Géométrie des interactions protéiques

Chloé Dequeker

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS
DEPUIS 1257

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Chloé DEQUEKER

Pour obtenir le grade de

DOCTEURE de SORBONNE UNIVERSITÉ

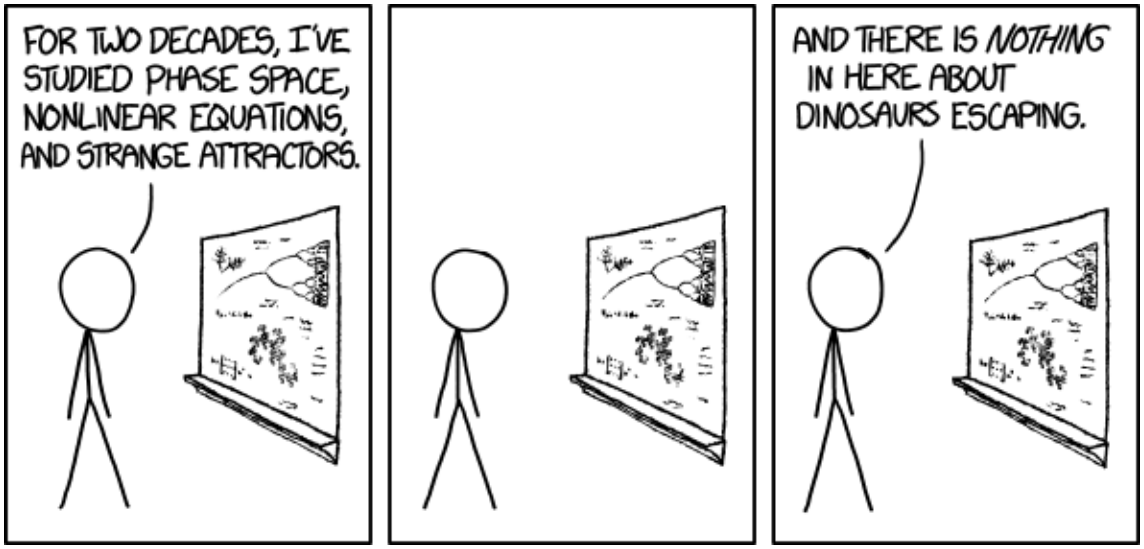
Sujet de la thèse :

Géométrie des Interactions Protéiques

Soutenance : le 21 septembre 2018

devant le jury composé de :

Mme. Alessandra Carbone	Directrice de thèse
Mme. Elodie Laine	Encadrante de thèse
Mme. Catherine Etchebest	Rapporteure
Mme. Anne Poupon	Rapporteure
M. Jacques Chomilier	Examinateur
Mme. Sophie Sacquin-Mora	Examinatrice
M. Jean-Daniel Zucker	Examinateur



Proteins are like dinosaurs, but better conserved. Fortunately, there is also plenty to say about them, especially the well conserved ones. Still nothing about dinosaurs escaping though.

https://xkcd.com/1399/

Thanks

I would like first and foremost to thank all the people that have surrounded me during these three years, to have accompanied me through the hardship that can be a doctorate. Sure, it can have its ups and downs, but this has overall been a wonderful experience which has opened me in so many different ways that I will definitely not have enough space to write here. I would like to thank my dad. He has been here for 26 years and will sadly miss the finishing line, but what matters is the journey, not the end. "Tiens bon, camarade" would he have said. I would like to thank my mum and sister who have borne me for quite some time now and supported me all the way. Then thank you to all the people, friends, mentors, colleagues, that have supported me (and my puns) for now 3 years (or more for some). To Alessandra and Elodie to have guided me through this work, and to Sophie Sacquin-Mora and Jean-Daniel Zucker for being part of my "Comité de suivi" and providing me with useful insights and ideas. To Anne Poupon and Catherine Etchebest for accepting to review my thesis. To the other members of the committee as well, Jacques Chomilier, Sophie Sacquin-Mora, Jean-Daniel Zucker, thank you for accepting of being a part of my committee. We thank Thom Vreven for providing his source code that is used in INTBuilder for reconstructing PDBs from the ZDOCK output.

To Andréa, Julien, Perrine, Colin, Elin, Diego, Tristan, Riccardo, Guillaume, Flavia, Francesco, Laurent, Nika, Louise, Mark: you have all changed my life during these past three years. And the last, but far from the least, I would like to mention my violin which I have grown to love and play during my doctorate.

Contents

List of Figures

List of Tables

Abstract

Protein-Protein Interactions (PPI) are at the centre of many biological processes, and their understanding is therefore of the utmost importance. This work focuses on two different aspect: the first is the prediction of interacting surfaces of the protein through computational means, relying on features such as the conservation of residues, their physico-chemical properties, the local geometry of the protein or a score derived from the protein's behaviour in a crowded environnement, inferred from Complete Cross-Docking (CC-D) calculations. The second part of this work focuses on the detection of interacting partners from a large scale CC-D; this part uses a combination of methods to score the likelihood for two proteins to interact with one another and present how it might be possible to apply such methods for even larger scale.

Chapter 1

Résumé en français

Contents

1.1 Présentation du contexte

La plupart des processus biologiques sont régulés par des interactions protéine-protéine (PPI). Une protéine est un polypeptide formé par un nombre variable de résidus d'acides aminés chaînés (de quelques dizaines à plusieurs milliers). Chacun de ces résidus possède une chaîne auxiliaire qui va définir ses propriétés. L'ensemble des résidus définissant une protéine détermine sa structure 3D ainsi que sa fonction biologique. Les protéines interagissent entre elles en se liant l'une à l'autre. Plusieurs types d'interactions sont présents et présentent différentes forces : certaines interactions instables seront de courte durée tandis que d'autres vont maintenir la protéine liée au sein d'un complexe biologique. Les processus biologiques sont le plus souvent régulés par une chaîne d'interactions entre protéines, appelée voie de signalisation et pouvant impliquer des centaines voire des milliers d'acteurs différents.

1.1.1 Le projet Help Cure Muscular Dystrophy

Le projet Help Cure Muscular Dystrophy (HCMD, ou "Aidons à Guérir la Dystrophie Musculaire") a pour but d'étudier les interactions de 2246 protéines humaines impliquées dans la dystrophie musculaire et pour lesquelles les structures 3D sont connues. Le but final est de pouvoir être capable de décrire de façon computationelle leurs interactions et ainsi d'aider à comprendre le rôle qu'elles jouent dans les différentes voies de signalisation impliquées. Ce projet est constitué de deux parties principales :

Phase 1 consiste en l'analyse de 84 complexes protéiques provenant d'un jeu de données benchmark de docking [76]. Chaque complexe de cet ensemble représente une interaction binaire (d'une protéine en particulier vers une autre protéine), résultant ainsi en 168 protéines différentes. L'équipe du laboratoire a effectué une expérience de docking asymétrique en utilisant le logiciel de docking MAXDo [95] qui y a été développé. L'expérience, qui a duré 7 mois et a fini en juin 2007, a été lancée sur la World Community Grid[1] (WCG), une organisation publique permettant à des personnes du public volontaires de participer à des projets de recherche en donnant du temps de calcul de leur ordinateur. Le rôle principal de cette phase est d'avoir un jeu sur lequel développer des algorithmes et obtenir un retour sur leur performance.

Phase 2 implique les 2246 chaînes protéiques pour lesquelles nous ne connaissons pas leur(s) partenaire(s). Un second CC-D a été réalisé sur cette seconde phase sur la WCG également, et a duré plus de quatre années entre mai 2009 et automne 2013.

[1] www.worldcommunitygrid.org

La première publication sur ces jeux de données dans le cadre de ce projet [67] a présenté notre capacité de détection des partenaires interagissant vis-à-vis des partenaires non-interagissants.

1.1.2 But de la thèse

Dans ce cadre, mon travail de doctorat a pour but d'améliorer les algorithmes pré-existants et d'en apporter de nouveaux pour le passage à l'échelle de la phase 2 du projet. Une première analyse [95] antérieure au lancement des deux phases précédemment décrites avait établi qu'il était possible de détecter les partenaires interagissant au sein d'un jeu de données en combinant les résultats de docking avec une description précise des sites d'interaction. Cette étude a motivé et été confirmée par la suivante, réalisée à plus grande échelle [67]. Cette dernière étude décrit nos capacités de discrimination à grande échelle vis-à-vis du jeu de données obtenu en première phase du projet, en utilisant les interfaces expérimentales (connues) des complexes ainsi que les prédictions réalisées à l'aide du programme développé dans l'équipe, JET [29]. Cette étude présente des résultats prometteurs, et mon travail consiste à apporter de nouveaux concepts ainsi que d'adresser les points faibles du projet sur les différents points suivants :

- Analyser les différentes façons des protéines d'interagir entre elles
- Développer de nouvelles méthodes et pipelines pour permettre une meilleure compréhension et une meilleure exploitation de l'interface de liaison entre deux protéines
- Combiner les connaissances et concepts ainsi acquis pour fournir une méthode efficace d'identification des partenaires dans le cadre d'un CC-D

Difficulté de la problématique

Bien que la prédiction des sites d'interaction soit un domaine très étudié, la prédiction à grande échelle de partenaires en interaction à travers un CC-D est encore à l'état de travail pionnier, ouvrant la voie à de futures études. La capacité à comprendre comment les protéines interagissent, en plus de fournir une meilleure compréhension de la régulation des processus biologiques, apporte de nombreuses applications pratiques pour la conception de petites molécules et pour la recherche contre de nombreuses maladies.

Comprendre comment fonctionnent les réseaux à grande échelle nous permettra ainsi de mettre en œuvre une automatisation où la plupart du travail est aujourd'hui effectué manuellement. Ne serait-ce qu'être capable de réduire la taille potentielle des partenaires en interaction pourrait permettre de considérablement réduire un

travail laborieux. L'amarrage moléculaire (docking) a jusqu'à présent été principalement utilisé pour discriminer les conformations natives d'un complexe protéique parmi un ensemble de leurres. Cependant, l'équipe du laboratoire a montré dans une étude précédente [95] que combiner des interfaces connues avec les conformations de docking (et leur énergie associée) était suffisant pour discriminer les interactions des partenaires interagissant par rapport aux non-interagissants. Cela a largement motivé le développement du logiciel de prédiction d'interface protéine-protéine par le laboratoire [29, 57]. Les conformations de docking ont rarement été analysées sous un tel angle, ce qui représente un défi supplémentaire d'un point de vue méthodologique. De telles études à grande échelle impliquent également le développement de nouvelles méthodes pour les analyser : le CC-D du jeu de données HCMD2 a généré plus de cent milliards de conformations de docking. Un autre important défi, en plus de la complexité combinatoire importante, est le grand espace des partenaires négatifs par rapport aux positifs.

1.1.3 Avancements réalisés

Beaucoup de logiciels actuels de prédiction d'interface protéine-protéine tentent maintenant d'évaluer les interfaces en considérant les interactions binaires avec une autre protéine. Cependant, la cellule est un environnement peuplé (voir Fig. 1.1) et la multiplicité d'interactions qu'une protéine fait et donc sa surface en interaction est largement sous-estimée. Les protéines font continuellement des interactions : certaines courtes et d'autres plus persistantes (voir Section 2.3.2). Elles peuvent ainsi interagir en compétition ou en coopération les unes avec les autres [64]. De fait, je souligne à quel point il est important de changer de paradigme d'une recherche de paires de partenaires vers une recherche de multiples interacteurs, potentiellement simultanément. Ce nouveau changement ouvre avec lui de nombreuses questions : quelles sont les limites d'une interface partagée entre plusieurs partenaires ? L'interaction binaire a-t-elle encore un sens ?

Plan de la thèse

Je présente dans cette thèse les avancements que j'ai réalisés par rapport aux points abordés ci-dessus.

Dans la première partie, je fournis une meilleure compréhension de l'interaction entre protéines dans un environnement peuplé, lorsque plusieurs interactions sont possibles. J'introduis ainsi les concepts de sites d'interaction multiples (spécifiques à un partenaire) et de régions (non spécifique) et comment nos prédictions pourraient nous guider vers une meilleure compréhension de ces derniers. Ceci est réalisé par une analyse d'un ensemble de données original de 262 chaînes protéiques. Cette section couvre mes objectifs d'analyse de la façon dont les protéines peuvent interagir

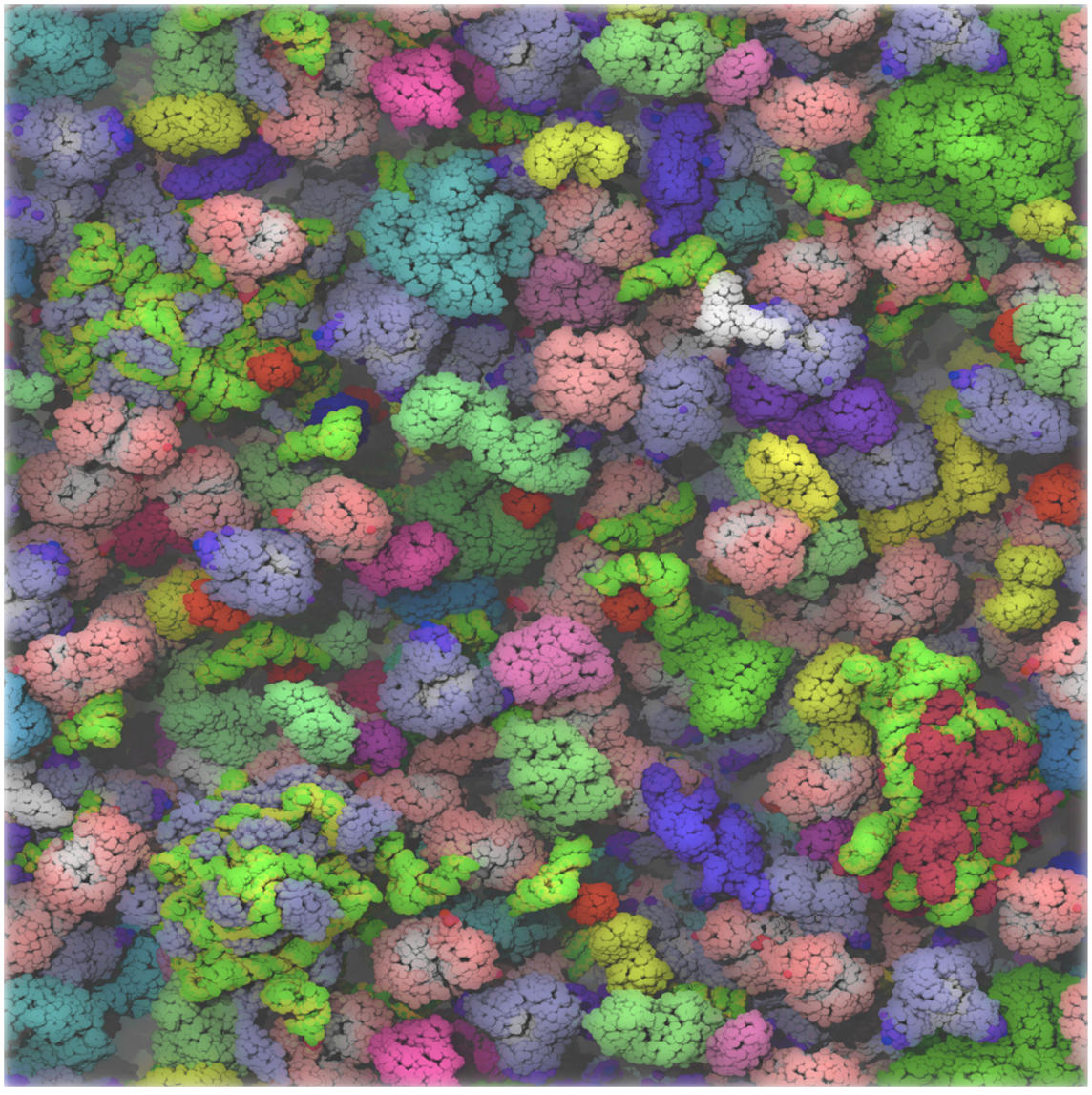


Figure 1.1: Représentation schématique d'un environnement cellulaire peuplé [73].
Crowded cell environment [73].

entre elles et sur le développement de nouveaux concepts et pipelines pour interpréter leurs interactions.

La deuxième partie met l'accent sur les progrès accomplis dans l'identification des partenaires interagissant dans le cadre d'un CC-D à large échelle. En combinant les connaissances jusque-là acquises et la meilleure compréhension des scores des résidus à l'interface et leurs rôles dans les interactions protéine-protéine, je montre à quel point il est essentiel de séparer les protéines par leurs fonctions respectives. J'apporte également une analyse sur ces différentes classes fonctionnelles indiquant comment certaines d'entre elles réagissent différemment à certains scores.

La quantité énorme de données générées lors de l'expérience de CC-D pour HCMD2 a nécessité le développement d'un logiciel rapide et adapté pour obtenir les interfaces de docking correspondantes : j'ai donc développé INTerface Builder (INTBuilder; voir Chapitre 6, [25]) pour répondre à ce problème. Dans la troisième partie je présente son développement ainsi que le nouvel algorithme pour la réduction d'espace de recherche qu'il apporte avec lui.

1.2 Méthodes

Afin de pouvoir correctement aborder les différents points sur lesquels j'ai travaillé, il est nécessaire de présenter les différentes méthodes avec lesquelles j'ai travaillé. Un grand nombre de ces méthodes reposant sur l'analyse de deux jeux de données de protéines, ce seront donc ceux-ci que je présenterai premièrement.

1.2.1 Jeux de données

`P-262`, un jeu de données de chaînes protéiques

Ce nouvel ensemble de données, nommé `P-262`, est un sous-ensemble du plus grand jeu de données de 2246 chaînes protéiques étudié dans le cadre du projet HCMD2. L'analyse de ce dernier nous a montré que certaines de ses structures appartenaient à des complexes connus et qu'il était ainsi possible de construire un sous-ensemble de protéines. Les chaînes de `P-262` sont celles restantes après avoir exclu : (a) les structures uniquement α-carbonnées (b) les chaînes pour lesquelles les résultats n'étaient pas disponibles (c) les chaînes formant des complexes coiled-coil (d) les complexes ayant des codes PDB obsolètes (e) les chaînes pour lesquelles aucune interface de 5 résidus ou plus n'a pu être trouvée dans le complexe PDB associé (f) les chaînes pour lesquelles aucune interface impliquée dans une interaction fonctionnelle biologique ne pouvait être trouvée parmi l'ensemble des homologues de la chaîne protéique dans la PDB (en considérant 90% d'identité de séquence). Étant donné que `P-262` est un nouveau jeu de données que nous avons décrit, il n'y a pas d'autres études

l'ayant analysé. Sur la base des informations récupérées des complexes PDB et suivant la classification de [43], les 262 chaînes de protéines ont été classées en sept classes fonctionnelles différentes : 6 Inhibiteurs (I), 7 G-protéines (G), 13 protéines Récepteurs (R), 17 Anticorps (AB), 10 Enzymes Régulatrices (ER), 56 autres Enzymes (E) et 136 Autres (O) protéines que nous n'avons pu classer dans aucune des autres sous-classes fonctionnelles.

Unité biologique

Les unités biologiques ou assemblages biologiques décrivent des interactions fonctionnelles. De telles unités biologiques sont soit déterminées par l'auteur(e) ou déterminées par un logiciel (PISA [54]) et nous avons choisi de considérer les deux méthodes. Cela garantit que les interfaces calculées dans le complexe à l'aide du logiciel INTBuilder [25] représentent une interaction biologique. Nous avons ainsi défini l'ensemble de données de 262 chaînes différentes provenant de 107 complexes composés de deux ou plusieurs chaînes.

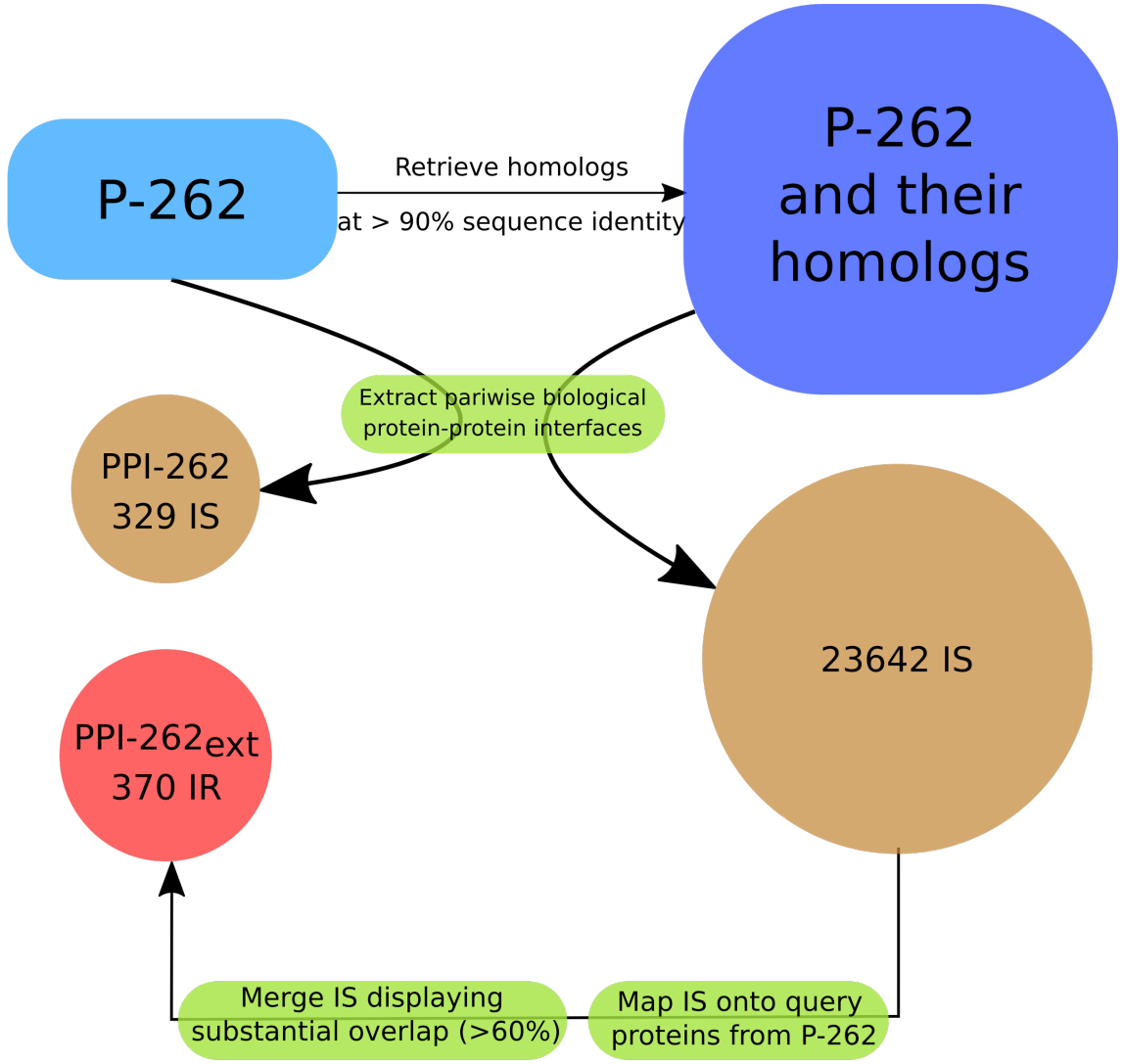
`PPI-262`, un ensemble d'interfaces expérimentales

Pour chaque chaîne, nous avons calculé chaque interaction expérimentalement connue au sein du complexe auquel elle appartient. Nous avons ainsi obtenu `PPI-262`, un ensemble de 329 sites d'interaction expérimentaux (IS, spécifiques à un seul partenaire).

$\texttt{PPI-262}_{\texttt{ext}}$, une extension de l'ensemble des interfaces expérimentales `PPI-262`

Partant de l'observation [69] que les interfaces fonctionnelles sont souvent conservées parmi les homologues proches, nous avons pu définir un ensemble de surfaces expérimentales depuis ceux-ci que nous avons par la suite reportées sur la protéine étudiée par alignement de séquence. En fusionnant ces surfaces, nous avons pu définir des régions expérimentales (IR, utilisées par un ou plusieurs partenaires) et avons ainsi obtenu $\texttt{PPI-262}_{\texttt{ext}}$, un jeu de données de 370 IR sur l'ensemble des chaînes de `P-262`. L'ensemble du pipeline est décrit en Figure 1.2. Nous montrons que dynJET^2 est utile pour détecter les IS ainsi que les IR. On note que souvent, la définition d'IR est plus biologiquement pertinente dans le cadre où elle rend mieux compte de la multiplicité des interactions.

Dans le cadre de ce travail de détection de multiple sites ou régions d'interactions, j'ai également développé un programme permettant d'effectuer une recherche des homologues de la protéine et d'obtenir leur(s) interface(s) expérimentale(s), avant d'effectuer un alignement de séquence pour traduire ces interfaces sur la protéine étudiée. Le développement de cet outil répond à un réel besoin et son développement



IS: Interaction Sites
IR: Interaction Regions

Figure 1.2: Représentation schématique du pipeline suivi pour obtenir les différents sets PPI-262 et $\texttt{PPI-262}_{\texttt{ext}}$.
Schematic representation of the pipeline we followed to obtain the different sets PPI-262 and $\textit{PPI-262}_{ext}$.

a nécessité un effort important afin qu'il soit par la suite facilement accessible et distribuable.

1.2.2 Jeu de données PPDBv2

Le jeu de données "Protein-Protein Dataset Benchmark" (PPDBv2) comprend 84 complexes protéiques connus qui ont chacun été séparé en tant que récepteur et ligand dans leur forme non liée. Ces complexes ne se réfèrent pas toujours à une seule chaîne, mais peuvent regrouper plusieurs d'entre elles en tant qu'unité biologique multimérique.

Une description antérieure [76] de l'ensemble de données ne le divisait que dans quatre sous-ensembles différents: Enzyme-Inhibitor (EI), Anticorps-Antigènes (forme non liée ; AA), Antigènes-Anticorps (forme liée ; ABA), Autres (OX). Tous les complexes sont sous la forme non liée (état qu'ils adoptent quand ils ne sont liés à aucun autre partenaire) mis à part le sous-ensemble ABA (pour lequel la structure représente les changements conformationnels subis lors de la liaison). Cette description, bien qu'elle fût celle considérée au début de mon travail de thèse, a été mise à jour dans [43]. Cette mise à jour du jeu de données fournit de nouvelles classifications séparant les protéines en classes fonctionnelles plus raffinées ainsi que de nouvelles structures protéiques à analyser. Bien que nous considérions la classification la plus précise pour les 168 protéines précédentes, nous n'avons pas pris en compte les nouvelles structures apportées par la mise à jour ; un cross docking complet a été réalisé sur les 168 premières protéines et nous n'avons pas la capacité de calcul pour réitérer la même expérience de CC-D en utilisant le même logiciel de docking MAXDo pour les nouvelles protéines.

En utilisant les nouvelles classifications de protéines de [43], nous obtenons donc le nombre suivant de protéines pour chacune des classes fonctionnelles : 20 anticorps-antigènes (forme non liée ; AA), 24 anticorps-antigènes (forme liée ; ABA), 38 enzyme-inhibiteurs (EI), 6 enzymes (avec une chaîne régulatrice ou accessoire ; ER), 12 enzyme-substrat (ES), 14 Autres contenant des G-protéines (OG), 14 Autres contenant des récepteurs (OR), 30 Autres ne pouvant être classifiées autre part (OX).

1.2.3 Amarrage Moléculaire

Des méthodes expérimentales basées sur la Résonance Magnétique Nucléaire (NMR) ou par cristallographie par rayons-X ont été utilisées pour obtenir la structure 3D de nombreux complexes protéiques. Cependant, l'accroissement important de nouvelles séquences protéiques découvertes chaque année continue d'augmenter et il est clairement apparent que de telles méthodes (NMR, rayons-X) ne permettent pas la résolution de structures 3D de complexes protéiques à une vitesse suffisante.

Pour pallier ce problème, de plus en plus d'efforts ont été employés à développer des méthodes computationelles pour simuler le processus d'interaction entre deux protéines (docking, ou amarrage moléculaire). Docker deux protéines consiste à prendre deux structures protéiques (une en tant que récepteur, l'autre en tant que ligand) et d'échantillonner l'espace autour du récepteur avec différentes positions du ligand. On obtient ainsi environ 300 000 conformations par couple de protéines. Une fois les échantillons obtenus, l'algorithme de docking va utiliser une fonction d'énergie pour évaluer chacune des différentes conformations. Cette fonction d'énergie va permettre de déterminer la stabilité de l'interaction entre les deux protéines. Le principe de la fonction d'énergie dans le domaine du docking moléculaire est de pouvoir discriminer une conformation favorable (où le récepteur et le ligand interagissent réellement ensemble dans un complexe biologique) par rapport à des conformations qui présenteraient moins de stabilité. Il existe de nombreux algorithmes de docking moléculaire [98, 95, 115, 116, 35, 19, 109, 104, 26] qui se basent sur différentes propriétés telles que la distance des atomes ou leurs propriétés physico-chimiques.

Plusieurs classes d'algorithmes de docking existent : les algorithmes avec une approche rigide (rigid-body docking), les algorithmes flexibles et des algorithmes hybrides. La première grande classe considère les protéines comme des objets immuables et échantillonne les différentes orientations possibles sans tenir compte des changements conformationels pendant l'étape de fixation d'une protéine à l'autre partenaire. Cette méthode présente une certaine modélisation de la réalité mais permet ainsi un temps de calcul bien inférieur à une approche entièrement flexible. Les algorithmes hybrides docking peuvent appliquer une étape de minimisation de l'énergie pour chacune des conformations obtenues. Cette étape permet d'effectuer des changements de conformations mineurs mais qui peuvent se révéler cruciaux afin de correctement évaluer la conformation obtenue. Afin de réduire le temps de calcul de docking, certaines approches ont été explorées telles que la modélisation en gros-grain des protéines où plusieurs atomes sont fusionnés en un seul, ou bien où la surface est approximée par un ensemble de fonctions gaussiennes à plus ou moins haute résolution.

1.2.4 Cross-Docking Complet

Le Complete Cross-Docking (CC-D, ou docking "tous contre tous") d'un jeu de données consiste à effectuer le docking de tous les couples possibles de protéines dans le jeu de données. Ainsi, pour un jeu de données de n protéines, on obtiendra n^2 couples différents. Dans une expérience de docking asymétrique classique, chaque protéine prend le rôle du récepteur (fixé dans l'espace) ainsi que celui du ligand (qui va orbiter autour du récepteur) ; dans un docking symétrique en revanche les deux protéines vont orbiter simultanément, il n'y aura ainsi pas de rôle tel que récepteur

ou ligand, réduisant ainsi effectivement le temps de calcul de moitié.

1.3 INTerface Builder: Un outil rapide de reconstruction d'interfaces

L'accroissement des ressources et de la puissance de calcul disponible ainsi que le développement des algorithmes de docking [35, 95, 87] ont permis d'étudier à grande échelle les PPI, où des dizaines à des milliers de protéines sont dockées les unes aux autres [95, 67, 56]. De ce fait, les calculs de CC-D génèrent des quantités très importantes (plus de 100 milliards) de conformations qui doivent être examinées afin d'en extraire les informations pertinentes. Plusieurs types d'analyses peuvent être effectuées, parmi lesquelles le calcul de la propension des résidus à se trouver à l'interface dans les conformations de docking. Cette propriété en particulier peut être exploitée afin de mieux prédire les sites d'interaction protéine-protéine [33, 95, 56] ainsi que les fonctions de ces dernières [107]. De plus, les interfaces des conformations de docking peuvent être analysées pour sélectionner les plus plausibles afin de détecter les partenaires interagissant dans la cellule [95, 67, 56]. Dans chaque cas, les analyses nécessitent une détection rapide et précise des résidus d'interface dans la conformation de docking.

Les approches les plus performantes identifient les résidus en interaction en fonction d'un critère de distance les séparant, des changements de surface accessible au solvant (SASA) au cours de l'interaction [60] ou selon une modélisation de l'interface par une triangulation de Voronoi [15]. Ces méthodes, bien que précises, ne sont pas suffisamment rapides pour la très importante quantité de données qu'il nous est nécessaire de traiter. Puisque le nombre de conformations peut atteindre plusieurs milliards sur des expériences de docking a grande échelle, l'algorithme utilisé doit donc être rapide et efficace. D'une part, les approches basées des grilles [102, 78] détectent efficacement les interactions entre les particules sur un critère de distance en complexité linéaire. D'autre part, le modèle Voronoi fournit une description plus détaillée de l'interface au détriment du temps de calcul plus important. Un autre goulot d'étranglement est l'entrée/sortie (I/O) requise. L'analyse des fichiers en utilisant les outils demande aujourd'hui l'écriture et la lecture de chaque fichier PDB, pour chaque conformation. Ce processus résulte en un très important I/O et il s'agit d'un point qu'il est nécessaire d'adresser dans le développement de cette nouvelle méthode. Les deux questions sont cruciales pour l'analyse des grands ensembles de docking. Spécifiquement pour les résoudre, j'ai développé INTerface Builder (INTBuilder), qui combine un nouvel algorithme réduisant l'espace de recherche avec une capacité de lire directement les fichiers de sorties des logiciels de docking les plus utilisés. En effet, l'algorithme d'INTBuilder (détaillé dans le Chapitre 6) peut atteindre une complexité de $\mathcal{O}(n)$ en réduisant considérablement l'espace de

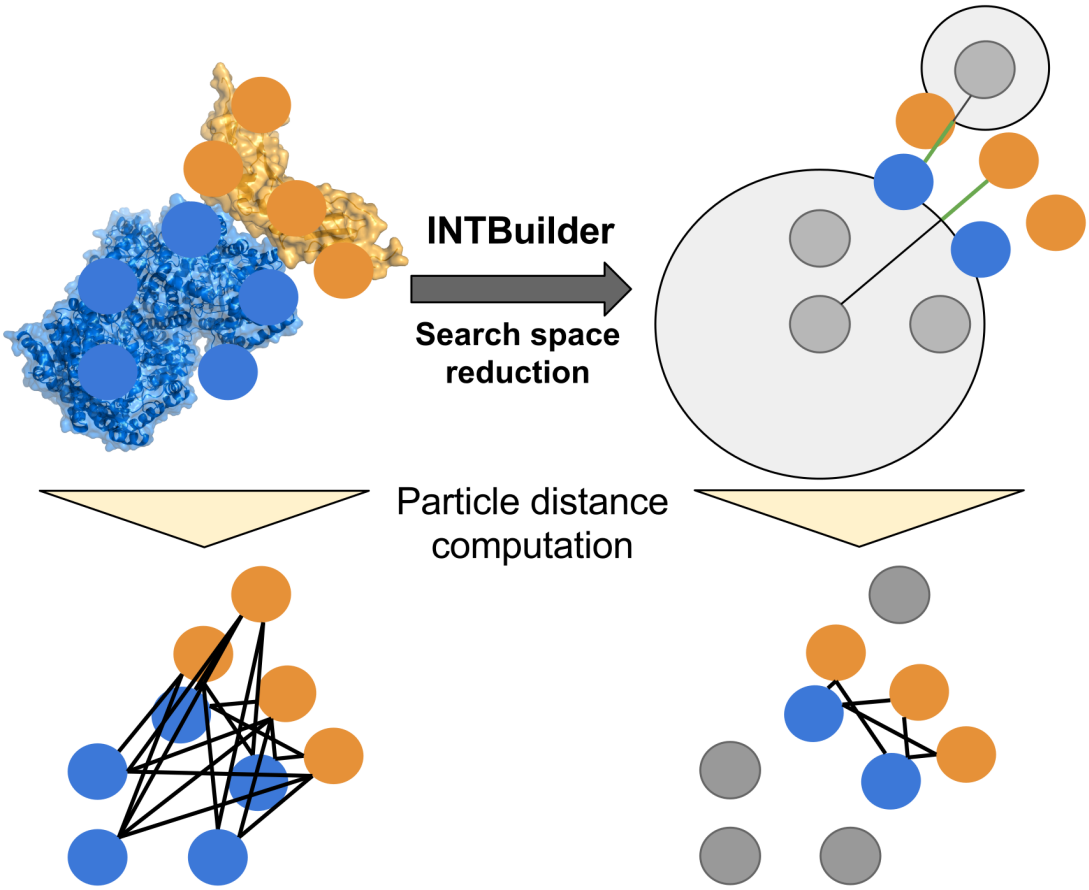


Figure 1.3: Ce schéma permet de mieux comprendre le fonctionnement de la méthode de réduction de l'espace de recherche de INTBuilder. On passe d'un calcul entre tous les atomes des deux protéines vers un calcul d'un ensemble réduit de points.
This schema helps understand the working of the INTBuilder software. We go from a full atom-atom computation involving many distance calculations to a reduce ensemble which is faster to compute.

recherche lors de l'analyse des distances inter-résidus des protéines considérées. De plus, INTBuilder considère explicitement la description des conformations docking par une transformation vectorielle et un ensemble d'angles d'Euler représentant la translation et les rotations à appliquer au ligand par rapport au récepteur.

Afin de faciliter l'utilisation de la fonction de rotation, la sortie de plusieurs algorithme de docking (iATTRACT [98], HEX [35], ZDOCK [19] et MAXDo [95]) est directement lue, contournant ainsi le problème d'écriture des résultats intermédiaires. Cela permet ainsi à INTBuilder de traiter des millions de conformations en quelques minutes. Autres logiciels (Rosetta [109], GRAMM-X [104]) génèrent directement les fichiers PDB résultants correspondant à chaque conformation, ce qui permet à INTBuilder de les analyser sans effectuer les rotations spécifiques aux conformations.

Bien qu'INTBuilder ait été conçu pour détecter les interfaces protéine-protéine, il peut facilement être utilisé pour identifier les sites d'interaction des petites molécules à partir de conformations obtenues par filtrage virtuel.

1.4 Détections et prédictions d'interfaces protéine-protéine

La conservation, les propriétés physico-chimiques et la géométrie locale autour des résidus ont été utilisées pour prédire les surfaces en interaction [105, 59, 11, 47, 36, 17, 81, 84, 29, 57, 30]. Au cours des 15 dernières années et sur la base (non exhaustive) de ces propriétés, un certain nombre d'outils de prédiction de sites d'interaction ont été développés [57, 113, 31, 106] (voir [30, 4] pour les reviews). Ces outils classent les résidus de surface comme interagissant ou non-interagissant, ou prédisent des patches d'interaction, généralement un ou deux par protéine. Un patch d'interaction est un groupe de résidus de surface géométriquement proches et susceptibles de participer ensemble à une ou plusieurs interaction(s). Une récente étude a souligné que bien que la plupart des études évaluent leur méthode contre des sites d'interactions couvrant généralement entre 25% et 30% de la surface de la protéine, jusqu'à 75% de cette surface pourrait en réalité être impliquée dans des interactions protéine-protéine [103]. Ce nombre a été estimé en copiant, pour une protéine donnée, toutes les interfaces protéiques à partir de structures de complexes dans la banque de données de protéines (PDB [9]) ayant un repli structural, indépendamment de leur identité de séquence. Bien que toutes les interfaces ainsi copiées ne soient pas susceptibles d'être fonctionnelles pour la protéine étudiée, cette estimation suggère que le pourcentage de la surface en interaction serait largement sous-estimé par la majorité des études.

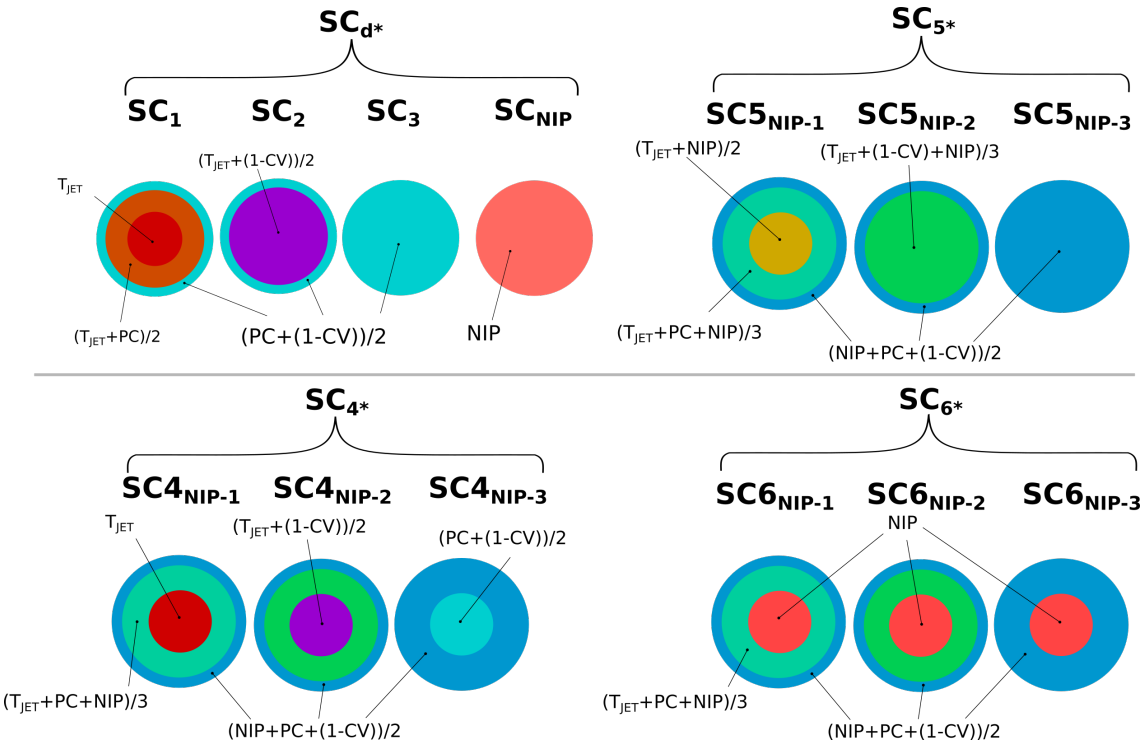


Figure 1.4: Méthodes de scoring décrites par dynJET^2 (SC_1, SC_2, SC_3, SC_{NIP}) et leurs dérivations SC_{4*}, SC_{5*} et SC_{6*}.
Scoring schemes described by dynJET^2 (SC_1, SC_2, SC_3, SC_{NIP}) and their derivations as SC_{4*}, SC_{5*} and SC_{6*}.

1.4.1 Le développement de dynJET^2

Au cours de ma thèse, j'ai apporté des améliorations aux logiciels de prédiction de sites d'interactions protéine-protéine pré-existants (JET et JET^2, [29, 57]). dynJET^2 permet d'intégrer un score arbitraire aux trois autres scores déjà pré-existants. De précédentes études [33, 55] ont montré qu'il était possible d'inférer un score issu des conformations de docking (NIP) pour aider la prédiction des sites d'interaction. J'ai ainsi pu intégrer ce score à différentes étapes de la prédiction des interfaces de dynJET^2. La Figure 1.4 illustre les stratégies de scoring que j'ai apporté avec dynJET^2.

L'étude de [95] a motivé à l'origine le développement de méthodes de prédiction d'interfaces protéine-protéine afin, finalement, d'être en mesure de remplacer les interfaces expérimentales par des prédictions. La première version de l'algorithme, JET [29], était uniquement basée sur la séquence de la protéine et utilisait l'annotation des propriétés physico-chimiques couplée au calcul d'une trace évolutive pour prédire les sites d'interaction. Cependant, avec la quantité croissante de structures 3D de protéines disponibles, une version plus récente a été développée afin d'incorporer ces données géométriques dans la méthode de prédiction. Plus précisément, cette nouvelle version JET^2 [57] fait usage de la variance circulaire des résidus pour repérer les régions protubérantes à la surface de la protéine. Trois propriétés sont donc ainsi considérées pour la prédiction des sites d'interaction par JET^2 : La conservation des résidus, leurs propriétés physico-chimiques et leur variance circulaire. Contrairement à de nombreux algorithmes de prédiction d'interfaces protéine-protéine, JET^2 se base sur un modèle pré-établi de la définition de l'interface [62] et oriente ainsi sa méthode de prédiction selon une approche seed-extension-outer layer, reproduisant le modèle décrit du Support-Core-Rim. Certains scores seront donc utilisés plutôt que d'autres pour prédire certaines régions de l'interface. On observera ainsi par exemple une conservation plus élevée en particulier dans la région centrale (la plus enfouie) de l'interface. Cette information va ensuite être utilisée par JET^2 pour prédire ces régions en particulier. De plus, JET^2 possède différentes méthodes de détection d'interfaces (Scoring Schemes, SC) qui sont adaptées à plusieurs types d'interactions protéine-protéine : SC_1, SC_2 et SC_3. Une grande base de données regroupant les predictions de JET^2 sur plus de 20 000 chaînes protéiques a également été publiée par l'équipe [91].

1.5 Caractérisation des interactions multiples entre protéines

La majeure partie des interactions protéiques se déroulant dans la cellule, elles peuvent être représentées sous forme de graphe où chaque nœud représente une molécule

et chaque arête une interaction. Notre connaissance des réseaux d'interactions reste cependant en grande partie incomplète, et l'évaluation expérimentale de l'ensemble des interactions possibles d'une protéine reste très difficile [42, 93]. Une protéine peut interagir avec plusieurs partenaires en même temps, avec deux (ou plus) partenaires interagissant à des endroits différents de sa surface, voire avoir plusieurs partenaires partageant de façon compétitive un même site d'interaction [64]. Afin d'avoir une vue globale sur la multiplicité des interactions protéiques, nous devons être en mesure de déchiffrer la complexité de leur surface vers une définition de sites (interaction spécifique à un partenaire) et de régions (interactions non spécifiques) d'interactions et une caractérisation de leurs propriétés. Une telle description, décrite au niveau des résidus, permettrait également de prédire l'impact des mutations sur les interactions protéiques et par conséquent leurs fonctions.

Comme expliqué précédemment, une stratégie alternative pour prédire les résidus en interaction consiste à exploiter calculs de docking. Les méthodes de docking ont été conçues à l'origine pour prédire la conformation native d'un complexe à partir des structures connues de ses unités. Les conformations candidates sont générées et évaluées sur la base de propriétés reflétant la force de l'association, par exemple leur complémentarité de surface, le champ électrostatique, la désolvatation or l'entropie conformationnelle. De ces ensembles de conformations, on peut dériver des statistiques pour estimer la propension de chaque résidu de la protéine à appartenir à une interface [33, 55]. Cela a été réalisé dans des études de docking binaire [37, 61, 44, 24, 45] où l'on sait déjà *a priori* que les deux candidats interagissent, dans des études de docking arbitraire [72] où les protéines d'un ensemble de référence sont ancrées à des protéines choisies arbitrairement. Enfin, on peut dériver ces statistiques d'un CC-D [95, 67, 107, 56, 55] qui impliquent des calculs de docking sur toutes les paires de protéines possibles au sein d'un jeu de données.

Il a été montré dans [64] que les protéines présentent des sites d'interaction pouvant être ciblés par une multitude de différents partenaires. Au cours de l'analyse du jeu de données `P-262`, nous combinons des propriétés calculées par l'analyse de séquence et de structure des protéines, à savoir la conservation des résidus, leurs propriétés physico-chimiques, la géométrie locale autour de ceux-ci et du score de propension inféré des simulations de docking. Ces propriétés sont ainsi utilisées pour nous aider à comprendre comment les protéines réussissent à interagir les unes avec les autres dans un environnement aussi encombré que la cellule. À l'aide de dynJET^2, nous prédisons des sites d'interaction en combinant les quatre caractéristiques précédemment citées. dynJET^2 identifie d'abord un petit groupe de résidus localisés à la surface de la protéine, appelé la graine, puis la prolonge avec deux couches successives de résidus. Les patchs prédits par les différents scores peuvent être soit complètement distincts ou alors peuvent se chevaucher partiellement, ainsi reflétant la multiplicité des interactions qu'une protéine peut établir pendant

sa durée de vie. Ces prédictions sont comparées à un ensemble de 329 interfaces connues expérimentalement parmi les 262 protéines.

1.5.1 Analyses de prédictions des interfaces d'interaction basées sur dynJET^2

Les surfaces protéiques sont utilisées de multiples façons par une protéine. Nous avons analysé un ensemble de protéines avec différentes fonctions et nous avons montré qu'un site d'interaction pour un partenaire peut en réalité être partagé avec plusieurs autres partenaires, de manière complète ou partielle. La prédiction des sites d'interaction a été réalisée avec dynJET^2, en tenant compte de quatre propriétés basées sur la conservation, les propriétés physico-chimiques des résidus à l'interface, la géométrie locale de surface de la protéine et un score inféré des conformations de docking.

Nous avons montré que dans certains cas, cette quatrième propriété est complémentaire aux trois premières. En outre, bien que certains IR ne pouvaient pas être prédits par une seule propriété, la combinaison de l'ensemble des quatre propriétés ont dans la plupart des cas permis de correctement définir les régions expérimentales. En prenant en compte l'ensemble des protéines homologues connues et leur complexes, nous pouvons fournir une description très précise de la surface en interaction pour notre ensemble de données de chaînes protéiques. Selon nos analyses, le pourcentage de la surface couverte par des surfaces expérimentales (biologiquement fonctionnelles) connues est de 48% sur $\texttt{PPI-262}_{\texttt{ext}}$ contre seulement 29% sur `PPI-262`, ce qui indique que la surface d'interaction des protéines est très largement sous-estimée, et qu'il est important de les prendre en compte lors de l'analyse des prédictions. Ces résultats sont aussi en accord avec une étude publiée récemment sur le sujet [103]. Les interfaces récupérées des protéines homologues ont été fusionnées et ont ainsi permit de mieux définir les IR. Je porte notamment l'attention sur l'importante quantité de IS expérimentaux initiaux (23642) distribués sur l'ensemble des protéines du jeu de données qui a été réduit à un faible nombre (1.4 IR par chaîne protéique). Par conséquent, nous avons pu constater dans l'évaluation des prédictions de dynJET^2 un important nombre d'entre elles qui se sont révélées décrire de réels sites biologiquement fonctionnels. On obtient ainsi une valeur moyenne de F1-score de 0.41 en comparant l'union des prédictions de dynJET^2 avec les IS de `PPI-262`, mais cette valeur augmente à 0.57 sur l'union des IR présentes dans $\texttt{PPI-262}_{\texttt{ext}}$. En particulier, le pourcentage de protéines pour lesquelles nous obtenons des prédictions avec F1-score > 0.6 augmente de 18% à 46% en considérant l'union de `PPI-262` et $\texttt{PPI-262}_{\texttt{ext}}$ respectivement. De plus, le nombre de mauvaises prédictions (F1-score < 0.2) diminue de $\sim 25\%$ à 4% entre `PPI-262` et $\texttt{PPI-262}_{\texttt{ext}}$.

Nous avons également essayé de comprendre les raisons à l'origine des moins

bonnes prédictions en regardant la Root Mean Square Deviation (RMSD) entre l'IR expérimentale et les interfaces expérimentales dont elle provient chez les homologues. Cette différence peut être très importante et se trouve être corrélée à la difficulté de prédire certains sites ou régions d'interactions. Bien que dynJET^2 reste résistant aux petits réarrangements, sa performance diminue progressivement au fur et à mesure que nous observons une augmentation de la valeur de RMSD. La moyenne du RMSD pour les bonnes prédictions (F1-score > 0.6) est de 4.7Å, 8.3Å pour les IR avec un F1-score intermédiaires ($0.3 \leq$ F1-score ≤ 0.6), 12.7Å pour les IR moins bien prédites (F1-score < 0.3) et 13.3Å pour les IR qui ont été complètement manquées.

Nous avons montré que les capacités de dynJET^2 pouvait nous aider à prédire si une prédiction pouvait cibler un ou plusieurs partenaires. En effet, alors qu'un faible nombre de partenaires est observé pour les IR possédant 0 ou 1 seed (prédite par dynJET^2), nous montrons ce signal disparait à mesure que le nombre de seeds dans l'IR augmente. Il pourrait être possible d'affiner cette approche vers un compte plus précis dans le futur.

L'un des principaux défis restants serait de diviser les interfaces prédites en IR ou, éventuellement, en IS. Cela nous permettrait de déduire le nombre de partenaire(s) avec lesquels la protéine considérée pourrait interagir avec, ainsi que décrire combien de régions fonctionnelles elle possède.

1.6 Détections de partenaires interagissant à grande échelle

La prédiction du site d'interaction de la protéine a longtemps été un sujet très étudié, ainsi que l'identification de la conformation native pour un complexe protéique. Cependant, les études à grande échelle essayant de décrire les partenaires interagissant en combinant ces deux approches restent très rares. De nombreux obstacles et difficultés en sont la cause : bien que de nombreux progrès aient été réalisés à l'égard du docking au cours des dernières décennies, cela reste une expérience coûteuse et son application à large échelle nécessite la mobilisation d'importantes ressources. De plus, identifier les partenaires corrects à l'échelle de plusieurs centaines de protéines demande une incroyable précision étant donné l'écrasante majorité de solutions négatives. En raison de ces entraves, la prédiction des interactions protéine-protéine à grande échelle via l'utilisation de méthodes de CC-D en est encore à ses débuts. À notre connaissance, il n'y a eu que trois études ([67], [110] et plus récemment [70]) employant la méthode d'un CC-D pour l'identification des partenaires à large échelle. Bien que [70] utilise des méthodes d'apprentissages (en particulier un classificateur Random Forest), il est intéressant de noter que le pipeline global suit le nôtre en combinant les méthodes de scoring des conformations

par le logiciel de docking avec des prédictions de sites d'interaction.

Les interactions entre protéines étant au centre de la plupart des processus biologiques, il est donc crucial comprendre comment et avec quels autres partenaires ces dernières interagissent [39]. Il y a une demande croissante, en pharmacologie par exemple, de pouvoir cibler certaines protéines en particulier [40]. Les premières analyses telles que [95] ont effectué un CC-D à petite échelle afin de répondre à la question de prédiction du partenaire. Cette tentative, tout comme d'autres études [51, 52], montre que l'énergie seule ne suffit pas à prédire les partenaires interagissant par rapport aux partenaires non interagissant. Cependant, [95] a montré qu'il était possible de prédire les partenaires avec une très bonne précision en combinant l'énergie de docking à une interface bien définie (dans ce cas, expérimentale). Dans notre étude, nous suivons les premiers pas de [67] dans cette direction qui a analysé les même résultats d'un CC-D sur la PPDBv2 de 84 complexes protéiques avec le logiciel de prédiction d'interfaces JET [29]. Depuis, plusieurs améliorations ont été apportées au pipeline, du point de vue de prédiction d'interface par le développement de JET^2 [57] puis de dynJET^2, mais également du point de vue méthodologique sur les méthodes d'évaluation des conformations de docking.

1.6.1 Résultats et nouveaux horizons

Dans cette partie de la thèse, j'effectue une analyse du PPDBv2 de 168 protéines pour comprendre les méthodes disponibles pour analyser les conformations de docking, et observer leur impact sur notre capacité de discrimination.

Je montre notamment comment nos prédictions sont suffisamment précises pour détecter les partenaires interagissant à grande échelle, atteignant parfois les limites fixées par les interfaces expérimentales. Ainsi, nous obtenons des AUC décrivant bien mieux les sous-ensembles que lors de la dernière étude ([67]). On présente en Figure 1.6b les AUC obtenues en utilisant le pipeline qui a été développé dans cette étude. On observe ainsi une très large augmentation dans les deux groupes liés aux anticorps, vérifiant les nouvelles capacités de dynJET^2 à utiliser le score NIP avec les méthodes de SC_{d*} pour mieux prédire leurs interfaces.

Avec ces résultats, cette étude ouvre la possibilité de porter sa méthode pour l'analyse d'un protéome complet, créant ainsi un réseau d'interactions de nombreuses protéines dans une cellule, potentiellement impliquées dans les même voies fonctionnelles.

Nous avons également apporté des éclaircissements sur l'importance de la séparation des protéines dans différentes classes fonctionnelles, nécessitant ainsi le développement de méthodes pour automatiquement les analyser et les trier selon leur fonction.

L'étude appelle également à de nouvelles recherches pour affiner les différentes

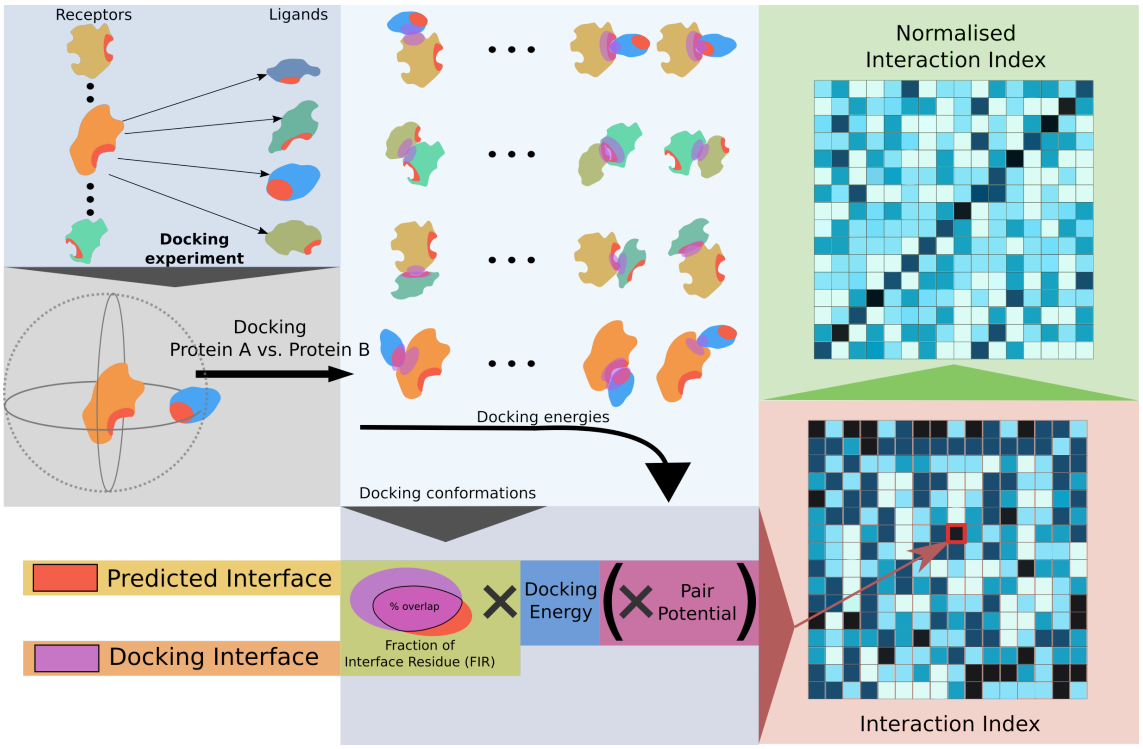
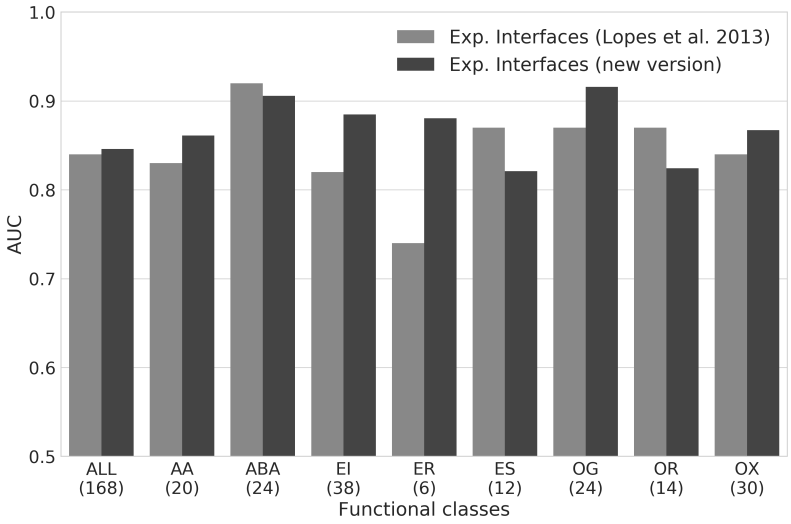
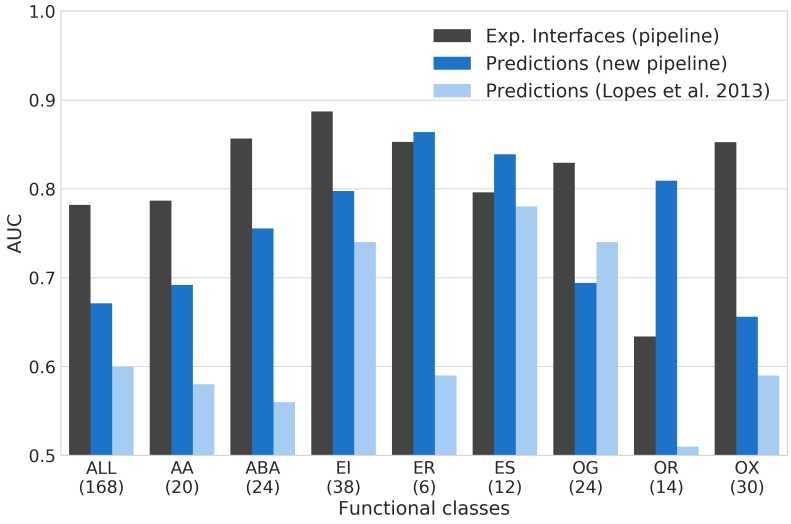


Figure 1.5: Nous présentons ici le schéma représentant le pipeline menant à calculer le NII score. Ce schéma ne prend pas en compte les différentes façons qui sont appliquées pour certaines classes fonctionelles. Tous les paramètres que nous avons étudiés et analysés ont été mis en couleur. Les matrices représentées ne correspondent pas à des valeurs réelles et sont là uniquement pour donner une example du pipeline.
We represent here the global scheme used to represent the pipeline leading us to compute the NII score. This scheme does not take into account the discrepancies of the different ways to compute the functional classes. All the different parameters which we were able to compare are highlighted in different colours. The matrices represented do not represent real values and are here for providing a clear example of the pipeline.



(a) Barplot comparant les AUC obtenues en utilisant la méthode dans [67] et avec la méthode pour laquelle nous obtenons les meilleures valeurs d'AUC en utilisant les interfaces expérimentales. Cela correspond à un threshold de distance de 4.5Å pour calculer les interfaces de docking, en ne considérant pas le CIPS [79] et en utilisant les fonctions d'énergie PISA et iATTRACT pour les sous-ensembles EI et ER respectivement.
Barplot comparing the AUC obtained using the method in [67] and the method for which we obtain the best AUC values, using experimental interface. This corresponds to choosing a 4.5Å distance threshold for computing the docking interfaces, not considering the pair potential and using PISA and iATTRACT energy functions for the subsets EI and ER respectively.



(b) Barplot comparant les résultats obtenus auparavant en utilisant les prédictions [67], les résultats que l'on obtient en utilisant le pipeline défini dans 5.4.1 (voir Figure 1.5) et les interfaces expérimentales avec le même pipeline que les prédictions.
Barplot comparing the results previously obtained using the predictions ([67]), newly obtained using the pipeline with the parameters defined in Section 5.4.1 (see Fig. 1.5) and the experimental interfaces with the same pipeline as the predictions.

Figure 1.6: Les deux figures comparent les AUCs obtenues en utilisant de différentes méthodes. Les prédictions de dynJET^2 sont utilisées avec la meilleure combinaison de patches selon le site expérimental étudié; The processus est décrit plus en détail en Section 5.2.2.
The two figures compare the AUCs obtained using different methods. The dynJET^2 predictions used were the best combination according to the target experimental site; this process is further explained in Section 5.2.2.

prédictions de dynJET^2 dans les plusieurs régions d'interaction des protéines. Ici, nous nous sommes appuyés sur la connaissance des sites expérimentaux afin de pouvoir localiser la prédiction d'intérêt de dynJET^2. Comme le montre [103] et comme le confirment l'étude que j'ai précédemment réalisée sur la multiplicité des sites d'interaction, une très large portion de la surface de la protéine pourrait être impliquée dans des interactions fonctionnelles alors que les sites spécifiques étudiés ici ne représentent qu'environ 25% de la surface. Cela signifie qu'afin de pouvoir nous libérer complètement des connaissances expérimentales pour prédire les partenaires interagissant et définir des sites d'interaction, nous devons être capables de séparer les prédictions des différentes méthodes de scoring dans des régions distinctes. Une nouvelle matrice représentant la force d'interaction entre les protéines du jeu de données pourrait alors être calculée non pas sur la base de protéines, mais sur chaque région de chaque protéine.

1.7 Conclusion

Le titre de ma thèse est "Géométrie des interactions protéiques" et son but était d'analyser différents ensembles de données de protéines et d'élargir l'échelle analyses existantes. Plus précisément, j'ai travaillé sur deux domaines: La détection et l'interprétation des sites de liaison aux protéines et l'identification des partenaires en interaction dans le cadre d'un Cross-Docking Complet à large échelle. L'analyse des sites d'interaction protéique a apporté de nombreuses informations, notamment le concept émergeant de sites d'interactions multiples et comment les protéines interagissent dans un environnement peuplé. Ce sujet (décrit en détail au chapitre 4) montre que la surface interagissante des protéines serait beaucoup plus grande que ce qui est actuellement pris en compte dans la plupart des cas. L'analyse apporte avec elle un nouvel outil qui pourrait être facilement utilisé lors d'une analyse plus approfondie des interfaces biologiques entre homologues d'une protéine, ainsi que dynJET^2 (développé à partir du logiciel JET^2 [57]), un logiciel de prédiction de sites d'interaction capable de prendre en compte toute notation à l'échelle des résidus dans sa méthode de prédiction. L'analyse apporte les concepts de sites d'interaction (IS) et Régions d'interaction (IR). Ces deux définitions sont essentielles pour comprendre comment nous pourrions interpréter les interfaces à la surface des protéines. De plus, l'étude montre comment il serait possible d'inférer le nombre de partenaires ciblant un IR spécifique, et combien de régions fonctionnelles une protéine possède.

La deuxième analyse, centrée sur l'identification des protéines en interaction dans un CC-D à grande échelle apporte de nombreux résultats prometteurs. Nous montrons ici comment le développement d'une méthode de prédiction d'interfaces plus avancée combinée à l'utilisation adaptée de méthodes d'évaluation nous a permis de faire de grands progrès en termes d'identification de partenaires. Le logiciel

INTBuilder (Chapitre 6, [25]) a été développé dans le cadre de cette étude pour répondre aux besoins spécifiques d'un logiciel performant pour le calcul des interfaces de docking. INTBuilder apporte avec lui un algorithme nouveau permettant de réduire l'espace de recherche d'un ensemble de particules.

Cette analyse indique à quel point il est primordial de prendre en compte la classe fonctionnelle à laquelle une protéine appartient afin de pouvoir correctement identifier son partenaire. De plus, nous montrons aussi que dans de nombreux cas, nos capacités d'identification des partenaires ont atteint une limite qui semble fixée par la qualité et la précision de nos prédictions. La recherche de meilleures et plus précises prédictions devraient être la prochaine étape, mais il faut également souligner que de telles prédictions ne peuvent être spécifiques à un seul partenaire. Cela implique qu'il ne serait théoriquement pas possible d'atteindre les capacités de discrimination aujourd'hui obtenues avec l'utilisation d'interfaces expérimentales. Plusieurs voies sont possibles : l'une serait de essayer de développer des méthodes automatiques pour la prédiction d'interfaces spécifiques à un partenaire, l'autre pourrait être de changer la façon dont nous regardons la question avec la méthode actuelle. Au lieu de caractériser comment une protéine interagit avec les autres à travers une seule interface prédite, nous pourrions regarder simultanément l'ensemble des interfaces prédites d'une protéine et voir comment chacune d'entre elles interagissent avec les différentes interfaces prédites d'autres protéines.

Chapter 2

Introduction to proteins and their interactions

Contents

2.1 From DNA to proteins

I describe here the different biological objects that I manipulated throughout my work, and which are necessary for the comprehension of the approaches developed along the thesis.

DNA or Deoxyribonucleic Acid, is constituted of four different nucleobases: Adenine (A), Cytosine (C), Guanine (G), Thymine (T). These nucleobases bind themselves together (A binds with T and C binds with G) to form DNA strands, and we have in each chromosome two DNA strands bound together in a helix shape. The chromosomes are located in the cell nucleus, as can be seen in Fig. 2.1. The sequence of bases in the DNA forms the genome, which carries the genetic information from one generation to the next one. A gene is a sequence of the DNA that can be transcripted to the mRNA and then translated to a protein to execute a function. In *Homo Sapiens*, there are approximately 3 billion base pairs.

mRNA are known as the messenger Ribonucleic Acid. Like the DNA, the mRNA is constituted of four different nucleic bases. These bases are the same save for the Thymine which is replaced by Uracile (U). As shown in the Fig. 2.2, the mRNA is a single strand base pair sequence created from the DNA and as its name indicates, carries the message copied from the DNA outside the nucleus to the *ribosomes*. It is less stable than the DNA and is only meant to transfer the necessary information for the ribosomes to construct a protein before being degraded. Ribosomes read the mRNA sequence and aggregate amino-acids residues in a chain. Ribosomes read the mRNA by steps of three base called *codons*, each corresponding to a residue; the lecture begins with the *start codon* AUG corresponding to the methionine. As the ribosomes read the mRNA sequence, they will aggregate amino-acids residues into a chain until they encounter either one of UAA, UAG or UGA, also known as the *stop codons*, which will cause them to release the protein chain.

Amino-acids residues are the monomers chained by the ribosomes reading the mRNA. Each Amino-acid possesses an amine (-NH_2) and carboxyl (-COOH) functional groups, linked together by a C_α atom. Twenty different amino-acids can be translated from the genetic code (DNA); each of them has a common part (called backbone) by which they bind to each other and a specific side-chain attached to the C_α atom which will determine their properties. Amino-acids residues are bound to each other through a covalent chemical bounds, and we refer to them as "residues" since their binding causes the loss of their acid groups. Polypeptides (long chain of residues) thus have a single N-terminal and C-terminal located at the two extremities of the chain.

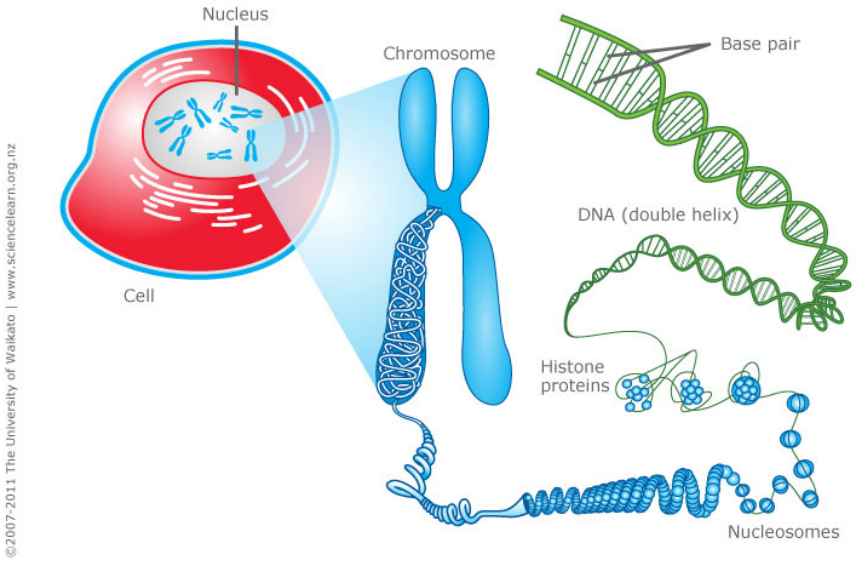


Figure 2.1: Schematic representation of the DNA as a chromosome in a cell.

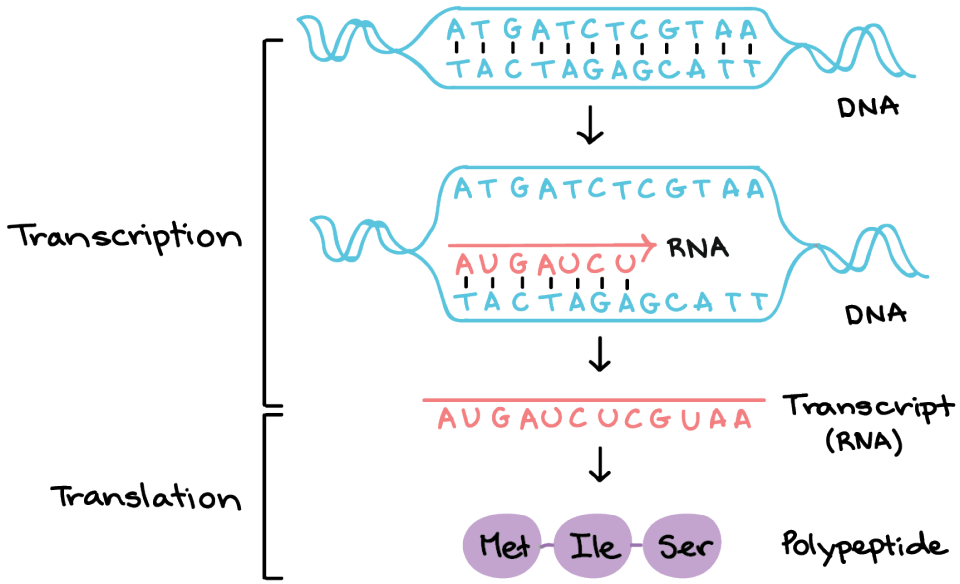


Figure 2.2: Scheme representing the different stages from DNA to protein. Picture made on https://www.khanacademy.org.

Proteins are the main objects I consider through my work. As previously described, they are formed by a chain of residues aggregated by the ribosomes. During the aggregation, the chain will fold itself and thus achieve a specific form. The protein may then undergo several changes known as post-translational modifications, which may cleave the chain or alter some of its residues. We consider four different structures for the proteins, which are essential to understand. The *primary structure* is the residue sequence, this does not take into account the 3D shape of the protein and thus does not need for it to be solved. The *secondary structure* represents the α-helix and β-strands determined from the torsion angles of the residues' backbone and stabilised through the hydrogen bonds between the backbone atoms. The *tertiary structure* represents the folding itself of the protein, usually burying the hydrophobic residues at its centre and constituting a hydrophilic surface accessible by the solvant. It is stabilised by non-bonding interactions between the backbone and side-chains atom (*e.g.* hydrogen bonds, van der Waals). Protein-Protein Interactions might also play a role in the tertiary structure stabilisation, introducing the *quaternary structure* which is the aggregation of multiple proteins chains, then called subunits. This aggregation forms a single functional unit called multimer.

2.2 Evolution and conservation

During the replication of the DNA, some "errors" might occur in the new sequence. They are known as mutations and may present varying degrees of impact (beneficial, neutral, deleterious or highly deleterious). The DNA code is degenerate, meaning that multiple codons code for the same residue; on the one hand, a mutation changing a codon but not its corresponding residue will therefore be neutral, having no ultimate effect on the protein sequence. On the other hand, mutations inducing change in the protein sequence can affect its function, depending on the change. Such a change could result in the complete loss of function for the protein (in the case of a highly deleterious mutation). Mutations are part of what makes each individual unique.

The principle of evolution has been simultaneously and independently introduced by Charles Darwin and Alfred Russel [23, 94]. The notion of evolution describes that all organisms originate from a single common ancestor. From there, the theory explains that the fittest individual will prosper more than others, if it possesses the genetic material the most adapted to its environment. At some points during history, mutations in the population will occur and cause it to split into two groups. From this theory, we can recreate a tree representing the dividing of populations where each leaf of the tree would represent a different *species*. Two species therefore possess

a last common ancestor, which is represented by a node in the tree corresponding to the last individual before this ancestral species was split into two branches.

Since different species originate from a single ancestor, it is therefore possible to find similarities in their genome. For that, a high number of sequences are retrieved (among homologous species, or even inside the same species) to obtain a set of homologous sequences, as in Fig. 2.4. In this figure, we observe that for each column there may be different amount of variation between the residues. The more a residue is consistent among homologous sequences, the more it is conserved; this notion helps determine key residues for the protein. For instance, a residue with a high conservation among the homologous sequences might have a crucial role for the protein function; hydrophobic residues for instance are essential to the protein folding process and therefore are highly conserved. It was also shown that conservation plays an important role in protein interactions [28, 5].

2.3 Proteins Interactions

Proteins achieve different function with a wide variety of other biological objects such as DNA, RNA, small ligands molecules or other proteins. This multitude of interactors make for very different types of interactions for the protein: essential characteristics for the binding of a small ligand molecule may not be relevant for binding to other proteins. Even inside a single interaction set (Protein-Protein for instance) and as we will further demonstrate, some characteristics show various degrees of importance depending on the specific type of proteins. Antibodies for instance show a very specific interface location and is often considered separately from the rest [30].

Such complexity accounts for the difficulty current methods encounter to understand interactions. In this work we focus on the protein-protein study case. To better understand the different characteristics specific to this case, I present the different context that might lead proteins to bind one another and describe the properties of the resulting interfaces. With these two notions, I will next advance to present the protein docking concept which endeavours to computationally reproduce the binding of two proteins.

2.3.1 Energies at the interface

Proteins fold and interact on the basis of free energy: the lowest the free energy is, the more stable the structure. The native structure of a protein for instance will correspond to the lowest free energy possible for it. The free energy difference between the native state and the ensemble of denatured conformations is 5-15 kcal/mol. In the same fashion, proteins will bind to one another because doing so will lower the

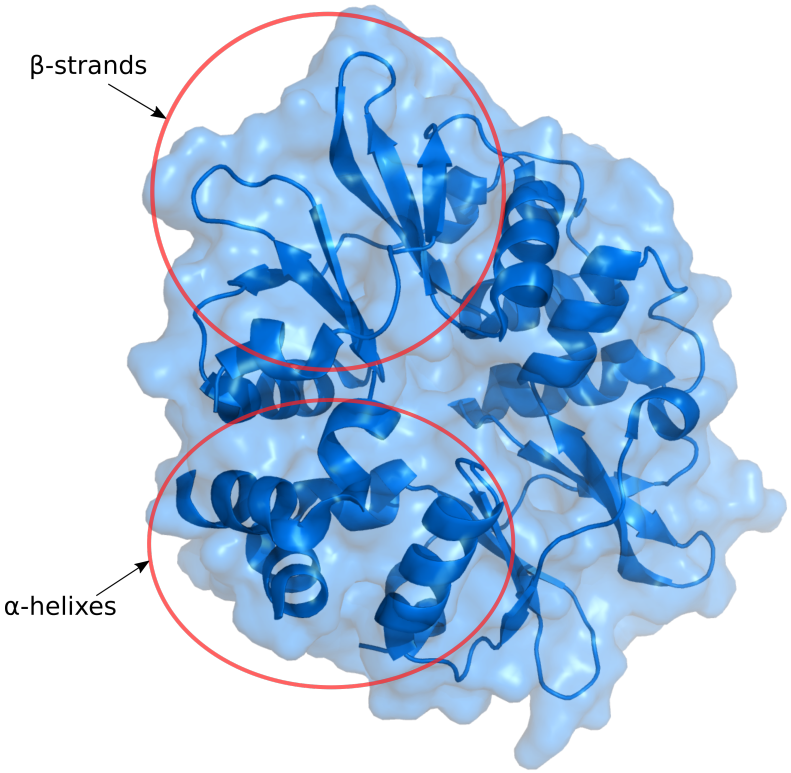


Figure 2.3: Representation of the 2rih protein complex in bound form.

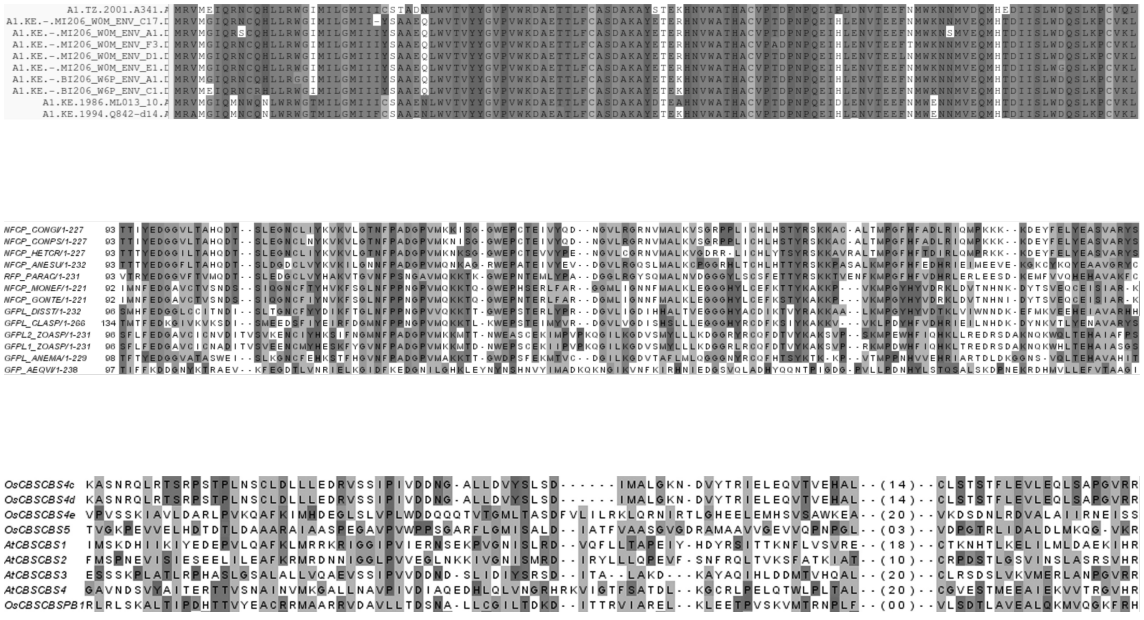


Figure 2.4: Example of multiple sequence alignments. We can observe how some columns (*i.e.*, positions on the sequence) consistently have the same residue (usually marked with a single colour) while other positions have largely varying residues. We also note that there may be changes among the individuals of the same species.

overall free energy of the complex. We define ΔG the Gibbs or Helmholtz free energy as:

$$\Delta G = \Delta H - T\Delta S$$

with ΔH the internal energy (from internal interactions) and $T\Delta S$ the entropy bu temperature. The tertiary structure, or the protein folding process is stabilised through non-bonding interactions:

- Electrostatic Interactions (5 kcal/mol)
- Hydrogen-bond Interactions (3-7 kcal/mol)
- Van Der Waals Interactions (1 kcal/mol)
- Hydrophobic Interactions ($<$ 10 kcal/mol)

2.3.2 Strength of the interactions

Proteins accomplish their function by binding to other proteins, in a more or less permanent fashion [77, 53, 86]. On the one hand, the fleeting *transient* interactions are not meant to last and usually have a low binding affinity. A protein interacting with another to regulate its function for a given time could make an example of a transient enzyme-inhibitor complex. On the other hand, *obligate* interactions present very strong binding affinity and are meant to be permanent interactions, or not easily breakable (see Fig. 2.5). Both of these types of interactions present different kinds of properties. Typically, larger complexes grouping themselves together to form a quaternary structure bind according to an obligate interaction (the interaction is stable and intended to last).

2.3.3 Proteins interfaces

Several methods are used to define the interface of two interacting partners. One of such methods is to look at the relative Accessible Surface Area (rASA) of the residues of each protein. Upon binding to one another, the residues at the interface become buried and their rASA changes (ΔASA$>$ 0). This variation of the ASA can therefore be used to characterise interface residues. Another method is to look at the distance separating the residues from the two proteins. Finally, a definition of the interface using a Voronoi model can also be used, as in [15].

Furthermore, it has been also shown that using features such as rASA it was possible to define multiple regions of the interface [22, 6, 62]. The definition of [62] brings a new area to the interface on top of the Rim and Core previously defined. We describe here his definition of the interface, onto which we will rely on for our analysis:

- The *Support* represents the central region of the interface, which is the most buried of the three. It is defined from the residues having less than 25% rASA in the unbound form of the protein.
- The *Rim* is the bordering region of the interface and is defined from residues with more than 25% rASA in the complexed state (bound to this other protein).
- The *Core* contains the residues that could not be classified into either the support or the rim category. The rASA of these residues shrinks from more than 25% rASA in the unbound form to less than 25% rASA in the bound form.

A typical 1000Å^2 interface involves an average of $\sim$ 28 residues, with $\sim$ 10 residues forming the core, $\sim$ 8 residues forming the support and $\sim$ 10 residues forming the rim [62]. The study suggests this definition of the interface general among globular proteins. It shows that the rim and support composition of residues are very similar, contrary to the core which present ones. It shows as well that this statement holds true for different interfaces size and different types of complexes.

2.3.4 Residues properties and protein-protein interface prediction

Proteins' residues present different characteristics used by protein-protein interface prediction software. At first, due to the lack of protein structural data, most early protein-protein interface prediction software (predictors) were sequence based (see Fig. 2.6A). Such predictors include residues scores such as the conservation level [3, 88, 68, 85], the propensity to belong to an interface (based on physico-chemical properties [114]) or hydrophobicity [34]. First methods using these features achieved a 64% accuracy score [30]. However, one of the main drawbacks from sequence based predictors is their inability to determine if a residue is at the surface or not. It is also why the introduction of 3D structures has drastically increased the global accuracy of all methods. As no new improvement of performance has been observed among new or existing tools, the combination of sequence based characteristics seems to have reached its limits. The research of protein-protein interaction site prediction has now shifted toward a more geometrical view of the issue [30].

The introduction of geometrical data has opened to door to a wide range of new features, such as secondary protein structure, residue Accessible Surface Area (ASA) or the overall shape of the protein; this new dimension also brought with it different methods of prediction algorithms. Such predictors can be **3D mapping based predictors** (Fig. 2.6B) which use biological information about the protein structure or the sequence and try to map it onto the protein [46, 21]. Another approach

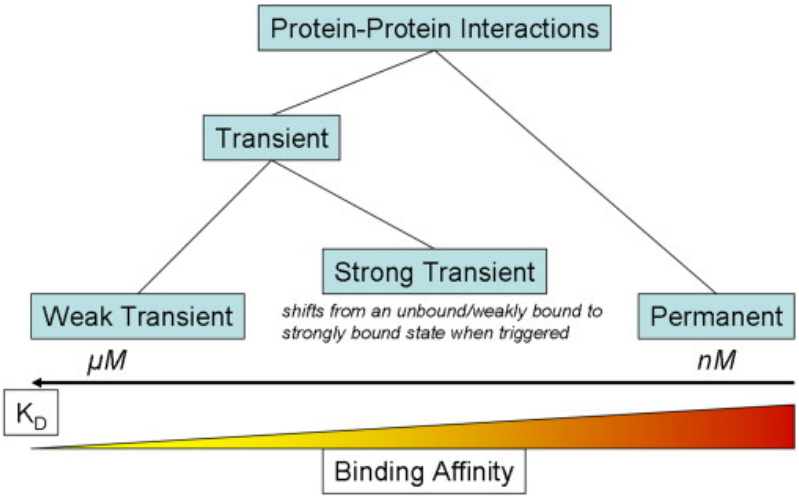


Figure 2.5: Representation of different ways to classify Protein-Protein Interactions in terms of binding affinity [86].

are **machine-learning based** classifiers (Fig. 2.6C). These classifiers regroup the most decisive features to build a classification model to assign to each residue its probability to belong to the interface. With the development of machine learning approaches, many such classifiers have been developed in the past years [14, 12, 27, 65, 119, 32, 18, 20, 99, 100, 89, 63, 8]. A number of probabilistic predictors were also used using Bayesian methods, Hidden Markov Model or Conditional Random Field [13, 82]; these probabilistic models can as well use such information as co-evolution analysis.

While a very broad set of different feature have been used, a previous study [117] showed that they could compare to state-of-the-art methods by using only four different features: solvent accessible surface area (SASA), hydrophobicity, conservation and propensity of the surface amino acids. On top of greatly reducing the algorithm complexity, the study also suggests that this will help reduce the risk of over fitting. It has been shown as well how docking a protein against non interactors could provide a meaningful score to predict interfaces [33, 67, 55]; however, this method suffer from the need to perform extensive docking calculations among many proteins to obtain sufficient data.

Overall, many features have come into play with the introduction of more protein structural data, and while many of these features demonstrated an important role in protein-protein interactions, it has also been clearly shown that not one single feature could predict all interactions. Combining the large variety of features to obtain relevant predicted interface has therefore been a major challenge in this field.

We summarise the following main properties that can be relied upon for protein-protein interface prediction (partially aggregated by [30]):

- Relative Surface Excluded solvent Area
- Solvation energy
- Electrostatic potential
- Conservation
- Physico-Chemical
- Circular Variance

2.3.5 Protein docking

To obtain the 3D structure, experimental methods based on Nuclear Magnetic Resonance (NMR) or X-ray crystallography have been used to determine many protein complexes. However, the past 20 years have seen an increasing number of new protein sequences being released every year, and it clearly appears that such methods

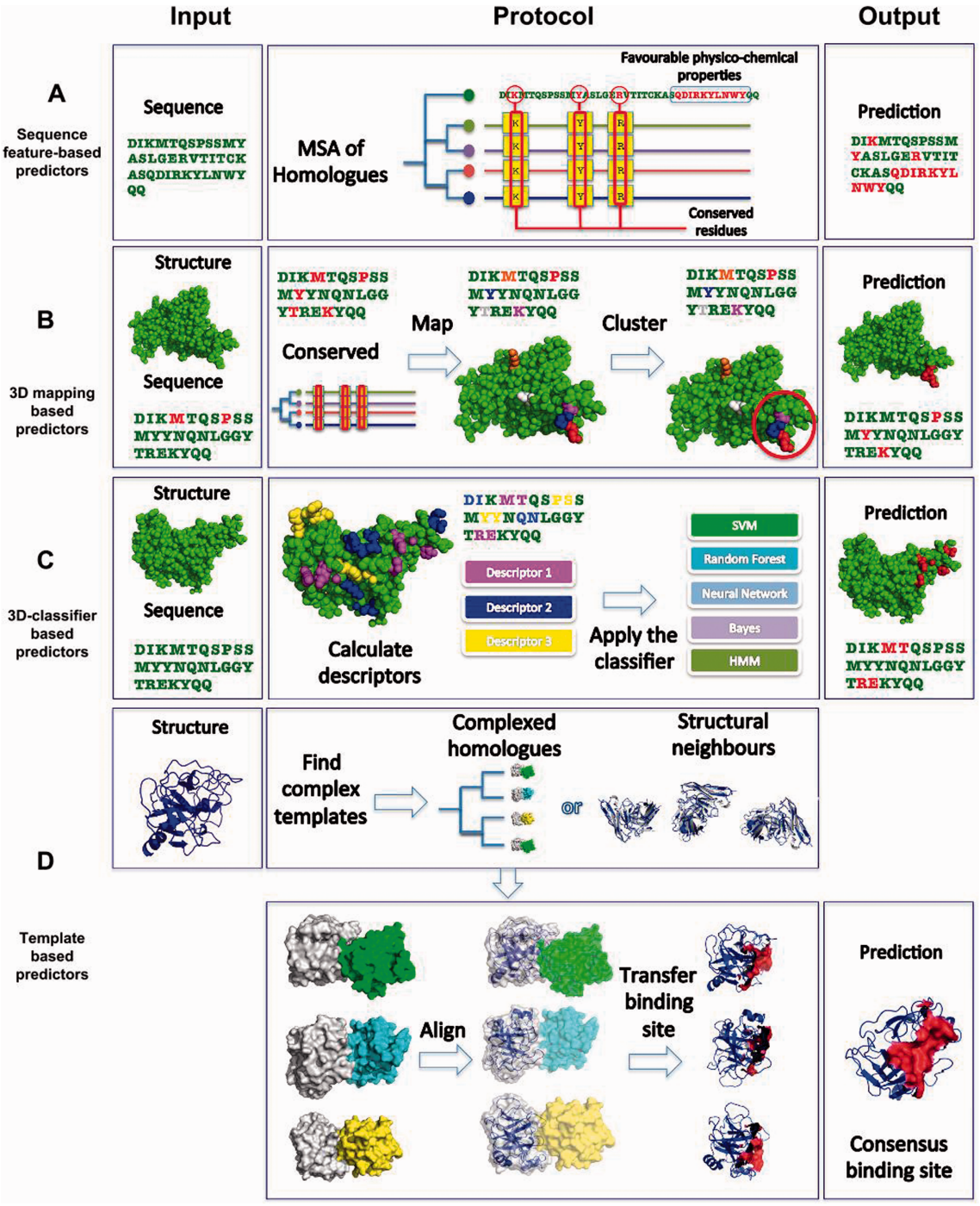
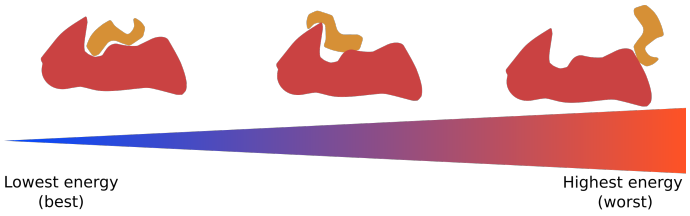
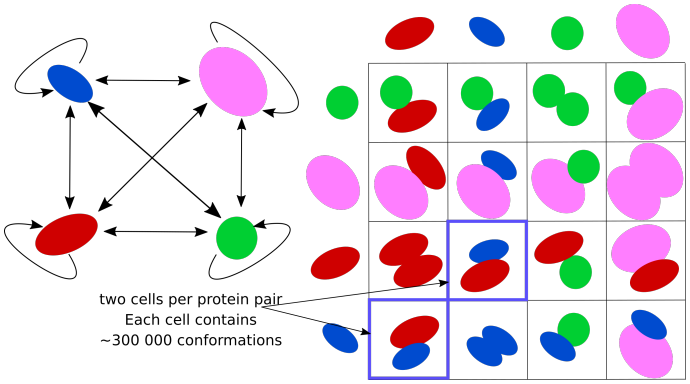


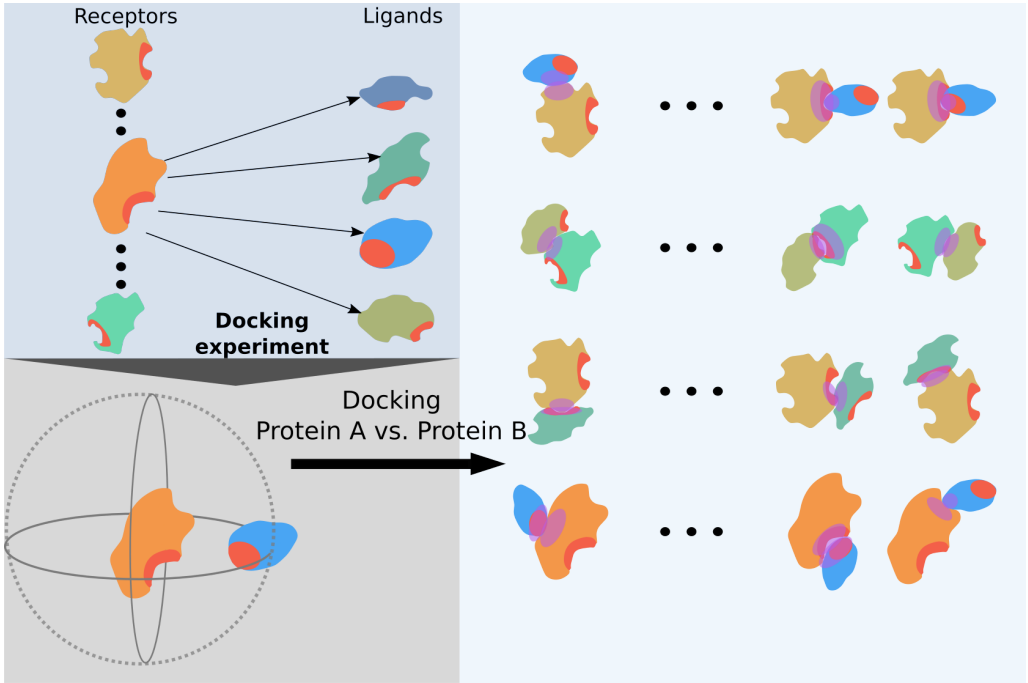
Figure 2.6: Schema representation of the different descriptors [30].



(a) Schema representing the docking energy for each conformation. As it is described, we have a better (thus lower) energy for favourable conformations than for the unfavourable ones.



(b) Representation of a Complete Cross-Docking, showing how we obtain a matrix representing all the different protein pair. For each cell, there are $\sim$ 300 000 possible conformations. We show in Fig. 2.7c how the docking process can output different conformations for a protein pair.



(c) Representation in more detail how the proteins are used during the docking experiment. We present here an asymmetric docking, where we can see a clear distinction between receptors and ligands.

Figure 2.7: Presentation of the docking concepts

(NMR, X-ray) are unable to keep up with the sequence release speed. To tackle this issue, more and more effort have been put into developing computational methods to simulate the binding process.

Protein docking consists in taking two protein structures (one as a receptor, the other as a ligand) and evaluating different structural conformations they can achieve. This is done by sampling the space around the receptor and simulating the binding process from the ligand to the receptor. For one pair of protein, about 300 000 different conformations are obtained.

For each conformation, the docking software then proceeds to evaluating it using an energy function. The resulting energy value indicates how stable the interaction is, and how likely the two proteins are to interact in this way (see Fig. 2.7a). Different protein docking algorithm exist [98, 95, 115, 116, 35, 19, 109, 104, 26] and are based on different properties such as atoms distance, biochemical or biophysical information. They can be classified in two groups: *rigid-body* and *flexible* docking. The rigid-body approach considers proteins as immutable objects and try out the different conformations without accounting for any conformational changes while the flexible docking tries to take into account those structural changes during the docking step. Although the rigid-body approach does not take into account structural changes. Overall, a rigid-body docking approach will under perform the greater the conformational change is upon binding. This trade of is made for the sake of performance, thus being much quicker than the flexible ones. Some hybrid docking algorithms also are capable combining a rigid-body approach with a flexible capacity.

Other means to reduce the computation time have been approached, such as using a coarse-grain reduced protein model, as developed in [115]. This coarse-grain representation (as fully described in [67]) places one pseudo-atom at the C_α position and either one or two pseudo-atom representing the side-chain (except for Gly). Ala, Ser Thr, Val Leu, Ile, Asn, Asp and Cys have a single pseudo-atom located at the geometrical centre of the side-chain heavy atoms. For the remaining amino acids, a first pseudo-atom is located midway between the C_β and C_γ atoms, while the second is placed at the geometrical centre of the remaining side-chain heavy atoms. This description, which allows different amino acids to be distinguished from one another, has already proved useful in protein-protein docking [115, 116, 7] and protein mechanics studies [97, 96].

Different docking software compute the interaction energy using by modelling the global energy of a complex of chosen inter-atomic interactions (usually found at the interface). Such energy functions take into account the complementarity of the interface between two proteins, adding a penalty in case of clash. Moreover, the many (see Fig. 2.8) physico-chemical properties of the amino-acids should be

compatible. The general form of these models can be written as:

$$E = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{non-bonded}} + E_{\text{others}}$$

A physical potential aims at maintaining a complementarity between the two surfaces, and can be found under the form:

$$\sum_{i<j} [\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}]$$

A statistical potential is also used to evaluate the likelihood of two residues interacting with one another. Such a likelihood can be described as:

$$\sum_{r_ij<r_c} \frac{P_{aa(i),aa(j)}}{n} \qquad P_{a,b} = -\log \frac{f(a,b)}{f(a)f(b)}$$

with $P_{aa(i),aa(j)}$ the probability of the Amino-Acides (aa) to interact with one another and n the total of possible interactions. Using such mesures allows docking software to evaluate the different conformations between two proteins.

2.3.6 Complete Cross-Docking

The Complete Cross-Docking (CC-D) of a dataset consists in docking every protein of this dataset against all the other proteins of the dataset, thus resulting in n^2 different possible pair with n being the number of proteins. As shown in Fig. 2.7b, 2.7c, we obtain a matrix where each cell represents the conformations calculated for a given protein pair. In a classical asymmetric docking computation, each protein assumes the role of either the receptor (fixed in space) or the ligand (which is orbiting around the receptor and testing the different conformations); in a symmetric docking the two proteins are orbited at the same time and there is therefore no such role as receptor and ligand, thus effectively cutting the amount of different pairs by half. In the HCMD project context (described below; Sec. 2.4.1), the laboratory team performed two CC-D on two different datasets.

2.4 Context of the thesis

2.4.1 The Help Cure Muscular Dystrophy project

The Help Cure Muscular Dystrophy (HCMD) project[1] aims at investigating 2246 human proteins which structures are known, with a particular focus on the proteins playing a role in neuromuscular disease. The goal is to be able to computationally

[1] http://www.ihes.fr/~carbone/HCMDproject.htm

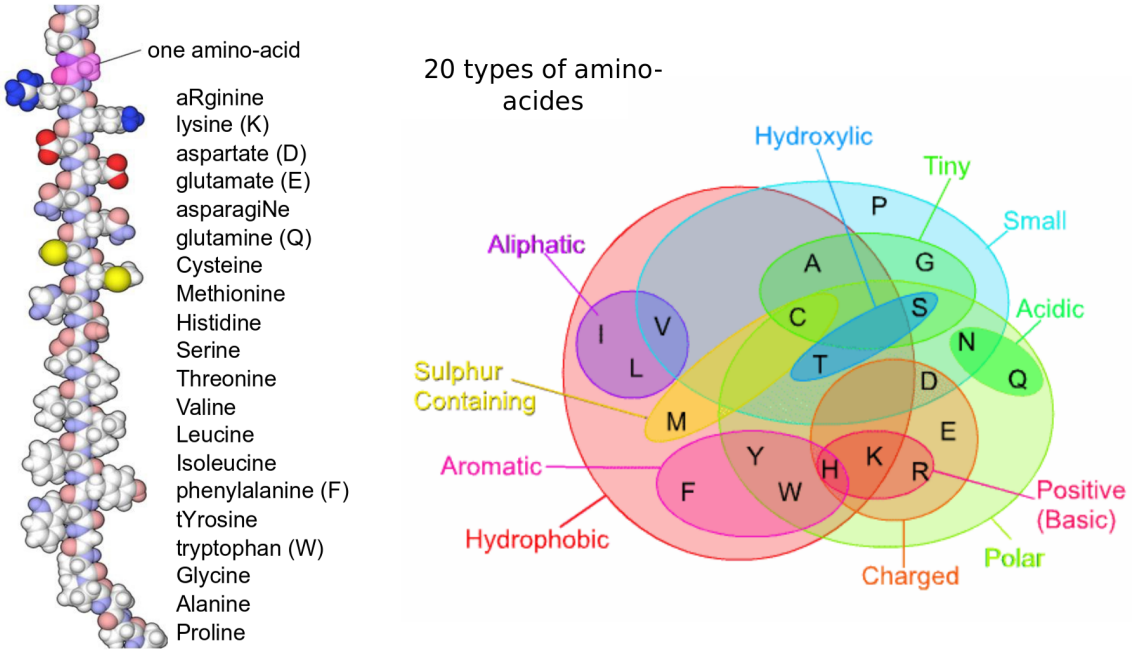


Figure 2.8: Amino-acids properties

describe the interactions between the 2246 proteins, thus understanding their interactions and being able to infer the different pathways involved. Two main phases separate this project:

- **Phase 1** consists in the analysis of a 84 protein complexes docking benchmark dataset (PPDBv2, see Section 3.4) assembled by [76]. Each complex represents a binary interaction, thus representing 168 proteins. The laboratory team performed an asymmetric CC-D using the MAXDo [95] software on this dataset using the World Community Grid[2] (WCG), a public computing grid letting people over the world participate in research computations. Using the WCG, the computation lasted 7 months and ended in June 2007. The role of this phase is to have a testing set, one onto which we will be able to develop algorithm and get a feedback on how well they are performing.

- **Phase 2** involves the actual 2246 proteins chains dataset, for most of which we don't know the interacting partner(s). This second phase, a CC-D experiment was also run on the WCG and lasted for more than four years from May 2009 to fall 2013.

The first publication on these datasets in the scope of this project [67] was an early study of the 168 proteins dataset, which I will henceforth refer to as the PPDBv2 dataset. It showed promising results, and presented how protein-protein interface predictions could be used to predict interacting partners. The continuing investigations on protein-protein interactions are now made within the framework of the MAPPING project which targets two crucial issues:

- What are the regions at a protein surface that interact with partner?

- Which proteins interact with which in the cell?

During this doctorate, I will essentially use the data produced during this project (CC-D calculations) along with the previous methodological developments done by the laboratory in this context.

2.4.2 Goals

Protein-protein interactions (PPIs) are essential to all biological processes and their misregulation is associated to many human diseases [10, 38]. Targeting PPIs with small molecule drugs has become increasingly popular in the treatment of diseases [2, 111, 39, 118]. Hence, it is important to determine which protein interacts with which one in the cell and in what manner. Although a lot of work and effort has already been done to understand the governing rules of these interactions [112, 1, 113, 90,

[2] www.worldcommunitygrid.org

49, 47, 83, 57, 29], we are not yet able to perfectly develop global methods that would explain all interactions.

This is where my work comes in: Refining the previous methods developed by the team in terms of understanding protein interactions and bringing new concepts through with original methods and analysis. The main goals can be described as:

- Analysing how the different proteins interact with one another
- Developing new methods and pipelines to understand and exploit the interacting surface of two proteins
- Joining the different aspects and concepts developed to provide an efficient way to identify interacting partners in a crowded environment

Difficulties of the questions

Although the prediction of interaction sites is a heavily studied field, large scale predictions of interacting partners through a CC-D is still at the state of pioneer work, paving the way for future studies. The ability to understand how proteins interact will, on top of providing a deep insight on how many of the biological processes are regulated, bring many practical applications for small-ligand design and research against many diseases. Understanding how large-scale networks function will allow us to implement an automation where most of the work done today is manual. Being able to reduce the potential size interacting partners to manually test could drastically reduce a dull, painstakingly long manual work.

Docking has up to this point mainly been used to discriminate the correct native conformation from a set of decoys for two proteins. However, the laboratory team has shown in a previous study [95] how combining known interfaces with docking conformations (and their associated energy) was sufficient to discriminate interacting partners from non-interacting ones. This study has largely motivated the development of protein-protein interface prediction software by the laboratory [29, 57]. Docking conformations have scarcely been analysed under such an angle before, which presents an additional challenge from a methodological point of view. Such large scale studies also imply developing new, adapted methods to analyse them: the HCMD2 CC-D has generated more than a *hundred billion* docking conformations. Another remaining important challenge is (on top of the important combinatory complexity) the large space of negative partners compared with the positive ones. In a pairwise analysis of partners, we have for instance in the PPDBv2 dataset (see Section 3.4) only 168 correct partner prediction amongst 28224 possible protein pairs (and thus 28056 negative pairs).

2.4.3 Advancements made

Many of the current protein-protein interface predictions software now aim to evaluate interfaces considering binary interactions with another protein. But the cell is a crowded environment (see Fig. 1.1) and the multiplicity of interactions of a protein makes and thus its interacting surface is largely underestimated. Proteins continuously make interactions: fleeting ones and more persistent alike (see Section 2.3.2). They interact in competition or in cooperation with each other [64] and we highlight how important it now is to shift the paradigm from looking for pairwise interaction to looking at multiple, potentially simultaneously ones. This new shift opens with it many questions: what are the limits of an interface shared among multiple partners, where does it start and stop? Do binary interaction still really make sense?

A plan for tackling these issues

In this thesis I present the advancements we made in the previously introduced context. As mentioned, two main parts will focus on the prediction and analysis of interaction sites and the large scale prediction of interacting partners respectively. In a third part, I will also present a tool that I developed, INTBuilder [25].

In the first part, I present the background of what has been done in terms of predicting the protein-protein interaction sites as well as presenting the major features describing them. Here, I provide a better understanding of the interaction of proteins in a crowded environment, when several interactions are possible. I introduce the concept of multiple interaction sites (partner-specific) and regions (non-specific), and how our predictions might guide us to a better understanding of them. This is done through an analysis of an original 262 protein chains dataset. This section largely covers my goals of analysing how proteins might interact with one another and developing new pipelines to interpret their interactions.

The second part focuses on the advancements at discriminating interacting partners from non-interacting ones. Joining the knowledge and the better understanding of scores at the interface and how they might play a role in protein-protein interactions, I will show how essential it is to separate proteins into their respective functional classes to evaluate them using different approaches. We also provide an analysis of how these different classes might respond differently to different approaches.

The humongous amount of data generated during the HCMD2 CC-D experiment ($\sim$ 100 billion conformations) required the development of a swift, high throughput and adapted software: I consequently developed INTerface Builder (INTBuilder; [25]) to answer this issue. In the third part I present its development and the novel algorithm of search space reduction it brings with it.

Chapter 3

Methodology

Contents

This chapter focuses on bringing the already existing definitions and methods onto which I have relied upon throughout my doctorate. These methods may have been used or developed for previous studies, or are essential to the understanding of the new concepts and methods we develop in this thesis. The methods we developed are detailed on the respective chapters of the analysis they were established for.

3.1 Interface residues

3.1.1 Surface residues

In the scope of this work, as well as in previous ones [67, 57], we have considered residues to belong to the surface of a protein if they were displaying at least 5% of relative Accessible Surface Area (rASA) computed using the Naccess [41] software. This definition is especially important as most of the methods used to analyse the residues involved in protein interactions only consider surface residues.

3.1.2 Experimental residues

Experimental residues are residues known to interact with a partner. To obtain them, the 3D structure of a protein complex must be resolved. As presented in Section 2.3.3 and for previous studies [67], a change in the rASA upon binding was used using the Naccess software.

3.1.3 Protein-Protein interface prediction

A predicted region is a cluster of residues which we call a *patch*, potentially describing an interaction with a specific partner or a cluster of residues covering an extended surface of the protein and potentially describing several interactions.

The development of protein-protein interface prediction algorithm resulted in a first version, JET [29], which used only sequence based features (although it still required the 3D structure to detect surface residues and to merge clusters) and used the scoring of the physico-chemical properties coupled with an evolutionary trace. The predictions of JET not being precise enough, a newer version JET^2 [57] was next released which made use of the circular variance property of the residues (reflecting the local geometry around them). Below, we describe the different residue-based scores computed and which were used to predict protein-protein interfaces.

Evolutionary Trace

T_{JET} reflects the *evolutionary conservation* level of a residue, and is computed from phylogenetic trees constructed by using sequences, homologous to a query sequence

and sampled by a Gibbs-like approach [29]. The Gibbs-like approach extracts N representative subsets of N sequences [29] in a way that, within each subset, the proportions of sequences sharing $[20-39]\%$, $[40-59]\%$, $[60-79]\%$, and $[80-98]\%$ sequence identity with the query sequence are similar (ideally, about one quarter for each group of identity). Sequences in a subset are then aligned using CLUSTALW2 [58] and a distance tree is constructed from the alignment based on the Neighbor Joining algorithm [101]. From each tree T, a *tree trace level* is computed for each position in the query sequence: it corresponds to the level n in the tree T where the amino acid at this position appeared and remained conserved thereafter (see [29] for a more precise definition). Let us recall that this definition of evolutionary trace is notably different from the measure defined in [66, 75] to rank protein residues.

Then, tree trace levels are averaged over the N trees to get statistically significant values, which we denote *relative trace significances*, or T_{JET}, and which are calculated as follows [29]:

$$T_{JET}(j) = \frac{w_I \times (\frac{1}{|I|} \sum_{h \in I} d_h) + w_j \times d_j}{w_I + w_j} \tag{3.1}$$

where I is the set of residue positions which are neighbours of a_j (*i.e.*, with at least one atom distant by less than 5Å to at least one atom of a_j) and where d_j is the relative trace significance of a_j. The weights were fixed at $w_I = 3$ and $w_j = 4$, as in [29]. T_{JET} values are scaled between 0 (least conserved residue of the protein) and 1 (most conserved residue of the protein) for the calculation of residue scores.

Physico-Chemical properties

PC indicates the physico-chemical propensity specific to amino acids located at a protein interface. The original values, taken from [80], range from 0 to 2.21 and are scaled here between 0 and 1 for the calculation of residue scores.

Circular Variance

CV is the circular variance, a measure of the vectorial distribution of a set of neighbouring points around a fixed point in 3D space [16]. For a given residue, CV reflects the density of the surrounding residues: residues buried within the protein will display high CV values, while exposed or protruding residues will display low CV values. Compared to solvent accessibility, CV changes more smoothly from the surface to the interior of the protein [74], and is thus less sensitive to small conformational changes. CV can be applied equally well to atomic or coarse-grain

representations [16]. The CV value of an atom i is computed as:

$$CV(i) = 1 - \frac{1}{n_i} \left| \sum_{j \neq i, r_i \leq r_c} \frac{\vec{r_{ij}}}{||\vec{r_{ij}}||} \right| \tag{3.2}$$

where n_i is the number of atoms distant by less than r_cÅ from atom i. The CV value of a residue j is then computed as the average of the atomic CVs, over all atoms of j. A low CV value indicates that a residue is located in a protruding region of the protein surface. CV values are scaled between 0 (most protruding residues) and 1 (least protruding residues) for the calculation of residue scores.

Normalised Interaction Propensity

It has previously been shown in [33, 107, 56] that it is possible to exploit the docking procedure to compute a propensity for each residue to belong to an interacting surface. This Interface Propensity (IP) value represents the probability for residue i of protein P to belong to an interaction site.

Here, IP is inferred from CC-D calculations using the MAXDo software [95] (see Section 3.2), where each query protein is docked against many protein partners, that are not necessarily partners in the cell [95, 67]. To compute the IP in earlier works [95, 67], we used a Boltzmann weighting factor which favours docked interfaces with low energies. As a consequence, for a given protein pair PQ, all interfaces with a 2.7kcal/mol or more energy difference from the lowest energy docked interface has a Boltzmann weight lower than 1% (see [67] for more details). This is meant to limit the propensity computations to only the most favourable conformations.

Here, as in [55], we limit the number of docked interfaces that would have to be reconstructed for determining the interface residues, and we choose to calculate residues' IP values using only the lowest energy docking poses satisfying the 2.7kcal/mol condition, we therefore have:

$$IP_P(i) = \frac{N_{int,P}(i)}{N_{pos,P}} \tag{3.3}$$

where $N_{pos,P}$ is the total number of energy-based filtered conformations of protein P docked against some protein Q in the dataset, and $N_{int,P}(i)$ is the total number of energy-based filtered conformations of protein P docked against some protein Q in the dataset having residue i occurring at the interface.

NIP (Normalised Interaction Propensity) is defined in Eq. (3.3) and reflects the propensity of a residue to be found at the interface. The normalisation process, as done in [67], is necessary to compare the IP scores among proteins: a positive NIP value indicates that the residue i is favoured to occur at potential binding sites, and

a negative NIP value indicates that it is disfavoured. NIP is defined as:

$$NIP_P(i) = \frac{IP_P(i) - < IP_P(j) >_{j \in P}}{\max\left(IP_P(j)\right)_{j \in P} - < IP_P(j) >_{j \in P}} \tag{3.4}$$

where $< IP_P(j) >_{j \in P}$ and $\max\left(IP_P(j)\right)_{j \in P}$ are the average IP and the maximum IP, respectively, computed over all the residues j in P. The NIP value represents how often a residue is docked on the retained conformations (that is, those conformations that have less than $2.7 kcal/mol$ energy difference from the best one, as explained above).

Combining residue-based scorings to predict interfaces

Unlike many prediction algorithms, JET^2 strives to reproduce the interface defined by [62] as close as possible. Indeed, looking at Fig. 3.1 we clearly see how the different parts (Support, Core, Rim, See Section 2.3.3) present various attributes. We find as expected for protein-protein interfaces a higher conservation particularly in the most buried region upon interaction (Support), conservation which falls short the farthest we get from the centre of the interface. Based on the Support-Core-Rim model, JET^2 implemented a seed-extension-outer layer model and derived it into three different scoring methods: SC_1, SC_2 and SC_3 (see Fig. 1.4). Each of these scoring methods targets different types of interfaces. A large database regrouping the JET^2 predictions on more than 20 000 protein chains has also been published by the team[1] [91].

I present here the three scoring schemes previously developed in the JET^2 software.

SC_1 targets very conserved residues (identified by the T_{JET} score) to form a seed which is then extended using both T_{JET} and PC scorings. An outer layer is added considering both PC and CV scorings. SC_1 is intended to detect diverse protein binding sites. This step, essentially unchanged compared to the original JET version, was extensively described in [29].

SC_2 detects both seed and extension layers using a combination of T_{JET} and CV scorings. It aims at detecting highly conserved residues that are not buried too deeply beneath the surface of the protein. The outer layer is defined based on PC and CV, as in SC_1. SC_2 specifically distinguishes protein interfaces from small ligand binding sites.

SC_3 disregards evolutionary information and solely employs PC and CV for detecting all three layers of the interface. SC_3 yields consistent predictions for interfaces displaying very low conservation signal, *e.g.* antigen binding sites.

[1] http://www.jet2viewer.upmc.fr/

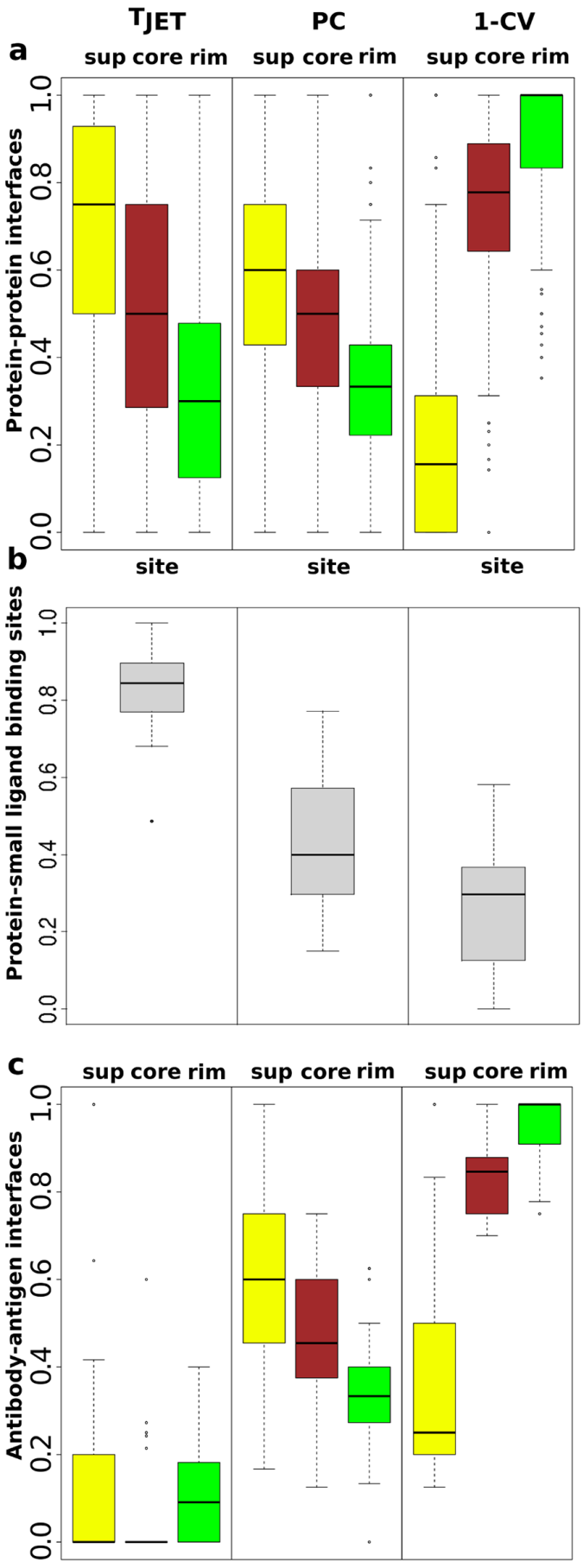


Figure 3.1: Boxplots describing the different properties used by JET^2 at the different areas of the interface. [57]

3.1.4 Evaluating the interface predictions

Interface evaluation

To evaluate our capacity to predict experimental surfaces, we use several well-known statistical measurements that we define below. We consider as a True Positive (TP) an experimental residue rightfully predicted, as True Negative (TN) a non predicted residue which is not experimental, as False Positive (FP) a predicted residue which is experimental and as False Negative (FN) an experimental residue which is predicted. From these definitions, we present below the different scores:

$$\begin{aligned}
\text{Recall} &= \frac{TP}{TP+FN} \\
\text{PPV} &= \frac{TP}{TP+FP} \\
\text{F1-score} &= 2 \times \frac{\text{Recall} \times \text{PPV}}{\text{Recall} + \text{PPV}} \\
\text{Specificity} &= \frac{TN}{FP+TN} \\
\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}
\end{aligned}$$

3.2 Protein docking and conformations scoring

Below, I give an overview of the MAXDo docking software (providing different conformations for each pair of proteins) and the different scoring methods (only providing a mean to evaluate a conformation) I used during my work. MAXDo uses a rigid-body docking approach which can be either symmetric or asymmetric. The asymmetric approach fixes one of the two proteins in space (receptor) and samples a set of starting points around it for the second (ligand). Each starting position and orientation of the ligand is described by a set of Euler angles (see Fig. 3.2) respectively to the receptor. The ligand next approaches the latter as close as possible. This has been the docking method used for performing the CC-D on the PPDBv2 dataset [76] (see Section 3.4). Each docking conformation can then be described using a set of Euler angle, as in Fig. 3.2.

The second method, the one used for the CC-D of the HCMD2 (2246 dataset) is the symmetrised docking. With this method, a set of starting positions is sampled around the receptor as well, but during the docking approach the ligand does not have to close in straight to the receptor and can instead deviate from its axis. During this CC-D, the starting positions were filtered out using a cone from the JET [29] predictions in order to reduce the computation time. This deviation helped avoid overlooking conformations presenting a good energy.

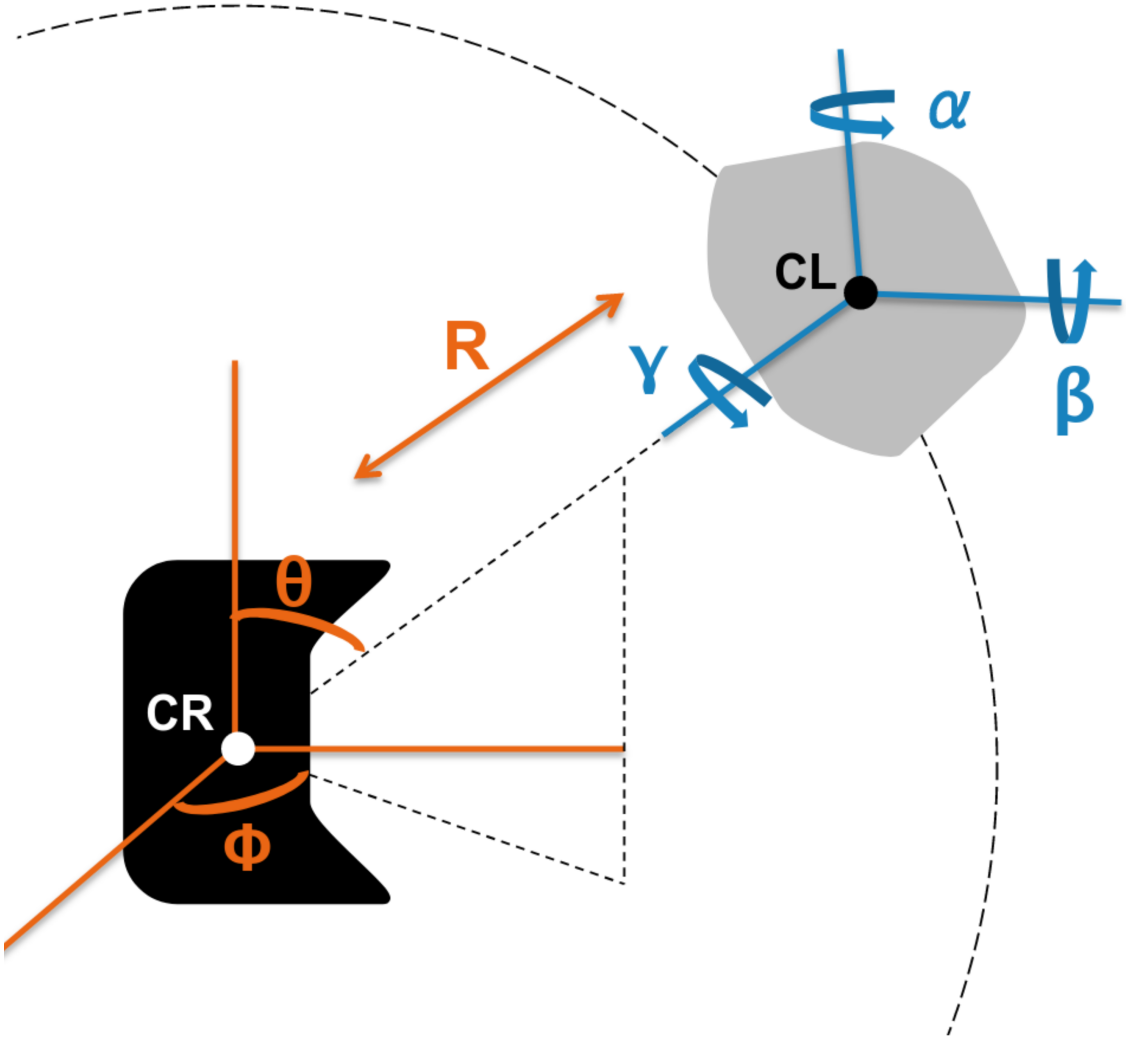


Figure 3.2: Docking process using Euler angles to describe the conformation. Available at https://github.com/meetU-MasterStudents/2017-2018_partage/blob/master/Docs/Meet-U_opening_2018_P6P7P11.pdf.

MAXDo [95] is the docking algorithm used to perform the Complete Cross-Docking computations, explained below (See 2.3.6, 2.4.1); it was developed by the laboratory and reimplements the docking method developed by Zacharias [115]. This algorithm is a rigid-body docking software we used on a coarse-grain reduced protein representation [115]. This coarse-grain representation was necessary in order to reduce the computation time of the whole docking process. To compute the energy, the interactions between the pseudo-atoms of the Zacharias representation [115] are treated using a soft LJ-type potential with appropriately adjusted parameters for each type of side-chain. In the case of charged side-chains, electrostatic interactions between net point charges located on the second side chain pseudo-atom were calculated by using a distance-dependent dielectric constant $\epsilon = 15r$, leading to the following equation for the interaction energy of the pseudo-atom pair i, j at distance r_{ij}:

$$E_{ij} = \left(\frac{B_{ij}}{r_{ij}^8} - \frac{C_{ij}}{r_{ij}^6}\right) + \frac{q_i q_j}{15 r_{ij}^2}$$

where B_{ij} and C_{ij} are the repulsive and attractive LJ-type parameters respectively, and q_i and q_j are the charges of the pseudo-atoms i and j.

iATTRACT [98] is a newer docking software that mixes a rigid-body approach with flexibility. The rigid-body first provides an ensemble of conformations which did not take into account any conformational change. iATTRACT then performs 2500 minimisation steps allowing simultaneously a large rotation of the protein with local deformations of the interface. This algorithm is the follow-up from its previous versions [115, 116]. In this study, we only used MAXDo's docking algorithm to obtain the docking poses. To each of these docking poses was applied iATTRACT's minimisation process before proceeding to use iATTRACT's energy function to score the conformation. The energy function of iATTRACT is described as:

$$V_{\text{protein}} = \sum_{ij} \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6} + \frac{q_i q_j}{\epsilon r_{ij}}$$

where the dielectric constant ϵ is set to 10.

PISA [54] is a scoring method developed to discriminate between biological and non biological complexes. PISA is based on the dissociation free energy to evaluate a complex stability. On top of the dissociation free energy, PISA considers larger assemblies more probable than the smaller ones and considers that single-assembly sets take preference over multi-assembly sets. As such, it can be used to evaluate the likeliness of a conformation to be biological (and thus can be used to score the conformations from a docking algorithm).

CIPS [79] is a pair potential scoring method developed in the laboratory. CIPS was trained using 230 bound structures from the Protein-Protein Docking Benchmark 5.0 [108]. CIPS is meant to be used as a high throughput technique able to largely filter out most of the non-native conformations with a low error rate. This will help when combined with our predictions to determine which partners interact together.

3.3 Detection of interacting partners

3.3.1 Interactions evaluation

To perform the discrimination, we score each conformation of a docked pair of proteins (P_1, P_2) to represent how likely these two proteins are to interact with each other in this conformation. This gives us a set of values for each protein pair, from which we select the best (the minimum, as the energy is negative). The previous study [67] used the following formula to compute the Interaction Index II_{P_1,P_2}:

$$II_{P_1,P_2} = \min(FIR_{P_1} * FIR_{P_2} * E^{MAXDo}_{P_1,P_2}) \tag{3.5}$$

where $E^{MAXDo}_{P_1,P_2}$ is the corresponding energy computed with MAXDo to each conformation of P_1 and P_2, FIR_{P_1} and FIR_{P_2} the Fraction of Interface Residues representing the overlap between the docking interface and the known or predicted interface. This method lets us define a unique II for each pair of proteins, from which we are able to define a matrix. This II value is then normalised using the Equations 3.6 and 3.7. The idea behind this normalisation is explained below.

Previous studies from the laboratory [67, 56] showed that taking into account how a protein interacts in the dataset is crucial to correctly assess how it interacts with a given partner. Thus, we define here for every protein pair P_1, P_2 a Normalised Interaction Index (NII) as:

$$NII_{P_1,P_2} = \frac{\min(II'_{P_1,P_2}, II'_{P_2,P_1})^4}{\min_P(II'_{P_1,P}) \min_P(II'_{P_2,P}) \min_P(II'_{P,P_1}) \min_P(II'_{P,P_2})} \tag{3.6}$$

where II'_{P_1,P_2} is a symmetrised version of the interaction index II_{P_1,P_2} and is defined as:

$$II'_{P_1,P_2} = \frac{II_{P_1,P_2}}{\sqrt{M_{P_1} M_{P_2}}} \qquad M_{P_i} = \frac{1}{2|\mathcal{P}|} \sum_{P_j \ in \mathcal{P}} II_{P_i,P_j} + II_{P_j,P_i} \tag{3.7}$$

where $\mathcal{P}$ are the 168 proteins of our dataset. NII values vary between 0 and 1. Values close to zero imply that two proteins cannot form an interface involving a significant fraction of the experimentally identified residues, or that interfaces involving these residues have poor interaction energies. Values close to one indicate

predicted interfaces with good energies and composed of experimentally identified residues.

For each protein P_1, we define as predicted partner of P_1 the protein P_i that leads to $NII_{P_1,P_i} = 1$.

3.3.2 Partner identification evaluation

To evaluate our capacity to identify interacting protein partners, we define here as TP the predicted protein pairs interacting with one another, as TN the protein pairs correctly predicted as non-interacting partners, as FP the non-interacting protein pairs predicted as interacting and FN the interacting protein pairs not predicted as interacting. Using these values, we define the False Positive Rate (FPR) and the True Positive Rate (TPR) as follows:

$$FPR = \frac{FP}{FP + TN} \quad TPR = \frac{TP}{TP + FN}$$

The computation of FPR and TPR for various thresholds enables the Receiver Operating Characteristics (ROC) curve to be drawn. The performance of our partner identification capacity is given by the resulting AUC (Area Under Curve) value. An AUC of 0.5 would correspond to a random prediction whereas an AUC of 1 would represent a perfect prediction.

3.4 Datasets

3.4.1 PPDBv2 dataset

The Mintseris Protein-Protein Benchmark Dataset v2 (PPDBv2, see Section 2.4.1, [76]) comprises 84 protein known complexes which where each separated in a receptor and a ligand protein in its unbound form. The average size of the protein in residues in this dataset is 287, the minimum 29, the maximum 1979 and the standard deviation 230. Those complexes do not always refer to a single chain, but can also regroup several of them, as a multimeric biological unit.

Part I

Protein-Protein Interface predictions

Chapter 4

Multiple binding site analysis

Contents

4.1 The question

I show in this chapter how we developed the dynJET^2 algorithm, a structure based interface prediction algorithm providing different scoring methods depending on the type of surface.

Furthermore, I present as well the analysis and a new insight of protein-protein interactions on an original dataset `P-262`. The publication (soon to be submitted) performs an analysis of this dataset, brings a new concept of Interacting Site (IS) versus Interacting Regions (IR) (see Chapter 3, Subsection 4.2.2) and explains how crucial the multiple sites' concept is for proteins interaction analysis and what we can do to interpret them. We also show how it might be possible to analyse our predictions to infer if a surface might interact with one or more partner.

First, I will present the different methods that we developed in the scope of this study, then I will present the context that led us to pursue this direction and finally the results obtained.

C. Dequeker, E. Laine, A. Carbone, "Multiple binding sites of protein-protein interactions predicted by combining sequence analysis and molecular docking", to be submitted, 2018

4.2 Methods

4.2.1 `P-262`, a dataset of protein chains

This original dataset, named `P-262`, is a subset of the larger one studied in the HCMD2 project (see Section 2.4.1). Starting the analysis of the 2246 proteins dataset showed us that some complex structures in the dataset were experimentally resolved, which allowed us to build the sub-dataset. The chains in `P-262` are those that remain from the larger one after excluding: (a) only α-carbon structures (b) chains for which results were missing (c) chains forming coiled-coils complexes (d) deprecated PDB code (e) chains for which no interface of 5 residues or more could be found in the associated PDB file (see Section 3.1) (f) chains for which no biological interfaces (of more than 5 residues) could be found for their homologs in the whole PDB (considering 90% sequence identity, see Section 4.2.5).

Biological Unit

Biological units or biological assemblies describe functional interactions. Such biological assemblies are either "author provided" or "software determined" (using the PISA software [54]), and we choose to consider both. This ensures that the interfaces computed in the complex using the INTBuilder software [25] (see Chapter 6) carry

a biological meaning. We thus defined the dataset of 262 different chains coming from 107 complexes comprised of two or more chains.

4.2.2 Experimental residues

The experimental residues presented in Section 3.1.2 are computed in this study using a distance based definition to determine which residues belong at the interface. To accomplish this, we used the INTBuilder software[1] (See Chapter 6; [25]) with a distance threshold of 5Å and considered the resulting set of experimental residues as defining the interface of the complex. Note that we only considered the experimental interfaces of at least 6 residues and computed from a complex known to be a biological unit (considered to be involved in a functional, biological interaction).

Interactions in non-binary complexes

Proteins might interact in several manners. Let A, B and C be three proteins; we consider as a *single interaction* the case where A and B exclusively and solely bind to one another, excluding C. Note that a single interaction may involve several partners: if B and C both bind to A, and B has at least one residue at less than 5Å from a residue of C, then we consider proteins B and C to be in contact, describing a single interaction with A. However, if B and C are not in contact, then we refer to the interactions between B and A, and C and A as *separate interactions*. We refer to a *multiple partners interaction*, if two or more proteins bind to another protein to form a complex, as for instance B and C binding to A. A more schematic representation of these definitions may be found in Fig. 4.1.

Interacting regions and sites of a protein surface

Protein surfaces can be decomposed in *Interaction Regions* (IR) or *Interaction Site* (IS). To define these IR and IS, we consider clusters of residues, either experimentally defined or predicted as in Section 3.1.3 (a cluster of residues is made of multiple residues separated by $\leq$ 5Å from one another). An experimental IS is an interacting surface specific to a single pairwise interaction between two proteins. An experimental IR describes a cluster of residues known to be involved in more than one interaction. Experimental regions are identified by using the approach described in Section 4.2.5, which gathers a set of IS retrieved from close homologs of the query protein (sequence identity $> 90\%$). To obtain a region from residues clusters, we merge two clusters of residues C_1, C_2 at the surface of a protein if the maximum proportion of their overlap with respect to their size ($\max\{\text{overlap}(C_1, C_2), \text{overlap}(C_2, C_1)\}$) is below a threshold (we used a threshold of 0.6 as it gave us the most realistic regions compared to the experimental information). Additionally, we

[1] www.lcqb.upmc.fr/INTBuilder/

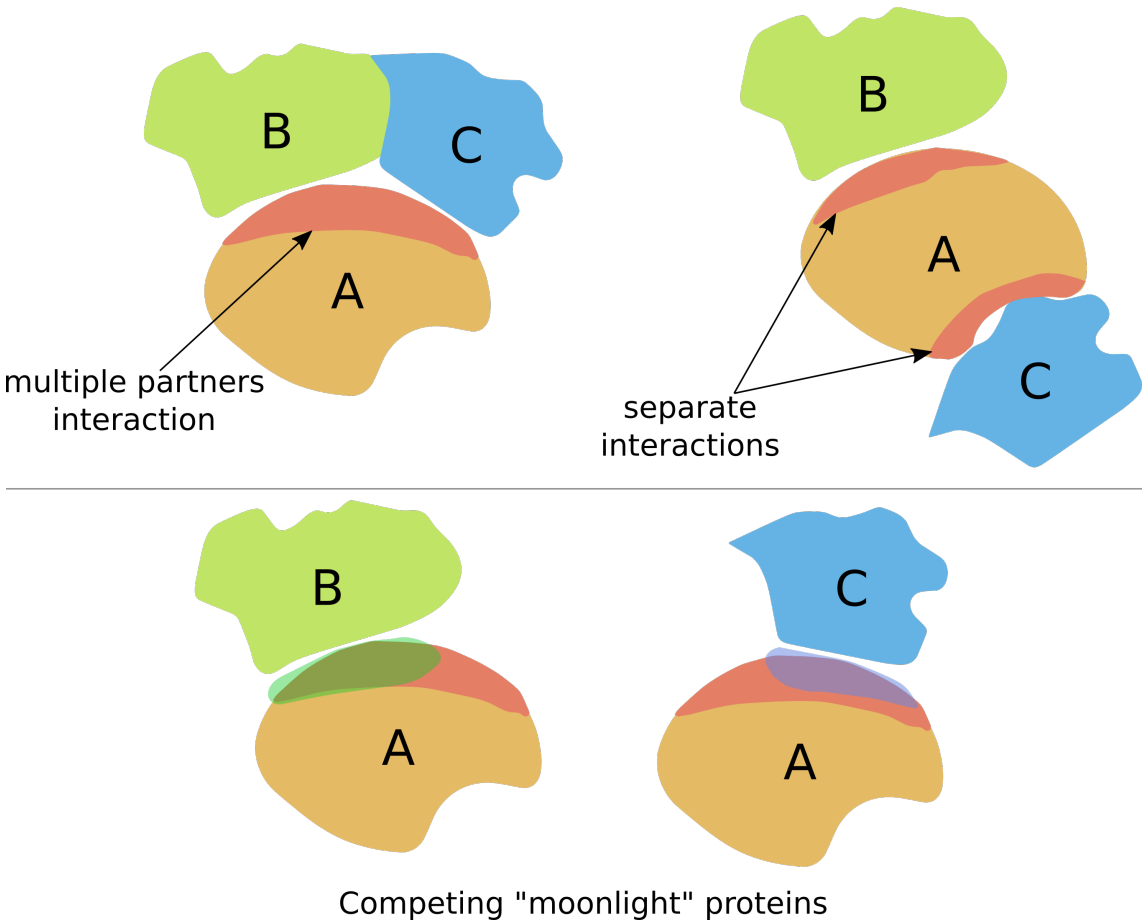


Figure 4.1: Schematic representation of the different types of interactions possible involving non-binary complexes.

merge any small cluster of at most 5 residues with another cluster if they overlap by at least one residue, regardless of the overlap percentage. This process is iterated over all interaction sites for a single protein until no more interaction site is left.

PPI-262, a dataset of experimental interfaces

We computed every experimentally known interaction (as defined below) between proteins of P-262 and obtained PPI-262, a set of 329 experimental interaction sites. The median size in residues for the 262 protein chains set is 192.5, the average is 200.5 and its standard deviation is 131.2, indicating a large variation of protein size inside the dataset. Indeed, the smallest protein comprises 21 residues against 789 for the largest. Based on the information recovered from the PDB complexes, the 262 protein chains have been classified in seven different functional classes, following [108]: 6 Inhibitors (I), 7 G-proteins (G), 13 Receptor-proteins (R), 17 Antibodies (AB), 10 Enzymes Regulatory (ER), 56 other Enzymes (E) and 136 Others (O) that we were not able to classify in any of the other functional subclasses.

PPI-262$_{ext}$, an extended dataset of experimental interfaces

To obtain a most accurate evaluation of our predictions, we extended the PPI-262 dataset to a wider range of interfaces coming from the known homologs of each protein in the dataset. We worked under the hypothesis that homologs with a sequence identity of 90% to the corresponding protein in PPI-262 describe the same protein. Hence, the interaction of the homologs with other partners could be used as extra information on the interactions of the protein in the dataset. To construct the extended set of interaction sites PPI-262$_{ext}$, we first searched for homologs in the Protein Data Bank (PDB) that have at least a 4Å resolution. These homologs were pre-computed by the PDB using BLASTClust (with the arguments -p T -b T -S 90) for clustering their sequences at 90% identity, and we downloaded them[2]. We then retrieved the homologs, retained only the ones known to belong to biological assemblies and computed their experimental interaction sites. This step provides a number of new IS. In order to map interacting residues from the homologous structure to the original protein, we perform a global pairwise sequence alignment using a blosum62 matrix between the protein and its homolog. The large number of IS (23642) thus obtained represent the totality of known functional interactions defined throughout the entirety of the PDB for the 262 query proteins. Once all the IS were mapped to each query protein, we merged them into IR, as described in Section 4.2.2. 370 IR total were obtained for the proteins in P-262. The whole processes leading to compute the PPI-262 and PPI-262$_{ext}$ sets is shown in Figure 1.2.

[2] ftp://resources.rcsb.org/sequence/clusters/

4.2.3 Predicted residues

As mentioned in the introduction, the main focus of this part is to provide a new insight, a better understanding of protein-protein interactions and an improvement of the already existing methods. In an effort to provide with more refined and accurate interfaces, we developed dynJET^2 (available at[3]). This enhanced version of the previously used protein-protein interface predictions software JET and JET^2 (see Section 3.1.3, [29, 57]) can incorporate any residue-based scoring on top of the prediction capacities of the JET^2 software.

With dynJET^2's ability to incorporate any residue-based score to the prediction algorithm, we were able to include the NIP score (see Section 3.1.3) in its computations method. This has allowed me to create a different scoring scheme for this score: SC_{NIP} (see below). This combination brings together the efficient clustering algorithm from JET^2 with the NIP value.

We note as well on Fig. 3.1c how poorly conserved are the AA interactions and other studies such as [30, 48, 50] confirm how different Antibodies interfaces are. As said in [48], "antibody protein interactions are relatively "happenstance" and are selected principally by the strength of the binding constant, without being subject to evolutionary optimisation over many years"; indeed, the capacity for docking-based scores such as NIP (further described below) show far greater capacity at defining the antibodies' interaction site. so match the possibility of dynamically adding new values to the interface prediction as opposed to the intrinsic values previously, we named this new extension of the software dynJET^2.

Definition of the scoring schemes

We computed the NIP for this dataset over $\sim 50\,000$ energy filtered conformations (see Section 3.1.3) per pair of protein docked, accounting for more than 1.6 billion interfaces in total for `P-262`; we therefore computed the IP value over $13\,100\,000$ different conformations on average for each residue.

To include the NIP score alongside SC_1, SC_2 and SC_3, we added it in dynJET^2 at different stages of the clusterisation (see Fig. 1.4). Each different strategy for introducing the NIP derives the three scoring schemes SC_1, SC_2 and SC_3. I refer to them below as SC_{4*}, SC_{5*} and SC_{6*}. This notation is further used in my work to refer as the best combination of the three derivations when considering predictions (See Section 4.2.4). In the same fashion, I refer to the set of patches from SC_1, SC_2, SC_3 and SC_{NIP} as SC_{d*}.

SC_{NIP} applies the NIP score of the residues to all three layers (core, extension and outer layer). The usage of NIP is motivated by the observation that proteins

[3] www.lcqb.upmc.fr/dynJET2/

tend to dock to their cognate partners and also to non-interactors via the same region at their surface [95, 71, 67, 107].

SC$_{4*}$ Merges the NIP score with the JET^2 scoring for all three layers.

SC$_{5*}$ Uses only the NIP for the seed detection, and combines it with JET^2 for the two remaining layers. This scoring scheme aims at picking up the seeds using the NIP information while still relying on JET^2 to extend further.

SC$_{6*}$ Keeps the JET^2 scoring for the seed detection and combines it with NIP for the remaining layers. Unlike SC$_5$, SC$_6$ relies on JET^2 to find the signals necessary to detect the seeds. Its two next layers are detected by combining both the JET^2 and NIP scores.

4.2.4 Best combination of predictions

To properly assess our interfaces predictions quality (using dynJET^2), we chose to take the combination of interfaces that would best match the experimental targeted interface. This is done by taking each set of predictions from either SC$_{d*}$, SC$_{4*}$, SC$_{5*}$ or SC$_{6*}$ (Fig. 1.4 for a definition of the sets) and merging the different predicted patches to obtain the best F1-score value against the targeted experimental interface. This process gives us a single predicted patch for each experimental interface.

Comparison with Multi-VORFFIP

To compare dynJET^2 to Multi-VORFFIP[4] (in Fig. 4.4a; [100]), we considered 252 protein chains, instead of the 262 comprising the `PPI-262` dataset from which we eliminated the chains used for Multi-VORFFIP's training and those for which it provided no answer. We then considered the residues as being predicted if Multi-VORFFIP gave them a probability of > 0.5 to belong to an interface, as in [57]. For each complex and each prediction method, the union of predicted residues was compared to the union of experimental IR's residues and the associated F1-score was computed.

4.2.5 Homology

Conformational variability of IRs

For each IR, the Root Mean Square Deviation (RMSD) of its backbone atoms (or, if not possible, its C-α atoms) was computed between the query structure from `PPI-262` and each of the homologous structures on which the IR was detected. For each homologous structure, only the subset of residues detected on this structure

[4] www.bioinsilico.org/cgi-bin/SUPER_VORFFI/htmlVORFFI/home

were considered to compute the RMSD. RMSD values were then averaged over the homologous structures (including the query structure if the IR was also detected on it). This gives us a single RMSD value for each IR.

Counting the number of partners

To count how many different partners a protein has, we consider all known homologs of the protein in the PDB and their partners. We cluster the partners depending on their sequence homology: two partners are different if they share less than 90% sequence identity. This threshold in agreement with the criteria we applied to protein chains and their homologs. The number of protein classes provides an estimation of the number of partners for the protein.

4.3 Multiple interactions

4.3.1 Background

A detailed description of the protein interactions with other proteins, nucleic acids and small molecules is expected to provide direct information on the biological processes they regulate and on the way to interfere with them. The ensemble of protein interactions taking place in a living cell can be represented as a graphical network, where each node stands for a molecule and each edge stands for an interaction. Our knowledge of interaction networks is largely incomplete, as the experimental assessment of all possible interactions of a protein is very challenging [42, 93]. A protein may interact with several partners at the same time each partner binding to a different site at its surface, or its surface may present a shared binding region that will be used by different partners at different moments of its lifetime [64]. In order to get a comprehensive view of the multiplicity of protein interactions, we need to be able to decipher the complexity of protein surfaces toward identifying binding sites and binding regions and characterising their specific properties. Such a description, provided at the residue level, would also permit to predict the impact of mutations on protein interactions and hence functions.

Prediction and coverage of the interacting surface

Conservation, physico-chemical properties, and local geometry have been used to predict interacting surfaces [105, 59, 11, 47, 36, 17, 81, 84, 29, 57, 30], and, based on these properties, in the past 15 years, a number of tools have been developed [57, 113, 31, 106] (see [30, 4] for surveys). These tools either classify surface residues as interacting or non-interacting, or predict interaction patches, generally one or two per protein. A recent study highlighted that while most prediction methods typically predict $25-30\%$ of the protein surface as interacting, as much as 75% of

the protein surface could potentially be used for protein-protein interactions [103]. This number was estimated by copying, for a given protein, all protein interfaces from complex structures in the Protein Data Bank (PDB [9]) having a similar fold irrespective of their sequence identity. Although not all copied interfaces are likely to be functional for the query protein, this estimation suggests that the interface to surface ratio is underestimated by most predictors.

An alternative strategy to predict interacting residues consists in exploiting molecular docking calculations. Docking methods were originally designed to predict the structure of a complex starting from the known structures of its components. Candidate conformations, called docking poses, are generated and evaluated based on properties reflecting the strength of the association, *e.g.* shape complementarity, electrostatics, desolvation, conformational entropy. By deriving statistics from the collection of docking poses, one can estimate the propensity of each protein surface residue to be found at a docked interface and use these propensities to identify binding sites [33, 55]. This has been realised in single docking studies [37, 61, 44, 24, 45], where the docking involves two protein partners already known to interact, in arbitrary docking studies [72], where proteins from a benchmark set are docked to arbitrarily chosen proteins, and in complete cross-docking (CC-D) studies [95, 67, 107, 56, 55], which involve performing docking calculations on all possible protein pairs within a given dataset.

Competition, cooperation and prediction of multiple binding sites

It has been shown in [64] that proteins present binding sites targetable by a multitude of different interactors. In the present analysis, we combine residue based properties inferred from protein sequence and structure analysis, namely evolutionary conservation, physico-chemical properties and local geometry, with residue propensities to be found at an interface derived from docking simulations to demonstrate how these features can help to decipher how such "hub" proteins might interact in a crowded environment such as the cell. We predict patches at protein surfaces with the dynJET^2 algorithm, an updated version of the existing tool JET^2 [57, 91] integrating the four features in four different scores (see Methods 3.1). Each dynJET^2 patch reproduces the support-core-rim model of interacting surfaces (see Section 2.3.3, [62]). To do so, dynJET^2 first identifies a small group of residues localised on the protein surface, called the "seed" of the patch, and then extends it with two successive layers of residues. The patches predicted by the different scores may be distinct or partially overlapping, reflecting the multiplicity of interactions a protein may establish during its lifetime. They are compared with a set of experimentally known protein interfaces detected at the surface of 262 protein chains.

These protein chains are part of a larger set of 2246 proteins involved in muscular dystrophy, on which we performed complete cross docking (see Methods 4.2.1).

Starting from the observation [69] that functional interfaces are conserved across closely related homologs, we retrieved all interacting surfaces described by complexes in the PDB involving either a protein from the dataset or a close homolog. By coupling these interacting surfaces, we were able to define experimental interacting sites (IS, used by a single partner) and interacting regions (IR, used by one or several partners) for each protein, recovering as much information as possible on the multiple interactions that the protein might have in the cell.

We show that dynJET^2 is useful to detect both IS and IRs. We demonstrate that the evaluation of protein-protein interface prediction algorithms cannot be correctly assessed by relying on one single complex for a given protein. In most cases, IS cannot be precisely defined based on their properties and it is more pertinent to consider IR instead. Moreover, by exploiting the three layer structure of the predicted patches, we are, in some cases, able to estimate the number of interacting partners.

4.3.2 Complexity of the multiple interfaces

Docking calculations and dynJET^2 predictions were performed on `P-262`, representing 262 protein chains. The predictions were assessed against two sets of experimental interfaces, `PPI-262` and $\texttt{PPI-262}_{\texttt{ext}}$. `PPI-262` comprises 329 IS detected on `P-262` and $\texttt{PPI-262}_{\texttt{ext}}$ 370 IR. The two examples in Fig. 4.2 illustrate the complexity of the experimental interaction surfaces. Binding sites may be disjoint, overlapping or included in others (Fig. 4.2, on the left), and they may be defined by the interaction with other proteins or small ligands (Fig. 4.2, on the right). The two examples show 5 IS (3 on the left and 2 on the right), which were merged into 3 distinguished IR (2 on the left and 1 on the right, contoured by thick forest green lines).

4.3.3 Estimation of the protein surface involved in functional interactions

A proper estimation of the protein surface involved in functional interactions is necessary to correctly assess protein interface prediction algorithms. On average, the union of experimental IS detected on `PPI-262` cover 29% of the protein surface (Fig. 4.3a). Hence, by looking at `PPI-262`, one may infer that the residues involved in functional interactions generally represent less than a third of the protein surface. However, when looking at the extended dataset $\texttt{PPI-262}_{\texttt{ext}}$ (Fig. 4.3b), the coverage increases up to 48% and a significant number of proteins (32) have their surface completely or almost completely covered by functional interactions (coverage $\geq$80%). This suggests that most of the proteins from `P-262` engage in multiple interactions with different partners. Considering only one complex for each protein leads to underestimating interacting surfaces.

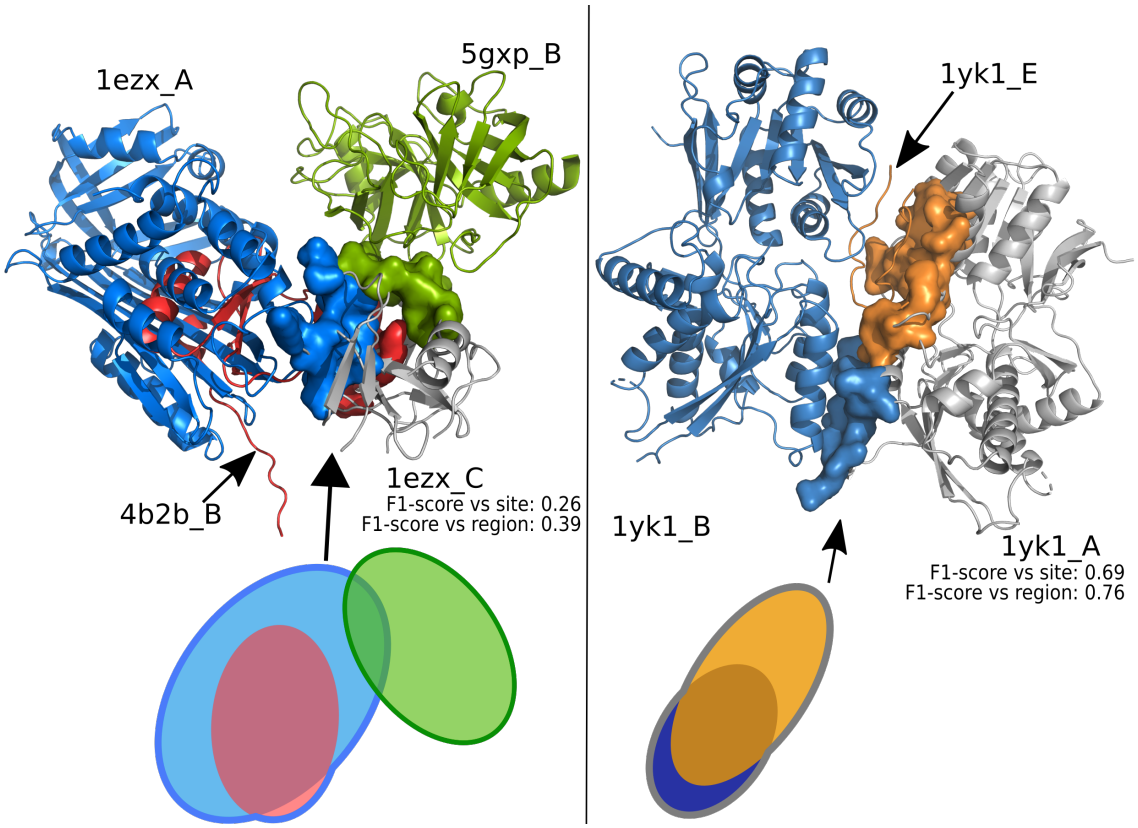


Figure 4.2: **Two examples of the usage of the protein surface by different partners.** The query proteins are displayed as grey cartoons, their interacting sites as opaque coloured surfaces and their partners as coloured cartoons and transparent surfaces. Left: protein chain 1ezx_C (in grey) interacts with its partner 1ezx_A (in blue) and two other partners, 4b2b_B (in red) and 5gxp_B (in green). The 3 corresponding IS lead to the definition of 2 IRs, as depicted on the schema at the bottom, where each IR is contoured by a thick forest green line. Notice that the green and blue IS are not merged because they overlap by less than 60% of their respective surfaces. Right: the complex 1yk1 is composed of two proteins (in grey and blue) and an interposed ligand (in orange). The 2 IS detected at the surface of the 1yk1_A chain are merged into an IR. F1-scores computed for dynJET^{2}predictions (best matching combination of predicted patches) against two IS (1ezx_A-1ezx_C and 1yk1_A-1yk1_E) and the associated IR are reported.

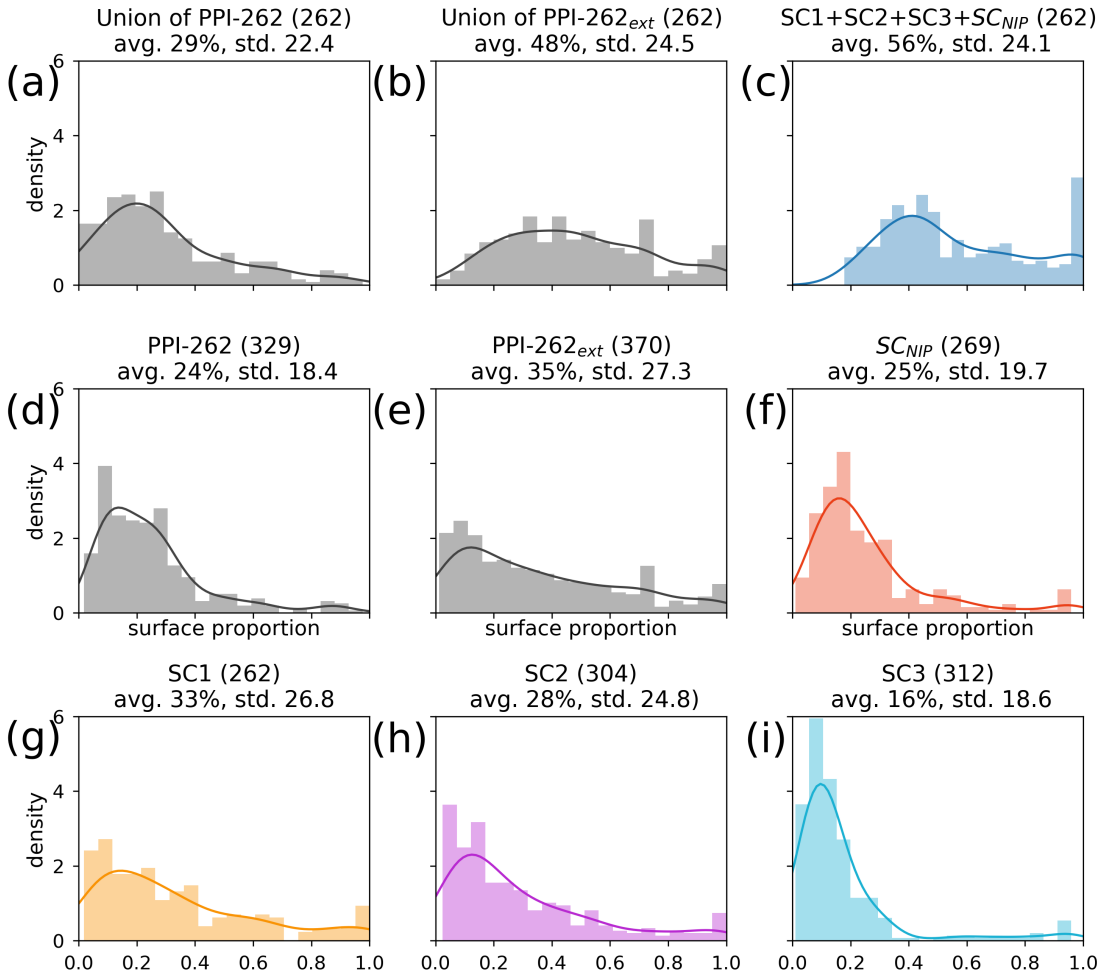


Figure 4.3: bf Proportion of protein surface covered by experimental interfaces and predicted patches. Distribution are reported for: **(a)** the union of IS from `PPI-262`, **(b)** the union of IR from $\texttt{PPI-262}_{\texttt{ext}}$, **(c)** the union of patches predicted by dynJET^2, **(d)** individual IS from `PPI-262`, **(e)** individual IR from $\texttt{PPI-262}_{\texttt{ext}}$,**(f-i)** individual patches predicted by each dynJET^2's scoring schemes (SC_1: yellow, SC_2: purple, SC_3: cyan, SC_{NIP}: red). The union of IS, IR or predicted patches is realised for each protein. Notice that the sizes of the predicted patches do not add up when considering their union, since several of them overlap.

The estimation provided by the union of dynJET^2 predictions is slightly higher, 56% on average (Fig. 4.3c). The associated distribution is similar to that of experimental interfaces (compare Fig. 4.3c with 4.3b), except for two notable differences at the extremities: the minimum coverage is higher for predictions than for experimental interfaces (18% versus 6.2%), and there are more proteins completely or almost completely covered ($\geq 80\%$) by predictions than by experimental interfaces. The first difference can be explained by the specifics of dynJET^2 clustering algorithm, which discards very small predictions (see Methods and [29]). The second difference suggests that all functional interfaces have not been yet experimentally characterised.

We also evaluated the relative sizes of individual experimental interfaces, namely IS and IR (Fig. 4.3de), and of individual patches predicted by dynJET^2 (Fig. 4.3fghi, and see Methods for a precise definition of predicted patches). Experimental IS and docking-based (SC_{NIP}) predicted patches represent about one quarter of the protein surface, on average, and display very similar distributions (compare Fig. 4.3d and 4.3f). Experimental IR and conserved (SC_1, SC_2) predicted patches are bigger, covering about one third of the protein surface, on average (Fig. 4.3e,g,h). They display much larger standard deviations, in the $[24-28]\%$ range, denoting their great variability. Finally, the predicted patches that are protruding and not conserved (SC_3) are the smallest (Fig. 4.3i), with an average size of 16% of the protein surface (almost twice as small as SC_1 predictions). These results suggest that SC_{NIP} and SC_3 are suited to detect binary binding sites whereas SC_1 and SC_2 rather describe generic binding regions.

4.3.4 Assessment of the overall predictive performance of dynJET^2

The identification of a protein's set of interacting residues is important to understand the determinants of molecular association. For each protein, we compared the union of all patches predicted by dynJET^2 with the union of all IS (respectively IRs) from `PPI-262` resp. $\texttt{PPI-262}_{\texttt{ext}}$). To do so, we relied on the F1-score, which reflects the balance between precision (or positive predictive value) and recall (or sensitivity). The average F1-score on `PPI-262` is 0.41 ± 0.24 and it increases up to 0.57 ± 0.19 on $\texttt{PPI-262}_{\texttt{ext}}$ (Fig. 4.4a). This increase reflects a global shift of the F1-score distribution toward higher values (`p-value`$=10^{-4}$ with the Mann-Whitney U test) In particular, the proportion of proteins with very good predictions (F1-score > 0.6) increases from 18 to 46% while the proportion of proteins with very poor predictions (F1-score < 0.2) drastically reduces from about one quarter to 4%. These results highlight the importance of considering all available experimental information to properly evaluate protein interface predictions. Predicted residues that would be

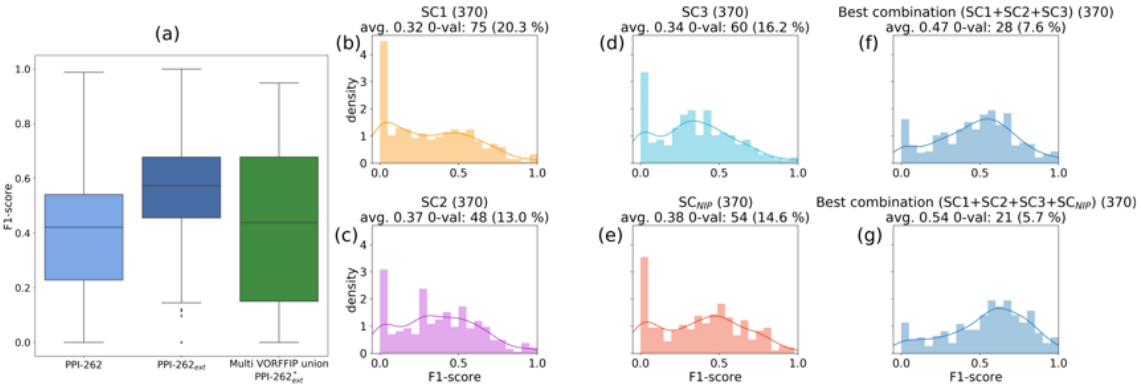


Figure 4.4: **Agreement between experimental interfaces and predicted patches.** **(a)** Distribution of F1-scores computed for the union of dynJET^2 predictions (in tones of blue) and for the union Multi-VORFFIP predictions (in green), for each protein. dynJET^2 predictions were assessed on the union of residues from `PPI-262` (in light blue) and from $\texttt{PPI-262}_{\texttt{ext}}$ (in dark blue), while Multi-VORFFIP predictions were assessed on a subset from $\texttt{PPI-262}_{\texttt{ext}}$ involving 252 protein chains (see Methods) **(b-e)** Distributions of F1-scores computed for individual patches predicted by dynJET^2 scoring schemes (SC_1: orange, SC_2: purple, SC_3: cyan, SC_{NIP}: red) against the best matching combination of IR from $\texttt{PPI-262}_{\texttt{ext}}$. Distributions of F1-scores computed for the best matching combination of predicted patches against each IR from $\texttt{PPI-262}_{\texttt{ext}}$.

considered as false positives when looking only at the restricted dataset, PPI-262, are actually involved in interactions with other partners as revealed by the extended dataset, $\texttt{PPI-262}_{\texttt{ext}}$. dynJET^2 predictions are more sensitive and more precise on this dataset.

We compared dynJET^2 predictions to those of Multi-VORFFIP [100], a state-of-the-art machine learning method, integrating a broad set of residue descriptors including solvent accessibility, energy terms, sequence conservation, crystallographic B-factors and Voronoi Diagrams derived contact density, in a two steps random forest ensemble classifier. Against a subset of $\texttt{PPI-262}_{\texttt{ext}}$(see Methods 4.2.4), Multi-VORFFIP predictions display a distribution of F1-scores much wider than that obtained for dynJET^{2}predictions (Fig. 4.4a, compare the blue and green boxes). Moreover, the average F1-score is of 0.42 ± 0.28, significantly lower than the average value of 0.57 ± 0.19 computed for dynJET^{2}on the same dataset.

4.3.5 Contribution of different scores in the detection of interacting regions

We further investigated to what extent the partitioning of protein surfaces into patches predicted by dynJET^2 matches experimental IRs. By definition, an IR is the result of merging several IS (see Methods 4.2.2). Two IS being merged into an IR may represent two binary interactions with two different partners targeting overlapping areas on the protein surface, as illustrated on Fig. 4.2, or a single binary interaction with a single partner whose binding mode slightly differs from one PDB structure to another. Hence, IR provide a way to account for multiple interactions and also for the binding mode variability of one single interaction. The multiplicity and diversity of interactions and associated binding modes support the definition of IRs, in addition to IS.

The distributions of F1-scores computed for each scoring scheme (Fig. 4.4bcde) display broad spectra of values, showing that none of the scores is sufficient on its own to detect all IRs. This observation is also illustrated by the two examples of Figs 4.5a and 4.5c, where several scores are necessary to capture the entirety of the experimental signal. Combining SC_1, SC_2 and SC_3 enables increasing the average F1-score by about 10 points and drastically reducing the number of completely missed to IR 28 over 370 (7.6%) (Fig. 4.4f). This is indicative of the complementarity of the three scoring schemes in their coverage of the protein surface, as already observed in [57]. Accounting for SC_{NIP} patches further enhances the quality of the predictions (compare Fig. 4.4f and 4.4g).

To better characterise the contribution of docking-based information, we compared the predictive performance of SC_{NIP} with those of SC_1, SC_2, SC_3 (Fig. 4.5b), either considered individually ($JET^2_{\max}$, on top) or altogether (JET^2_{comb}, at the bot-

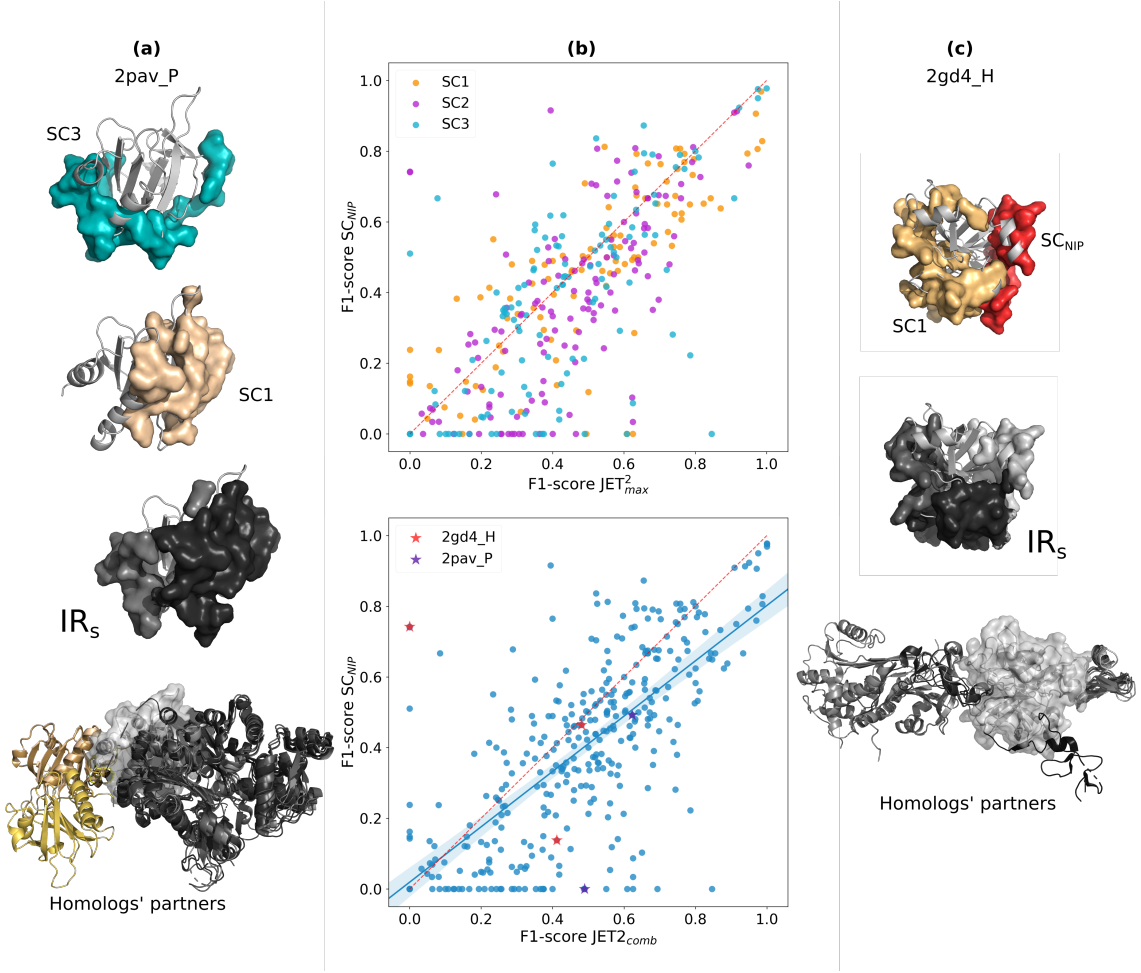


Figure 4.5: **Examples and comparison of dynJET^2 predictions.** (a) Protein structure 2pav_P (light grey cartoon) displayed with the patches predicted by SC_1 (in beige) and SC_3 (in cyan), the two experimental IR from $\texttt{PPI-262}_{\texttt{ext}}$ (in grey tones) and the corresponding partners (beige, yellow and black cartoons); (b) Scatter plot of F1-scores computed for SC_1, SC_2, SC_3 (x-axis) and for SC_{NIP} (y-axis) against experimental IR from $\texttt{PPI-262}_{\texttt{ext}}$. For each IR, the best matching patch or combination of patches is considered. Top: scone, SC_2 and SC_3 are considered individually and the best matching scoring scheme, JET^2_{max}, is retained. Bottom: SC_1, SC_2 and SC_3 are combined together to define JET^2_{comb}. (c) Protein structure 2gd4_H (light grey cartoon) displayed with the patches predicted by SC_1 (beige) and SC_{NIP} (red), the three experimental IR from $\texttt{PPI-262}_{\texttt{ext}}$ (in grey tones) and the corresponding partners (medium grey, dark grey and black cartoons).

tom). We observed that the vast majority of IR (68%, respectively 75%) are better or equally detected by JET^2_{max} (resp. JET^2_{comb}) than by SC_{NIP}. Hence, evolutionary conservation, physico-chemical properties and local geometry are generally able to better capture protein interface signals than the coarse-grained empirical energy function used in the docking experiment. Nevertheless, there are a number of cases where docking-based predictions are more accurate (Fig. 4.5b, points above the diagonals). Protein 2gd4_H provides a good example for this (Fig. 4.5c): among the three IR displayed at its surface, one (in white) is very well detected by SC_{NIP} (in red, F1-score = 0.74), while it is completely missed by JET^2_{comb}. In cases like this, docking-based data provide valuable information to improve predictions by unveiling interfaces that could not be detected otherwise.

4.3.6 From an interacting region to the prediction of multiple protein interactions

94% of the IR from $\texttt{PPI-262}_{\texttt{ext}}$ could be detected, at least partially, by using all $dynJET^2$ scoring schemes (Fig. 4.4g). Some of these IR display a very good match with a predicted patch (see SC_1in Fig. 4.5a and SC_{NIP}in Fig. 4.5c). It may also happen that a predicted patch covers several IRs, as illustrated on Fig. 4.5a and 4.5c, where the patches predicted by SC_3 and SC_1 , respectively, extend over 2 IR (in dark grey and black).

Fig. 4.5a shows a SC_3 prediction of an IR extending over two IS. While this prediction is correct in the sense that it covers a known interacting surface, it lacks precision when considering each one of the two sites individually. The same observation is illustrated in Fig. 4.5c where SC_1 covers two experimental IS. In some of these ambiguous cases, it is possible to infer the existence of multiple interacting sites within a region by crossing the information gathered from predictions coming from different scores. Indeed, the presence of SC_1 in Fig. 4.5a (middle), shows us that an experimental interacting surface is present at this location. Coupling this information with the SC_3 prediction (Fig. 4.5a, top) could be an indicator of the existence of two IS within the IR predicted with SC_3.

More generally, we looked into the process that leads $dynJET^2$ to identify IR and explicitly considered the seeds that $dynJET^2$ extends to propose a prediction. These seeds correspond to the support, that is the central layer, of the Levy geometrical model of protein interfaces [29, 57] and we want to use them to test whether they are good indicators of IS. To evaluate the number of seeds lying in experimental IRs, we merged the seeds of SC_1, SC_2, SC_3, SC_{NIP} predictions that were in contact and in Fig. 4.6 we report the number of resulting seeds for predicted IR and experimental ones. We observe that SC_3 and SC_{NIP} generate predictions containing one or two seeds at most indicating that they tend to identify binary interactions. SC_1 and

SC_2 show roughly the same counting on IR with one or two seeds but also on IR with three or four seeds, as displayed by experimental interfaces of the dataset `PPI-262`, suggesting that SC_1 and SC_2might be good indicators for determining the presence of multiple interactions in a predicted IR.

We can also observe that a non negligible number of IR in the dataset $\texttt{PPI-262}_{\texttt{ext}}$ is associated to the existence of 3 or 4 seeds. This seems to suggest that only a combination of scores can identify these IR and that the characteristics of the seeds can be different in the same IR.

4.3.7 Number of interacting partners

For each protein, we retrieved from the PDB all the homologs and their partners, and identified the associated experimental IRs. Then, we compared the number of partners targeting an IR to the number of seeds predicted by dynJET^2 in that IR. We wanted to test the hypothesis that different seeds might be associated to different partners. Fig. 4.6b shows that the number of seeds can indeed be used as an indicator of the number of partners a protein has.

We noticed that IR with a few partners are sometimes difficult to predict; we find 38 experimental IR for which no seeds were predicted, although 17 of them are at least partially covered by dynJET^2 predictions, including a seed but also its extensions (see Methods).

While predictions of one seed in the experimental IR indicate a small amount of partners on average, we observe that this assumption becomes less and less sharp while the number of seeds increases. A precise estimation of the number of partners cannot be correctly realised for two main reasons: first, the finite size of an IR can only admit a limited number of seeds within it, and second, the intrinsic nature of the protein might render impossible the estimation. For instance, we could retrieve up to 405 different partners for the antibody chain 3C08_L. This protein chain has many homologs (1273), and one can expect its homologs to be other antibodies targeting different proteins. A precise counting of the variability is impossible.

Hence, in order to improve the evaluation of the number of seeds in experimental IRs, we merged overlapping seeds (of at least one residue) from SC_1, SC_2, SC_3 and SC_{NIP}. We observe a sharp signal where having two merged seeds or more correlates with a high number of partners (Fig. 4.6b).

In conclusion, although the number of seeds does not strictly correlate with the number of partners, we observe that it can be used as an indicator for a protein to have a high or a low number of partners. In particular, interfaces for which no seeds are detected consistently display a low number of partners.

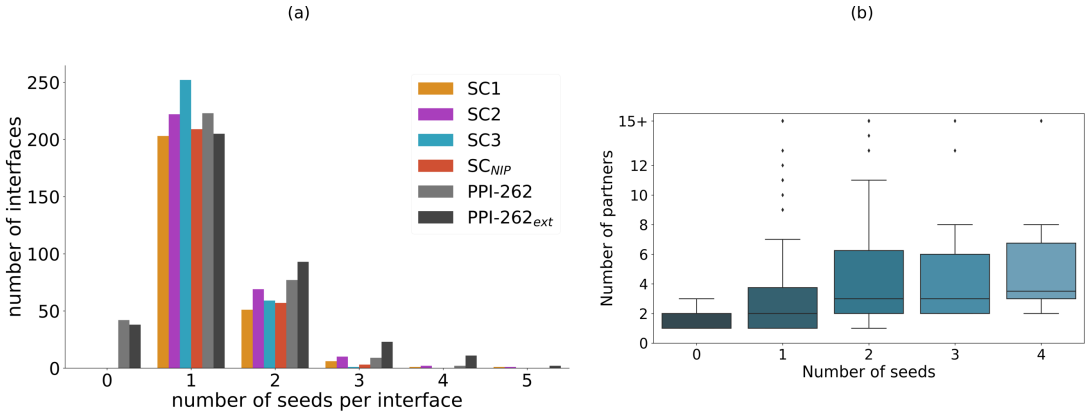


Figure 4.6: **Comparing number of partners versus number of seeds.** (a) Number of seeds corresponding to dynJET^2 predictions based on different scores (SC_1 in yellow, SC_2 in violet, SC_3 in cyan and SC_{NIP} in red) and experimental interfaces in the datasets `PPI-262` (grey) and `PPI-262`$_{\texttt{ext}}$ (black). (b) Number of partners for each experimental IR in `PPI-262`$_{\texttt{ext}}$, and number of seeds predicted by dynJET^2 in the IR.

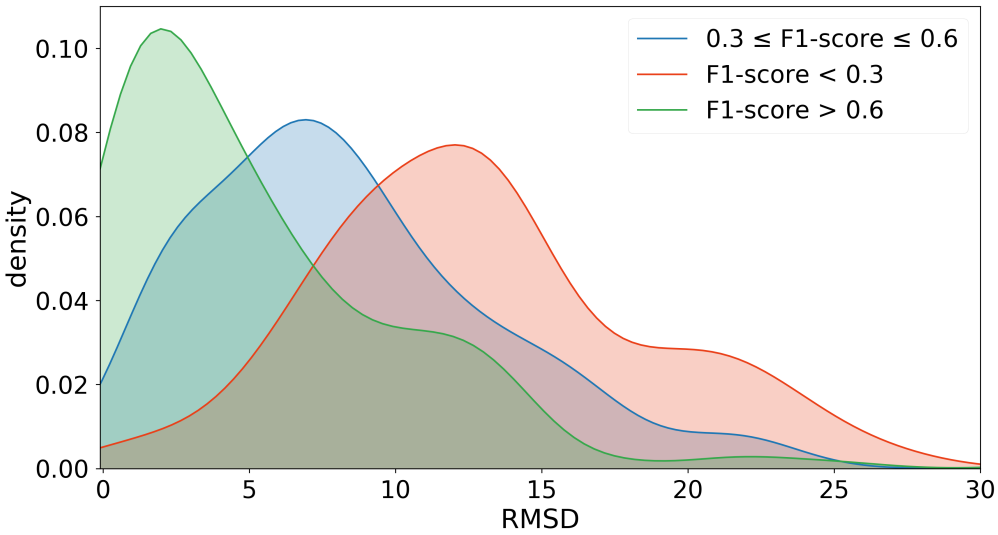


Figure 4.7: **Conformational deviations computed on IR between query structures and homologs' structures.** Distribution of the RMSD for 370 experimental IR from `PPI-262`$_{\texttt{ext}}$ computed between each of the 262 protein structures from `P-262`and the structures of its homologs. The IR are split into three groups based on the $F1$-scores computed for the best-matching dynJET^2 predictions: F1-score > 0.6 in green (153 IRs), F1-score < 0.3 in orange (80) and $0.3 \leq$ F1-score ≤ 0.6 in blue (139). Note that the orange curve includes 22 IR which were completely missed by dynJET^2.

4.3.8 Influence of conformational changes

Docking calculations and dynJET^2 predictions were performed on the crystallographic structures from P-262, while the experimental interfaces from PPI-262$_{ext}$ were detected on a much larger set of structures displaying various degrees of conformational deviations. To assess the influence of such conformational changes on the quality of the predictions, we computed the Root Mean Square Deviation (RMSD) of the IR backbone atoms between each query structure from P-262 and the structures of its homologs (see Methods). We observe that the quality of the predictions deteriorates with increasing conformational deviations (Fig. 4.7). The average RMSD is of 4.7Å for well detected IR (F1-scores > 0.6), 8.3Å for IR detected with intermediate sensitivity and/or precision ($0.3 \leq$ F1-score ≤ 0.6), 12.7Å for poorly detected IR ($0 <$ F1-scores < 0.3) and 13.3Å for completely missed IR (F1-score $= 0$). Given that 8Å represents a substantial conformational rearrangement, this analysis also shows that dynJET^2 is able to detect binding interfaces even when they are deformed.

4.4 Perspectives

Protein surfaces are used in multiple ways by a protein. We have analysed a pool of proteins with different functions and showed that an interaction site for a partner might be shared with several other partners, in either a complete or partial way.

Protein binding site prediction has been realised with dynJET^2, a modified version of JET^2, taking into account three scoring schemes based on conservation, physico-chemical properties of residues at the interface and local geometry of the protein surface, together with a fourth scoring scheme based on docking propensity. We have shown that, in some cases, the fourth schema is complementary to the first three. Also, some IR could not be predicted by one single scoring scheme, but a combination of them was able to accurately describe the experimental interface.

By taking into account all known homologous proteins and their crystallographic complexes, we could provide the most accurate description of the interacting surface for our dataset of proteins. The percentage of the surface covered by known interactions is 48% on PPI-262$_{ext}$, compared to 29% on PPI-262. It is important to notice that experimental patches do not simply add supplementary interfaces, but they help to better identify interacting regions that adjust partners in alternative complexes. By merging together these alternative sites, we could synthesise over 370 patches, spread over different homologs, into a relatively small number of regions (1.4 per protein chain). As a consequence, in the evaluation of dynJET^2 predictions, we could appreciate that a large amount of predicted regions proved to be accurate with respect to experimental regions identified by homologs and describing real biological and functional interfaces.

We also tried to understand the reasons behind poor predictions by looking at

the amount of structural difference among experimental interfaces across homologs. This difference can be very important and it correlates with the difficulty of accurately predicting a binding site. Although dynJET^2 remains resistant to small rearrangements, its performance steadily decreases as we observe an increase in the conformational changes of a protein during complexification.

We showed how reproducing the support-core-rim model could help us predict the tendency of a protein to be partner specific or to bind to many partners. It seems plausible to refine the approach towards a more accurate count. With the help of future PPI data, it also seems achievable to associate functions to the partners binding on different surface areas, described by different seeds on a region.

One of the main remaining challenges would be to split the predicted interfaces into IR or possibly into IS. Being able to do so would allow us to infer the number of partner the considered protein might interact with, as well as describing how many functional regions it has.

Part II

Partners discrimination

Chapter 5

Partners discrimination

Contents

5.1 The question

Prediction of the interactions sites of the protein has long been a heavily studied subject, as well as identifying the correct native conformation for two proteins among a set of decoys. However, large scale studies trying to identify interacting partners through a CC-D experiment remains at the pioneer stage. Many difficulties lay ahead: although much progress has been made in this regard over the past decades, protein docking remains a resource intensive experiment and applying it to an all-to-all situation requires extensive computational resources. Then, identifying correct partners at the scale of several hundreds of proteins requires an incredible accuracy as the number of interacting partners only represents a fraction of the possible solutions. For instance, the PPDBv2 dataset contains 168 proteins and thus 28224 (168×168) possible protein pairs with only 168 correct interactions against 28056 incorrect interactions.

Due to these shortcomings, predicting protein-protein interactions at large scale using CC-D methods is still in its early days. To our knowledge, there has only been three studies (the first being ours [67], then [110, 92] and the latest published recently; [70]). However, [56] shows that geometrical docking alone does not carry sufficient information to distinguish cognate partners from non-interactors in an unbiased CC-D experiment. Although [70] uses machine learning methods (specifically a Random Forest classifier), it is interesting to see that the global pipeline proceeds in the same way as ours, combining scoring methods with binding site predictions to evaluate the conformations.

C. Dequeker, E. Laine, A. Carbone, "Protein partners discrimination reached with coarse-grain docking and binding sites predictions", in preparation, 2018.

5.2 Methods

5.2.1 Towards a better description of the PPDBv2 dataset

An early description of the dataset (version 2) released by [76] (see Section 3.4.1) only split it into four different subsets: Enzyme-Inhibitor (EI), Antibodies-Antigens (AA), Antibodies-Bound Antigens (ABA), Others (OX). All complexes are in the unbound form (state which they adopt when they are not binding to any other partner), except for the ABA subset, which is in the bound form (the structure represents the conformational changes they may have undergone upon binding). This description, while it has been considered at the beginning of my work, has been updated (version 5, PPDBv5) in [108]. This dataset update provides new classifications separating the proteins into more refined functional classes as well as new protein structures to analyse. Although we did consider the functional classifica-

tion refinement of the PPDBv2 168 proteins, we did not take into account the new protein structures brought by the update; a complete cross docking was performed on the first 168 proteins, and we did not have the computational power to redo the experiment using the same docking software to the new ones.

Using the new protein classifications of PPDBv5 [108], we obtain the following number of proteins for each functional classes (see Fig. 5.1): 20 unbound Antibodies-Antigenes (AA), 24 Antibodies-Bound Antigens (ABA), 38 Enzymes-Inhibitors (EI), 6 Enzymes (with a regulatory or accessory chain) (ER), 12 Enzymes-Substrates (ES), 24 Others G-protein containing (OG), 14 Others Receptor containing (OR), 30 Others miscellaneous (OX).

A diverse protein-size dataset

We represent in Fig. 5.1 the different subsets thus defined. We can clearly see a large difference among the different functional classes in terms of variability. It is important to note that some subset inherently show a low variability due to their limited size. However, it is clear when comparing similarly sized subsets the differences observed; for instance, OX presents a much higher standard deviation (246) compared to OG (117). This surface size variability sheds some light on how different proteins are, and how difficult it might be to find a rule able to predict how they interact.

5.2.2 Interface residues

In this study we use the same definition of the experimental residues as the one described in Section 4.2.2 and we consider as well the same predictions from dynJET^2 (see Section 4.2.3).

Interface predictions and why combine them together

JET^2 provided three different scoring schemes which could be used to detect different type of interfaces. With dynJET^2's ability to include another score to the interface prediction, we added the NIP score at different stages of the clustering process. This is illustrated on Fig. 1.4. For each different stages of NIP inclusion tested, we derived all three scoring methods SC_1, SC_2 and SC_3 into SCX_{NIP-1}, SCX_{NIP-2}, SCX_{NIP-3} respectively.

Since a predicted patch does not always precisely match an experimental site (see Chapter 4, [57]) and in order to compare two proteins in terms of partner discrimination, we use the combination of the predicted patches for which we obtain the highest F1-score (see Methods 5.2.2). In this process, the set of predictions considered comes from a single group SCX of predictions. We therefore only consider

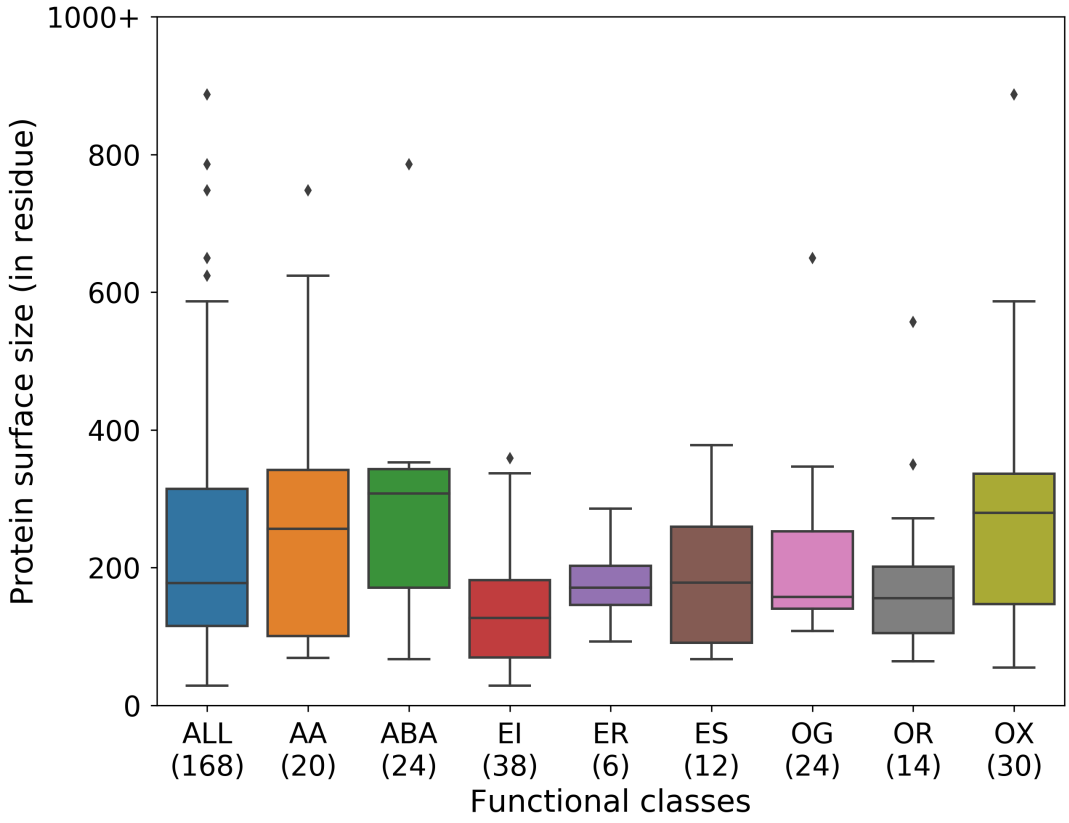


Figure 5.1: Boxplot representation of the surface size of the 168 proteins functional classes defined in [108].

predictions coming from either $SC4_{NIP-1-2-3}$, $SC5_{NIP-1-2-3}$, $SC6_{NIP-1-2-3}$ or SC1-2-3-NIP and refer to them as SC_{4*}, SC_{5*}, SC_{6*} and SC_{d*}respectively (see Fig. 1.4).

Shifts of the experimental interface

In order to generate the necessary data for Fig. 5.7 (see Section 5.4.4), we performed a gradual shift of the experimental interfaces. The percentage of the shifted surface is rounded to get the amount of residues to shift (number of iteration of the following process). Then for each residue to shift, we pick a bordering residue of the interface r_b (being part of the interface, and in contact with a surface residue not being part of the interface), then take at random a neighbour r_n of the farthest interface residue from r_b. We then consider r_n as being part of the interface and r_b not being part of the interface anymore. The residue r_b cannot be picked again to be part of the interface in the following iteration.

5.2.3 Detection of interacting partners

Interactions evaluation

We consider as interacting partners (True Positive) the proteins known to form a complex, and as non-interacting partners (True Negative) the other proteins. This definition of true positive and true negative is used to compute the AUC in order to evaluate our interaction predictions.

In order to score the likelihood for a protein pair to interact, the laboratory team developed in earlier studies [95, 67] an Interaction Index (II, see Section 3.3). We modified this interaction index which now takes into account a reference interface (predicted or experimentally known, see Section 4.2.3) coupled with the docking interface computed using INTBuilder (see Chapter 6, [25]), a docking energy (from iATTRACT, PISA or MAXDo) and a pair potential score (CIPS [79]). To combine our reference interface with the docking interface, we compute the Fraction of Interface Residues (FIR) of the docking interface contained in the reference (experimental or predicted) interface. This gives us for each conformation a FIR value ranging from 0 to 1 for both proteins of the pair. We describe the II_{P_1,P_2} as:

$$II_{P_1,P_2} = FIR_{P_1} \times FIR_{P_2} \times E_{P_1,P_2} \times PP_{P_1,P_2} \tag{5.1}$$

where FIR_{P_1} and FIR_{P_2} are the FIR assigned to the proteins P_1 and P_2 respectively for each conformation, E_{P_1,P_2} the energy computed using an energy function (MAXDo, iATTRACT, PISA) and PP_{P_1,P_2} a pair potential score (CIPS) assigned to the conformation.

5.2.4 Functional classes specific scores

To compute the Interaction Index for all the different functional classes, we tried to stick to one single method to avoid over fitting the results with too many different computation ways. Thus, we define the following default parameters that were applied for all functional classes, save for EI, ER and OR (described next): all residue-residue is considered a contact from 5Å distance threshold. The default combination of interface was SC_{6*}. The MAXDo energy function was used, and to compute the II we multiplied by the CIPS pair potential.

For three functional classes we modified some of these default values. In EI and ER, we use the PISA and iATTRACT energy functions without CIPS respectively. In OR, we use the MAXDo energy function alone without multiplying it with the CIPS.

5.3 Background

There is an increasing demand, in pharmacology for instance, to be able to target specific proteins among many [40].

Early studies such as [95] performed small scale CC-D in order to answer the partner prediction question. This attempt, along other studies [51, 52], shows that energy alone is not sufficient in predicting interacting partners. However, [95] shows that it is possible to predict partners with a high precision by combining a well defined interface (in this case, the experimental interface) with the given docking energy associated to the conformation. In this study, we follow the first steps made by [67] which analysed a large scale CC-D of 84 protein complexes with the interface prediction software JET [29]. Since, several improvements were made to the pipeline, including the development of JET^2 [57], then more recently dynJET^2 (see Section 4.2.3) as well as the integration of new developed scores.

Proteins bind to each other through a number of properties; conservation, physico-chemical properties of residues, geometry of the protein, phosphorylation [105, 59, 11, 47, 36, 17, 81, 84, 29, 57, 30] being of the most important ones. We show here how the dynJET^2 prediction software is able to tackle the complexity of predicting the multitude of different protein interfaces through its different scoring methods, and how it is able to help in the identification of interacting partners.

5.4 Scores used and their impact on partner identification

The PPDBv2 dataset (see Methods 3.4.1, Section 5.2.1) forms 84 binary complexes known to interact, and we strive to discriminate interacting partners from non in-

teracting ones.

The CC-D experiment was realised on the full dataset on unbound structures, leading to 28224 docking Simulations. For each couple of proteins, about 300 000 ligand-receptor orientations were explored (which we refer to as conformations) corresponding to ligand and receptor complete surfaces; this experiment required more than 7 months of computation time on the WCG in 2007, as mentioned in Sec. 2.4.1. The docking algorithm simulates the actual docking process in which ligand-receptor pairwise interaction energies are calculated. In this study we used several different energy functions to evaluate the docking conformations, as mentioned in Methods 3.2.

II and NII computation

We now consider for the present analysis four main components in the II computation: The predicted interface, the docking interface (computed for each docking conformation), the energy score computed with an energy function (MAXDo [95], iATTRACT [98] or PISA [54]) and the presence or absence of a pair potential scoring (CIPS [79]). The II formula which is now used for the study is described at Eq. 5.1.

The pair potential scoring evaluates the likelihood of the observed residue-residue interactions. The whole process that we now use is described in the Fig. 1.5 pipeline. We show there how from a set of receptors and ligands, the laboratory team performed a CC-D experiment to obtain an ensemble of conformations. Then, we present how we use the known/predicted interface with the docking interface to obtain the FIR and combine it with the docking energy (from MAXDo, iATTRACT or PISA) along with the pair potential CIPS. We show in Fig. 5.6a and Fig. 5.6c the obtained II matrices obtained using experimental interface to compute the FIR (Fig. 5.6a) and our binding site predictions (Fig. 5.6c).

In a previous study [56], we showed the importance of taking into account the proteins' behaviour among the dataset to really be able to interpret their II. Particularly, it has been shown that proteins may adopt a sticky behaviour (*i.e.*, consistently producing high II) while others may show a reluctance to bind to other partners (globally low II). Thus, as in [67] and further explained in *Methods* in equation 3.6 and 3.7, we perform a normalisation step on the II matrix. This normalisation step is crucial to take into account the behaviour of a protein among the studied dataset. We further show in Fig.5.6 the impact such normalisation can have on the noise reduction using the experimental and the predicted interfaces respectively; the transition from Fig. 5.6a to Fig. 5.6b shows how applying the normalisation reduces the noise and make the diagonal (representing interacting partners) come out. In more quantitative terms, the normalisation improves the discrimination AUC from 0.74 to 0.82. The transition from Fig. 5.6c to Fig. 5.6d (using predicted interfaces)

describes an increase in the AUC value from 0.33 to 0.67.

5.4.1 Predicting the interacting partners

As mentioned above, we rely on four main parameters. The predicted surface is determined through the scoring schemes used by our prediction algorithm dynJET^2. The predictions considered are the best combination from either SC_{5*}, SC_{6*}, SC_{4*} or SC_{d*} (See Fig. 1.4, Methods) this process is further explained in Section 5.4.3. The docking interface is computed using a distance threshold with the INTBuilder software [25]. This distance threshold can therefore be tweaked to vary the docking interfaces size. The energy function is either one of iATTRACT, PISA or MAXDo. The conformations were obtained using the MAXDo docking software. The scoring performed by PISA or MAXDo relies on the conformations computed using MAXDo. In contrast, the scoring done by iATTRACT involves a minimisation (using its own provided tool) of the conformations generated by MAXDo. As we further show, our different functional classes respond differently to these parameters and we rigorously compared their effect separately to determine if they should or not be included in our partner discrimination pipeline.

In order to avoid the risk of over fitting, we strove to define a single default method that would match most functional classes, and considered altering a parameter for a class if it consistently brought improvement to our partner discrimination capacity using the dynJET^2 predictions. For this, we computed the resulting AUC of every possible combination of the parameters, for each functional class. We choose as default parameters the parameters providing the globally best partners discrimination capacity.

We ranked every possible combination of parameters according to their average AUC values (see Tables 5.5, 5.9, 5.6, 5.7, 5.8). We then define as default parameters values those for which we globally obtained the best results. To decide for each class if one parameter value should be used instead of another, we ranked the ten best parameters combinations by their outcome AUC and plotted them in Fig. 5.2, 5.4, 5.5. For each barplot, we present the 10 best results of each class and divide them using the studied parameter for its possible values. To decide if one parameter value should be specifically used for a functional class, we perform Mann Whitney U-test of two distributions: the first regroups every AUC values (for all parameters combinations) for this functional class with the default parameter value while the second distribution fixes the considered value. Under a `p-value` of 0.05, we consider the studied parameter value to significantly improve our discrimination potency and decide to use its value for the given class. Below, we present our observations following this pipeline for each of the four parameters:

Distance threshold is represented in Fig. 5.2. We show that the threshold dis-

tance impacts very little on the different functional classes, except the 6.0Å threshold which is deleterious for AA and ABA functional classes. We therefore choose the 5Å distance threshold and use it next to compute the following AUC tests.

Predictions show that the SC_{6*} (using solely NIP to detect the seeds of the interface) method consistently provides equally or better results than other predictions methods for all functional classes, except ER. We therefore performed a Mann Whitney U-test to compare the AUC distributions for ER using on one set the SC_{5*} predictions and the other the SC_{6*} predictions. We obtained a `p-value` of 0.24 and therefore decided to keep the SC_{6*} prediction method for all functional classes.

Energy function is represented in Fig. 5.5a and Fig. 5.5b. We observe that for all functional classes except EI and ER, MAXDo performs equally well or better than iATTRACT or PISA in the detection of interacting partners in a CC-D. Thus, we performed a Mann Whitney U-test for EI and ER fixing the two MAXDo and PISA parameters. We thus obtain two `p-value` of 2.55×10^{-6} and0.21 respectively. We performed a third test for ER between the MAXDo and iATTRACT distributions and obtained a `p-value` of 3.12×10^{-6}. We therefore decided to choose the PISA energy function for EI, and the iATTRACT one for ER (as the `p-value` with PISA wasn't sufficient to declare it different from MAXDo).

Pair Potential is represented in Fig. 5.3. We separated this plot among the different energy functions. We observe that using the pair potential using the iATTRACT or PISA energy functions degrades our discrimination capacity. However, we observe that using the MAXDo function on pair with the CIPS pair potential provided equally or better results for all functional classes except OR. Thus, similarly as before, we performed a test for OR fixing the MAXDo with CIPS pair potential for one set and with the MAXDo energy function without CIPS for the other and obtained a `p-value` of 0.01. We therefore considered this distribution as different and did not use the CIPS pair potential for the OR functional class.

5.4.2 Difference between predictions and experimental results

Overall, we present in Fig. 1.6b the AUC obtained using this method with experimental interfaces along with our predictions SC_{6*}. The barplot also shows the previously attained results to show the improvements made [67]. It is interesting to

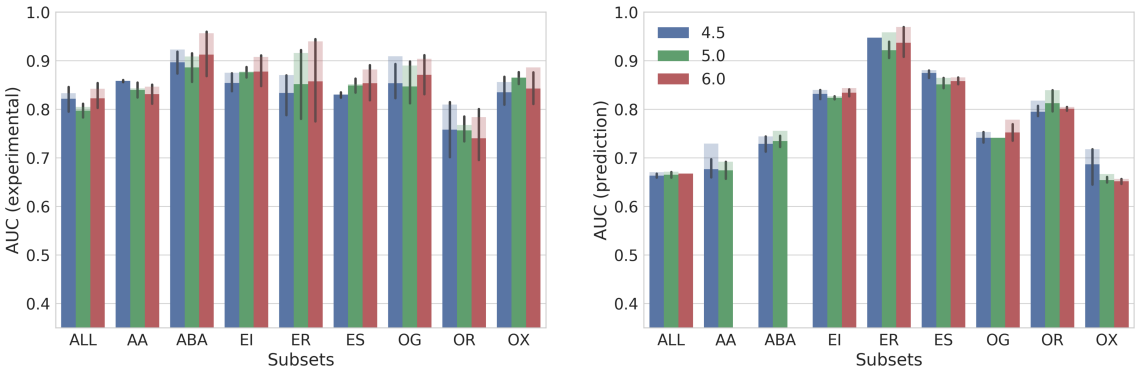


Figure 5.2: Barplot representation of the AUC values when separating by threshold (in Å). For each subset, its top 10 methods were considered and we then separated them according to the distance threshold used to compute the docking interfaces. The opaque bar represents the average of the AUC values and the transparent one represents the maximum value achieved among the different methods. If one of the parameters is not selected in the 10 best combinations, it is possible to not appear on the plot (which is the case for the threshold of 6.0Å for AA and ABA in this plot).

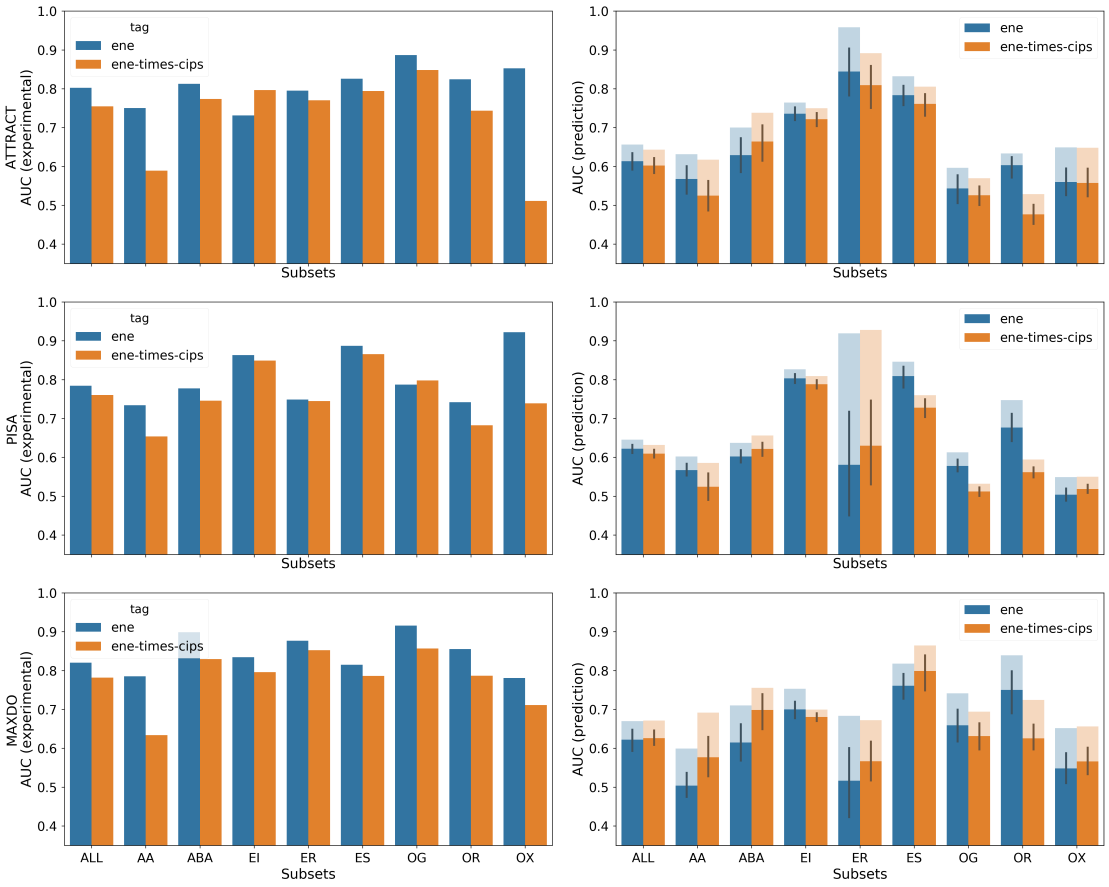


Figure 5.3: Barplot representation of the AUC values when separating by energy, and adding or substituting the pair potential (CIPS). For each subset, its top 10 methods were considered and we then separated them according to the presence or absence of the CIPS pair potential. The opaque bar represents the average of the AUC values and the transparent one represents the maximum value achieved among the different methods. We bring the attention on the fact that this plot only show a single possible combination of parameters for the experimental interfaces (right), thus explaining why no transparent bar are shown.

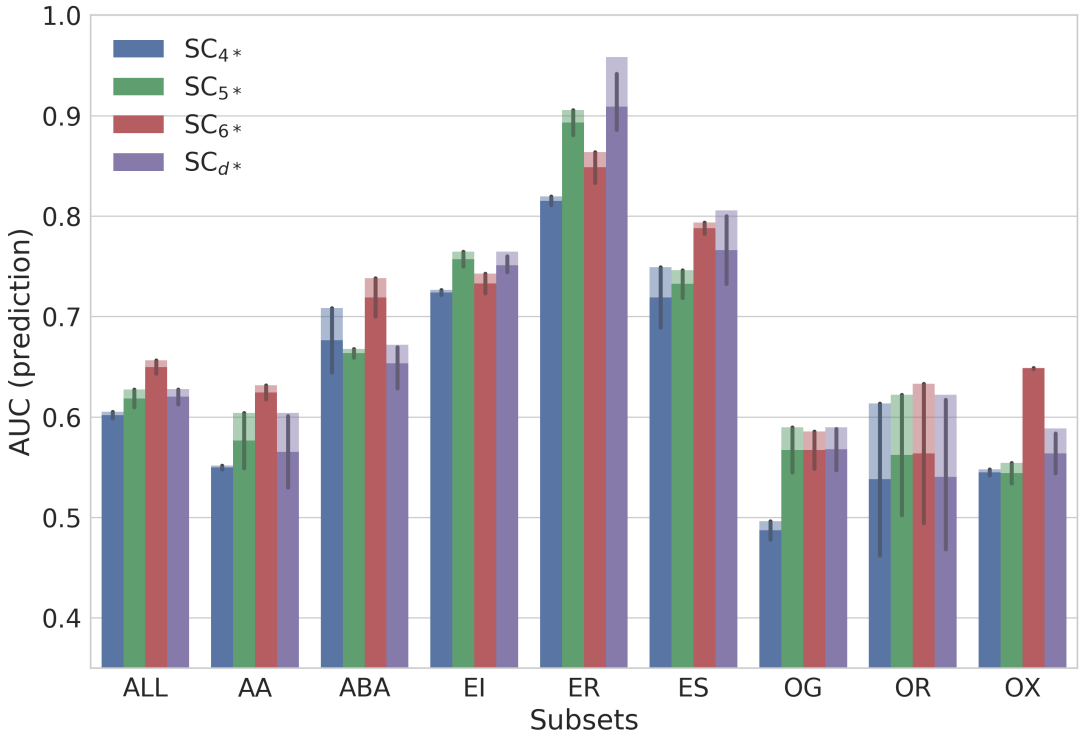


Figure 5.4: Barplot representation of the AUC values when separating by predictions. For each subset, its top 10 methods were considered and we then separated them according to the prediction used. The opaque bar represents the average of the AUC values and the transparent one represents the maximum value achieved among the different methods.

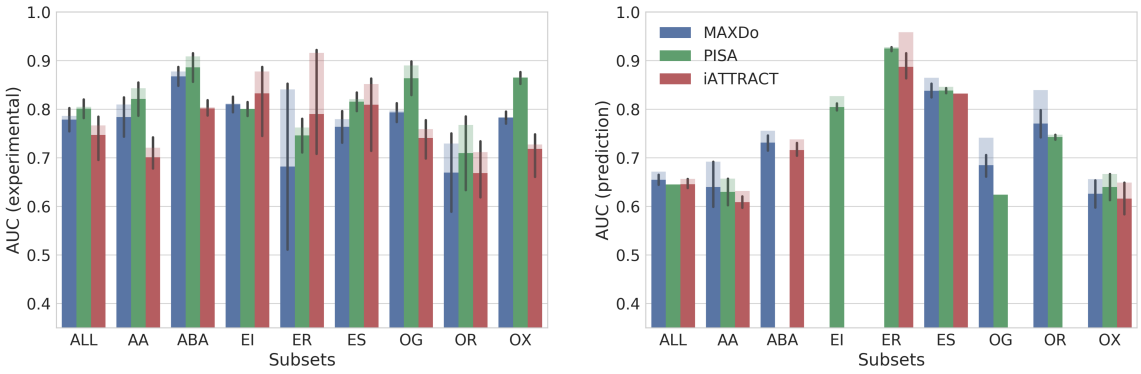


Figure 5.5: Barplot representation of the AUC values when separating by energy functions. For each subset, its top 10 methods were considered and we then separated them according to the energy function used. The opaque bar represents the average of the AUC values and the transparent one represents the maximum value achieved among the different methods.

look at the AA and ABA groups, which reflects how adding 3D information about the protein improved the results. We observe as well that adding the CIPS pair potential score to the experimental values decrease their partner discrimination efficiency. CIPS is a high throughput software meant to swiftly reduce the search space of possible native conformations with a high precision. CIPS is especially helpful when using the predictions as it restricts a large number of possible conformations from being considered. Coupling it with other features let us balance their individual advantages and shortcomings. However, I think that while the CIPS filters many wrong conformations, it sometimes underrates a near native conformation. Although this would not affects much the predictions as its filtering of wrong conformations is more effective than the few errors it makes removing near native conformations; however, experimental interfaces do not need any external guidance to evaluate the "right" conformation. Thus, CIPS' effectiveness in removing the wrong conformations would be redundant and the few errors it would make impact more the discrimination potency. With more time, I would further validate this hypothesis by analysing the conformations. In Fig. 1.6a, we show how the best method for the experimental interfaces improves over the previously obtained AUCs.

5.4.3 Predictions using dynJET^2

Previously in Chapter 4, we showed how multiple interactions regions exist at the proteins' surfaces and how dynJET^2 predictions, if matched against a single of these regions, could at first present many false positive. In this dataset, considering the many predictions made by dynJET^2, we must find a way to evaluate specific predictions: those that would best match the experimental IS. Therefore to set ourselves in the context where we are analysing specific IS, we consider for each IS the best matching combination of our predictions (according to the F1-score; see Sec. 5.2.2). Looking at Fig. 5.5, we note that while the same trends are globally maintained between experimental results and predicted ones, some scoring methods are far more forgiving than others of the lower accuracy of the interfaces. For instance, The difference observed for the ER subset shows that the iATTRACT scoring scheme alone is able to compensate the dynJET^2 predictions (which performs poorly in terms of F1-score, see Table 5.1). We also show that we are able to better predict interacting partners when considering smaller classes. We even perform as well using the dynJET^2 predictions as when using the experimental interface (for example ER: AUC of 0.81 using predictions against 0.79 using experimental values).

Interestingly, the combination for the predictions working the best is SC_{6*}, thus relying solely on the NIP to detect the seeds, which the dynJET^2 software will next expand. It is interesting as well to see that the combinations SC_{d*}, SC_{4*} or SC_{5*} do not always bring similar results. Looking more closely at Table 5.1, we note that the functional classes for which SC_{6*} performs substantially better are AA and

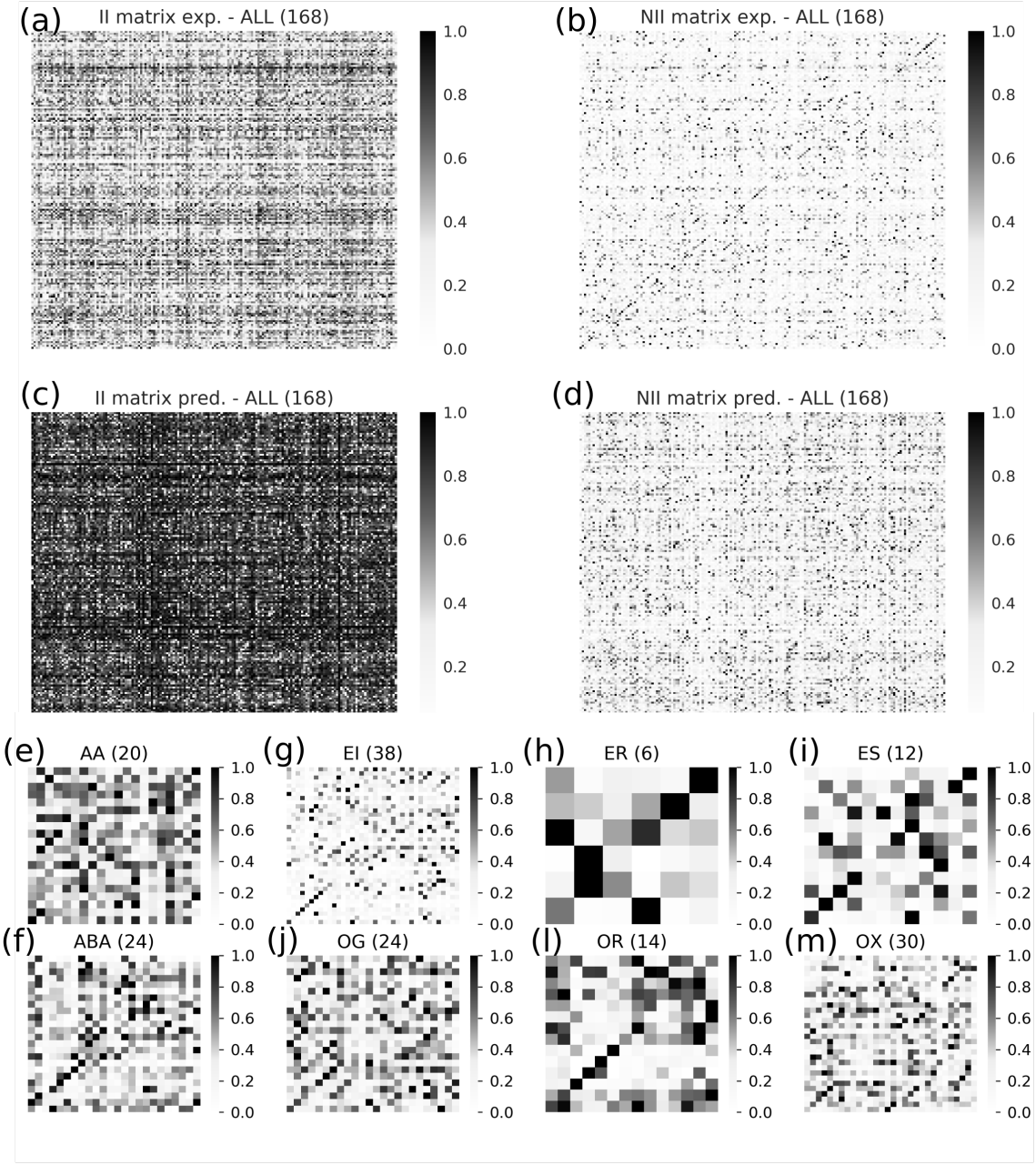


Figure 5.6: Partner prediction matrices using the experimental and predicted interfaces. The pipeline used to compute the II is the one described in Section 5.4.1 for both experimental and predicted matrices. For every line of the matrix, we represent a protein when it was considered as a receptor during the CC-D experiment and for every column, we represent a protein when it was considered as a ligand during the CC-D experiment. We ordered the matrix by putting on the diagonal the complexes known to interact. The parameters described in Methods for each dataset were used for the subset matrices. The ALL matrix was computed using only the default parameters. We present here the different matrices **(a)** the II experimental matrix. **(b)** the NII experimental matrix **(c)** the II predicted matrix (using default parameters, see Methods) **(d)** the NII predicted matrix (with default parameters) **(e)** the NII AA predicted matrix **(f)** the NII ABA predicted matrix **(g)** the NII EI predicted matrix **(h)** the NII ER predicted matrix **(i)** the NII ES predicted matrix **(j)** the NII OG predicted matrix **(k)** the NII OR predicted matrix **(k)** the NII OX predicted matrix.

Table 5.1: Table representing the average **F1-values** obtained for each functional classes using the best combination according to the experimental interface, for each set of predictions. The best combination is made as in Methods 5.2.2.

Predictions	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_{4*}	0.48	0.39	0.48	0.58	0.43	0.51	0.51	0.40	0.40
SC_{5*}	0.48	0.39	0.47	0.58	0.43	0.50	0.49	0.42	0.42
SC_{6*}	0.49	0.45	0.52	0.58	0.43	0.49	0.49	0.41	0.42
SC_{d*}	0.46	0.40	0.46	0.57	0.37	0.49	0.47	0.40	0.40

Table 5.2: Table representing the average **recall** values obtained for each functional classes using the best combination according to the experimental interface, for each set of predictions. The best combination is made as in Methods 5.2.2.

Predictions	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_{d*}	0.64	0.63	0.65	0.73	0.68	0.70	0.64	0.55	0.51
SC_5	0.63	0.60	0.66	0.72	0.64	0.69	0.63	0.58	0.51
SC_6	0.62	0.59	0.61	0.72	0.60	0.70	0.65	0.56	0.51
SC_4	0.62	0.62	0.64	0.73	0.56	0.68	0.61	0.57	0.50

Table 5.3: Table representing the average **PPV** values obtained for each functional classes using the best combination according to the experimental interface, for each set of predictions. The best combination is made as in Methods 5.2.2.

Predictions	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_6	0.43	0.39	0.47	0.51	0.33	0.40	0.41	0.35	0.39
SC_4	0.42	0.32	0.41	0.52	0.37	0.43	0.46	0.34	0.40
SC_5	0.41	0.31	0.40	0.51	0.33	0.41	0.43	0.36	0.40
SC_{d*}	0.39	0.33	0.38	0.48	0.26	0.40	0.40	0.33	0.38

Table 5.4: Table representing the average **accuracy** values obtained for each functional classes using the best combination according to the experimental interface, for each set of predictions. The best combination is made as in Methods 5.2.2.

Predictions	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_4	0.85	0.81	0.89	0.82	0.85	0.87	0.88	0.82	0.85
SC_6	0.85	0.82	0.90	0.83	0.83	0.85	0.86	0.82	0.86
SC_5	0.85	0.81	0.88	0.83	0.82	0.86	0.87	0.82	0.86
SC_{d*}	0.83	0.81	0.87	0.81	0.76	0.85	0.86	0.81	0.84

ABA; the F1-score value of only the antibodies predictions using SC_{6*} present is 0.13 greater than the second best performing prediction (SC_{5*}) with a value of 0.66 over 0.53 for SC_{6*} and SC_{5*} respectively for the antibodies. We also know that the SC_{NIP} alone is not sufficient to fully predict the interfaces. From this statement, and knowing that the SC_{6*} scoring method is performing well, we can assume that for most interaction site, the NIP value is able to pick up the centre of it, and that making use of the intrinsic properties such as evolutionary trace or physico-chemical properties are crucial to fully define the protein-protein interfaces.

5.4.4 Interface sensitivity

To assess how sensitive the interfaces were, and how was the AUC impacted by small and larger variation of the interface, we shifted different amounts of the experimental interfaces using the process described in *Methods* 5.2.2. For each shift we ran a prediction experiment using the shifted interfaces, which result we then reported on Fig. 5.7 along their F1-score compared to the non shifted experimental interface. We shifted for 10 different percentage of the surface (by step of 10%), and for each different percentage we ran the partner discrimination experiment 10 times to ensure consistent results.

We note that the functional classes with the fewest proteins also present the most varying results, which is an expected outcome. We highlight how some subsets react very differently than others. A striking drop in the AUC value is observed for the functional classes AA, ABA, OG, OR, OX starting from the very first shift (only 10% of the interface being shifted). The biggest drop occurs for the OX group, which the proteins could not be placed in any of the other functional classes and which is therefore also the most difficult group for us to predict since we cannot rely on any specific measure. Conversely, the EI group does not show any difference in partner prediction performance for the first shift, and is also the group for which the best results could be easily achieved (as early as in the previous study [67]). These plots also show that our predictions seem to fit in the same range of here achieved AUC, maybe indicating the limit of our partner prediction method considering the quality of our predictions. They also state that the limiting factor now to better predict interacting partners are our interface predictions. A third point of the Fig. 5.7 shows that all three enzyme classes show very good resistance to light modification of the experimental interface, which is not the case of other subsets (AA, ABA, OG, OR, OX).

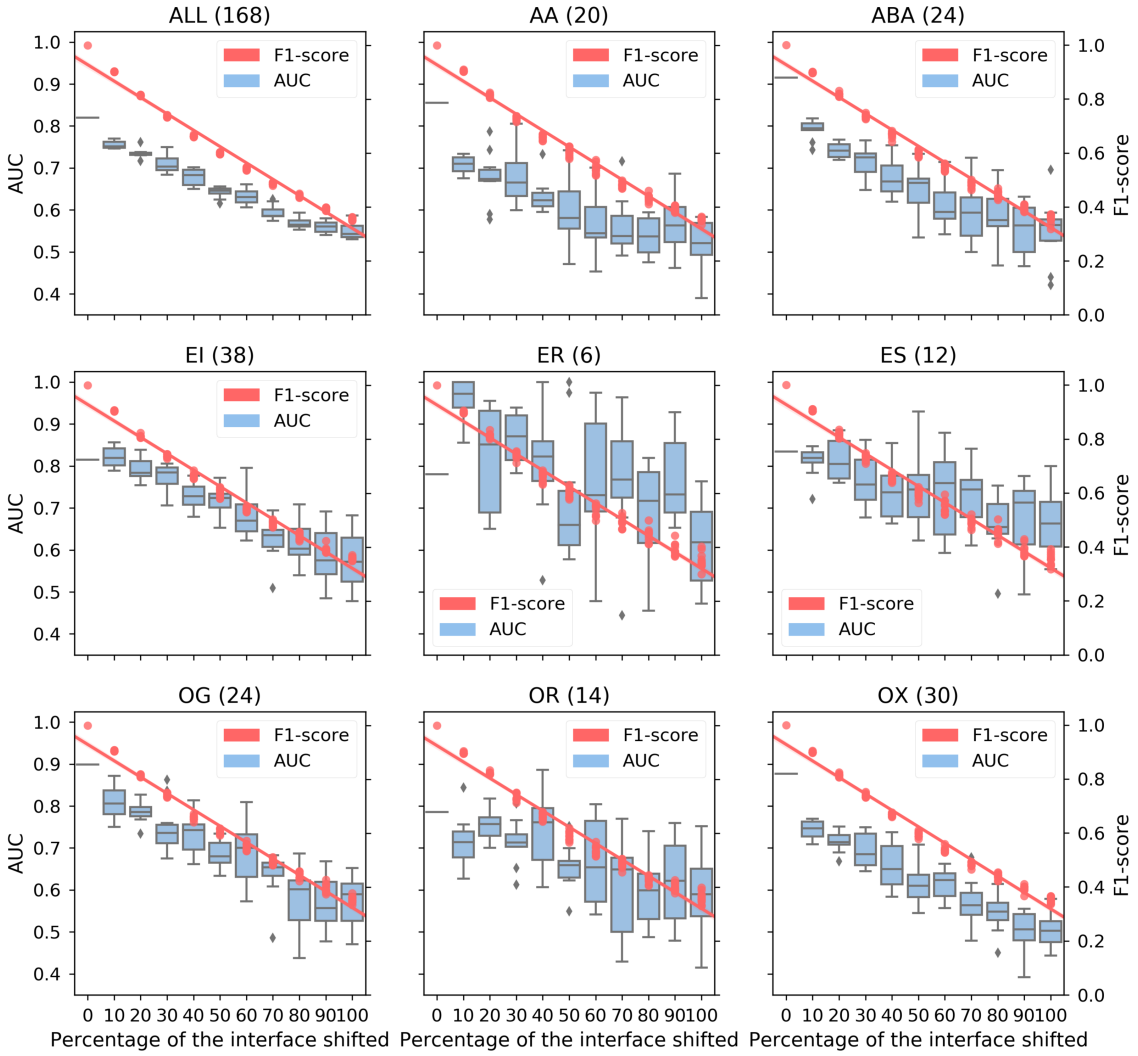


Figure 5.7: Shifts of the experimental interfaces for every functional classes and for the main dataset. The AUCs were computed using the energy from MAXDo only, with the experimental interfaces and using docking interfaces computed with a threshold of 5Å. The F1-score values correspond to the F1-score of the shifted experimental interfaces compared to the non-shifted ones.

5.5 Perspectives

In this study, we have shown how our predictions are precise enough to detect interacting partners in a large scale study, sometimes reaching the limits set by the experimental interface. We have also put in evidence how some functional classes are tackled more efficiently by different energy function, suggesting that this could help identify these specific features better captured by PISA and iATTRACT. This study opens up the possibility on running full proteome analysis, thus building an interaction network of many proteins in a cell, or involved in a same functional pathway. We also have shed some light on how important separating proteins into different functional classes, thus requiring the development of methods to automatically analyse and sort their functions.

The study also calls for new methods to refine the different predictions of dynJET^2 in separated interacting regions. Here, we relied on the knowledge of the experimental site we where looking for to locate the prediction region of interest to us. In [103], they show that as much as 75% of the protein surface might be experimentally active, while single interacting sites as studied here only represent about 25% of the surface. This means that to be able to fully unshackle ourselves from the experimental knowledge to predict interacting partners and define interaction sites, we should be able to separate the prediction patches into separate ones. A new interaction matrix could then be computed not based on each protein, but with each line describing a interacting region, targeting potentially a different set of proteins.

Table 5.5: Table representing the AUC values obtained for each functional classes using the experimental interfaces, for each combination of parameters possible. Lines were sortes according to the average of AUC computed over each functional class, and weighed according to their number of proteins. The red value refers to the AUC obtained for the matrix in Fig. 5.6b.

Interface	Energy	CIPS	Distance Threshold	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
Exp. Interface	MAXDo	No	6.0	**0.85**	0.84	**0.96**	0.84	0.73	**0.89**	0.91	0.80	**0.90**
Exp. Interface	MAXDo	No	5.0	0.82	**0.86**	0.92	0.81	0.78	0.83	0.90	0.79	0.88
Exp. Interface	MAXDo	No	4.5	**0.85**	**0.86**	0.91	0.83	0.84	0.82	**0.92**	**0.82**	0.87
Exp. Interface	MAXDo	Yes	6.0	0.80	0.81	0.85	0.85	0.53	0.80	0.87	0.64	0.88
Exp. Interface	iATTRACT	No	4.5	0.82	**0.86**	0.93	0.84	0.88	0.76	0.85	0.81	0.81
Exp. Interface	MAXDo	Yes	4.5	0.80	**0.86**	0.86	0.78	0.71	0.83	0.82	0.69	0.83
Exp. Interface	iATTRACT	No	6.0	0.81	**0.86**	0.91	0.83	0.85	0.78	0.83	0.72	0.80
Exp. Interface	iATTRACT	Yes	4.5	0.78	**0.86**	0.89	0.79	0.65	0.81	0.83	0.69	0.79
Exp. Interface	MAXDo	Yes	5.0	0.78	0.79	0.86	0.79	0.71	0.80	0.83	0.63	0.85
Exp. Interface	iATTRACT	No	5.0	0.80	0.82	0.89	0.83	0.85	0.73	0.81	0.75	0.80
Exp. Interface	iATTRACT	Yes	6.0	0.77	0.81	0.87	0.83	0.64	0.71	0.81	0.65	0.82
Exp. Interface	iATTRACT	Yes	5.0	0.75	0.74	0.85	0.79	0.51	0.80	0.77	0.59	0.77
Exp. Interface	PISA	No	6.0	0.77	0.74	0.79	0.90	**0.94**	0.85	0.78	0.70	0.77
Exp. Interface	PISA	Yes	6.0	0.77	0.73	0.76	**0.91**	0.78	0.82	0.75	0.69	0.82
Exp. Interface	PISA	No	5.0	0.78	0.74	0.79	0.89	0.92	0.86	0.78	0.73	0.75
Exp. Interface	PISA	No	4.5	0.77	0.74	0.79	0.88	0.84	0.84	0.77	0.71	0.73
Exp. Interface	PISA	Yes	4.5	0.76	0.75	0.78	0.86	0.55	0.83	0.74	0.68	0.74
Exp. Interface	PISA	Yes	5.0	0.76	0.68	0.80	0.87	0.74	0.85	0.75	0.65	0.74

Table 5.6: Table representing the AUC values obtained for each functional classes using the best combination of SC_{4*} according to the experimental interface. The best combination is made as in Section 5.2.2. The AUC values are represented for each combination of parameters possible. Lines were sortes according to the average of AUC computed over each functional class, and weighed according to their number of proteins.

Interface	Energy	CIPS	Distance Threshold	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_{4*}	MAXDo	Yes	4.5	0.63	0.58	**0.71**	0.69	0.60	0.76	0.68	0.72	0.52
SC_{4*}	MAXDo	Yes	5.0	0.61	0.59	0.74	0.66	0.67	0.68	0.61	0.62	0.52
SC_{4*}	iATTRACT	Yes	4.5	0.62	**0.73**	0.67	0.74	0.64	0.73	0.52	0.56	0.52
SC_{4*}	MAXDo	No	6.0	0.63	0.49	0.64	0.68	0.61	**0.79**	**0.70**	0.78	0.52
SC_{4*}	iATTRACT	No	4.5	**0.64**	0.67	0.67	0.78	0.87	**0.79**	0.51	0.63	0.52
SC_{4*}	MAXDo	No	5.0	0.63	0.48	0.65	0.67	0.68	0.74	**0.70**	**0.84**	0.53
SC_{4*}	MAXDo	Yes	6.0	0.62	0.48	0.64	0.73	0.60	0.67	0.69	0.59	**0.56**
SC_{4*}	MAXDo	No	4.5	0.63	0.50	0.64	0.72	0.65	0.76	0.68	0.82	0.53
SC_{4*}	iATTRACT	No	6.0	0.59	0.57	0.60	0.69	0.72	0.75	0.58	0.57	0.55
SC_{4*}	PISA	No	4.5	0.63	0.57	0.62	**0.84**	0.78	0.76	0.56	0.70	0.50
SC_{4*}	iATTRACT	Yes	5.0	0.60	0.55	**0.71**	0.72	0.82	0.69	0.48	0.46	0.54
SC_{4*}	PISA	No	6.0	0.63	0.56	0.59	0.80	**0.97**	0.76	0.62	0.64	0.49
SC_{4*}	iATTRACT	No	5.0	0.61	0.55	0.64	0.73	0.81	0.75	0.50	0.61	0.55
SC_{4*}	PISA	No	5.0	0.63	0.54	0.61	0.83	0.92	**0.79**	0.59	0.68	0.48
SC_{4*}	PISA	Yes	4.5	0.62	0.58	0.62	0.81	0.55	0.70	0.54	0.59	0.50
SC_{4*}	iATTRACT	Yes	6.0	0.58	0.53	0.59	0.72	0.62	0.58	0.60	0.55	0.53
SC_{4*}	PISA	Yes	5.0	0.61	0.48	0.64	0.80	0.93	0.69	0.51	0.56	0.51
SC_{4*}	PISA	Yes	6.0	0.61	0.48	0.55	0.83	0.91	0.66	0.58	0.55	0.51

Table 5.7: Table representing the AUC values obtained for each functional classes using the best combination of SC_{5*} according to the experimental interface. The best combination is made as in Section 5.2.2. The AUC values are represented for each combination of parameters possible. Lines were sortes according to the average of AUC computed over each functional class, and weighed according to their number of proteins.

Interface	Energy	CIPS	Distance Threshold	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_{5*}	iATTRACT	Yes	4.5	0.64	**0.67**	0.67	0.75	0.78	0.79	0.59	0.58	0.60
SC_{5*}	iATTRACT	No	4.5	**0.65**	0.63	0.65	0.79	0.88	0.80	0.57	0.64	**0.64**
SC_{5*}	MAXDo	Yes	5.0	0.62	0.58	**0.70**	0.68	0.46	0.83	0.68	0.63	0.51
SC_{5*}	MAXDo	Yes	4.5	0.64	0.52	**0.70**	0.72	0.56	0.77	0.73	0.72	0.52
SC_{5*}	MAXDo	Yes	6.0	0.62	0.54	0.63	0.73	0.56	0.71	0.71	0.62	0.54
SC_{5*}	MAXDo	No	4.5	0.64	0.47	0.64	0.72	0.52	0.82	**0.74**	**0.79**	0.51
SC_{5*}	iATTRACT	No	5.0	0.63	0.60	0.66	0.76	**0.91**	0.72	0.59	0.62	0.55
SC_{5*}	iATTRACT	No	6.0	0.62	0.57	0.63	0.72	0.86	0.71	0.62	0.60	0.56
SC_{5*}	PISA	No	4.5	0.64	0.56	0.64	0.81	0.49	0.82	0.56	0.71	0.54
SC_{5*}	MAXDo	No	6.0	0.63	0.50	0.57	0.70	0.57	0.73	0.76	0.72	0.50
SC_{5*}	iATTRACT	Yes	5.0	0.61	0.55	0.67	0.75	0.88	0.75	0.54	0.50	0.53
SC_{5*}	PISA	No	5.0	0.63	0.55	0.64	0.81	0.42	**0.84**	0.56	0.66	0.51
SC_{5*}	MAXDo	No	5.0	0.62	0.45	0.63	0.69	0.52	0.74	0.71	**0.79**	0.49
SC_{5*}	iATTRACT	Yes	6.0	0.60	0.51	0.59	0.73	0.71	0.64	0.61	0.57	0.57
SC_{5*}	PISA	Yes	4.5	0.62	0.56	0.64	0.78	0.47	0.69	0.54	0.57	0.53
SC_{5*}	PISA	No	6.0	0.63	0.53	0.64	0.81	0.67	0.83	0.56	0.71	0.47
SC_{5*}	PISA	Yes	6.0	0.63	0.49	0.60	**0.84**	0.80	0.75	0.56	0.57	0.55
SC_{5*}	PISA	Yes	5.0	0.61	0.49	0.66	0.78	0.62	0.74	0.52	0.56	0.53

Table 5.8: Table representing the AUC values obtained for each functional classes using the best combination of SC_{6*} according to the experimental interface. The best combination is made as in Section 5.2.2. The AUC values are represented for each combination of parameters possible. Lines were sortes according to the average of AUC computed over each functional class, and weighed according to their number of proteins. The red value refers to the AUC obtained for the matrix in Fig. 5.6d.

Interface	Energy	CIPS	Distance Threshold	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_{6*}	MAXDo	Yes	5.0	**0.67**	0.69	**0.76**	0.70	0.58	0.84	0.69	0.63	0.66
SC_{6*}	iATTRACT	Yes	4.5	0.66	**0.71**	0.70	0.73	0.63	0.82	0.59	0.50	0.70
SC_{6*}	MAXDo	Yes	4.5	**0.67**	0.61	0.73	0.71	0.55	0.81	0.73	0.72	0.61
SC_{6*}	MAXDo	No	5.0	**0.67**	0.60	0.70	0.70	0.51	0.80	0.71	**0.81**	0.65
SC_{6*}	iATTRACT	No	4.5	0.66	0.65	0.65	0.74	0.80	0.83	0.58	0.63	**0.72**
SC_{6*}	MAXDo	No	6.0	**0.67**	0.56	0.62	0.71	0.53	0.84	**0.78**	0.69	0.60
SC_{6*}	MAXDo	No	4.5	0.66	0.55	0.65	0.69	0.48	0.83	0.70	0.79	0.65
SC_{6*}	iATTRACT	No	5.0	0.66	0.63	0.70	0.74	**0.86**	0.78	0.59	0.63	0.65
SC_{6*}	iATTRACT	Yes	5.0	0.64	0.62	0.74	0.72	0.83	0.79	0.55	0.49	0.65
SC_{6*}	MAXDo	Yes	6.0	0.65	0.61	0.62	0.74	0.66	0.71	0.74	0.61	0.62
SC_{6*}	iATTRACT	No	6.0	0.63	0.57	0.63	0.73	0.78	0.71	0.70	0.65	0.62
SC_{6*}	iATTRACT	Yes	6.0	0.62	0.52	0.62	0.74	0.69	0.64	0.64	0.55	0.66
SC_{6*}	PISA	No	4.5	0.65	0.57	0.61	0.81	0.51	0.78	0.57	0.78	0.56
SC_{6*}	PISA	Yes	6.0	0.64	0.59	0.57	**0.82**	0.73	0.81	0.55	0.63	0.57
SC_{6*}	PISA	Yes	4.5	0.62	0.62	0.62	0.78	0.48	0.77	0.54	0.61	0.53
SC_{6*}	PISA	No	6.0	0.63	0.59	0.59	0.80	0.53	**0.86**	0.53	0.80	0.53
SC_{6*}	PISA	No	5.0	0.63	0.57	0.61	0.80	0.56	0.83	0.56	0.75	0.50
SC_{6*}	PISA	Yes	5.0	0.62	0.59	0.63	0.77	0.72	0.76	0.52	0.59	0.51

Table 5.9: Table representing the AUC values obtained for each functional classes using the best combination of SC_{d*} according to the experimental interface. The best combination is made as in Section 5.2.2. The AUC values are represented for each combination of parameters possible. Lines were sortes according to the average of AUC computed over each functional class, and weighed according to their number of proteins.

Interface	Energy	CIPS	Distance Threshold	ALL	AA	ABA	EI	ER	ES	OG	OR	OX
SC_{d*}	MAXDo	No	6.0	**0.65**	0.57	0.61	0.71	0.40	0.84	**0.78**	0.71	0.58
SC_{d*}	MAXDo	Yes	5.0	0.64	0.54	**0.73**	0.70	0.48	0.86	0.67	0.61	0.56
SC_{d*}	iATTRACT	Yes	4.5	0.63	**0.66**	0.65	0.76	0.54	0.81	0.58	0.45	**0.63**
SC_{d*}	MAXDo	Yes	6.0	**0.65**	0.56	0.68	0.75	0.56	0.75	0.73	0.60	0.57
SC_{d*}	MAXDo	Yes	4.5	0.64	0.61	0.65	0.71	0.39	0.82	0.69	0.67	0.54
SC_{d*}	MAXDo	No	5.0	0.64	0.48	0.63	0.71	0.37	0.82	0.74	0.78	0.57
SC_{d*}	MAXDo	No	4.5	0.64	0.49	0.57	0.72	0.42	0.82	0.75	0.75	0.60
SC_{d*}	PISA	No	4.5	0.66	0.62	0.61	0.83	0.46	0.80	0.62	0.77	0.58
SC_{d*}	iATTRACT	No	4.5	0.64	0.63	0.60	0.78	**0.95**	0.81	0.55	0.62	**0.63**
SC_{d*}	PISA	No	6.0	**0.65**	0.62	0.60	0.81	0.52	**0.87**	0.57	**0.80**	0.58
SC_{d*}	iATTRACT	No	5.0	0.63	0.60	0.62	0.75	0.96	0.79	0.59	0.60	0.58
SC_{d*}	PISA	No	5.0	**0.65**	0.60	0.58	0.82	0.44	0.85	0.61	0.74	0.55
SC_{d*}	iATTRACT	No	6.0	0.62	0.58	0.61	0.74	0.93	0.74	0.65	0.61	0.54
SC_{d*}	iATTRACT	Yes	5.0	0.62	0.51	0.67	0.74	0.89	0.81	0.55	0.43	0.59
SC_{d*}	PISA	Yes	6.0	**0.65**	0.54	0.58	**0.84**	0.70	0.78	0.57	0.61	**0.63**
SC_{d*}	iATTRACT	Yes	6.0	0.61	0.57	0.62	0.75	0.68	0.67	0.62	0.50	0.57
SC_{d*}	PISA	Yes	4.5	0.63	0.65	0.63	0.80	0.45	0.78	0.54	0.56	0.53
SC_{d*}	PISA	Yes	5.0	0.63	0.55	0.63	0.81	0.56	0.76	0.53	0.57	0.55

Part III

A tool for computing interfaces

Chapter 6

INTerface Builder

Contents

INTerface Builder (INTBuilder) is a fast, easy-to-use software to compute protein-protein interfaces. It is designed to retrieve interfaces from molecular docking software outputs in an empirically determined linear complexity. INTBuilder directly reads the output formats of popular docking programs like ATTRACT, HEX, MAXDo and ZDOCK, as well as a more generic format and Protein Data Bank (PDB) files. It identifies interacting surfaces at both residue and atom resolutions. This work has been published in [25].

6.1 Background and presentation of the question

The increasing amount of computing resources and the development of efficient molecular docking algorithms [35, 95, 87] have made possible large-scale studies of PPIs, where tens to thousands of proteins are docked to each other [95, 67, 56]. These cross-docking calculations generate millions to billions of conformations that must be screened in order to extract pertinent information. Several types of analysis can be performed, among which the calculation of the residues' propensity to be found at the interface in the docking poses. This property can be exploited toward protein binding sites [33, 95, 56] and functions [107] prediction. Also, docking interfaces can be analysed to select those that resemble the most known or predicted protein interfaces toward the identification of the cellular partners [95, 67, 56]. Both types of analysis require the fast and accurate detection of interacting residues in the docking conformations.

State-of-the-art approaches identify interacting residues based on inter-atomic distances, changes in residue Solvent Accessible Surface Area (SASA) upon binding [60] or a Voronoi model of the interface [15]. These methods suffer issues stemming from the large amount of data they need to handle. The first one is the speed of their algorithm. Since the number of conformations can go up to several billions on large-scale docking experiments, the algorithm used should be both fast and accurate in its computation of the interface. On the one hand, approaches based on grid-boxing or zoning [102, 78] efficiently detect interactions between particles based on a distance criterion in linear complexity. On the other hand, Voronoi model provides a more detailed description of the interface at the expense of more computation time. Another bottleneck is the input/output (I/O) required. To be able to analyse docking ensembles with current tools, one has to write and read the PDB file corresponding to each docking pose before actually computing the interface with the various software available today, the whole process resulting in a very high I/O.

Both issues are crucial to the analysis of large docking ensembles. To specifically address them, we have developed INTerface Builder (INTBuilder), which combines a new, efficient algorithm with the ability to directly read the output of rigid-body

docking software. Indeed, the algorithm of INTBuilder (detailed below) can achieve a complexity of $\mathcal{O}(n)$ by drastically reducing the search space when scanning protein surfaces for interface residue. INTBuilder explicitly considers the description of the docking pose by a scalar and a set of Euler angles representing the translation and rotations to be applied to the ligand relative to the receptor. To facilitate the usage of the rotating feature, the output of several rigid-body docking algorithm (iATTRACT [98], HEX [35], ZDOCK [19] and MAXDo [95]) is directly read with the effect of bypassing the I/O need. This allows INTBuilder to treat millions of conformations in a few hours. Other software (Rosetta [109], GRAMM-X [104]) directly outputs the resulting PDB files corresponding to each conformation, which allows INTBuilder analyse them without performing the rotations.

Although INTBuilder was designed to detect protein-protein interfaces, it can also readily be employed to identify the binding sites of small molecules (chemical compounds) from conformations obtained by virtual screening.

6.2 Algorithm

INTBuilder defines interfaces as sets of atoms or of residues, depending on the chosen scale, that are close to each other in a protein complex. It uses only one parameter (customisable by the user), that is the threshold distance under which two particles (residues or atoms) will be considered as interacting; we refer to this distance as $d-thresh$. A naive algorithmic approach would be to consider the two sets of particles $\mathcal{P}_1$ and $\mathcal{P}_2$ of each partner respectively and compute all the inter-atomic distances, thus leading to an $\mathcal{O}(n^2)$ complexity, n being the number of particles.

The idea behind the INTBuilder algorithm is to reduce the search space of particles before actually computing the inter-atomic distances (Fig. 6.1 and Algorithm 1). To do so, INTBuilder first selects the geometric centre $p-I$ of the ensemble of particles from the partner 1, $\mathcal{P}-1$. It then selects the farthest particle from it among of the ensemble of particles for the partner 2, $\mathcal{P}-2$, and name it $p-I$. From $p-I$, it computes the minimum distance to any particle belonging to $\mathcal{P}-1$ and subtracts to it $d-thresh$. We call the result of this subtraction $d-cut$. Any particle of $\mathcal{P}-2$ that is strictly closer to $p-I$ than $d-cut$ is removed from $\mathcal{P}-2$. Next, the algorithm selects the farthest particle of $\mathcal{P}-1$ from $p-I$, names it $p-I$ in turn and operates the same process. These steps are looped over while at least one particle has been removed with each iteration. The second step of the algorithm simply consists in computing all inter-atomic distances between the remaining candidate particles. We define two sets $\mathcal{I}-1$ and $\mathcal{I}-2$ representing interface particles of partner 1 and partner 2 respectively. As such, any pair of particles from partner 1 and partner 2 are added to $\mathcal{I}-1$ and $\mathcal{I}-2$ respectively if they are separated by a distance lower than $d-thresh$. To ascertain that the algorithm does not erroneously remove any

interface particle, we reason as follows.

We want to show that at each iteration (cycle `do` at line 4 in Algorithm 1), INTBuilder reduces the number of particles in $\mathcal{P}_1, \mathcal{P}_2$ while keeping those lying at the interface. We denote $d_{i,j}$ the distance between particles p_i and p_j.

Each iteration comprises two "internal iterations" (cycles `for` at line 8 and 16 in Algorithm 1), the first eliminating some particles in $\mathcal{P}_2$ and the second in $\mathcal{P}_1$. At the beginning of each internal iteration, INTBuilder defines a particle p_I (lines 6 and 14 in Algo 1). At the first iterative step, INTBuilder takes, as p_I, the farthest particle of the partner 2 from the centre of mass of the partner 1.

If p_I belongs to the interface, notice that $min\{d_{I,j} - d_{thresh} \,|\, p_j \in \mathcal{P}_2\} < 0$ by definition. This implies that no particles' deletion will be realised by INTBuilder at the first internal iteration step, and the algorithm will go on by considering the particle in $\mathcal{P}_1$ that is most distant from p_I and will take this particle to be the new p_I.

If p_I does not belong to the interface, then let p_o be any particle of $\mathcal{P}_2$ belonging to the interface. We want to prove that p_o cannot be removed by INTBuilder. INTBuilder chooses a particle $p_m \in \mathcal{P}_1$ that is the closest to p_I. Then, it removes from $\mathcal{P}_2$ all particles p_j satisfying the equation:

$$d_{I,j} < d_{I,m} - d_{thresh} \tag{6.1}$$

Since p_o belongs to the interface of partner 2, by definition of particles at the interface, there is a particle $p_k \in \mathcal{P}_1$ belonging to the interface of partner 1 such as $d_{o,k} \leq d_{thresh}$. In order to show that p_o does not satisfy equation (1), we show:

$$d_{I,o} \geq d_{I,m} - d_{thresh} \tag{6.2}$$

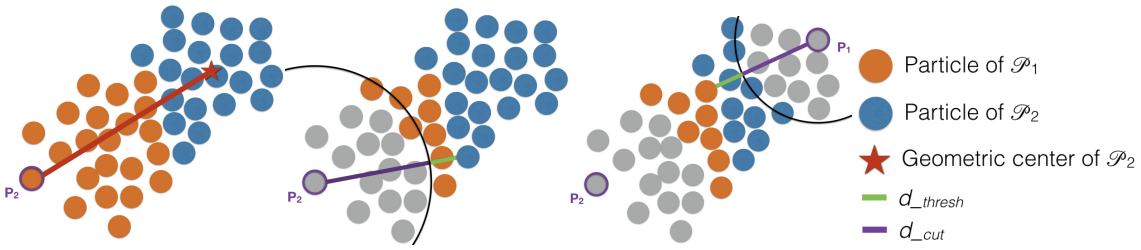
Notice that $d_{I,m} \leq d_{I,k}$ because of the way p_m was chosen, and since $d_{I,k} \leq d_{I,o} + d_{o,k}$, we have

$$d_{I,m} \leq d_{I,o} + d_{o,k} \tag{6.3}$$

Since $d_{o,k} \leq d_{thresh}$ then, by (3), we derive $d_{I,m} - d_{thresh} \leq d_{I,o}$, that is (2), as claimed above. To show that particles in the interface are not removed in $\mathcal{P}_1$ by the second internal iteration of the algorithm, we proceed in a similar way.

Although the worst case scenario could theoretically lead the algorithm to a complexity of $\mathcal{O}(n^2)$, that only happens if the whole surface of the protein is interacting (the complexity of INTBuilder is mainly linked with the size of the interacting surface itself more than the size of the protein).

To estimate the empirical complexity of the algorithm, we computed the inter-



(a) First step (b) Second step (c) Iteration of the second step

Figure 6.1: Scheme of the search space reduction algorithm. **(a)** The geometric centre of the blue partner (red star) is chosen as a starting point and the farthest particle p_2 of the orange partner is selected. **(b)** The minimum distance between p_2 and the blue partner is computed and d_{thresh} is subtracted to it to obtain d_{cut}. All the particles closer than d_{cut} (in grey) are removed from the orange partner. **(c)** The particle p_1 of the blue partner that is the farthest from p_2 is chosen and the reduction step is repeated.

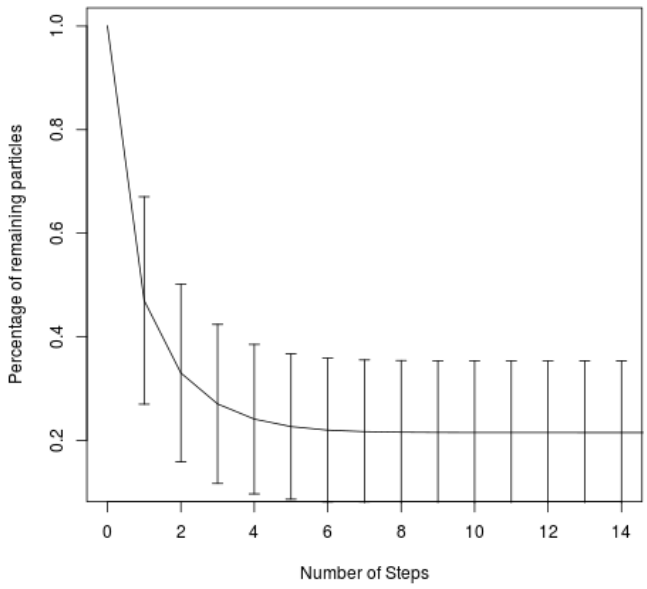


Figure 6.2: Percentage of remaining residues in two proteins P_1, P_2 given as entry to INTBuilder in regard to the number of steps performed by the algorithm. The plot is constructed from 10% of conformations randomly chosen from the PPDBv2 database [76, 67]. After the 6th step, the curve reaches a stable behaviour where only 22% of the residues are kept for most proteins.

Algorithm 1 Reducing the search space and pairwise detection

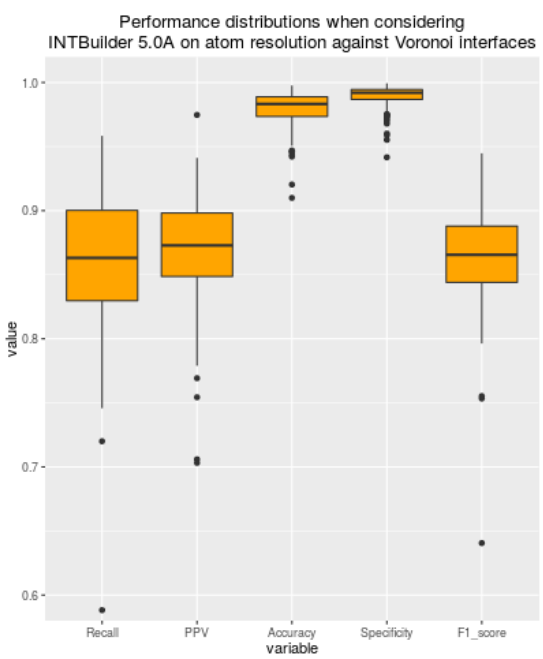
1: let $\mathcal{P}_1$ be the ensemble of particles for the partner 1
2: let $\mathcal{P}_2$ be the ensemble of particles for the partner 2
3: compute the geometric centre of $\mathcal{P}_1$ and call it p_I
4: **do**
5: choose p_2 such that $d_{p_2,p_I} \geq d_{p_j,p_I}$ for all $p_j \in \mathcal{P}_2$
6: let p_2 be called p_I
7: compute d_{cut} as $\min(d_{p_I,p_i} - d_{thresh})$ for all $p_i \in \mathcal{P}_1$
8: **for** $p_j \in \mathcal{P}_2$ **do**
9: **if** $d_{p_I,p_j} < d_{cut}$ **then**
10: remove p_j from $\mathcal{P}_2$
11: **end if**
12: **end for**
13: choose p_1 such that $d_{p_1,p_I} \geq d_{p_i,p_I}$ for all $p_i \in \mathcal{P}_1$
14: let p_1 be called p_I
15: compute d_{cut} as $\min(d_{p_I,p_j} - d_{thresh})$ for all $p_j \in \mathcal{P}_2$
16: **for** $p_i \in \mathcal{P}_1$ **do**
17: **if** $d_{p_I,p_i} < d_{cut}$ **then**
18: remove p_i from $\mathcal{P}_1$
19: **end if**
20: **end for**
21: **while** at least an element is removed in $\mathcal{P}_1$ or $\mathcal{P}_2$
22:
23: let $\mathcal{I}_1$ be the set of interface particles for the partner 1
24: let $\mathcal{I}_2$ be the set of interface particles for the partner 2
25: **for** $p_i \in \mathcal{P}_1$ **do**
26: **for** $p_j \in \mathcal{P}_2$ **do**
27: **if** $d_{p_i,p_j} \leq d_{thresh}$ **then**
28: add p_i to $\mathcal{I}_1$
29: add p_j to $\mathcal{I}_2$
30: **end if**
31: **end for**
32: **end for**

faces of about 50 million complex structure predictions, obtained from a complete cross-docking of 168 proteins [76] using the docking algorithm MAXDo [95]. Overall, we found that the do-while loop (Algorithm 1, lines 4-21) had an average of 5.8 iterations and a maximum number of iterations N_{max} of 23. Thus, the reduction of the search space algorithm is realised in $\mathcal{O}(n \times N_{max})$. Since N_{max} is constant, this step has a time complexity of $\mathcal{O}(n)$. The last part of the INTBuilder algorithm (from line 23 on) computes all the distances between the remaining candidate particles of $\mathcal{P}-1$ and $\mathcal{P}-2$ and stores them in $\mathcal{I}-1$ and $\mathcal{I}-2$ respectively if they are in contact with one another. Although the complexity of this last step is $\mathcal{O}(n^2)$, n holds only for roughly a quarter of its original value after the space reduction obtained in the first part of the algorithm (Fig. 6.2).

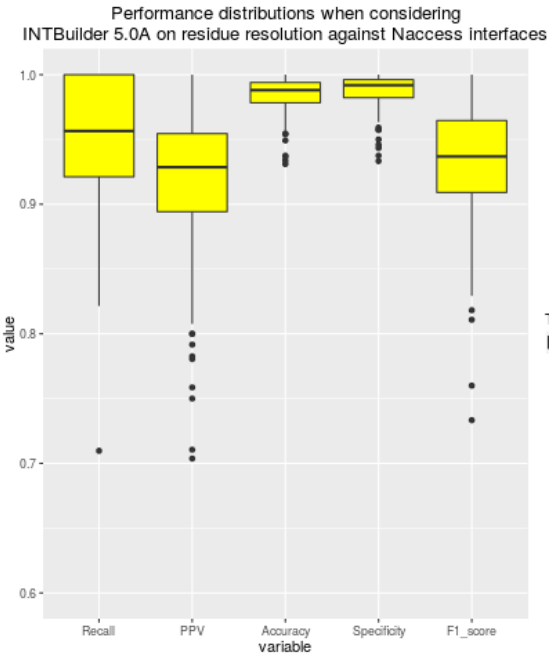
6.3 Comparison with other methods

INTBuilder is distance based, and as other similar methods its main challenge consists in reducing the search space before computing all pairwise distances between remaining candidates particles. As INTBuilder, boxing approaches [102] focus on reducing the search space and do so with a complexity of $\mathcal{O}(n)$. An important part of the boxing approaches consists in defining the grid size, which adds another parameter to the program. To the best of our knowledge, no tool is available to specifically detect protein-protein interfaces using a boxing approach. In contrast, INTBuilder has the advantage of its algorithmic simplicity, ease of implementation and of a single defined parameter (threshold distance). Overall, boxing approaches are applied to more general issues (Discrete Element Method, Molecular Dynamics) while INTBuilder focuses on a specific issue. We have measured the computation time required by INTBuilder and a naive approach (computing every inter-atomic distances) and specifically evaluated the computation time of INTBuilder's algorithm compared to the naive approach in Table 6.3. The results show a decrease of the computation time of the interface determination by a factor from ten to one hundred over the naive algorithm, depending on the size of the protein. INTBuilder's efficiency was also compared with Naccess [41] and the Voronoi model [15] when computing the interface for a single complex (Table 6.4). Since we do not read from a docking output, we do not use INTBuilder's perk of bypassing the I/O. This permits us to focus on the algorithm speed itself in its comparison to other software. When looking at several conformations however, INTBuilder's ability to bypass the I/O and allows it to outshine the other software in terms of computation speed. Indeed, both software require to write the PDB file corresponding to each conformation, which proved to be extremely hindering for treating the 50 million conformations of our set. Both tables show that INTBuilder is consistently faster than the other two software, its increase in speed ranging from twenty to more than one hundred times faster. We computed in the table 6.5 the interface for five hundreds conformations computed with HEX [35]. We show here the importance of the I/O ability implemented in INTBuilder (also present in the Naive approach). Naccess and Voronoi give a computation time in the same order of magnitude as the docking time itself. The naive approach, while benefiting from the I/O ability of INTBuilder also shows its lack of scalability when considering bigger complexes.

We compared the accuracy with which the different methods were able to define interfaces. All three of them yield similar interfaces (Table 6.1 and Fig. 6.3). On average, the detected interfaces comprise the same number of particles (atoms or residues), and they share more than 79% of particles in common (Table 6.1). We further evaluated the impact of the small differences between the interfaces detected



(a) Boxplot representation of performances distributions when comparing interfaces computed with INTBuilder (atom-resolution) against those computed using a Voronoi description [15] on 84 complexes from the PPDBv2 database [76]. The data used for the plot regroups 4 750 938 different conformations.



(b) Boxplot representation of performances distributions when comparing interfaces computed with INTBuilder (residue-resolution) against those computed using the Naccess software on 168 complexes from [76]. The data used for the plot regroups 49 192 401 different conformations.

Figure 6.3: Performance distributions when comparing INTBuilder 5.0Å to other methods

	Atom Voronoi	Residue Naccess
Recall	0.79	**0.90**
PPV	0.80	**0.83**
Accuracy	0.98	**1.00**
Specificity	0.99	**1.00**
F1-score	0.79	**0.86**
Naccess/Voronoi average interface size	78	16
INTBuilder average interface size	78	17

Table 6.1: Statistical values obtained when comparing INTBuilder with a 5Å distance cutoff to Naccess and Voronoi model. For the INTBuilder-Voronoi comparison, 4 750 938 conformations were treated and the interfaces were detected at the atomic scale. For the INTBuilder-Naccess comparison, 49 192 401 conformations were treated and the interfaces were detected at the residue scale. PPV stands for Positive Predictive Value.

by INTBuilder, Naccess and Voronoi (Table 6.1) on the discrimination of binding partners. We considered the 14 196 possible protein pairs of our dataset of 168 proteins and the goal was to single out the 84 experimentally validated pairs of interactors. The docking interfaces detected by INTBuilder, Naccess and Voronoi were compared to the experimentally known interfaces. For each protein pair, the docking pose with the interface resembling the experimental interface the most was selected, and the overlap between docking and experimental interfaces was used to compute an interaction index for the protein pair. All protein pairs were then ranked based on their interaction indices (see [67] for a detailed description of the protocol). The discrimination power of the approach was estimated by the Area Under the Curve (AUC). The AUC values obtained on the whole dataset and on the different functional classes are very similar between the three detection methods (Table 6.2). In other words, no significant advantage over INTBuilder could be gained from using another method. These results show that INTBuilder is accurate enough to be used in the context of partner discrimination.

6.4 Conclusion

We have presented INTBuilder, a new, easy-to-use and very efficient software which computes the interface between two proteins. The speed of its algorithm comes from a new way to reduce the search space before computing the interacting distances between remaining particles and is able to achieve an $\mathcal{O}(n)$ complexity. INTBuilder itself has been implemented in such a way that it can process millions of different conformations coming from docking software in a limited amount of time. Specif-

	Atom		Residue	
	INTBuilder	Voronoi	INTBuilder	Naccess
AA (20)	0.83	0.84	**0.86**	0.83
ABA (24)	0.86	0.91	**0.92**	**0.92**
EI (38)	0.84	**0.88**	0.81	0.82
ER (6)	**0.78**	0.72	**0.78**	0.74
ES (12)	0.87	**0.90**	0.83	0.87
OG (24)	0.93	**0.95**	0.90	0.87
OR (14)	0.81	0.82	0.79	**0.87**
OX (30)	0.87	**0.92**	0.88	0.84

Table 6.2: AUC values for the identification of interacting partners in the Protein-Protein Docking Benchmark v2 [76]. The complete cross-docking experiment is described in [67]. The AUCs were obtained by using experimental interfaces and docking interfaces computed according to the method described in the column. The dataset is divided into 8 functional classes: Antibody-Antigen (AA), Bound Antibody-Antigen (ABA), Enzyme-Inhibitor (EI), Enzyme-Regulator (ER), Enzyme-Substrate (ES), Other linked to G-protein (OG), Other regulatory (OR) and Other (OX).

Complexes	Size (atoms)	INTBuilder (s) (atom)	Naive approach (s) (atom)
7CEI	1724	**0.0004**	0.0027
1FC2	2010	**0.0002**	0.0024
1ACB	2291	**0.0004**	0.0034
1TMQ	4479	**0.0010**	0.0124
1JPS	4858	**0.0010**	0.0225
1IBR	4944	**0.005**	0.0261
1RLB	5171	**0.0004**	0.0206
2VIS	5337	**0.0006**	0.0294
1ML0	6221	**0.0020**	0.0112
1N2C	20058	**0.0042**	0.3607

Table 6.3: Computation time required to compute the interface for each bound complex using inter-atomic distances. We compare the time required for the computation of the interface only, and do not consider the I/O. The comparison is made when using INTBuilder's algorithm to reduce the search and when using a naive approach computing all inter-atomic distances. Calculations have been realised on a single core processor Intel Xeon E3-1271 v3 @ 3.60GHz.

Complexes	Size (atoms)	INTBuilder (s) (atom/residue)	Naccess (s) (residue)	Voronoi (s) (atom)
7CEI	1724	**0.003**	0.225	0.139
1FC2	2010	**0.006**	0.382	0.227
1ACB	2291	**0.004**	0.377	0.174
1TMQ	4479	**0.006**	0.784	0.284
1JPS	4858	**0.006**	0.707	0.291
1IBR	4944	**0.014**	0.575	0.387
1RLB	5171	**0.006**	0.757	0.297
2VIS	5337	**0.014**	1.271	0.338
1ML0	6221	**0.011**	1.057	0.382
1N2C	20058	**0.028**	3.413	1.18

Table 6.4: Computation time required to compute the interface of the bound complex. For the three tools, time is expressed in seconds (s). For Naccess and Voronoi, time includes external tools to perform the necessary rotations. Calculations were realised on a single core processor Intel Xeon E3-1271 v3 @ 3.60GHz.

Complexes	Size (residues)	HEX (s)	INTBuilder (s) (atom/residue)	Naive approach (s) (atom/residue)	Naccess (s) (residue)	Voronoi (s) (atom)
7CEI	1724	184	**0.404**	1.081	101.3	67.5
1FC2	2010	264	**0.498**	0.632	80.4	50.6
1ACB	2291	176	**0.551**	1.044	84.0	61.3
1TMQ	4479	168	**1.139**	3.500	115.8	92.9
1JPS	4858	192	**2.964**	6.507	146.3	120.5
1IBR	4944	152	**5.573**	8.453	177.8	123.7
1RLB	5171	176	**1.590**	7.392	161.7	130.5
2VIS	5337	200	**1.382**	8.851	156.2	121.9
1ML0	6221	176	**1.426**	2.676	111.0	105.1
1N2C	20058	256	**16.781**	59.157	470.8	473.9

Table 6.5: Computation time required to compute the interface of 500 conformations for each bound complex, using the docking algorithm HEX. Time is expressed in seconds (s). For Naccess and Voronoi, time includes external tools to perform the necessary rotations. Calculations have been realised on a single core processor Intel Xeon E3-1271 v3 @ 3.60GHz.

ically, it can directly read the output of known rigid-body docking software. This feature allows it to avoid any excess of I/O and thus brings a valuable gain of time when considering large set of docking conformations.

The data obtained from the interfaces of large-scale docking calculations can be exploited to identify cellular partners and/or compute propensities of residues to be found at the interface. Although INTBuilder was designed for PPIs, it can also be readily applied to small-molecule docking. The simplicity of INTBuilder's usage makes it a valuable tool to identify the binding sites of small molecules from conformations obtained by virtual screening.

Part IV

Conclusion

The title of the PhD. thesis is "Geometry of protein interactions" and its goal was to analyse different datasets of proteins and enlarge the scale of the existing analysis. Specifically, I worked on two fields: The detection and interpretation of the protein binding sites and the identification of interacting partners in a large scale Complete Cross-Docking study (CC-D).

The analysis of protein interaction sites has brought much information, including the emerging concept of multiple interaction sites and how proteins interact in a crowded environment. This topic (described in depth in Chapter 4) shows that the interacting surface of proteins would be far greater than expected and far greater than what is currently accounted for in most cases. The analysis brings with it a new tool which could be readily used for further analysis of biological interfaces among homologs of a query protein and dynJET^2 (developed from the JET^2 software [57]), a prediction software able to take into account any residue-based scoring into its prediction method. The analysis brings the concepts of Interaction Sites (IS) and Interaction Regions (IR). These two definitions are essential to understand how we might interpret the interfaces at the proteins' surface. Furthermore, the study shows how it might be possible for a protein to infer if an IR is targeted by several partners and how many functional regions a protein has. New work and effort should go in two directions: further investigating ways to separate the dynJET^2 predictions into matching IR and refining the precision with which we are able to determine if a predicted interface is actually an IS (specific to one partner) or an IR.

The second analysis, centred on the identification of interacting proteins in a large scale CC-D also brings many promising results. We show here how the development of a more advanced interface prediction method along the use of adapted scoring methods regarding the proteins' functions has allowed us to make great progress in terms of partner discrimination. To answer the need of high-performing software to compute the interfaces corresponding to docking conformations, I developed the INTBuilder software (Chapter 6, [25]) which brings an innovative way of reducing the search space of an ensemble of particles. This analysis brings an important message showing how crucial it is to take into account the functional class a protein belongs to. Moreover, we show as well that in many cases our capacities in terms of partners identification have reached a limit which seems set by the quality and the precision of our predictions. Searching for better and more accurate predictions should be the next goal, but it should also be stressed that such predictions will not be specific to a single partner. This implies that it would not be possible to attain experimental-like discrimination results. Several paths lay ahead: One would be to try to develop automatic methods for partner-specific interface prediction, the other could be to shift the way we look at the issue with the current method. Instead of characterising how a protein interacts with others through a single, well-defined predicted interface, we could look simultaneously at all predicted interfaces of a

protein and see how each of them interact with the different predicted interfaces of other proteins.

6.5 Work done

Articles

C. Dequeker, E. Laine, A. Carbone. INTerface Builder: A Fast Protein-Protein Interface Reconstruction Tool. *J Chem Inf Model,* 57(11):2613-2617, Nov 2017.
Data and software available at http://www.lcqb.upmc.fr/INTBuilder/

C. Dequeker, E. Laine, A. Carbone, "Multiple binding sites of protein-protein interactions predicted by combining sequence analysis and molecular docking", to be submitted, 2018
Data and software available at http://www.lcqb.upmc.fr/dynJET2/ (as soon as the work is published)

C. Dequeker, E. Laine, A. Carbone, "Protein partners discrimination reached with coarse-grain docking and binding sites predictions", in preparation, 2018

Posters

C. Dequeker, E. Laine, A. Carbone, Large scale analysis of protein interactions, *Journées Ouvertes de Biologie Informatique & Mathématiques (JOBIM 2018)*, Marseille, France, July 3-6, 2018.

C. Dequeker, E. Laine, A. Carbone, Large scale analysis of protein interactions, *Journée de Biologie Structurale*, Paris, France, October 2nd, 2017.

C. Dequeker, R. Raucci, E. Laine, A. Carbone. Large scale analysis of protein interactions. *15th European Conference on Computational Biology (ECCB 2016)*, The Hague, The Netherlands. September 3-7, 2016

F. Corsi, C. Dequeker, E. Laine, F. Nadalin, R. Raucci and A. Carbone. Large scale analysis of protein interactions. *Symposium du Réseau de Biologie des Systèmes de Sorbonne Universités*, Paris, France, June 6th, 2016.

Presentations

C. Dequeker, E. Laine, A. Carbone, Proteins and their multiple interaction sites. *UPMC Young Researchers' Meeting: Modeling Complex Biological Systems*, Paris, France, December 13th, 2017.

C. Dequeker, E. Laine, A. Carbone. Approaches and scorings for partner discrimination. *MAPPING meeting.* Paris, France, June 30th, 2017.

C. Dequeker, E. Laine, A. Carbone. Approaches and scorings for partner discrimination. *Internal Seminar at the Laboratory of Computational and Quantitative Biology (LCQB).* Paris, France, April 13th, 2017.

Bibliography

[1] Patrick Aloy and Robert B. Russell. Structural systems biology: modelling protein interactions. *Nature Reviews Molecular Cell Biology*, 7(3):188–197, March 2006.

[2] M. R. Arkin, Y. Tang, and J. A. Wells. Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality. *Chem. Biol.*, 21(9):1102–1114, Sep 2014.

[3] Aharon Armon, Dan Graur, and Nir Ben-Tal. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information11Edited by F. Cohen. *J. Mol. Biol.*, 307(1):447–463, Mar 2001.

[4] T. T. Aumentado-Armstrong, B. Istrate, and R. A. Murgita. Algorithmic approaches to protein-protein interaction site prediction, 2015.

[5] A. Selim Aytuna, Attila Gursoy, and Ozlem Keskin. Prediction of protein–protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12):2850–2855, Jun 2005.

[6] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joël Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins Struct. Funct. Bioinf.*, 53(3):708–719, Nov 2003.

[7] K. Bastard, C. Prevost, and M. Zacharias. Accounting for loop flexibility during protein-protein docking. *Proteins*, 62(4):956–969, Mar 2006.

[8] Calem J. Bendell, Shalon Liu, Tristan Aumentado-Armstrong, Bogdan Istrate, Paul T. Cernek, Samuel Khan, Sergiu Picioreanu, Michael Zhao, and Robert A. Murgita. Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor. *BMC Bioinf.*, 15(1):82, Dec 2014.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, Jan 2000.

[10] S. Betzi, A. Restouin, S. Opi, S. T. Arold, I. Parrot, F. Guerlesquin, X. Morelli, and Y. Collette. Protein protein interaction inhibition (2P2I) combining high throughput and virtual screening: Application to the HIV-1 Nef protein. *Proc. Natl. Acad. Sci. U.S.A.*, 104(49):19256–19261, Dec 2007.

[11] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, 280(1):1–9, Jul 1998.

[12] Andrew J. Bordner and Ruben Abagyan. Statistical analysis and prediction of protein–protein interfaces. *Proteins Struct. Funct. Bioinf.*, 60(3):353–366, May 2005.

[13] James R. Bradford, Chris J. Needham, Andrew J. Bulpitt, and David R. Westhead. Insights into Protein–Protein Interfaces using a Bayesian Network Prediction Method. *J. Mol. Biol.*, 362(2):365–386, Sep 2006.

[14] James R. Bradford and David R. Westhead. Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–1494, Apr 2005.

[15] F. Cazals, F. Proust, R. P. Bahadur, and J. Janin. Revisiting the Voronoi Description of Protein-Protein Interfaces. *Protein Sci.*, 15(9):2082–2092, Sep 2006.

[16] Nicoletta Ceres, Marco Pasi, and Richard Lavery. A protein solvation model based on residue burial. *Journal of Chemical Theory and Computation*, 8(6):2141–2144, 2012.

[17] P. Chakrabarti and J. Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–343, May 2002.

[18] Huiling Chen and Huan-Xiang Zhou. Prediction of interface residues in protein–protein complexes by a consensus neural network method: Test against NMR data. *Proteins Struct. Funct. Bioinf.*, 61(1):21–35, Oct 2005.

[19] R. Chen and Z. Weng. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, 47(3):281–294, May 2002.

[20] Yuehui Chen, Jingru Xu, Bin Yang, Yaou Zhao, and Wenxing He. A novel method for prediction of protein interaction sites based on integrated RBF neural networks. *Comput. Biol. Med.*, 42(4):402–407, Apr 2012.

[21] Gong Cheng, Bin Qian, Ram Samudrala, and David Baker. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.*, 33(18):5861–5867, Jan 2005.

[22] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites11Edited by A. R. Fersht. *J. Mol. Biol.*, 285(5):2177–2198, Feb 1999.

[23] Charles Darwin. *On the origin of species*. John Murray, 1859.

[24] Sjoerd J de Vries and Alexandre MJJ Bonvin. Cport: a consensus interface predictor and its performance in prediction-driven docking with haddock. *PloS one*, 6(3):e17695, 2011.

[25] C. Dequeker, E. Laine, and A. Carbone. INTerface Builder: A Fast Protein-Protein Interface Reconstruction Tool. *J Chem Inf Model*, 57(11):2613–2617, Nov 2017.

[26] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *J. Am. Chem. Soc.*, 125(7):1731–1737, Feb 2003.

[27] Qiwen Dong, Xiaolong Wang, Lei Lin, and Yi Guan. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC Bioinf.*, 8(1):147, Dec 2007.

[28] Adrian H. Elcock and J. Andrew McCammon. Identification of protein oligomerization states by analysis of interface conservation. *Proc. Natl. Acad. Sci. U.S.A.*, 98(6):2990–2994, Mar 2001.

[29] S. Engelen, L. A. Trojan, S. Sacquin-Mora, R. Lavery, and A. Carbone. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS Comput. Biol.*, 5(1):e1000267, 2009.

[30] Reyhaneh Esmaielbeiki, Konrad Krawczyk, Bernhard Knapp, Jean-Christophe Nebel, and Charlotte M. Deane. Progress and challenges in predicting protein interfaces. *Briefings Bioinf.*, 17(1):117–131, Jan 2016.

[31] Reyhaneh Esmaielbeiki and Jean-Christophe Nebel. Scoring docking conformations using predicted protein interfaces. *BMC Bioinf.*, 15(1):171, Dec 2014.

[32] Piero Fariselli, Florencio Pazos, Alfonso Valencia, and Rita Casadio. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.*, 269(5):1356–1361, Mar 2002.

[33] J. Fernandez-Recio, M. Totrov, and R. Abagyan. Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.*, 335(3):843–865, Jan 2004.

[34] Xavier Gallet, Benoit Charloteaux, Annick Thomas, and Robert Brasseur. A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.*, 302(4):917–926, Sep 2000.

[35] A. W. Ghoorah, M. D. Devignes, M. Smail-Tabbone, and D. W. Ritchie. Protein Docking Using Case-Based Reasoning. *Proteins*, 81(12):2150–2158, Dec 2013.

[36] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43(2):89–102, May 2001.

[37] Solène Grosdidier and Juan Fernández-Recio. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC bioinformatics*, 9(1):447, 2008.

[38] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell. Protein-protein interaction networks and biology–what's the connection? *Nat. Biotechnol.*, 26(1):69–72, Jan 2008.

[39] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):47–52, Dec 1999.

[40] Andrew L. Hopkins. Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, 4(11):682, Oct 2008.

[41] S. J. Hubbard and J. M. Thornton. Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*, 2, 1993.

[42] E. L. Huttlin, L. Ting, R. J. Bruckner, F. Gebreab, M. P. Gygi, J. Szpyt, S. Tam, G. Zarraga, G. Colby, K. Baltier, R. Dong, V. Guarani, L. P. Vaites, A. Ordureau, R. Rad, B. K. Erickson, M. Wuhr, J. Chick, B. Zhai, D. Kolippakkam, J. Mintseris, R. A. Obar, T. Harris, S. Artavanis-Tsakonas, M. E. Sowa, P. De Camilli, J. A. Paulo, J. W. Harper, and S. P. Gygi. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2):425–440, Jul 2015.

[43] H. Hwang, T. Vreven, J. Janin, and Z. Weng. Protein-protein docking benchmark version 4.0. *Proteins*, 78(15):3111–3114, Nov 2010.

[44] Howook Hwang, Thom Vreven, Brian G Pierce, Jui-Hung Hung, and Zhiping Weng. Performance of zdock and zrank in capri rounds 13–19. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3104–3110, 2010.

[45] Howook Hwang, Thom Vreven, and Zhiping Weng. Binding interface prediction by combining protein–protein docking results. *Proteins: Structure, Function, and Bioinformatics*, 82(1):57–66, 2014.

[46] C. Axel Innis. siteFiNDER|3D: a web-based tool for predicting the location of functional sites in proteins. *Nucleic Acids Res.*, 35(suppl_2):W489–W494, Jul 2007.

[47] S. Jones, A. Marin, and J. M. Thornton. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, 13(2):77–82, Feb 2000.

[48] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *PNAS*, 93(1):13–20, Jan 1996.

[49] Susan Jones and Janet M Thornton. Review Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, page 8, 1996.

[50] Susan Jones and Janet M. Thornton. Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.*, 272(1):133–143, Sep 1997.

[51] Panagiotis L. Kastritis and Alexandre M. J. J. Bonvin. Are Scoring Functions in Protein-Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *J. Proteome Res.*, 9(5):2216–2225, May 2010.

[52] Panagiotis L. Kastritis, Iain H. Moal, Howook Hwang, Zhiping Weng, Paul A. Bates, Alexandre M. J. J. Bonvin, and Joël Janin. A structure-based benchmark for protein–protein binding affinity. *Protein Sci.*, 20(3):482–491, Mar 2011.

[53] Roland Krause, Christian von Mering, Peer Bork, and Thomas Dandekar. Shared components of protein complexes—versatile building blocks or biochemical artefacts? *Bioessays*, 26(12):1333–1343, Dec 2004.

[54] Evgeny Krissinel and Kim Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, 372(3):774 – 797, 2007.

[55] Nathalie Lagarde, Alessandra Carbone, and Sophie Sacquin-Mora. Hidden partners: Using cross-docking calculations to predict binding sites for proteins with multiple interactions. *Proteins*, xx(xx):xx–xx, xx 2018.

[56] E. Laine and A. Carbone. Protein Social Behavior Makes a Stronger Signal for Partner Identification than Surface Geometry. *Proteins*, 85(1):137–154, Jan 2017.

[57] Elodie Laine and Alessandra Carbone. Local geometry and evolutionary conservation of protein surfaces reveal the multiple recognition patches in protein-protein interactions. *PLOS Computational Biology*, 11(12):1–32, 12 2015.

[58] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov 2007.

[59] T. A. Larsen, A. J. Olson, and D. S. Goodsell. Morphology of protein-protein interfaces. *Structure*, 6(4):421–427, Apr 1998.

[60] M. F. Lensink and S. J. Wodak. Docking and scoring protein interactions: CAPRI 2009. *Proteins*, 78(15):3073–3084, Nov 2010.

[61] Marc F Lensink and Shoshana J Wodak. Blind predictions of protein interfaces by docking calculations in capri. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3085–3095, 2010.

[62] Emmanuel D. Levy. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *J. Mol. Biol.*, 403(4):660–670, Nov 2010.

[63] Bi-Qing Li, Kai-Yan Feng, Lei Chen, Tao Huang, and Yu-Dong Cai. Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. *PLoS One*, 7(8):e43927, Aug 2012.

[64] Hong Li, Yuan Zhou, and Ziding Zhang. Competition-cooperation relationship networks characterize the competition and cooperation between proteins. *Sci. Rep.*, 5:11619, Jun 2015.

[65] Nan Li, Zhonghua Sun, and Fan Jiang. Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC Bioinf.*, 9(1):553, Dec 2008.

[66] O. Lichtarge, H. R. Bourne, and F. E. Cohen. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257(2):342–358, 1996.

[67] A. Lopes, S. Sacquin-Mora, V. Dimitrova, E. Laine, Y. Ponty, and A. Carbone. Protein-Protein Interactions in a Crowded Environment: an Analysis via Cross-Docking Simulations and Evolutionary Information. *PLoS Comput. Biol.*, 9(12):e1003369, 2013.

[68] Simon C. Lovell and David L. Robertson. An Integrated View of Molecular Coevolution in Protein–Protein Interactions. *Mol. Biol. Evol.*, 27(11):2567–2575, Nov 2010.

[69] Buyong Ma, Tal Elkayam, Haim Wolfson, and Ruth Nussinov. Protein–protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *PNAS*, 100(10):5772–5777, May 2003.

[70] Surabhi Maheshwari and Michal Brylinski. Across-proteome modeling of dimer structures for the bottom-up assembly of protein-protein interaction networks. *BMC Bioinf.*, 18:257, May 2017.

[71] J. Martin and R. Lavery. Arbitrary protein-protein docking targets biologically relevant interfaces. *BMC Biophys*, 5:7, 2012.

[72] Juliette Martin and Richard Lavery. Arbitrary protein- protein docking targets biologically relevant interfaces. *BMC biophysics*, 5(1):7, 2012.

[73] Sean R. McGuffee and Adrian H. Elcock. Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Comput. Biol.*, 6(3):e1000694, Mar 2010.

[74] Mihaly Mezei. A new method for mapping macromolecular topography. *Journal of Molecular Graphics and Modelling*, 21:463 – 472, 2003.

[75] I. Mihalek, I. Res, and O. Lichtarge. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, 336(5):1265–1282, 2004.

[76] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, and Z. Weng. Protein-Protein Docking Benchmark 2.0: an Update. *Proteins*, 60(2):214–216, Aug 2005.

[77] Julian Mintseris and Zhiping Weng. Structure, function, and evolution of transient and obligate protein–protein interactions. *PNAS*, 102(31):10930–10935, Aug 2005.

[78] B. K. Mishra. A review of computer simulation of tumbling mills by the discrete element method: Part i—contact mechanics. *Int. J. Min. Proc.*, 71:73–95, March 2003.

[79] Francesca Nadalin and Alessandra Carbone. Protein–protein interaction specificity is captured by contact preferences and interface composition. *Bioinformatics*, 34(3):459–468, Feb 2018.

[80] S. S. Negi and W. Braun. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. *J Mol Model*, 13(11):1157–1167, 2007.

[81] H. Neuvirth, R. Raz, and G. Schreiber. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, 338(1):181–199, Apr 2004.

[82] Hani Neuvirth, Ran Raz, and Gideon Schreiber. ProMate: A Structure Based Prediction Program to Identify the Location of Protein–Protein Binding Sites. *J. Mol. Biol.*, 338(1):181–199, Apr 2004.

[83] I. M.A. Nooren. NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. *The EMBO Journal*, 22(14):3486–3492, July 2003.

[84] Y. Ofran and B. Rost. ISIS: interaction sites identified from sequence. *Bioinformatics*, 23(2):e13–16, Jan 2007.

[85] Florencio Pazos, Manuela Helmer-Citterich, Gabriele Ausiello, and Alfonso Valencia. Correlated mutations contain information about protein-protein interaction 11Edited by A. R. Fersht. *J. Mol. Biol.*, 271(4):511–523, Aug 1997.

[86] James R. Perkins, Ilhem Diboun, Benoit H. Dessailly, Jon G. Lees, and Christine Orengo. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure*, 18(10):1233–1243, Oct 2010.

[87] B. G. Pierce, K. Wiehe, H. Hwang, B. H. Kim, T. Vreven, and Z. Weng. ZDOCK Server: Interactive Docking Prediction of Protein-Protein Complexes and Symmetric Multimers. *Bioinformatics*, 30(12):1771–1773, Jun 2014.

[88] Tal Pupko, Rachel E. Bell, Itay Mayrose, Fabian Glaser, and Nir Ben-Tal. Rate4Site: an algorithmic tool for the identification of functional regions. *Bioinformatics*, 18(suppl_1):S71–S77, Jul 2002.

[89] Zhijun Qiu and Xicheng Wang. Prediction of protein–protein interaction sites using patch-based residue characterization. *J. Theor. Biol.*, 293:143–150, Jan 2012.

[90] Dana Reichmann, Ofer Rahat, Mati Cohen, Hani Neuvirth, and Gideon Schreiber. The molecular architecture of protein–protein binding sites. *Current Opinion in Structural Biology*, 17(1):67–76, February 2007.

[91] H. Ripoche, E. Laine, N. Ceres, and A. Carbone. JET2 Viewer: a database of predicted multiple, possibly overlapping, protein-protein interaction sites for PDB structures. *Nucleic Acids Res.*, 45(7):4278, Apr 2017.

[92] David W. Ritchie and Graham J. L. Kemp. Protein docking using spherical polar Fourier correlations. *Proteins Struct. Funct. Bioinf.*, 39(2):178–194, May 2000.

[93] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S. D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B. E. Begg, P. Braun, M. Brehme, M. P. Broly, A. R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B. J. Gutierrez, M. F. Hardy, M. Jin, S. Kang, R. Kiros, G. N. Lin, K. Luck, A. MacWilliams, J. Menche, R. R. Murray, A. Palagi, M. M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J. M. Sahalie, A. Scholz, A. A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A. O. Tejeda, S. A. Trigg, J. C. Twizere, K. Vega, J. Walsh, M. E. Cusick, Y. Xia, A. L. Barabasi, L. M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M. A. Calderwood, D. E. Hill, T. Hao, F. P. Roth, and M. Vidal. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226, Nov 2014.

[94] Alfred Russel. *On the Tendency of Varieties to Depart Indefinitely From the Original Type. Paper on natural selection sent by Wallace to Darwin.* Linnean Society, 1858.

[95] S. Sacquin-Mora, A. Carbone, and R. Lavery. Identification of Protein Interaction Partners and Protein-Protein Interaction Sites. *J. Mol. Biol.*, 382(5):1276–1289, Oct 2008.

[96] S. Sacquin-Mora, E. Laforet, and R. Lavery. Locating the active sites of enzymes using mechanical properties. *Proteins*, 67(2):350–359, May 2007.

[97] S. Sacquin-Mora and R. Lavery. Investigating the local flexibility of functional residues in hemoproteins. *Biophys. J.*, 90(8):2706–2717, Apr 2006.

[98] Christina E. M. Schindler, Sjoerd J. de Vries, and Martin Zacharias. iATTRACT: Simultaneous global and local interface optimization for protein–protein docking refinement. *Proteins Struct. Funct. Bioinf.*, 83(2):248–258, Feb 2015.

[99] Joan Segura, Pamela F. Jones, and Narcis Fernandez-Fuentes. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinf.*, 12(1):352, Dec 2011.

[100] Joan Segura, Pamela F. Jones, and Narcis Fernandez-Fuentes. A holistic in silico approach to predict functional sites in protein structures. *Bioinformatics*, 28(14):1845–1850, Jul 2012.

[101] J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.*, 5(6):729–731, Nov 1988.

[102] C.-W. Hong T. Iwai and P. Greil. Fast particle pair detection algorithm for particle simulations. *Int. J. Mod. Phys. C*, 10(05), July 1999.

[103] S. Tonddast-Navaei and J. Skolnick. Are protein-protein interfaces special regions on a protein's surface? *J Chem Phys*, 143(24):243149, Dec 2015.

[104] A. Tovchigrechko and I. A. Vakser. GRAMM-X Public Web Server For Protein-Protein Docking. *Nucleic Acids Res.*, 34(Web Server issue):W310–314, Jul 2006.

[105] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.*, 6(1):53–64, Jan 1997.

[106] Manoj Tyagi, Ratna R. Thangudu, Dachuan Zhang, Stephen H. Bryant, Thomas Madej, and Anna R. Panchenko. Homology Inference of Protein-Protein Interactions via Conserved Binding Sites. *PLoS One*, 7(1):e28896, Jan 2012.

[107] L. Vamparys, B. Laurent, A. Carbone, and S. Sacquin-Mora. Great interactions: How binding incorrect partners can teach us about protein recognition and function. *Proteins*, 84(10):1408–1421, Oct 2016.

[108] Thom Vreven, Iain H. Moal, Anna Vangone, Brian G. Pierce, Panagiotis L. Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A. Bates, Juan Fernandez-Recio, Alexandre M. J. J. Bonvin, and Zhiping Weng. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *J. Mol. Biol.*, 427(19):3031–3041, Sep 2015.

[109] C. Wang, P. Bradley, and D. Baker. Protein-protein docking with backbone flexibility. *J. Mol. Biol.*, 373(2):503–519, Oct 2007.

[110] Mark Nicholas Wass, Gloria Fuentes, Carles Pons, Florencio Pazos, and Alfonso Valencia. Towards the prediction of protein interaction partners using physical docking. *Mol. Syst. Biol.*, 7(1):469, Jan 2011.

[111] J. A. Wells and C. L. McClendon. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature*, 450(7172):1001–1009, Dec 2007.

[112] Li C. Xue, Drena Dobbs, Alexandre M.J.J. Bonvin, and Vasant Honavar. Computational prediction of protein interfaces: A review of data driven methods. *FEBS Letters*, 589(23):3516–3526, November 2015.

[113] Li C. Xue, Drena Dobbs, and Vasant Honavar. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinf.*, 12(1):244, Dec 2011.

[114] Changhui Yan, Drena Dobbs, and Vasant Honavar. A two-stage classifier for identification of protein–protein interface residues. *Bioinformatics*, 20(suppl_-1):i371–i378, Aug 2004.

[115] M. Zacharias. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.*, 12(6):1271–1282, Jun 2003.

[116] Martin Zacharias. Attract: protein–protein docking in capri using a reduced protein model. *Proteins: Structure, Function, and Bioinformatics*, 60(2):252–256, 2005.

[117] Hermann Zellner, Martin Staudigel, Thomas Trenner, Meik Bittkowski, Vincent Wolowski, Christian Icking, and Rainer Merkl. Prescont: Predicting protein-protein interfaces utilizing four residue properties. *Proteins Struct. Funct. Bioinf.*, 80(1):154–168, Oct 2011.

[118] L. Zhao and J. Chmielewski. Inhibiting protein-protein interactions using designed molecules. *Curr. Opin. Struct. Biol.*, 15(1):31–34, Feb 2005.

[119] Huan-Xiang Zhou and Yibing Shan. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins Struct. Funct. Bioinf.*, 44(3):336–343, Aug 2001.