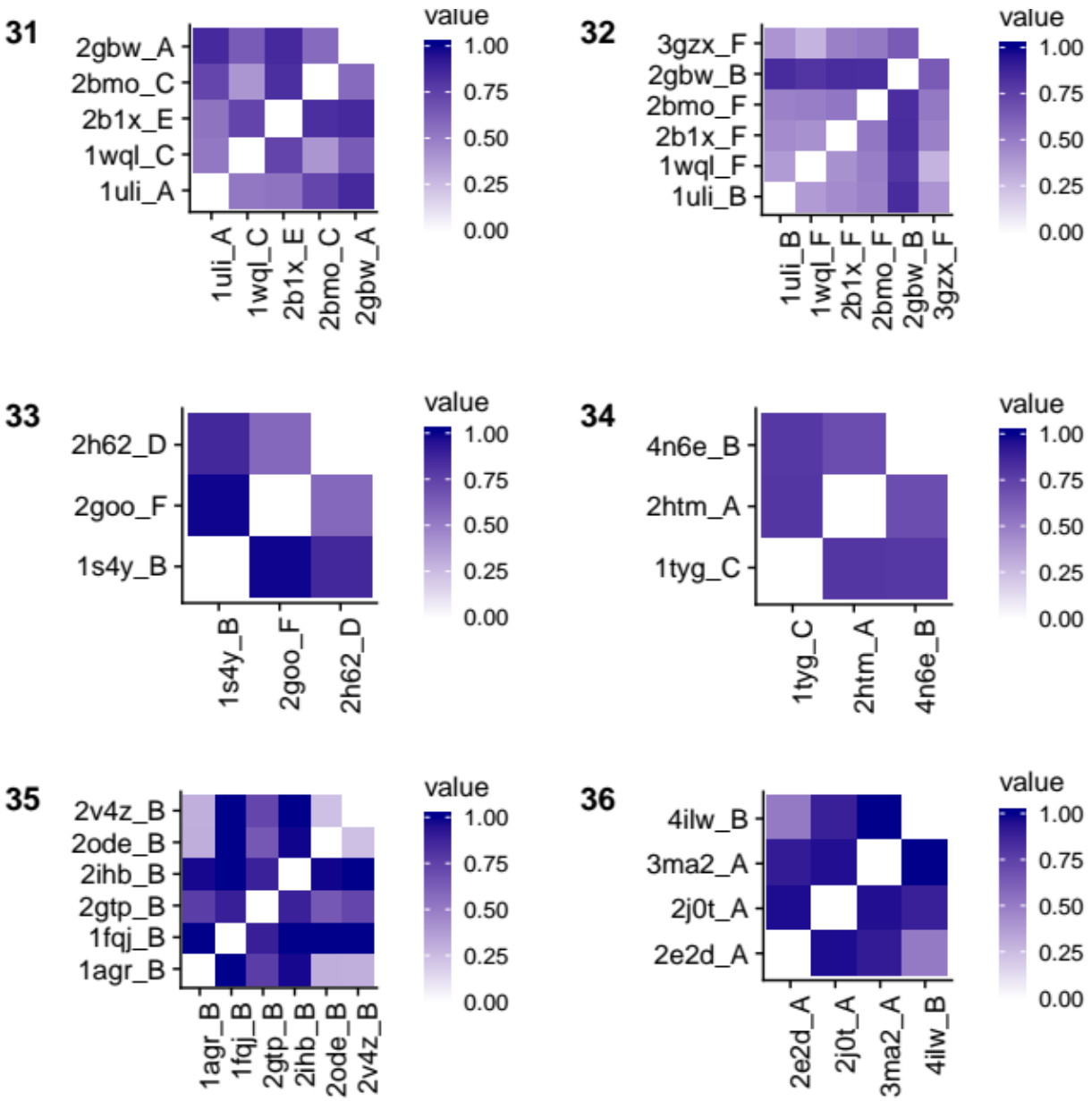


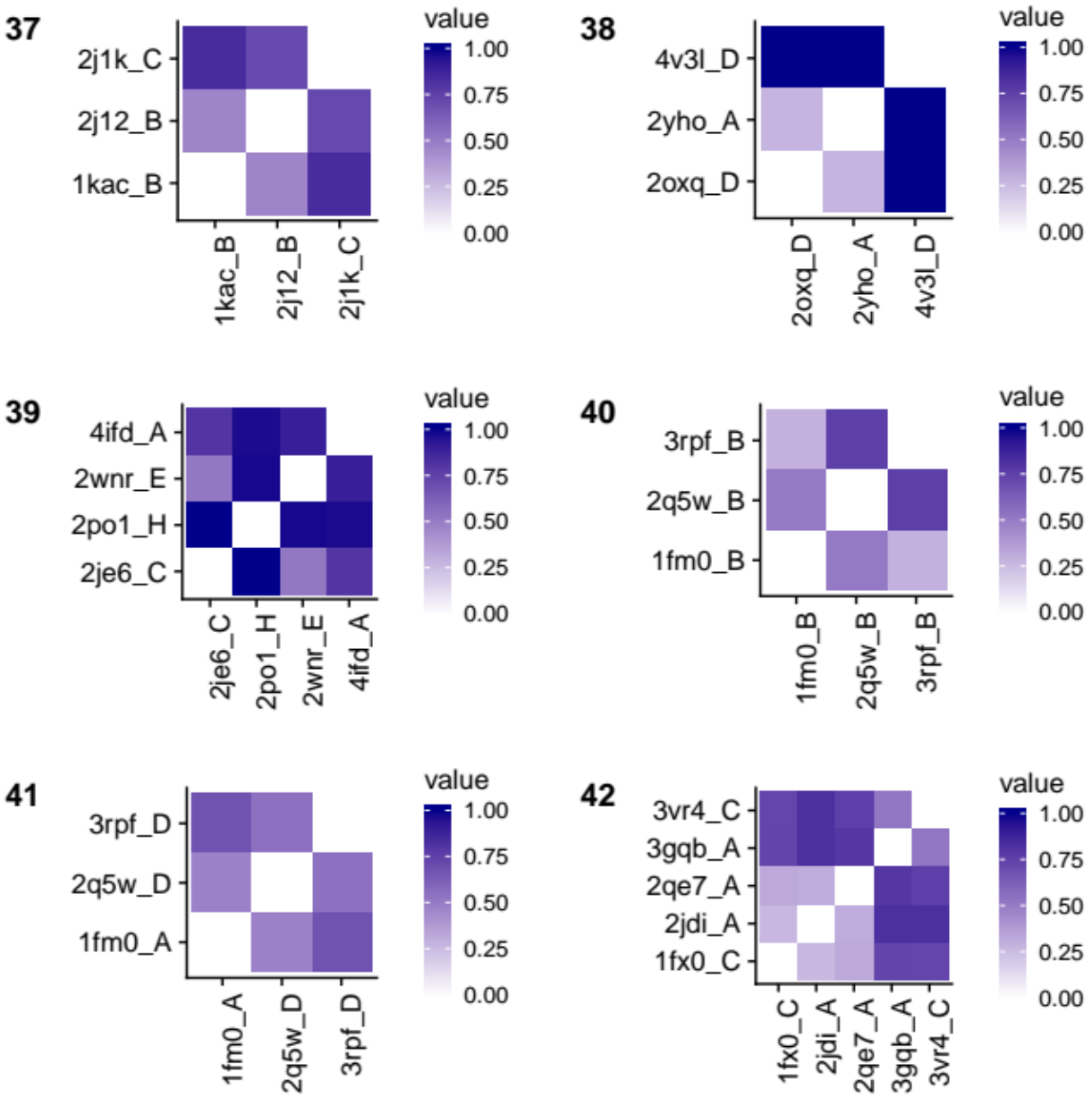

7
4cxa_B
3mi9_B
3g33_D
2w96_A
2jgz_B
2cch_B
1w98_B
1g3n_C
1f5q_B
1f5q_B
1g3n_C
1w98_B
2cch_B
2jgz_A
2w96_D
3g33_B
3mi9_B
4cxa_B
value
1.00
0.75
0.50
0.25
0.00
8
3n06_A
1f6f_A
1axi_B
1axi_B
1f6f_A
3n06_A
value
1.00
0.75
0.50
0.25
0.00
9
2xac_B
1rv6_B
1flt_B
1flt_B
1rv6_B
2xac_B
value
1.00
0.75
0.50
0.25
0.00
10
2v5q_B
1g3n_B
1blx_B
1blx_B
1g3n_B
2v5q_B
value
1.00
0.75
0.50
0.25
0.00
11
1m2t_B
1hwm_B
1ggp_B
1ggp_B
1hwm_B
1m2t_B
value
1.00
0.75
0.50
0.25
0.00
12
3e2l_A
2g2u_A
1jtg_A
1jtg_A
2g2u_A
3e2l_A
value
1.00
0.75
0.50
0.25
0.00

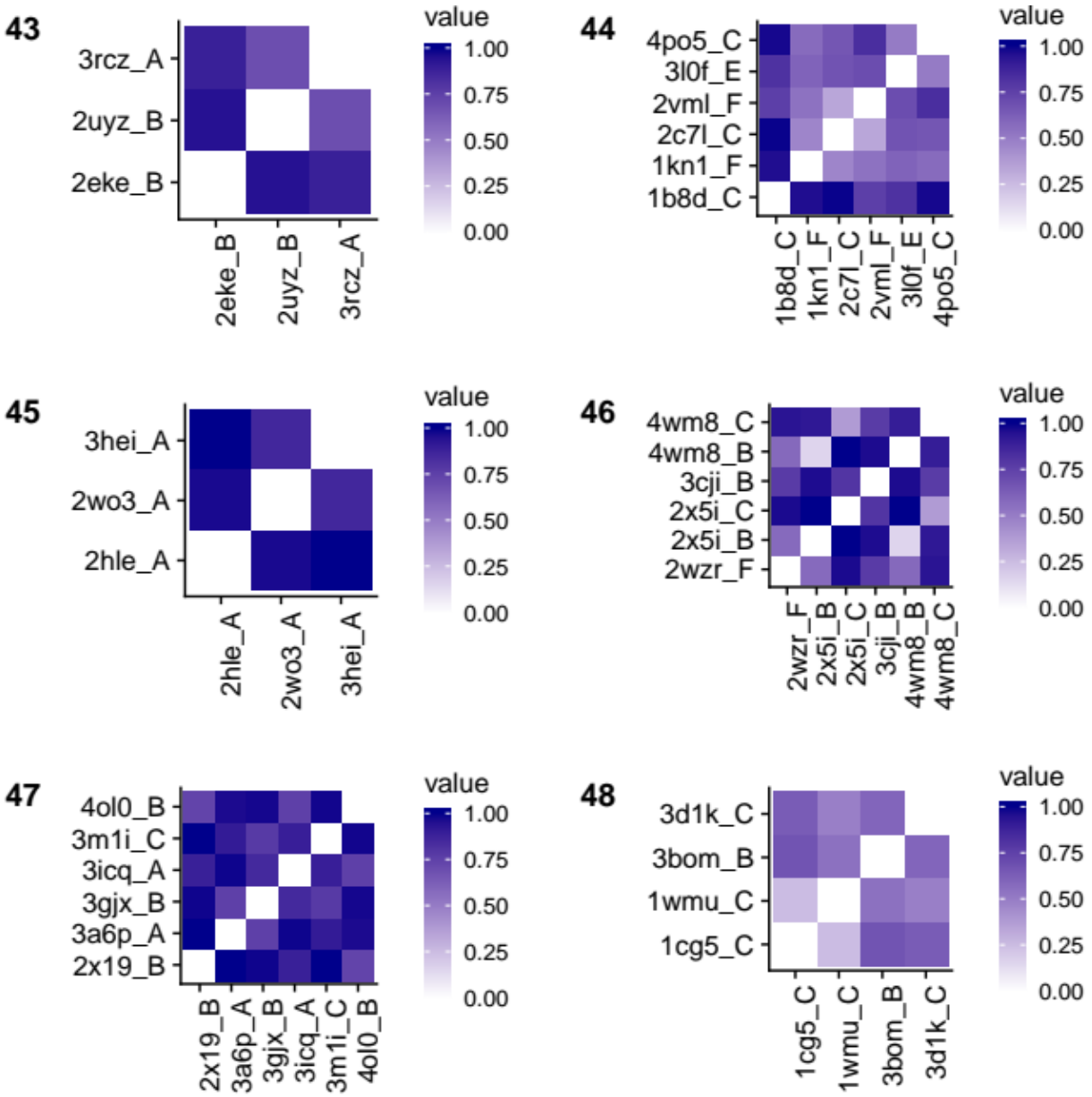

13
1m2t_A
1hwm_A
1ggp_A
1ggp_A
1hwm_A
1m2t_A
value
1.00
0.75
0.50
0.25
0.00
14
3ltf_D
1nql_B
1mox_D
1mox_D
1nql_B
3ltf_D
value
1.00
0.75
0.50
0.25
0.00
15
3pb1_A
1oph_A
1k9o_A
1k9o_A
1oph_A
3pb1_A
value
1.00
0.75
0.50
0.25
0.00
16
3mkb_C
1v4x_A
1spg_B
1out_B
1out_B
1spg_B
1v4x_A
3mkb_C
value
1.00
0.75
0.50
0.25
0.00
17
4bnr_A
3uou_A
3uir_A
3t62_A
3gym_A
1zid_A
1yc0_A
1p2j_A
1eaw_A
1bzx_A
1bzx_A
1eaw_A
1p2j_A
1yc0_A
1zid_A
3gym_A
3t62_A
3uir_A
3uou_A
4bnr_A
value
1.00
0.75
0.50
0.25
0.00
18
3bs5_A
1sv0_B
1pk1_A
1pk1_A
1sv0_B
3bs5_A
value
1.00
0.75
0.50
0.25
0.00

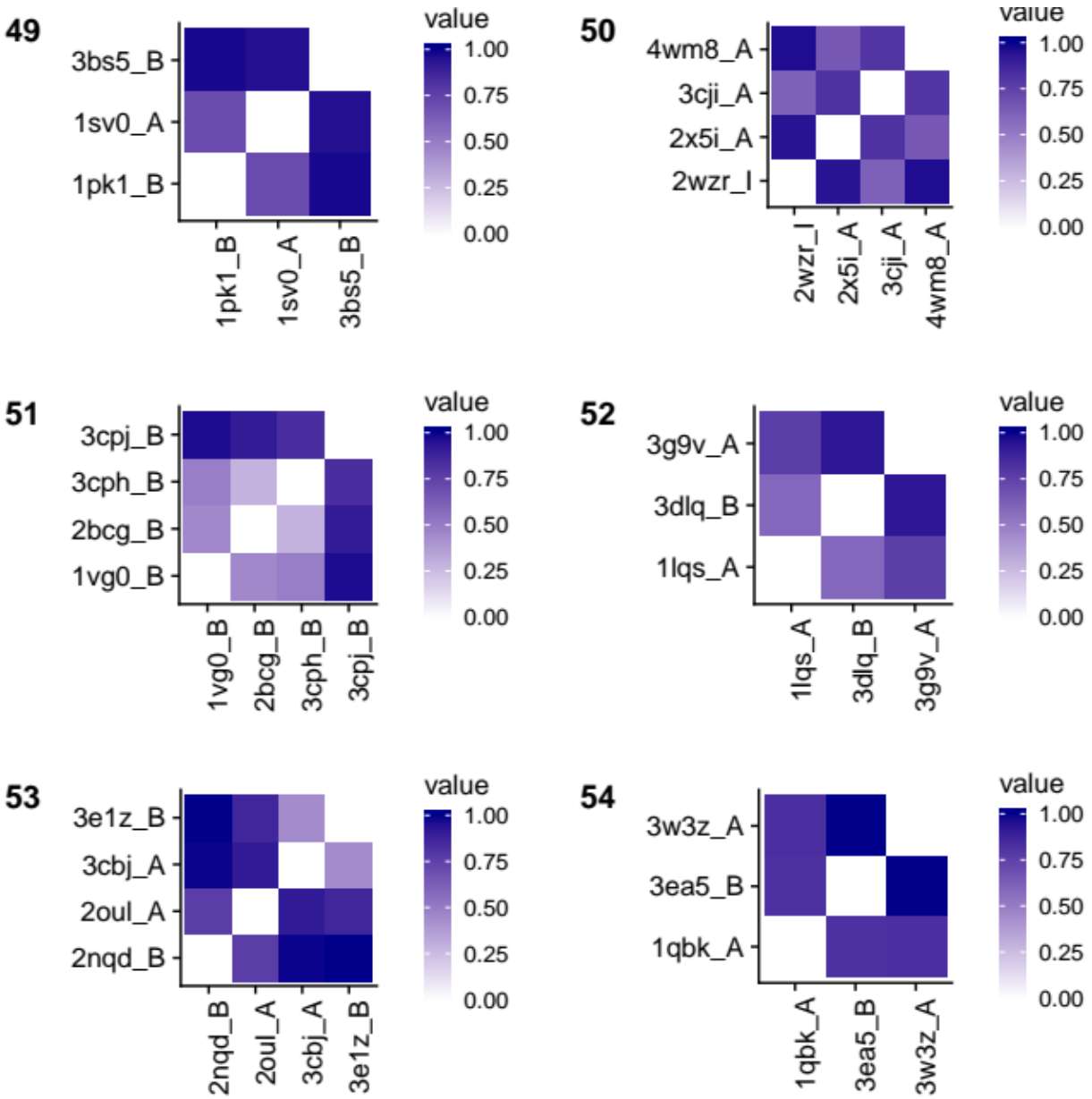

19
1sgp_A
1ppf_A
1ldt_A
1cgi_A
value
1.00
0.75
0.50
0.25
0.00
20
value
1.00
0.75
0.50
0.25
0.00
21
3mkb_B
1v4x_D
1spg_C
1out_C
1gcv_B
value
1.00
0.75
0.50
0.25
0.00
22
2j8x_A
1uug_A
1ugh_A
1udi_A
value
1.00
0.75
0.50
0.25
0.00
23
4d0n_B
2vrw_B
2dfk_A
1xcg_A
1x86_A
1lb1_A
1ki1_B
value
1.00
0.75
0.50
0.25
0.00
24
2bex_B
1z7x_A
1a4y_D
value
1.00
0.75
0.50
0.25
0.00

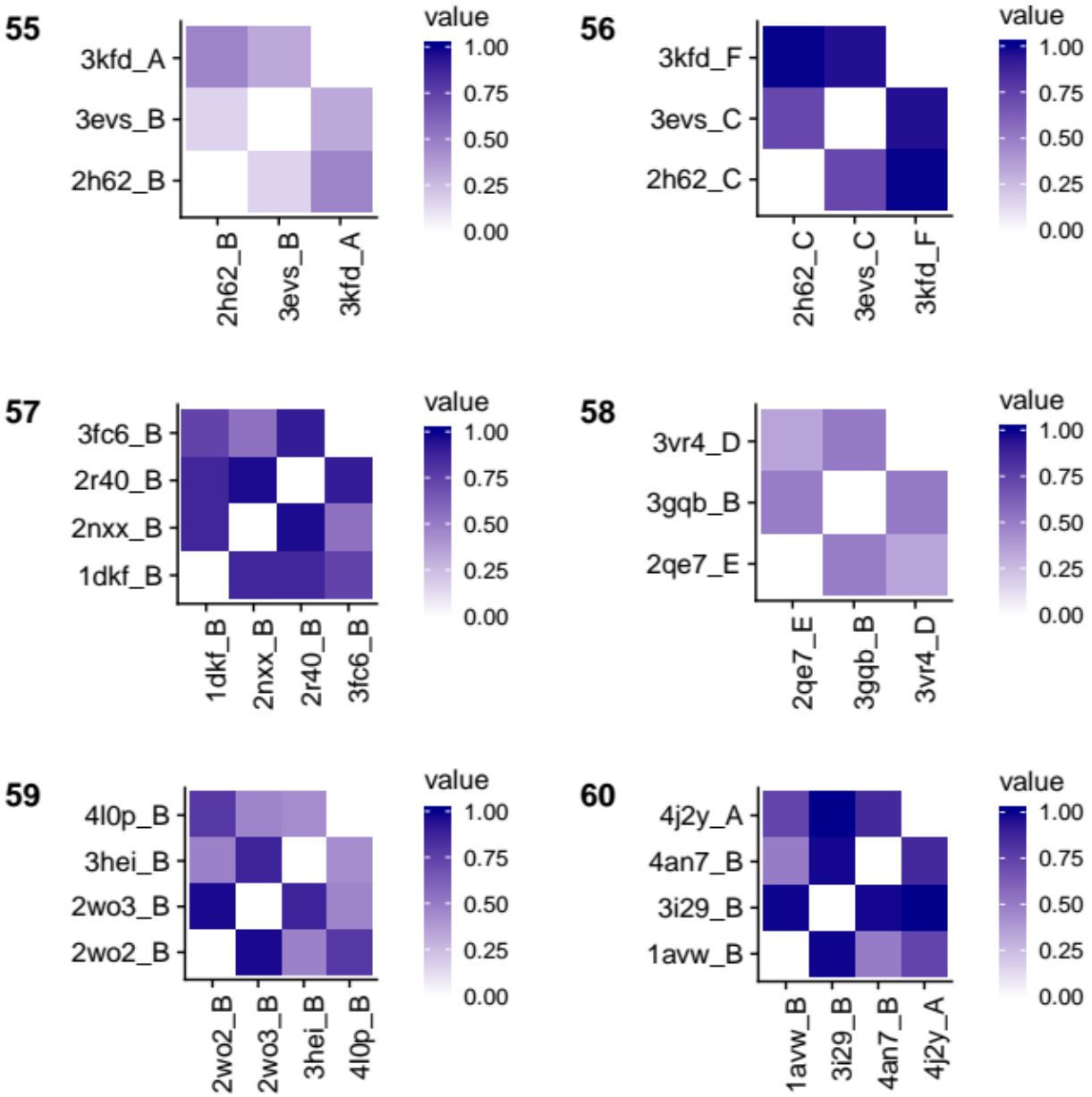

25
4ged_B
2bcn_B
1u75_B
1u75_B
2bcn_B
4ged_B
value
1.00
0.75
0.50
0.25
0.00
26
3l0f_G
2vml_G
2c7l_F
1kn1_L
1b8d_F
1b8d_F
1kn1_L
2c7l_F
2vml_G
3l0f_G
value
1.00
0.75
0.50
0.25
0.00
27
2ckh_A
2bkr_A
1tgz_A
1euv_A
1euv_A
1tgz_A
2bkr_A
2ckh_A
value
1.00
0.75
0.50
0.25
0.00
28
2ckh_B
2bkr_B
1tgz_B
1euv_B
1euv_B
1tgz_B
2bkr_B
2ckh_B
value
1.00
0.75
0.50
0.25
0.00
29
4n0h_A
3kfu_F
3h0l_A
2f2a_A
2f2a_A
3h0l_A
3kfu_F
4n0h_A
value
1.00
0.75
0.50
0.25
0.00
30
3kfu_C
3h0l_C
2f2a_B
2f2a_B
3h0l_C
3kfu_C
value
1.00
0.75
0.50
0.25
0.00

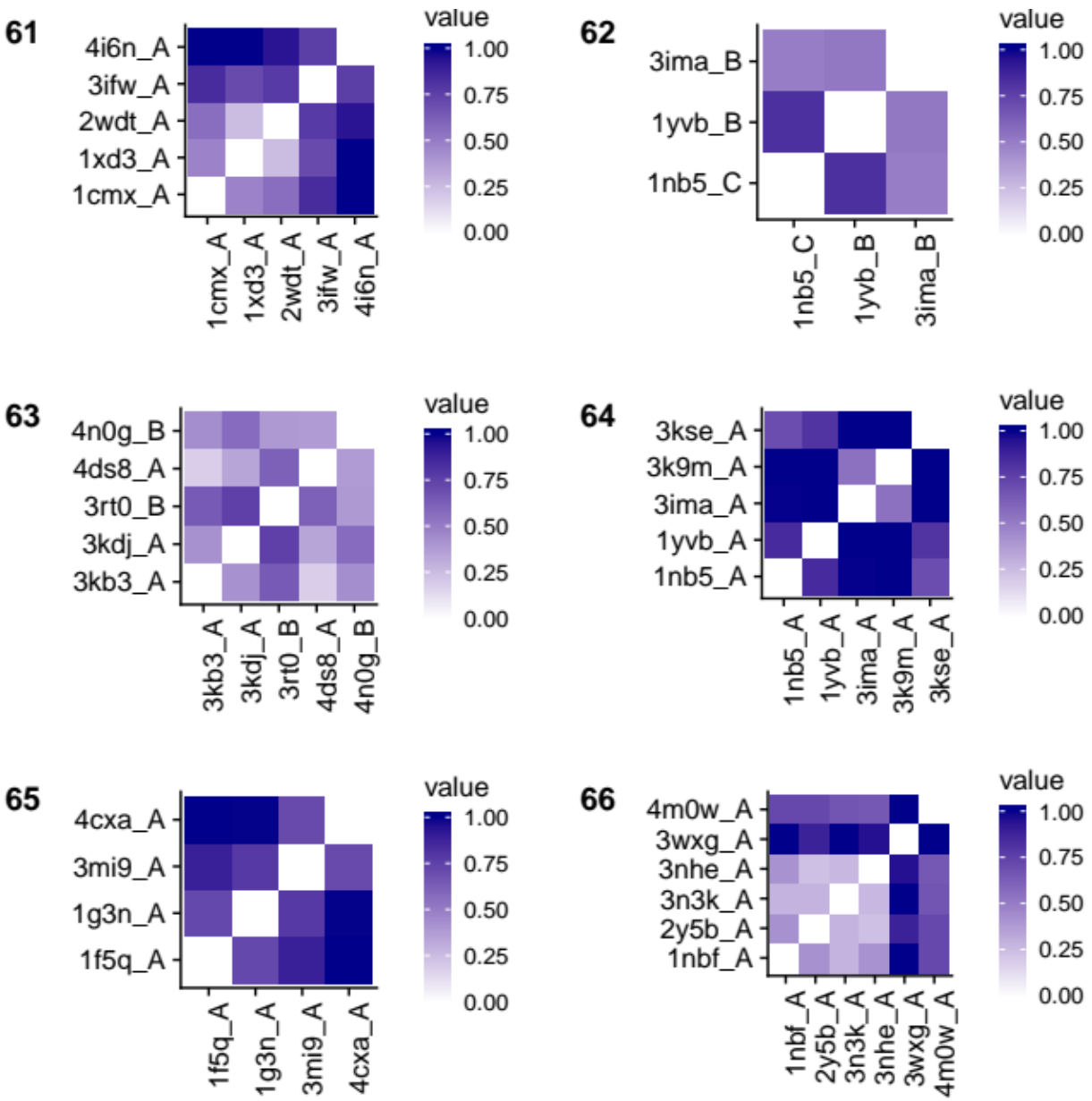

31
2gbw_A
2bmo_C
2b1x_E
1wql_C
1uli_A
1uli_A
1wql_C
2b1x_E
2bmo_C
2gbw_A
value
1.00
0.75
0.50
0.25
0.00
32
3gzx_F
2gbw_B
2bmo_F
2b1x_F
1wql_F
1uli_B
1uli_B
1wql_F
2b1x_F
2bmo_F
2gbw_B
3gzx_F
value
1.00
0.75
0.50
0.25
0.00
33
2h62_D
2goo_F
1s4y_B
1s4y_B
2goo_F
2h62_D
value
1.00
0.75
0.50
0.25
0.00
34
4n6e_B
2htm_A
1tyg_C
1tyg_C
2htm_A
4n6e_B
value
1.00
0.75
0.50
0.25
0.00
35
2v4z_B
2ode_B
2ihb_B
2gtp_B
1fqj_B
1agr_B
1agr_B
1fqj_B
2gtp_B
2ihb_B
2ode_B
2v4z_B
value
1.00
0.75
0.50
0.25
0.00
36
4ilw_B
3ma2_A
2j0t_A
2e2d_A
2e2d_A
2j0t_A
3ma2_A
4ilw_B
value
1.00
0.75
0.50
0.25
0.00

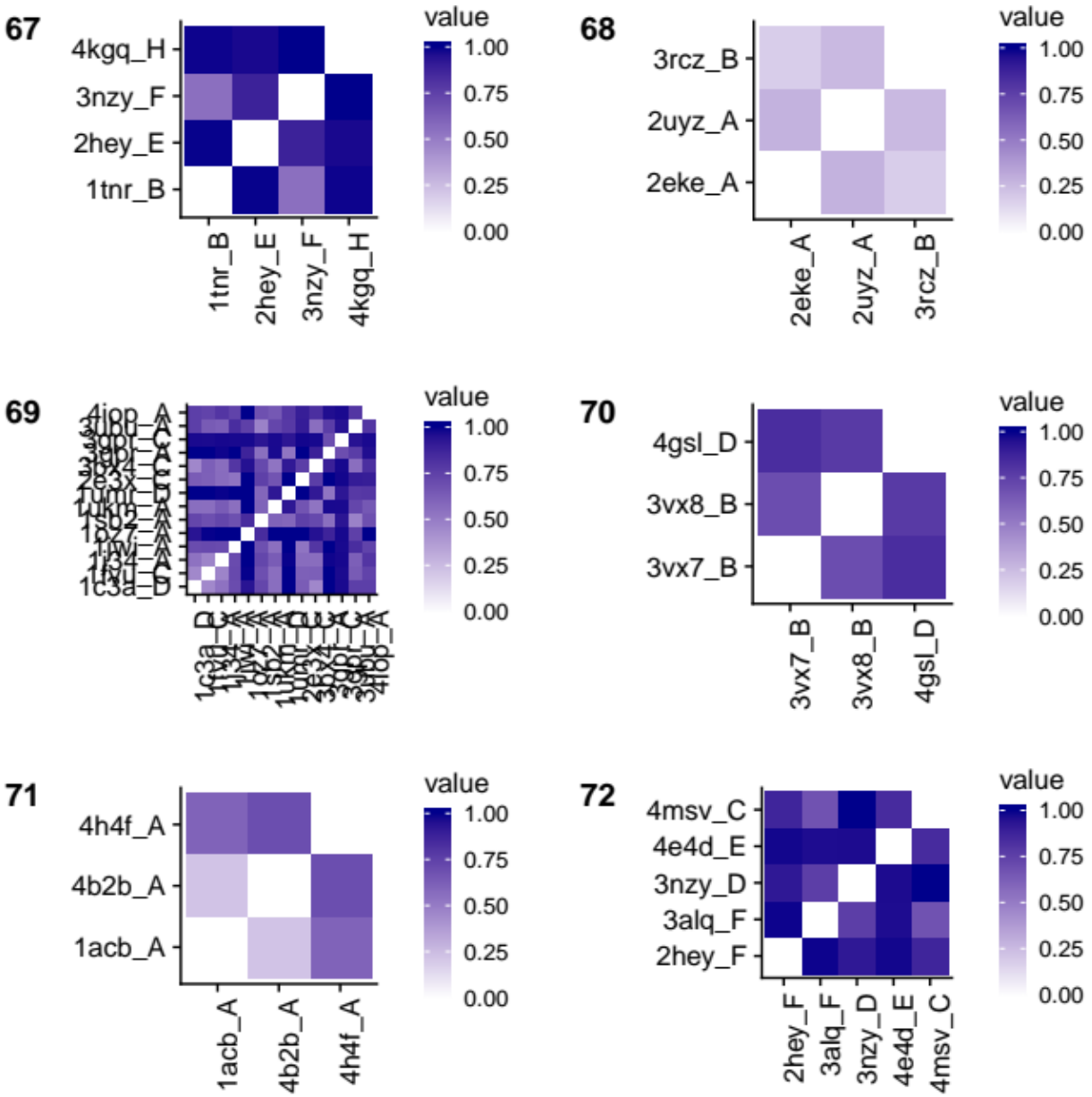

37
2j1k_C
2j12_B
1kac_B
1kac_B
2j12_B
2j1k_C
value
1.00
0.75
0.50
0.25
0.00
38
4v3l_D
2yho_A
2oxq_D
2oxq_D
2yho_A
4v3l_D
value
1.00
0.75
0.50
0.25
0.00
39
4ifd_A
2wnr_E
2po1_H
2je6_C
2je6_C
2po1_H
2wnr_E
4ifd_A
value
1.00
0.75
0.50
0.25
0.00
40
3rpf_B
2q5w_B
1fm0_B
1fm0_B
2q5w_B
3rpf_B
value
1.00
0.75
0.50
0.25
0.00
41
3rpf_D
2q5w_D
1fm0_A
1fm0_A
2q5w_D
3rpf_D
value
1.00
0.75
0.50
0.25
0.00
42
3vr4_C
3gqb_A
2qe7_A
2jdi_A
1fx0_C
1fx0_C
2jdi_A
2qe7_A
3gqb_A
3vr4_C
value
1.00
0.75
0.50
0.25
0.00

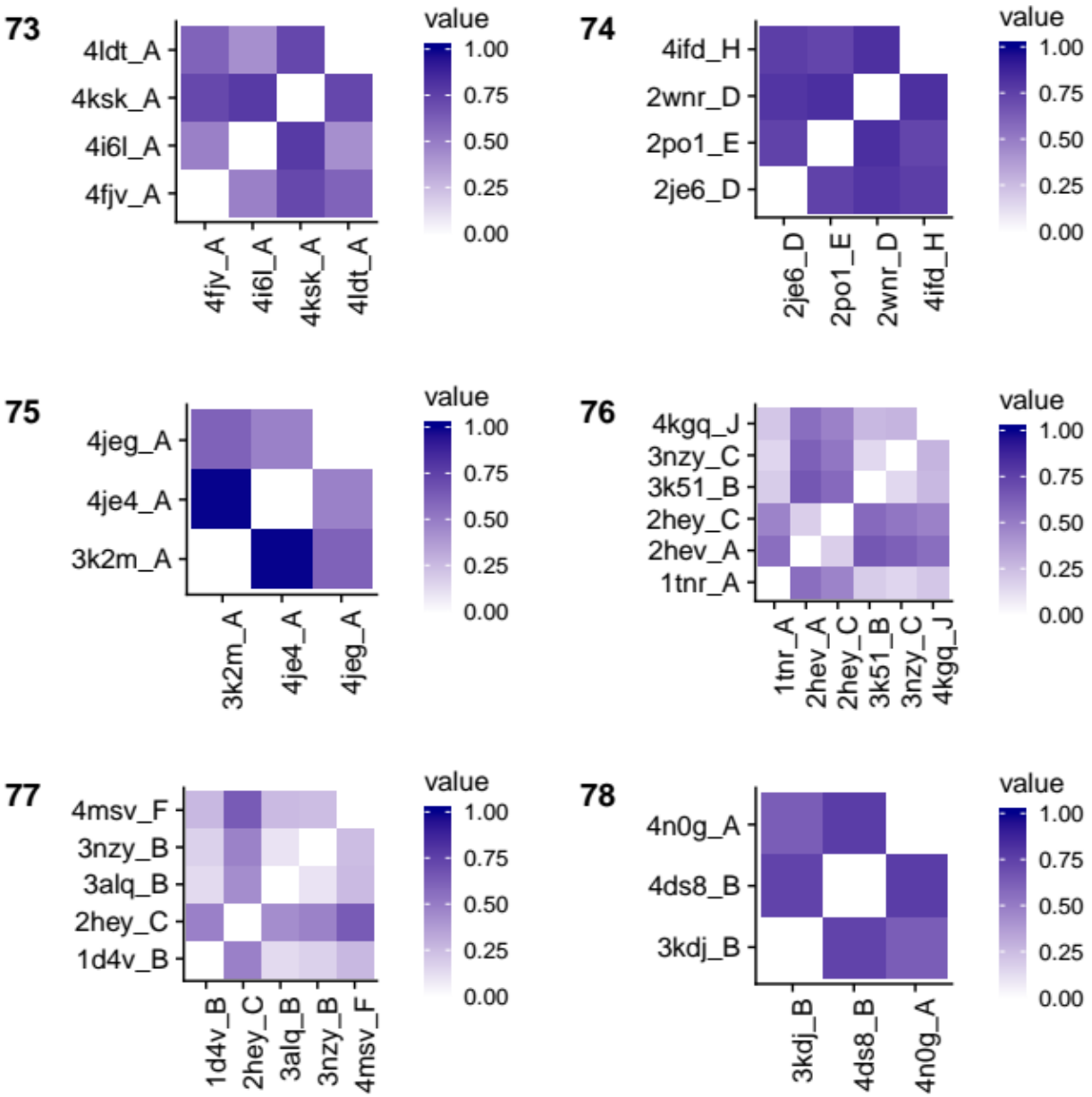

43
3rcz_A
2uyz_B
2eke_B
2eke_B
2uyz_B
3rcz_A
value
1.00
0.75
0.50
0.25
0.00
44
4po5_C
3l0f_E
2vml_F
2c7l_C
1kn1_F
1b8d_C
1b8d_C
1kn1_F
2c7l_C
2vml_F
3l0f_E
4po5_C
value
1.00
0.75
0.50
0.25
0.00
45
3hei_A
2wo3_A
2hle_A
2hle_A
2wo3_A
3hei_A
value
1.00
0.75
0.50
0.25
0.00
46
4wm8_C
4wm8_B
3cji_B
2x5i_C
2x5i_B
2wzr_F
2wzr_F
2x5i_B
2x5i_C
3cji_B
4wm8_B
4wm8_C
value
1.00
0.75
0.50
0.25
0.00
47
4ol0_B
3m1i_C
3icq_A
3gjx_B
3a6p_A
2x19_B
2x19_B
3a6p_A
3gjx_B
3icq_A
3m1i_C
4ol0_B
value
1.00
0.75
0.50
0.25
0.00
48
3d1k_C
3bom_B
1wmu_C
1cg5_C
1cg5_C
1wmu_C
3bom_B
3d1k_C
value
1.00
0.75
0.50
0.25
0.00

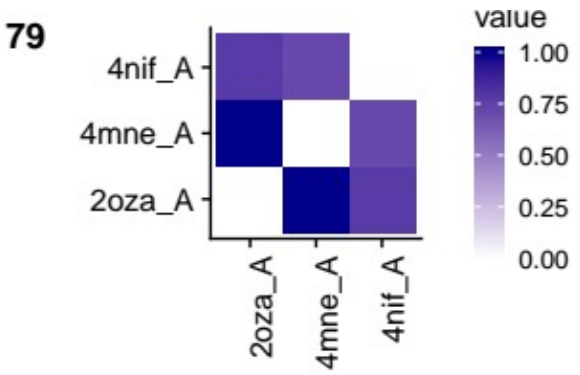

49
3bs5_B
1sv0_A
1pk1_B
1pk1_B
1sv0_A
3bs5_B
value
1.00
0.75
0.50
0.25
0.00
50
4wm8_A
3cji_A
2x5i_A
2wzr_I
2wzr_I
2x5i_A
3cji_A
4wm8_A
value
1.00
0.75
0.50
0.25
0.00
51
3cpj_B
3cph_B
2bcg_B
1vg0_B
1vg0_B
2bcg_B
3cph_B
3cpj_B
value
1.00
0.75
0.50
0.25
0.00
52
3g9v_A
3dlq_B
1lqs_A
1lqs_A
3dlq_B
3g9v_A
value
1.00
0.75
0.50
0.25
0.00
53
3e1z_B
3cbj_A
2oul_A
2nqd_B
2nqd_B
2oul_A
3cbj_A
3e1z_B
value
1.00
0.75
0.50
0.25
0.00
54
3w3z_A
3ea5_B
1qbk_A
1qbk_A
3ea5_B
3w3z_A
value
1.00
0.75
0.50
0.25
0.00

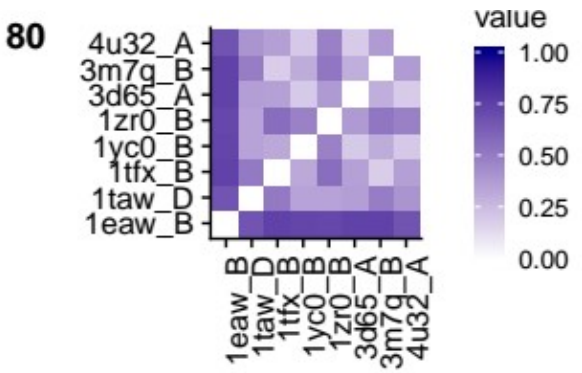

55
3kfd_A
3evs_B
2h62_B
2h62_B
3evs_B
3kfd_A
value
1.00
0.75
0.50
0.25
0.00
56
3kfd_F
3evs_C
2h62_C
2h62_C
3evs_C
3kfd_F
value
1.00
0.75
0.50
0.25
0.00
57
3fc6_B
2r40_B
2nxx_B
1dkf_B
1dkf_B
2nxx_B
2r40_B
3fc6_B
value
1.00
0.75
0.50
0.25
0.00
58
3vr4_D
3gqb_B
2qe7_E
2qe7_E
3gqb_B
3vr4_D
value
1.00
0.75
0.50
0.25
0.00
59
4l0p_B
3hei_B
2wo3_B
2wo2_B
2wo2_B
2wo3_B
3hei_B
4l0p_B
value
1.00
0.75
0.50
0.25
0.00
60
4j2y_A
4an7_B
3i29_B
1avw_B
1avw_B
3i29_B
4an7_B
4j2y_A
value
1.00
0.75
0.50
0.25
0.00

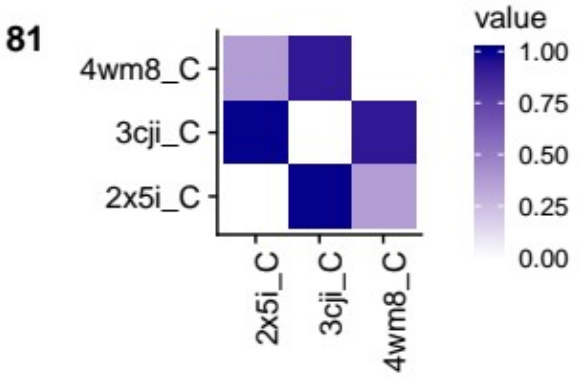

61
4i6n_A
3ifw_A
2wdt_A
1xd3_A
1cmx_A
1cmx_A
1xd3_A
2wdt_A
3ifw_A
4i6n_A
value
1.00
0.75
0.50
0.25
0.00
62
3ima_B
1yvb_B
1nb5_C
1nb5_C
1yvb_B
3ima_B
value
1.00
0.75
0.50
0.25
0.00
63
4n0g_B
4ds8_A
3rt0_B
3kdj_A
3kb3_A
3kb3_A
3kdj_A
3rt0_B
4ds8_A
4n0g_B
value
1.00
0.75
0.50
0.25
0.00
64
3kse_A
3k9m_A
3ima_A
1yvb_A
1nb5_A
1nb5_A
1yvb_A
3ima_A
3k9m_A
3kse_A
value
1.00
0.75
0.50
0.25
0.00
65
4cxa_A
3mi9_A
1g3n_A
1f5q_A
1f5q_A
1g3n_A
3mi9_A
4cxa_A
value
1.00
0.75
0.50
0.25
0.00
66
4m0w_A
3wxg_A
3nhe_A
3n3k_A
2y5b_A
1nbf_A
1nbf_A
2y5b_A
3n3k_A
3nhe_A
3wxg_A
4m0w_A
value
1.00
0.75
0.50
0.25
0.00


67
4kgq_H
3nzy_F
2hey_E
1tnr_B
1tnr_B
2hey_E
3nzy_F
4kgq_H
value
1.00
0.75
0.50
0.25
0.00
68
3rcz_B
2uyz_A
2eke_A
2eke_A
2uyz_A
3rcz_B
value
1.00
0.75
0.50
0.25
0.00
69
1c3a_D
value
1.00
0.75
0.50
0.25
0.00
70
4gsl_D
3vx8_B
3vx7_B
3vx7_B
3vx8_B
4gsl_D
value
1.00
0.75
0.50
0.25
0.00
71
4h4f_A
4b2b_A
1acb_A
1acb_A
4b2b_A
4h4f_A
value
1.00
0.75
0.50
0.25
0.00
72
4msv_C
4e4d_E
3nzy_D
3alq_F
2hey_F
2hey_F
3alq_F
3nzy_D
4e4d_E
4msv_C
value
1.00
0.75
0.50
0.25
0.00


73
4ldt_A
4ksk_A
4i6l_A
4fjv_A
4fjv_A
4i6l_A
4ksk_A
4ldt_A
value
1.00
0.75
0.50
0.25
0.00
74
4ifd_H
2wnr_D
2po1_E
2je6_D
2je6_D
2po1_E
2wnr_D
4ifd_H
value
1.00
0.75
0.50
0.25
0.00
75
4jeg_A
4je4_A
3k2m_A
3k2m_A
4je4_A
4jeg_A
value
1.00
0.75
0.50
0.25
0.00
76
4kgq_J
3nzy_C
3k51_B
2hey_C
2hev_A
1tnr_A
1tnr_A
2hev_A
2hey_C
3k51_B
3nzy_C
4kgq_J
value
1.00
0.75
0.50
0.25
0.00
77
4msv_F
3nzy_B
3alq_B
2hey_C
1d4v_B
1d4v_B
2hey_C
3alq_B
3nzy_B
4msv_F
value
1.00
0.75
0.50
0.25
0.00
78
4n0g_A
4ds8_B
3kdj_B
3kdj_B
4ds8_B
4n0g_A
value
1.00
0.75
0.50
0.25
0.00


79
4nif_A
4mne_A
2oza_A
2oza_A
4mne_A
4nif_A
value
1.00
0.75
0.50
0.25
0.00


80
4u32_A
3m7q_B
3d65_A
1zr0_B
1yc0_B
1tfx_B
1taw_D
1eaw_B
1eaw_B
1taw_D
1tfx_B
1yc0_B
1zr0_B
3d65_A
3m7q_B
4u32_A
value
1.00
0.75
0.50
0.25
0.00


81
4wm8_C
3cji_C
2x5i_C
2x5i_C
3cji_C
4wm8_C
value
1.00
0.75
0.50
0.25
0.00

Titre : Développement d'une méthode *in silico* pour caractériser le potentiel d'interaction des surfaces protéiques dans un environnement encombré

Mots clés : biologie structurale, amarrage moléculaire, interactions protéine-protéine

Résumé : Dans la cellule, les protéines évoluent dans un environnement très dense et interagissent ainsi avec un grand nombre de partenaires spécifiques et non-spécifiques qui entrent en compétition. L'objectif de ma thèse est de caractériser les propriétés physiques et évolutives des surfaces protéiques pour comprendre comment la pression de sélection s'exerce sur les protéines, façonnant leurs interactions et régulant ainsi cette sévère compétition.

Pour cela, j'ai développé une méthodologie permettant de caractériser la propension des protéines à interagir avec les protéines de leur environnement, par des approches de *docking*. La cartographie moléculaire permettant la visualisation et la comparaison des propriétés de la surface des protéines, j'ai donc mis en place un nouveau cadre théorique basé sur une représentation des paysages énergétiques d'interaction par des cartes d'énergies. Ces cartes (en deux dimensions) reflètent de manière synthétique la propension des surfaces protéiques à engager des interactions avec d'autres protéines. Elles sont donc d'un grand intérêt pratique pour déterminer les régions des surfaces protéiques les plus enclines à engager des interactions avec d'autres molécules.

Ce nouveau cadre théorique a permis de montrer que les surfaces des protéines comprennent des régions de différents niveaux d'énergies de liaison (régions chaudes, intermédiaires et froides pour les régions d'interaction favorables, intermédiaires et défavorables respectivement).

Une partie importante de la thèse a consisté à caractériser les propriétés physico-chimiques et évolutives de ces différentes régions. L'autre partie a consisté à appliquer cette méthode sur plusieurs systèmes : complexes homomériques, protéines du cytosol de *S. cerevisiae*, familles d'interologues. Ce travail ouvre la voie à un grand nombre d'applications en bioinformatique structurale, telles que la prédiction de sites de liaison, l'annotation fonctionnelle ou encore le design de nouvelles interactions.

En conclusion, la stratégie mise en place lors de ma thèse permet d'explorer la propension d'une protéine à interagir avec des centaines de partenaires d'intérêts, et donc d'investiguer le comportement d'une protéine dans un environnement cellulaire spécifique. Cela va donc au-delà de l'utilisation classique du *docking* "binaire" puisque notre stratégie fournit une vision systémique des interactions protéiques à l'échelle des "résidus".

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Development of an in silico method to characterize the interaction potential of protein surfaces in a crowded environment

Keywords : structural biology, molecular docking, protein-protein interactions

Abstract : In the crowded cell, proteins interact with their functional partners, but also with a large number of non-functional partners that compete with the first ones. The goal of this thesis is to characterize the physical properties and the evolution of protein surfaces in order to understand how selection pressure exerts on proteins, shaping their interactions and regulating this severe competition.

To do this I developed a framework based on docking calculations to characterize the propensity of protein surfaces to interact with other proteins. Molecular cartography enables the visualization and the comparison of surface properties of proteins. I implemented a new theoretical framework based on the representation of interaction energy landscapes by 2-D energy maps. These maps reflect in a synthetic manner the propensity of the surface of proteins to interact with other proteins. These maps are useful from a practical point view for determining the regions of protein's surface that are more prone to interact with other proteins.Our new theoretical framework enabled to show that the surface of proteins harbor regions with different levels of propensity to interact with other proteins (hot regions, intermediate and cold regions to favorable, intermediate and unfavorable regions respectively).

A large part of this thesis work consisted in characterizing the physico-chemical properties and the evolution of these regions. The other part of this thesis work consisted in applying this methodology on several study systems: homomeric complexes, cytosolic proteins from S. cerevisiae, families of interologs. This work opens the way to numerous practical applications in structural bioinformatics, such as binding site prediction, functional annotation and the design of new interactions.

To conclude, the strategy implemented in this work enable the exploration of the propensity of a protein to interact with hundred of protein partners. It thus enables the investigation of the behavior of a protein in a crowded environment. This application goes beyond the classical use of protein docking as a, because our strategy provides a systemic point of view of protein interactions at an atomic resolution.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France