Exploration and structural modelling of protein interactions using evolutionary information

Chloé Quignot

Thèse de doctorat

NNT : 2020UPASQ014

université PARIS-SACLAY

Exploration et modélisation structurale d'interactions protéiques guidées par l'information évolutive

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 569, Innovation thérapeutique : du fondamental à l'appliqué (ITFA)
Spécialité de doctorat: Biochimie et biologie structurale
Unité de recherche : Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.
Référent : Faculté de pharmacie

Thèse présentée et soutenue en visioconférence totale, le 11 Décembre 2020, par

Chloé QUIGNOT

Composition du Jury

Olivier LESPINET Professeur, Université Paris-Saclay	Président
Isabelle ANDRE Directrice de Recherche, CNRS, INSA	Rapporteur & Examinatrice
Jean-Christophe GELLY Maître de Conférence, HDR, Université de Paris	Rapporteur & Examinateur
Pablo CHACON Directeur de Recherche, Institute of Physical Chemistry « Rocasolano »	Examinateur
Annick DEJAEGERE Professeure, Université de Strasbourg	Examinatrice
Raphaël GUEROIS Directeur de Recherche, CEA Saclay	Directeur de thèse
Jessica ANDREANI Chargée de Recherche, CEA Saclay	Co-encadrante de thèse

Acknowledgements

Firstly and above all, I would like to thank both my supervisors, Jessica and Raphaël, for their patience, enthusiasm, encouragement, immense knowledge and continuous support throughout my thesis. I have really appreciated their invaluable advice and feedback whilst writing up this report and could not have had better supervisors. Together, they gave me the opportunity to further my masters research project and to extend my current thesis to December. Jessica, I really admire your organisational skills and efficiency! And Raphaël, I particularly enjoyed our many brainstorming sessions and discussions during our lunchtime runs (when I managed to keep up with you, that is)!

Of course, my PhD would not have been possible without the generous funding from the IDEX Paris-Saclay (IDI 2017). I am sincerely grateful for this.

I would like to thank the jury in advance for their patience, time and interest in my work and in particular, Pablo for his collaboration with the integration of our scoring function in his FRODOCK package. In that sense, I would also like to thank Pierre and his team at RPBS for their collaboration on updating our InterEvDock server, our two interns, Pierre and Merwan for their contributions to our team and projects as well as the teams that kept the many computing centres running that I had the honour to use during my PhD.

Many thanks to all my lab colleagues for accepting me as part of the team, their stimulating discussions and good humour shared during numerous coffee breaks and organised events. I will really miss our annual raclette sessions (even if the seminar room smelt like cheese for the following several weeks)! I also really enjoyed our insightful conversions in our regular journal clubs.

Special thanks goes to my former colleagues Seydou and Arun for pointing me along the path, for their insightful comments, pertinent questions and great advice. They both helped me integrate in the team and get accustomed to the team's many tools and

software (I'm still using *screen*, what a brilliant invention!). I am grateful to Arun for all his invaluable computer expertise (I would have been really stuck sometimes without his help!) and thank you so much for slowly introducing me to the CAPRI world! Both Arun and Seydou have become really good friends and I absolutely enjoy the numerous games weekends with Seydou, Damien and Jingqi that keep my brain ticking!

I am grateful to Hélène and Marie, my "rigid-buddy" teammates as Marie likes to call us, for their immense support and relaxing times and chats about everything and nothing! Even during corona times, we managed to keep in touch and just never seemed to run out of conversation topics.

I could not have completed this manuscript without the support of my best friends from the good old Masters days, Mel, Tiph and MaVi. I thank them for all the moral support and fun we've had in the last 5 years, their sympathetic ear and all the happy distractions that rested my mind (I wonder why I always think of food when I think of you?!). We'll have to fit in this famous roller session that we've been planning for a while now!

Last but not least, I would like to say a heartfelt thank you to my family. Thank you Mamie and Papy for your constant interest. I am grateful to my parents for making me who I am and believing in me, for their constant interest in my welfare and all their encouragement throughout my career and education. I could not have survived these last months without all the yummy home-made meals you prepared for me every weekend Mama! To finish, I would like to give a special thank you to my sister, my soulmate, for both her technical and moral support, her sunny smile and positive outlook throughout the whole thesis and my life in general.

Table of Contents

ACKNOWLEDGEMENTS ..... **I**
TABLE OF CONTENTS ..... **III**
TABLE OF FIGURES ..... **VII**
TABLE OF TABLES ..... **IX**
DICTIONARY OF ACRONYMS ..... **XI**
FRENCH SUMMARY ..... **XII**
CHAPTER 1 ..... **1**
1.1 Protein structure and protein interactions **6**
1.1.1 Protein composition 6
1.1.2 Hierarchical levels in protein structure 7
1.1.3 Acquiring protein structures 9
1.1.3.1 High-resolution techniques *9*
1.1.3.2 Complementary "low-resolution" techniques *12*
1.1.4 Protein databases 13
1.1.4.1 Sequence-related databases *13*
1.1.4.2 Structure-related databases *14*
1.1.5 Protein interactions and experimental detection methods 16
1.1.5.1 Characteristics of protein interfaces *16*
1.1.5.2 Experimental detection of protein interfaces *18*
1.1.6 Protein interaction databases 20
1.1.7 Protein networks 22
1.2 Protein evolution and co-evolution concepts **25**
1.2.1 Protein evolution 25
1.2.1.1 Mutations and epistasis *25*
1.2.1.2 Homology relationships *26*
1.2.1.3 Introduction to co-evolution *27*
1.2.2 Interface evolution 28
1.2.2.1 Tools to assess and score interface similarity *29*
1.2.2.2 Evolution of different interface regions *30*
1.2.2.3 Compensatory mutations in protein interfaces *31*

1.2.2.4 Insertions and deletions in protein interfaces *32*
1.2.2.5 Probing evolutionary paths of interface structures *33*
1.2.3 PPI network evolution 35
1.3 Computational structural prediction 40
1.3.1 Structural prediction of monomers 41
1.3.1.1 Homology modelling of individual protein structures *42*
1.3.1.2 Ab initio modelling of individual structures *43*
1.3.1.3 Evaluation of structure prediction methods for individual proteins *45*
1.3.2 Binding surface prediction 46
1.3.2.1 Conservation-based predictors *47*
1.3.2.2 Coevolution-based predictors *48*
1.3.2.3 Homology-based predictors *50*
1.3.2.4 Predicting binding modes in disordered regions using evolution *50*
1.3.3 Structural prediction of protein assemblies 53
1.3.3.1 Template-based docking *54*
1.3.3.2 Template-free docking *56*
1.3.3.3 Covariation-based interface structure prediction *66*
1.3.3.4 Evaluation *68*
1.4 Overview of this manuscript 74
CHAPTER 2 77
2.1 The InterEvDock2 server 85
2.1.1 Web interface 85
2.1.2 Molecular docking procedure 89
2.1.3 Docking from input sequences 89
2.1.4 User-defined constraints 90
2.1.5 Runtime 91
2.2 Results 92
2.2.1 Benchmarking on PPI4DOCK 92
2.2.2 Predictions of CAPRI targets 96
2.2.3 Description of docking case studies from input sequences and using constraints 98
2.3 Concluding remarks 100
CHAPTER 3 102

3.1 Methods 106
3.1.1 Docking benchmark 106
3.1.2 Scoring functions 106
3.1.2.1 Consensus scores 107
3.1.3 Homology-enriched docking pipeline 108
3.1.3.1 Subsampling homologs in the coMSAs 109
3.1.3.2 Threading models 110
3.2 Results 111
3.2.1 Consensus approach with implicit homology scoring 111
3.2.2 First steps towards an atomic version of InterEvScore 113
3.2.2.1 Atomic scoring without co-evolution 113
3.2.2.2 Adding co-evolution to the atomic InterEvScore 114
3.2.3 InterEvScore with explicitly modelled homologs 115
3.2.4 Homology-enriched SOAP-PP 116
3.2.5 Homology-enriched Rosetta interface score (ISC) 117
3.2.5.1 Using ISC to re-score homology-enriched decoys 118
3.2.6 Homology-enriched consensus scoring 119
3.3 Discussion 121
CHAPTER 4 125
4.1 Methods 133
4.1.1 Target preparation 133
4.1.2 Protein-protein docking challenge (T131-T132, T133, T136) 133
4.1.3 Protein-peptide docking challenge (T134-135) 134
4.2 Results 135
4.2.1 Protein-protein docking using *ab initio* free docking strategy (targets T131-T132) 136
4.2.1.1 Targets 131 & 132: Success of the InterEvDock server undermined by misleading biological information 137
4.2.2 Taking evolution into account in template-based docking strategies (targets T133, T136) 138
4.2.2.1 Target 133: Optimisation of an interface locally but drastically remodelled by design. 139

4.2.2.2 Target 136: Combining multi-domain and multi-subunit template-based modelling in a symmetric homomultimer. *139*

4.2.3 Evolutionarily conserved and recurrent structural motifs as guide for docking (targets T134-T135). 140

4.2.3.1 Targets 134 & 135: Evolution-driven recognition of Small Linear interaction Motifs in non-trivial cases. *141*

4.3 Discussion **143**

CHAPTER 5 ..... **145**

REFERENCES **153**

A. SUPPLEMENTARY MATERIALS FOR CHAPTER 1 ..... **167**

B. SUPPLEMENTARY MATERIALS FOR CHAPTER 2 ..... **170**

a. InterEvDock2 pipeline **170**

b. PPI4DOCK benchmark (Yu and Guerois 2016) **173**

c. Success rates for interface residue predictions **174**

d. Default constraint thresholds **176**

e. Performance according to sequence identity with PPI4DOCK template **177**

f. Performance comparison with the Weng benchmark **177**

C. SUPPLEMENTARY MATERIALS FOR CHAPTER 3 ..... **180**

a. Supplementary methods **180**

1. Docking parameters 180

2. Scoring functions 180

3. Details on coMSA calculation 181

4. Threading models 181

b. Supplementary results **182**

1. Supplementary tables 182

2. Supplementary figures 189

D. SUPPLEMENTARY MATERIALS FOR CHAPTER 4 ..... **192**

a. RosettaScript protocol for round CAPRI45 ..... **192**

Table of Figures

Figure 1-1: Protein structure – from primary to quaternary. 6
Figure 1-2: Yearly cumulative release of structures in the PDB for X-ray, NMR and cryo-EM methods. 10
Figure 1-3: Yeast-two-hybrid explained. 19
Figure 1-4: Schematic representation of protein sequence covariation. 27
Figure 1-5: Examples of binding compensations through multivalence. 35
Figure 1-6: Evolutionary fates of a duplicated gene pair in a protein interaction network. 37
Figure 1-7: Structure modelling flowchart when using evolutionary information. 41
Figure 1-8: Interface residue prediction of the RBBP5 protein using different programs. 48
Figure 1-9: Graphical summary of a selection of user-friendly methods used in structural protein interaction prediction. 53
Figure 1-10: Illustration of template-based docking. 55
Figure 1-11: Template-free docking pipeline. 57
Figure 1-12: Schematic representation of FFT and SFT. 60
Figure 1-13: CAPRI thresholds. 71
Figure 1-14: Scatter plot of the novel DockQ decoy evaluation criteria against IS-Score 72
Figure 2-1: InterEvDock2 pipeline. 86
Figure 2-2: Explanation of InterEvDock2 advanced options. 88
Figure 2-3: Venn diagram of prediction performances for the three scoring components in InterEvDock2. 95
Figure 2-4: Successful CAPRI target T95 prediction with InterEvDock2. 97
Figure 2-5: Successful prediction of a case in PPI4DOCK using InterEvDock2. 98
Figure 3-1: Docking pipeline with explicit modelling of decoy homologs. 109
Figure 3-2: Success rate as a function of the number of selected decoys for individual and consensus scores. 112
Figure 3-3: Performance of an atomic InterEvScore. 114
Figure 4-1: Representation of models predicted by free docking. 136

Figure 4-2: Template-based prediction of CAPRI targets T133 and T136. 138

Figure 4-3: Examples of recurrent anchoring patterns used to constrain docking models between the dynein light chain and its binding partner MAG (CAPRI target T134).....142

Figure B-1: Illustration of the residue prediction success rate of InterEvDock2, ZDOCK3.0.2 and a random reference. 175

Figure C-2: Venn diagrams between scores. 189

Figure C-3: Bootstrap performance distributions. 190

Table of Tables

Table 1-1: List of popular docking tools and their properties. 58

Table 2-1: InterEvDock2 performance on PPI4DOCK. 93

Table 3-1: Performance of consensus scores including InterEvScore implicit homology scoring. 111

Table 3-2: Performance of InterEvScore using coMSAs without or with threaded models. 116

Table 3-3: Performance of SOAP-PP against SPP-h[40]. 117

Table 3-4: Scoring performance of Rosetta homology-enriched ISC 118

Table 3-5: Performance of ISC and ISC-h[10] on 150 pre-selected decoys. 119

Table 3-6: Performance of homology-enriched consensus scores. 119

Table 4-1: Summary of CAPRI targets in rounds 42-45. 129

Table 4-2: Results for CAPRI targets in rounds 42-45. 135

Table 4-3: Assessment summary for our best submitted CAPRI targets 135

Table A-1: Links to web resources 167

Table B-2: InterEvDock2 performance according to target-template sequence identity in PPI4DOCK. 177

Table B-3: InterEvDock2 performance on 47 cases in common between PPI4DOCK and Weng benchmarks. 177

Table B-4: InterEvDock2 performance on 85 cases from the Weng benchmark. 178

Table C-5: List of the 752 docking cases used as a benchmark set in this study. 182

Table C-6: InterEvScore statistical potential 185

Table C-7: Scoring performance of homology-enriched SCORES. 185

Table C-8: Numbers and timescales (on one CPU) of various elements and programmes. 185

Table C-9: Top 1 and top 5 compared to top 10 success rates for consensus scores. 186

Table C-10: Performance of the repulsive term in Rosetta's score and ISC-h[10]/1k on the worst third or worst homologs 186

Table C-11: Performance over PPI4DOCK difficulty categories. 186

Table C-12: Performance with a more stringent near-native definition. 187

Table C-13: Performance in terms of top 150 nDCG. 187
Table C-14: Performance of consensus scores including InterEvScore implicit homology scoring. 188
Table C-15: Performances as reported in the InterEvDock2 paper. 188
Table C-16: Performances of InterEvScore with 2-body and 2/3-body potentials. 189

Dictionary of Acronyms

AP-MS (AC-MS)	Affinity Purification (Capture) – Mass Spectrometry
BioGRID	BIOlogical General Repository of Interaction Database
CAPRI	Critical Assessment of Prediction of Interactions
CASP	Critical Assessment of protein Structure Prediction
CATH	Class, Architecture, Topology and Homologous family
co-IP	co-Immunoprecipitation
cryo-EM	Cryogenic Electron Microscopy
DCA	Direct Coupling Analysis
DCG	Discounted Cumulative Gain
DMS	Deep Mutational Scanning
FCC	Fraction of Common Contacts
Fnat	Fraction of NATive contacts
HMM	Hidden Markov Model
IDR/IDP	Intrinsically Disordered Regions/Proteins
IED2	InterEvDock2
IES	InterEvScore
ISC	Rosetta's Interface Score
ITC	Isothermal Titration Calorimetry
MSA	Multiple Sequence Alignment
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PFAM	Protein FAMilies
PPI	Protein-Protein Interaction
PSSM	Position-Specific Scoring Matrix
RMSD	Root Mean Squared Deviation
SAXS	Small-Angle X-ray Scattering
SCA	Statistical Coupling Analysis
SCOP	Structural Classification Of Proteins
SPP	SOAP-PP
TAP	Tandem Affinity Purification
TBM	Template-Based Modelling
UniProt	UNIversal PROTein resource
XL-MS	Cross-link – mass spectrometry
Y2H	Yeast-two-hybrid

French summary

Les protéines et autres macromolécules jouent un rôle central dans une multitude de processus biologiques chez tous les êtres vivants. Leurs fonctions sont très diverses ; elles peuvent par exemple agir dans la défense immunitaire, jouer des rôles de messagers ou contribuer à la structuration de la cellule, ou orchestrer le transport et le stockage d'autres macromolécules ou leur dégradation. Les protéines agissent rarement seules ; de ce fait, l'étude de leurs interactions est primordiale pour mieux comprendre les mécanismes biologiques de la cellule. La structure tridimensionnelle de deux protéines en interaction peut nous donner une information précieuse sur leur façon de communiquer. Comme la détermination expérimentale de ces structures n'est pas toujours possible ou facile à mettre en œuvre, leur prédiction via des méthodes purement numériques/bioinformatiques, telles que l'amarrage moléculaire (plus connu sous le nom anglais de "free docking"), peut fournir une alternative utile dans l'étude de comment deux protéines (ou plus) interagissent.

Dans le free docking, nous générons de nombreux modèles d'interface possibles (étape d'échantillonnage) puis nous leur attribuons des scores afin de choisir les plus vraisemblables. Les critères de tri peuvent être basés sur des lois physiques, sur des règles statistiques ou sur l'information de (co-)conservation de certaines caractéristiques à l'interface. En effet, les protéines et leurs surfaces d'interaction sont souvent conservés dans différentes espèces car elles doivent maintenir leur(s) fonction(s) pour assurer la viabilité de la cellule. Les modes d'interaction (structures 3D du complexe protéique) sont également conservés et les surfaces moléculaires impliquées dans l'interaction présentent des traces de coévolution, c'est-à-dire de mutations corrélées permettant de maintenir le mode d'interaction. Cette information de conservation ou de coévolution peut donc s'avérer être très utile dans le choix de la (ou des) meilleure(s) prédiction(s).

Mon projet de thèse s'articule autour du développement et de l'amélioration de ces outils de prédiction, en particulier grâce à l'exploitation de l'information évolutive, une des thématiques phares de l'équipe. L'état de l'art en matière de méthodes prédictives et d'analyse s'appuyant sur l'information de conservation et de coévolution a été récemment résumé dans un article dont je suis co-auteur (Andreani, Quignot et al. 2020).

Dans le cadre de mon projet de thèse, j'ai participé à des développements majeurs de notre serveur de docking, InterEvDock2 (https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/). A partir des protéines fournies par l'utilisateur, InterEvDock2 propose 10 modèles d'interface les plus plausibles, sélectionnés en combinant des scores basés sur la physique, des potentiels statistiques et l'information co-évolutive. InterEvDock2 accepte aussi en entrée des structures oligomériques ou des séquences protéiques, pour lesquelles il peut automatiquement modéliser la structure monomère pour le docking. L'utilisateur peut également intégrer des connaissances a priori sur l'interaction sous la forme de contraintes sur les résidus ou paires de résidus afin d'éliminer toute solution non-pertinente. Le pipeline complet peut être exécuté de façon automatique ou plus contrôlée en utilisant des points d'arrêt stratégiques et/ou par ajustement de paramètres. J'ai validé les performances d'InterEvDock2 sur un large ensemble de 812 cas de docking hétérodimériques, pour lesquels les structures des complexes sont connues expérimentalement et les structures non-liées sont modélisées par homologie. InterEvDock2 a été capable de trouver une structure de complexe correcte dans 32 % de ces cas, ce qui représente une très bonne performance pour un pipeline automatique dans le domaine difficile de la prédiction structurale des complexes protéiques. La haute performance de ce serveur en matière de prédiction des résidus d'interface est très intéressante pour les biologistes souhaitant valider expérimentalement le mode d'assemblage prédit par des mutations, avec une probabilité de 75 % d'avoir au moins une prédiction correcte sur deux résidus prédits (un sur chaque partenaire). Ce travail a fait l'objet d'une publication dans l'édition serveur de *Nucleic Acids Research* (Quignot, Rey et al. 2018).

J'ai ensuite recherché un moyen plus explicite d'intégrer dans les fonctions de score l'information évolutive contenue dans les alignements de séquences. J'ai rendu cette information compatible avec l'utilisation de scores atomiques par la modélisation tridimensionnelle des interfaces homologues. En combinant cette approche avec le pipeline d'InterEvDock2, j'ai pu améliorer la performance prédictive de 32 à 40% sur notre large ensemble de cas tests. Ce travail a fait l'objet d'un pré-print dans BioRxiv et HAL (Quignot, Granger et al. 2020) et est en cours de publication. Les données et scripts ont été mis à disposition de la communauté (http://biodev.cea.fr/interevol/interevdata/).

De plus, durant ma thèse, j'ai pu participer à 14 défis de docking via le concours international de prédiction, CAPRI (Critical Assessment of Predicted Interactions), dont 4 cibles en Novembre 2020 impliquant des interactions entre les protéines humaines et celles du coronavirus SARS-Cov-2. Dans CAPRI, les équipes développant des méthodes de docking peuvent les tester à l'aveugle en prédisant les structures d'interactions protéine-protéine nouvellement résolues et pas encore publiées au moment de l'épreuve. La résolution de ces cibles très diverses et souvent difficiles, s'est faite par un important travail d'équipe. Les stratégies qui ont permis à notre équipe d'être classée première sur la période 2016-2019 ont été résumées dans une publication récente dans *Proteins* dont je suis co-auteur (Nadaradjane, Quignot et al. 2019).

Le travail effectué durant ma thèse vise à améliorer la prédiction structurale des interactions protéine-protéine dans leur ensemble afin d'aider les biologistes à étudier leurs protéines ou leurs voies biologiques d'intérêt. Dans un avenir proche, nous aimerions valoriser le travail sur l'intégration atomique de l'information évolutive et la forte augmentation associée des performances prédictives à travers une troisième version de notre serveur InterEvDock. Le travail de ma thèse constitue une étape vers l'objectif final de la prédiction des interactomes. L'intérêt croissant pour les techniques d'apprentissage automatique en biologie structurale et leur efficacité dans la prédiction de la structure des protéines laissent penser que des améliorations majeures pourraient

également être apportées à l'avenir en appliquant ces techniques au docking des protéines.

CHAPTER 1
Introduction

In this chapter, I will introduce basic concepts related to my PhD project. That is, I will give a brief introduction on proteins and their interactions, protein structure and ways to predict these structures using computational structural biology. I will put a particular emphasis on the use of evolutionary information in this task. This chapter is partly based on a review I co-authored published in May 2020, which focuses on the integration of evolutionary information in protein structure prediction in a user-orientated perspective (Andreani, Quignot et al. 2020). Parts of this chapter that are adapted from the review material include most protein and protein interaction evolution aspects in sections 1.2.1 and 1.2.2 as well as most evolution-guided prediction explanations and examples in sections 1.3.1.1, 1.3.2 and 1.3.3

Proteins and other macromolecules are central players in the myriad of cellular functions in all living organisms. Their functions are very diverse; they carry out important roles in the immune system, act as messengers or important structural components in the cell, or carry out transport and storage of other macromolecules or their degradation. Proteins mainly carry out their functions in networks, making the study of their interactions fundamentally important to probe and understand the mechanisms behind all the biological processes in the cell. The structure of interacting proteins can give us significant information on how they communicate and coordinate with each other. As the experimental determination of 3D complex structures is not always possible and can be too labour intensive, time consuming and/or costly, computational predictions of these interfaces with docking tools can provide a very helpful alternative or a complementary viewpoint to study how two (or more) proteins interact.

Protein complex structural prediction is a difficult problem to solve in most cases especially due to the inherent flexibility of proteins and the limited amount of experimental data to learn from. Protein-protein interactions (PPI) and the way they bind together are often conserved in many different species owing to functional constraints and share a common evolutionary history that may provide us with one or several structures (templates) to copy from, if these were previously resolved experimentally. Template-based docking is a computational method that makes use of this information to come up with reliable predictions. Unfortunately, as mentioned above, templates do not always exist and when they do, they are sometimes hard to identify and align with the interface we want to predict. An alternative computational method is free docking, which first involves a systematic search of the best complex structures. This usually generates a very large amount of possible solutions that have to be ranked and filtered efficiently according to one or several criteria, for example, the rules of physics, statistics, conservation information of individual protein interface features or co-evolution of these features.

The aim of my PhD project was to develop and improve docking tools, particularly by making use of co-evolution information, as it is one of the main focuses of our lab. I participated in the recent update of our free and automated molecular docking server InterEvDock2 to a

more complete and user-friendly version and validated its performance on a large set of test cases (Chapter 2). More recently, I have been working on improving the ranking step in docking by exploring a better and richer integration of co-evolution information (Chapter 3). Throughout my thesis, I have been able to participate and challenge myself in several blind testing rounds organised by the CAPRI community where labs such as ours can test the performance of their PPI prediction tools in real-case blind-test scenarios (Chapter 4).

Hereafter will follow a brief introduction to the concepts that I will use throughout this manuscript revolving around the importance of proteins and their interactions (section 1.1), their evolution and conservation within various species (section 1.2) and the benefit of knowing their structure and how it can be experimentally or computationally acquired (section 1.3). In a final section, I will outline the main points of my manuscript (section 1.4).

1.1 PROTEIN STRUCTURE AND PROTEIN INTERACTIONS

This section is dedicated to a brief introduction on protein composition and structure, the importance of studying their interactions and how they can be studied, ending with a brief overview of protein interaction networks.

1.1.1 Protein composition

Proteins are **polypeptides** made up of one or several chains of covalently linked amino acids (residues). Twenty different canonical amino acid types exist, unique by the composition of their side-chains, which also confers them specific physical chemical properties. Many possible classifications of amino acids exist, e.g. they can be grouped into four categories: apolar (Glycine, Alanine, Valine, Leucine, Isoleucine, Methionine, Phenylalanine, Tryptophan and Tyrosine), polar uncharged (Serine, Threonine, Cysteine, Proline, Asparagine and Glutamine), negatively charged (Aspartate and Glutamate) and positively charged (Arginine, Histidine and Lysine). Cysteine, Glycine and Proline are sometimes classified into a separate fifth group because of their special side-chain configuration (Cysteines are capable of forming covalent disulphide bonds whereas Glycine and Proline can both disrupt regular protein structure motifs).

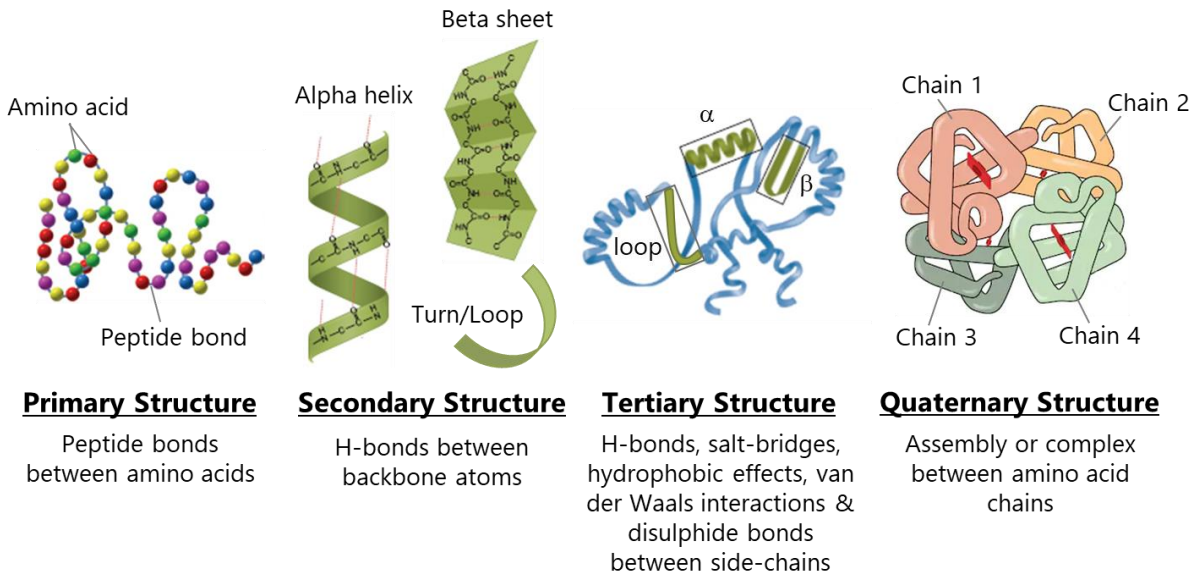


Figure 1-1: Protein structure – from primary to quaternary. *Protein structure has four hierarchical levels. Amino acids covalently link together with peptide bonds to form a chain of residues (Primary structure). This chain can fold into organised secondary structures such as alpha helices, beta-sheets or turns and loops*

through hydrogen bonds (H-bonds) between their backbone atoms. A more compact and stable structure arises when interactions between the side-chains of residues occur (Tertiary structure). Finally, proteins can form a quaternary structure when several chains are involved and interact together. This figure was adapted from Google Images.

1.1.2 Hierarchical levels in protein structure

The linear sequence of amino acids is called the protein's **primary sequence** (Figure 1-1) and is always listed from the N-terminus to the C-terminus (N-ter and C-ter; named after the amino and carboxyl groups of the first and last translated amino acid of the chain). Proteins fold into 3-dimensional (3D) objects while they are translated and the resulting shape is very much dependant on its amino acid composition. There are several levels to protein folding. Proteins can fold into regular **secondary structures** (Figure 1-1), namely α-helices, β-sheets or turns/loops in the 3-state classification system (Pauling, Corey et al. 1951), guided by hydrogen bonding (H-bonds) between backbone amino and carbonyl functional groups of two different residues. The 3-state classification is a good enough approximation in visual structural exploration but a more sophisticated and detailed classification is often required in the computational world for more accurate results. DSSP (Kabsch and Sander 1983) is an 8-state classification (three types of α-helices, two types of β-sheets and three types of turns/loops) and is considered a gold-standard in structural biology, and is used in programmes such as the hh-suite package (Steinegger, Meier et al. 2019) where DSSP assignments and/or predictions are used to better match protein sequence alignments or in SOAP-Loop to identify loops from protein structures (Dong, Fan et al. 2013).

More stable and compact **tertiary structures** occur when residue side-chains get involved with each other through H-bonds, salt bridges, disulphide bonds, hydrophobic effects and/or van der Waals interactions (Figure 1-1). Proteins can be composed of several well-packed globular domains linked by loosely structured or completely unstructured regions, called intrinsically disordered regions (IDR). In the cytosol, hydrophobic and van der Waals effects will tend to drive hydrophobic (apolar) residues towards the centre of the protein whereas hydrophilic (polar) or charged residues tend to be found on the surface where they can interact with the solvent. Unlike with secondary structures, domains of tertiary structure are more difficult to classify due to the large number of combinations that exist but can be

regrouped by general similarity of their secondary structure arrangements and/or following evolutionary relationships. SCOP (Structural Classification Of Proteins) (Andreeva, Kulesha et al. 2020) and CATH (Class, Architecture, Topology and Homologous superfamily) (Sillitoe, Dawson et al. 2019) and PFAM (Protein FAMilies) (El-Gebali, Mistry et al. 2019) are all examples of databases which perform domain classification and which will be described in more detail later on (section 1.1.3). The division of protein structures into domains is useful for an easier and more accurate structural and functional characterisation of proteins. The automatic identification of these domains remains a field of research in itself. Some algorithms predict domains from protein structure directly using "top-down" and/or "bottom-up" strategies (Guo, Xu et al. 2003, Zhou, Xue et al. 2007), of which SWORD provides an interesting multi-partitioning approach (Postic, Ghouzam et al. 2017). Other methods only use features deduced from the primary sequences, such as sequence profiles or secondary structure and accessibility predictions (Hong, Joo et al. 2019, Shi, Chen et al. 2019). Within these methods and in light of recent advances in the contact prediction field, FUpred uses an innovative strategy identifying domains based on contact map predictions (Zheng, Zhou et al. 2020).

The last structural level (**quaternary structure**) is formed when several chains of residues interact to form a multi-subunit structure, also known as protein assembly or protein complex. Depending on what proteins are implicated and how they cooperate in the cell, protein complexes can be given several labels. When two or more identical proteins interact, they form a homo-oligomeric complex; otherwise, they are classified as hetero-oligomers. According to interaction kinetics, complexes with a very short half-life (seconds or less) are labelled as transient whereas complexes that last minutes to hours are said to be permanent. By their short-lasting nature, transient complexes are harder to identify with certain experimental detection methods (section 1.1.5, page 16) but they are nevertheless very important in cells whenever a high turn-over is required (e.g. they might carry out post-translational modifications or participate in numerous cascades of reactions). Another existing terminology relies on the structural integrity of the subunits composing the complex: complexes are said to be obligate if the individual components of the complex cannot exist freely and functional in solution, and non-obligate otherwise.

Despite this rigid description of protein structure, it is important to note that proteins, just like other macromolecules, remain **dynamic objects** within the cell. Small changes constantly occur such as bond vibration and side-chain rotation and large conformational changes might come about when two proteins interact (Marsh, Teichmann et al. 2012) (see section 1.1.5.1, page 16). (Tripathi and Bankaitis 2017)

1.1.3 Acquiring protein structures

Structure and function are often correlated features in proteins. In most cases, proteins have to adopt their final structure (or lack thereof) in order to be able to carry out their cellular functions properly, often implying interactions with other proteins. For this reason, studying the structure of a protein or protein complex might give important insight as to what its function might be or might help elucidate why it does not perform as it should in certain pathologies, e.g. associated to mutations in the amino acid sequence. It might also help in the design of new drugs by discovering potential targets, their binding sites and how best to improve binding affinity.

In the following two subsections, I will detail three main **experimental methods** that can be used to decipher **high-resolution structures** of protein and protein complexes, namely X-ray crystallography, protein NMR (Nuclear Magnetic Resonance) and cryo-EM (cryogenic Electron Microscopy). I will also present methods that provide **complementary structural information**, especially for protein complexes, such as small-angle X-ray scattering (SAXS), cross-link - mass spectrometry (XL-MS) or deep mutational scanning (DMS). These approaches do not directly provide the structure of a protein or protein complex but provide enough information to model the protein structure when coupled with *in silico* techniques.

1.1.3.1 High-resolution techniques

X-ray crystallography is currently the most widely used technique, accounting for 90% of all structures in the Protein Data Bank (PDB (wwPDBconsortium 2019), Figure 1-2), whereas protein NMR and cryo-EM only occupy 7% and 2% respectively out of ~152,000 structures in total. In X-ray crystallography, data is collected by measuring the diffraction of an X-ray

beam, with a wavelength close to interatomic distances, on an ordered and regular sample. This results in a unique diffraction pattern from which we can deduce an electron density map and, finally, reconstruct the 3D structure and atomic detail of the sample by matching sample composition with map density. This challenging step can be simplified by basing the attribution on another similar structure when available (molecular replacement) – several workarounds exist when this is not the case. The quality of the end result is also highly dependent on the sample quality. Thus, another challenge lies within the preparation of the sample, which has to be properly crystallised into a regular and repetitive arrangement in order to create a clean diffraction pattern. Each protein has its own ideal and initially unknown crystallisation conditions that have to be optimised with a combinatorial approach. Nonetheless, the workload can nowadays be alleviated with the assistance of automated robotic systems. A non-negligible disadvantage in crystal structures is the risk of getting non-biological complexes formed between proteins that are, in reality, crystallisation artefacts.

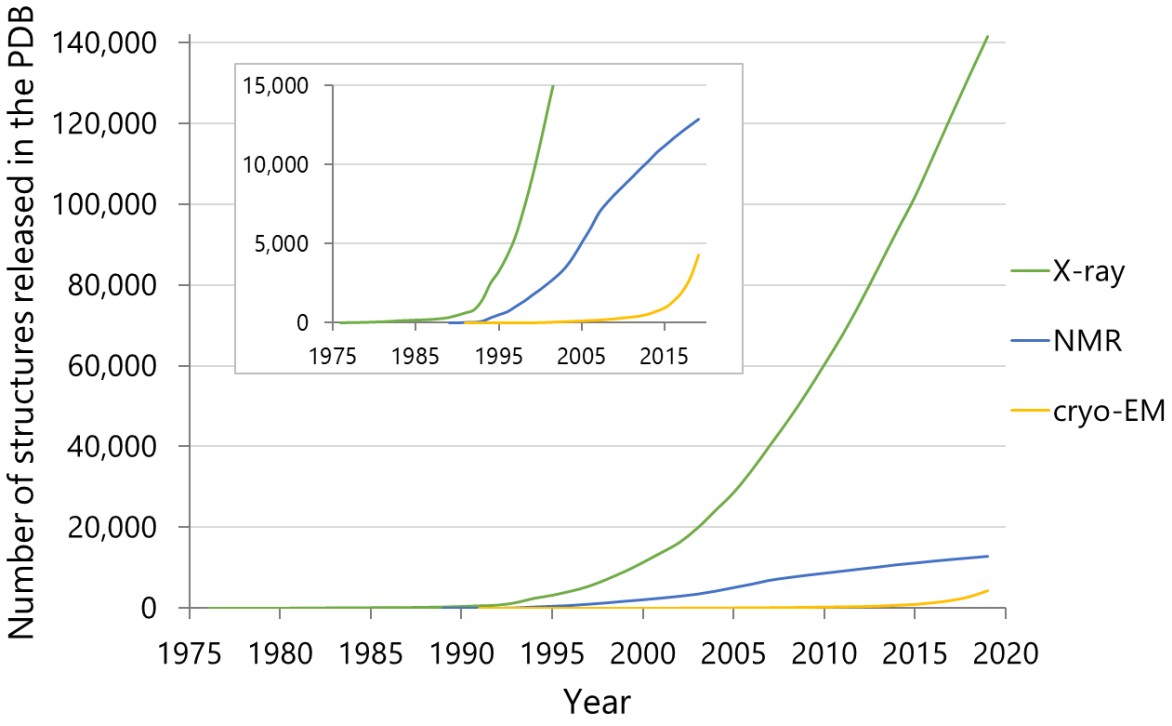


Figure 1-2: Yearly cumulative release of structures in the PDB for X-ray, NMR and cryo-EM methods. *Illustration of the cumulative number of structures released in the PDB every year since 1976 up to 2019 for the three main experimental methods of structural acquirement, namely X-ray crystallography (green), protein NMR (blue) and cryo-EM (yellow) with a zoomed-in version for a better visualisation of protein NMR and cryo-EM progress over the years.*

Although X-ray crystallography provides high-resolution structures, it also has the downside of presenting an out-of-context and rigid view of protein structure. In contrast, **solution NMR** allows proteins to stay in their physiological environment and can capture their dynamics. As in MRI (Magnetic resonance imaging), molecules are subjected to a powerful magnetic field in the face of which atoms behave differently according to their type and their neighbourhood. In protein NMR, this information can be used to deduce interactions and distances between atoms, which in turn can be used as constraints to fold 3D models that best fit the data. Protein NMR is an ideal contender for proteins that are difficult to crystallise such as intrinsically disordered proteins (IDP). Even though recent advances are pushing the upper molecular mass boundary, protein NMR remains best suited to relatively small proteins (Jiang and Kalodimos 2017). NMR can also be used to scan the interaction surface of a protein through the differential analysis of the chemical shifts between its bound and unbound states.

Cryo-EM has lately experienced a "resolution revolution" (Shoemaker and Ando 2018) and the yearly contribution of cryo-EM-resolved structures in the PDB has been increasing exponentially these last few years (Figure 1-2). This phenomenal jump in precision was possible thanks to recent technological advances (i.e. direct-electron detectors for less blurry images for instance) but also thanks to more powerful reconstruction algorithms (e.g. single-particle analysis with a Bayesian approach for parameter determination) (Bai, McMullan et al. 2015, Nakane, Kotecha et al. 2020, Yip, Fischer et al. 2020). In fact, cryo-EM recently managed to overtake the number of NMR depositions when looking at contributions on a yearly basis, a fact that is hidden when studying cumulative contributions. Cryo-EM is a type of transmission electron microscopy where the sample is frozen in solution and shone through by a beam of electrons. As for when a torch shines over an object, the proteins in the sample cast out a vast array of 2-dimentional (2D) "shadows" corresponding to a number of different orientations and from which algorithms can then deduce the 3D shape of the protein. Cryo-EM is especially popular for large and potentially more disordered molecular assemblies and is gradually closing the gap with X-ray crystallography in terms of structural resolution. For example, the best-resolved structure in the PDB using cryo-EM at the time of this manuscript

is the human apoferritin at 1.15 Å resolution deposited in August 2020 (7A6A) (Yip, Fischer et al. 2020). Cryo-EM is also less sensitive to sample purity and does not require as large an amount of proteins as the two traditional methods.

Each of these techniques have their own advantages and disadvantages. X-ray crystallography is the most commonly used and outputs structures with the best resolution but presents a very rigid view of protein structure. Solution NMR resolves proteins in their physiological environment but is often limited in protein size. Cryo-EM is getting increasing attention in the field thanks to recent advances in experimental and post-processing techniques. It enables the structural resolution of very large complexes with resolutions that are starting to compete with X-ray crystallography.

1.1.3.2 Complementary "low-resolution" techniques

SAXS provides a dynamic and low-resolution view of protein or protein complex size and shape and is compatible with a broad range of experimental conditions. After projecting an X-ray beam onto a sample, analysis of the resulting scattering pattern can help deduce the general shape of the macromolecules within it. In addition, combined with computational approaches, SAXS can provide structural models of protein-protein complexes at large scale (Xia, Mamonov et al. 2015, Jimenez-Garcia, Bernado et al. 2020).

In **XL-MS**, crosslinking reagents of constant size covalently bridge protein partners together making them more stable and thus easier to isolate and characterise. This approach can be performed *in vivo* as well *in vitro* and can be particularly interesting for short-lasting transient or weakly interacting protein complexes. Mass spectrometry analysis of the cross-linked peptides helps to identify the location of cross-linking sites on the proteins and as such, also the location of protein regions that are in close spacial proximity. As for SAXS, XL-MS data can be combined with computational tools to reconstruct the possible 3D structure of the protein complex (Orban-Nemeth, Beveridge et al. 2018).

Traditionally, protein interfaces are validated using one or several mutants, which are expensive to produce. With a **deep mutational scanning (DMS)** approach, mutagenesis can be

performed more systematically. DMS is a recent method in which high-throughput assays linking genotype to a measurable phenotypic property are coupled to next-generation sequencing in order to systematically quantify the effect of large numbers of mutations on biological systems (Fowler and Fields 2014). Even though challenging to decipher, DMS data can give valuable insights into protein structure and function. For example, important regions in proteins such as the hydrophobic core or the interface are expected to be more sensitive to mutations. In addition, co-varying positions might reflect spatial proximity in the 3D structure of a protein or protein complex and can be used to guide modelling approaches (Rollins, Brock et al. 2019, Schmiedel and Lehner 2019).

These "low resolution" techniques provide important information for modelling the structures of proteins and protein complexes. I will later discuss in section 1.3.3 how this type of information can be used as constraints e.g. in molecular docking.

1.1.4 Protein databases

Large-scale protein analysis is becoming increasingly popular thanks to new technologies and advances in protein science (e.g. whole genome projects, novel sequencing technologies, high-throughput assays). As such, impressive amounts of data are regularly generated, revolving around protein sequences, protein structures and their associated function. Various types of databases play an important role in centralising comprehensive resources of protein information. In this section, I will describe the two databases most commonly used for structural bioinformatics, UniProt and the PDB, which provide comprehensive data for protein sequences and structures respectively. I will also detail other databases, such as the NCBI Protein database for protein sequences and SCOP, CATH and PFAM for structures and their classification.

1.1.4.1 Sequence-related databases

One of the most widely used databases for protein sequence and functional information is the **Universal Protein Resource (UniProt) database** (UniProtConsortium 2019) providing

comprehensive non-redundant sequence data and regrouping human input with information from other databases. Each protein in UniProt is given a unique accession number which is becoming a gold standard in structural biology. UniProt is divided into four sections. UniProt Knowledgebase (UniProtKB) is the centrepiece of UniProt. UniProt Archive (UniParc) regroups all non-redundant protein sequences available with links to all underlying sources and versions of these sequences. UniProt Reference Clusters (UniRef) clusters sequences automatically across species according to different sequence identity thresholds and enables faster sequence search. Finally, UniProt Metagenomic and Environmental Sequences (UniMES) was specifically created to store metagenomic and environmental data directly recovered from environmental samples. UniProtKB is composed of two sub-sections called UniProt/Swiss-Prot and UniProt/TrEMBL (Translated EMBL Nucleotide Sequence Data Library) for manually annotated and reviewed data and automatically annotated data respectively, both listing over 563,000 and 195,000,000 proteins as of September 2020. UniProtKB provides data on protein sequence, name, taxonomy, structure, classifications, citations and cross-references by reliably fusing information taken from various databases.

The **National Centre for Biotechnology Information (NCBI) Protein database** (Ncbi Resource Coordinators 2018) is also famous but contains mainly raw data. This makes it noisier because of redundancy or contradictory or incorrect information but also possibly more enriched in information. Just like UniProt, NCBI protein records are stored with additional data (e.g. UniProt identifier, gene information, biological pathways and structure). Both UniProt and NCBI provide useful tools for protein manipulation and visualisation such as a BLAST homolog sequence search engines.

1.1.4.2 Structure-related databases

Protein structures are stored in databases for common use in the scientific community. Their classification can be especially useful for deducing protein function based on related experimental protein annotations. One of the most used international resources in structural biology is the **Protein Data Bank** (PDB, already discussed in section 1.1.3), which will celebrate its 50th anniversary in 2021 and is managed by the Worldwide PDB (wwPDB) consortium

(wwPDBconsortium 2019). The PDB stores not only 3D atomic protein coordinates but also those of nucleic acids and complexes with metals and small molecules as well as associated experimental data and metadata information(wwPDBconsortium 2019). As of September 2020, there are over 160,000 structures released in the PDB.

The **Structural Classification of Proteins (SCOP)** (Andreeva, Kulesha et al. 2020) and **CATH** (Sillitoe, Dawson et al. 2019) databases both provide a useful and reliable classification of proteins based on their structure and evolutionary relationships. Both are automatically as well as manually updated in order to minimise classification error. SCOP classification is hierarchical and based on protein domains. Domains are organised into classes, then folds, according to their tertiary structure similarities, then superfamilies and families according to their evolutionary history. One of the purposes of SCOP is to provide useful structural information to biologists that may be extrapolated to their own proteins of interest. Its classification is also widely used across computational biology tools and databases. In CATH, protein domains are clustered into Homologous superfamilies by evolutionary similarity and are given a Class, Architecture and Topology label according to their structural similarity.

The **Protein FAMily (PFAM)** database (El-Gebali, Mistry et al. 2019) heavily depends on multiple sequence alignments (MSAs) in order to classify protein domains and consequently provides a reliable MSA for each protein domain. MSAs are generated using hidden Markov models in a profile-sequence manner, meaning that a sequence profile, generated from an initial highly-reliable but small 'seed' alignment, is used to search for more remote homologs in order to build a more complete but trustworthy MSA. Related PFAM entries are grouped into clans. Together with structural information, entries are further tagged as domain, coiled-coil, disordered, motif or repeat and family if no clear subdivision can be made according to protein structure.

This is a non-exhaustive presentation of existing protein-related databases. A large amount of other protein-related databases exist (Xu 2012), focusing on other aspects of proteins such as protein-protein interactions (detailed in section 1.1.6), protein structure modelling, specific diseases or organisms, etc.

1.1.5 Protein interactions and experimental detection methods

When studying a protein, it is often important to find out its interacting partners in the cell. The term **protein-protein interaction** (PPI) can have several interpretations, from a loose definition of protein association without necessary physical contact (functional association) to a more stringent definition, where proteins have to be in **direct physical contact**. I will use the latter throughout this manuscript, as direct physical contact is especially important within our structural modelling goal. In this section, I will start with a brief overview of protein interface characteristics and then follow on with a description of a few main experimental methods to detect and study protein interactions.

1.1.5.1 Characteristics of protein interfaces

Biochemical analysis as well as the study of structural data from the PDB database provide essential information to identify the specific characteristics that define a protein-protein interface. Protein interfaces cover on average 1,200 to 5,000 $Å^2$ of the protein surface and contain on average 230 atomic contacts (about 61 residue-residue contacts) with an average of 2 salt-bridges, 9 hydrogen bonds involving side-chain atoms, and 35 apolar contacts according to a study on over 1,000 different interface structures (Andreani, Faure et al. 2012). Similar studies have also shown that interface stability can be attributed to only a few key interface residues at the interface, called **hotspots**, with a clear bias in composition towards tryptophan (21%) and arginine (13.3%) (Morrow and Zhang 2012). Interfaces can be divided into zones according to residue burial (Levy 2010) or contact count variation upon binding (Eames and Kortemme 2007). These zones also tend to have preferences in terms of composition, with mostly apolar residues in the core of the interface, providing the "stickiness" of the PPI, and polar and charged residues mostly on the rim, generally involved in interaction specificity.

Interface composition, together with interface size, is an important factor in protein **binding affinity**. Protein binding affinities vary a lot from one complex to another. They are usually expressed by the dissociation constant K_D (often in the nM to mM range with lower values

reflecting a higher affinity) or by the Gibbs free energy dissociation (ΔG). The prediction of these properties is an ongoing challenge in the bioinformatics world (Geng, Xue et al. 2019), partly due to complex phenomena such as structural rearrangements upon binding.

Not all proteins abide by a simple "lock and key" binding system, where the unbound states of both proteins are already complementary in shape and chemical composition. A large number of proteins bind with small to large **conformational changes,** following what is called the induced fit or conformational selection model (Csermely, Palotai et al. 2010). These are often terms that come up when studying enzyme-inhibitor complexes, which tend to co-evolve towards an interface with a high degree of surface complementarity (Tripathi and Bankaitis 2017). As can be deduced from the name, in the induced fit model, the ligand protein is thought to induce a conformational change in the receptor protein upon binding. In these processes, specific anchor residues at the interface might play an important role in stabilising the intermediate bound state in a similar way to boat anchors. An alternate and more popular model is the conformational selection model, in which the receptor is assumed to already exist as several conformations on its own and the ligand only tilts the balance towards the conformation that is best suited to its binding. Indeed, NMR studies show that conformations similar to the bound state already exist in the absence of the ligand.

Because of the crowded cellular context, true functional PPIs are in constant competition with non-relevant surfaces on other macromolecules. This leads to the question of **specificity** in PPIs. The interactions that govern interfaces are mostly carried out by direct interactions between residues of both partners but might also be bridged through water molecules (about 30% of an interface's contacts) (Rodier, Bahadur et al. 2005). Specific interfaces tend to be "dry" interfaces, where these water molecules form a ring around the interface core. Non-specific interactions, including those that arise from crystal packing, tend to have water molecules that permeate their interfaces and are thus called "wet" interfaces. And although most interfaces are driven by hotspot residues, a study in the context of protein docking concluded that the whole interface should be taken into account in order to better distinguish true interfaces rather than just the core elements (Nadalin and Carbone 2018).

1.1.5.2 Experimental detection of protein interfaces

Several experimental techniques exist to detect and study PPIs, some are classified as **high-throughput** as they help identify a large amount of potential protein partners at once and, thus, are often used as a first step in protein-protein interaction analysis. Commonly used approaches include yeast-two-hybrid (Y2H) and affinity purification/capture - mass spectrometry (AP-MS or AC-MS) screening. On the other hand, **low-throughput methods** enable the analysis of a protein or a complex of interest directly and as such are considered more reliable. Among these methods, Y2H, co-immunoprecipitation (co-IP) and isothermal titration calorimetry (ITC) are widely used. In this section, I will provide a brief description the principle behind each of these techniques.

Y2H was originally introduced to detect binary interactions in yeast. It makes use of the activation of downstream reporter genes, such as Gal4 in yeast, through binding of a transcription factor onto an upstream activating sequence in living yeast cells (Figure 1-3). This transcription factor is split into two separate parts, one binds the DNA (binding domain, BD) and the other activates the transcription (activation domain, AD). The transcription factor is fully functional only when both parts are close together. In Y2H, proteins are either called "bait" if fused with BD or "prey" if fused with AD. The idea is that when bait and prey physically interact, AD and BD are brought together leading to the transcription of the chosen reporter gene. The method has been adapted to an automated high-throughput screening strategy where several preys taken from a library of protein fragments or whole protein sequences are tested in individual cells. Identification of physical interacting partners can then be performed by PCR amplification and sequencing for example. Y2H has the advantage of being able to detect physiological (as *in vivo*) and weak PPIs and, above all, presents a strong indication of direct physical interaction. However, it can lead to a significant amount of false positives as it might detect non-specific interactions and, in the case of screening, interactions between proteins that might not usually coexist in time and/or in space (Bruckner, Polge et al. 2009). It also involves synthetic fusion of the bait and prey proteins which might affect the structural and functional integrity of the proteins.

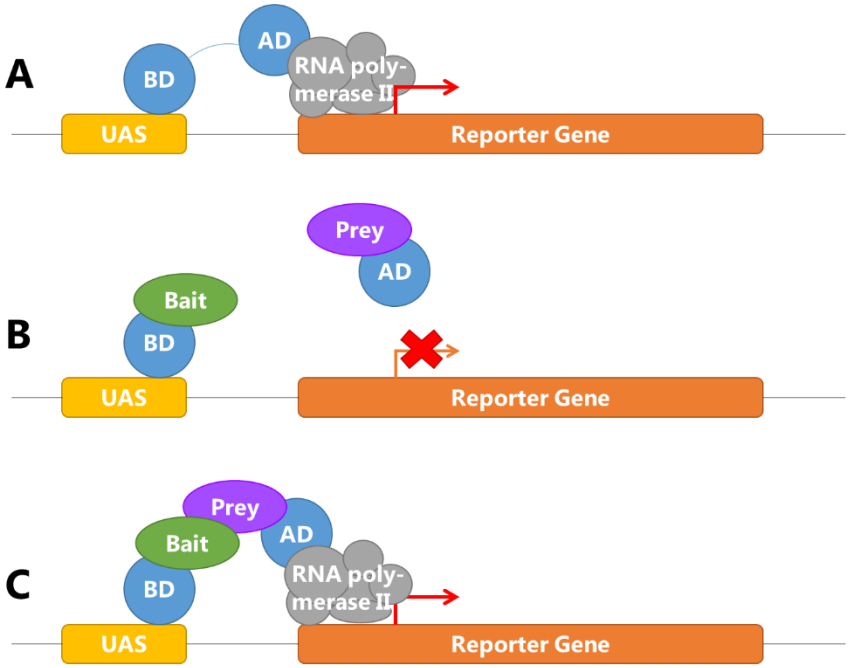


Figure 1-3: Yeast-two-hybrid explained. *Yeast-two-hybrid (Y2H) makes use of a reporter gene that is activated only if the DNA-binding domain (BD) and the activation domain (AD) of a transcription factor (in blue) are close together. The original complete transcription factor enables transcription of the reporter gene (A) by binding to the upstream activation sequence (UAS) on one end and recruiting the transcription machinery on the other. In Y2H, a bait protein (green) is fused to the BD and a prey protein (purple) to the AD. The interaction between bait and prey is confirmed only if the reporter gene is transcribed because bait and prey manage to bridge the gap between both parts of the transcription factor. (B) illustrates an example where there is no interaction and (C) where bait and prey do interact.*

AP-MS screening is another popular high-throughput approach in which cellular extracts containing tagged bait proteins are prepared then purified in order to retrieve all bait-interacting partner complexes. The purification is usually performed in two steps (tandem affinity purification, TAP) for cleaner results. Complex constituents are then characterised with mass spectrometry. The properties of this technique imply that indirect interacting partners (i.e partners that do not physically interact with the bait) are also eluted in the purification step.

Similar to affinity purification, **co-IP** relies on isolating a bait protein with a specific antibody and extracting with it all its potential direct or indirect interacting partners. These can then be identified through Western Blot.

Finally, **ITC** is a technique that relies on heat exchange measurements upon binding. As in a titration, increasing quantities of a protein A is added to a solution of protein B, all the while

measuring the temperature of the solution. ITC not only confirms protein-protein interactions but can also be used to deduce binding affinity, stoichiometry and other thermodynamic properties of the interaction without immobilisation or labelling of the proteins.

1.1.6 Protein interaction databases

As we just saw, large amounts of experimental information about protein interactions other than protein structure are regularly produced by a multitude of different techniques. Collecting and analysing this information constitutes a key step in constructing whole protein interaction networks. For an easier use of this data by the scientific community, it is important to centralise it in an intelligible and accessible fashion. Unlike for sequence and structural information, where UniProt and the PDB centralise the vast majority of the data for protein sequences and structures respectively, molecular interaction data curation is still mainly performed by numerous small-to-medium independent projects with different data-acquisition policies. Primary databases collect the data directly from peer-reviewed publications, meta-databases try to regroup information from several primary databases and predictive databases combine experimental information from primary databases with predictions of molecular interactions. Examples of common PPI databases, which I will describe in this section, include Biological General Repository for Interaction Datasets (BioGRID) (Oughtred, Stark et al. 2019), IntAct (Orchard, Ammari et al. 2014) and Search Tool for the Retrieval of Interacting Genes/Proteins (STRINGdb) (Szklarczyk, Morris et al. 2017).

BioGRID and **IntAct** are both primary databases. BioGRID centralises nearly 1.6 million protein and genetic interactions taken from scientific literature through controlled experimental vocabularies and text mining methods and from high-throughput datasets. In the same spirit, IntAct lists over a million binary PPIs collected from automatic deep literature curation or controlled and direct user submissions. Entries mainly focus on human or other main model organisms. BioGRID additionally includes post-translational modifications (PTMs) data as well as interactions between proteins or genes and small molecules. For each entry, both databases provide details on the experimental evidence supporting said data, references to corresponding publications, cross-references as well as tools for easier exploration

of the identified PPIs. Experimental evidence for physical interactions include high- and low-throughput methods (e.g. affinity purifications followed by different types of identification methods as well as Y2H, co-IP or co-crystallisation explained above).

The predictive database **STRINGdb** regroups known and predicted physical and functional PPI data for over 5,000 organisms as long as they are specific and biologically meaningful (Szklarczyk, Morris et al. 2017). Known interactions are collected and reassessed from high-throughput experimental PPI data, text-mining and various other curated databases. Predictions are highly-controlled and based on co-expression, genomic context and knowledge transfer between organisms. Thanks to these efficient methods, STRINGdb lists more than 2 billion protein-protein interactions. STRINGdb is known for its user-friendly interface and easy traceability with detailed explanation of evidence and associated quality estimate.

A systematic comparison of PPI databases can be found in (Bajpai, Davuluri et al. 2020). A multitude of more **disease- or organism-specific databases** also exists. For instance, the Online Mendelian Inheritance in Man (OMIM) database is a popular database that lists a number of human genes and genetic disorders and traits (McKusick-Nathans Institute of Genetic Medicine). The Human Protein Atlas also provides many sub-databases of proteins specific to certain human cells, tissues or organs as well as a Pathology Atlas regrouping human cancer-related mRNA and protein expression data (Thul, Akesson et al. 2017, Uhlen, Zhang et al. 2017). The Human Protein Reference Database (HPRD) (Keshava Prasad, Goel et al. 2009) lists most human proteins as well as their known PPIs, is manually curated and entries related to diseases are annotated and linked to OMIM. Host-pathogen interactions are a special type of interaction for which coevolution might be present through co-adaptation mechanisms between the two species (Woolhouse, Webster et al. 2002) HIV-1-human is an example of database hosted by NCBI regrouping virus-host PPIs (Ako-Adjei, Fu et al. 2015). Many more databases and web services exist to provide information about protein-protein interactions, from large, comprehensive databases (Miryala, Anbarasu et al. 2018) down to smaller databases focusing on specific interaction characteristics such as the structural details and energetics of protein interfaces (Gromiha, Yugandhar et al. 2017, Barradas-Bautista, Rosell et al. 2018). It is important to note that since experimentally-acquired data

only partially covers all existing proteins and PPIs, PPI databases present an incomplete view of PPI networks.

1.1.7 Protein networks

As mentioned earlier, proteins usually perform their function in groups, thereby forming an entire **network of interactions**. The whole set of these interactions in a particular context (e.g. in an organism or in a cell) constitutes a protein **interactome**. The richness and accessibility of PPI data accumulated over the years, for example through high-throughput assays such as Y2H or AP-MS described above (section 16, page 16), gene ontology or gene expression, enables the mapping of PPI networks. These networks can be particularly useful to predict the biological function of an uncharacterised protein by assuming that it has the same function as the proteins it clusters with ("guilt by association"). Protein networks follow the network modularity principle suggesting that highly connected groups of proteins constitute the building blocks of the network. These blocks indicate functional modules or protein complexes. Functional modules are made of proteins that participate in a same biological pathway but that might interact at different moments and places (e.g. transcriptosome, signalling cascades, cell-cycle regulation). Protein complexes, however, consist of proteins interacting at the same time and place thereby forming a single multimolecular machine (e.g. splicing machinery, transcription factors). PPI networks provide a **draft map** bringing together all the details centred on biological pathways of interest which might help elucidate the complex mechanisms that are behind them.

PPI networks can be analysed thanks to tools borrowed from mathematical network theory. A PPI network can be assimilated to a graph in which proteins are **nodes** and physical contacts are represented by **edges** between the nodes (Yamada and Bork 2009). Contacts are specific and serve a precise function. PPI networks have several properties. They follow the **small-world effect** meaning that the maximum number of connections separating any two proteins in the network is relatively small. A recent study on 12 different protein networks (7 eukaryotes and 5 prokaryotes), reconstructed through an extensive mining of the major PPI databases, showed that the average number of steps tended to be between 3.8 and 4.8, with

A. thaliana having an exceptionally high average path length of 8.5 (Xu, Bezakova et al. 2011). This high connectivity enables an efficient and rapid flow of signals within the system. PPI networks are also **scale-free networks**, meaning that most proteins only have a few partners (~5-6, a value termed the node degree in graph theory) whereas a very small number of proteins, called hubs, have over 100 connections. Party hubs are connected to many proteins at the same time and space whereas date hubs connect with their many partners at different times and spaces. The scale-free architecture of PPI networks enables individual paths to be switched on and off more easily and makes the network generally more **stable to perturbations** of single proteins. Indeed, when a random protein is disrupted, there is a higher chance of it affecting a protein with fewer connections than a hub due to their relevant frequencies, thereby limiting network disruption. Even when hubs are affected, other hubs are sometimes able to compensate for the lost connectivity. Hubs are important components of a protein network as they connect together groups of proteins that would otherwise be isolated from the rest of the interactome. Many cancer-related proteins are hub proteins, the tumour suppressor protein P53 being a famous example (Collavin, Lunardi et al. 2010).

Many tools exist to make the study of networks more accessible to users. Cytoscape (Shannon, Markiel et al. 2003) is a popular tool for network analysis for which there are several apps specific to PPI network analysis. It is important to note that the view that we have of protein interactomes today is usually **incomplete, noisy and quite often biased** – noisy, because of a large fraction of identified false positive and false negative complexes linked to the data-acquisition techniques used, and biased, simply because some proteins or pathways are preferentially studied. There is additionally a bias linked to the data-acquisition method as different approaches detect a largely complementary set of interactions. This highlights the importance of putting together data resulting from different experimental assays. An additional difficulty in studying PPI networks lies within their dynamic property as connections between proteins vary at any moment in time and are highly dependent on the cellular context.

Structural information is one of the many types of information about PPI that can be integrated into networks and serve PPI prediction. Structural modelling of whole PPI networks is only a long-term goal of computational structural biology.

1.2 PROTEIN EVOLUTION AND CO-EVOLUTION CONCEPTS

The evolution of proteins is linked to the evolution of the genome and as such, also to the evolution of the species. In this section, I will first introduce protein evolution at the level of individual proteins, then describe the evolution of protein interfaces and finally give an overview of PPI network evolution.

1.2.1 Protein evolution

Proteins evolved overtime by a series of successive changes affecting protein-encoding genes explaining their huge diversity and complexity observed today. Gene evolution includes four main evolutionary events, namely speciation, where new species are created followed by independent divergence of each species' genes; duplication of a gene within a same species also followed by independent divergence; gene loss; and horizontal gene transfer between species, a common process in prokaryotes (Kolodny, Pereyaslavets et al. 2013). Gene modifications imply mutations of nucleic acids. However, although there are potentially many to be made, only those providing a selective advantage (i.e. contributing to a better global fitness) or neutral mutations are kept overtime. A large part of these mutations is detrimental to cell survival (e.g. impaired gene transcription or loss of function or structure of the resulting protein).

1.2.1.1 Mutations and epistasis

As proteins evolve over long timescales and under constraints to maintain essential roles for the purpose of survival, complex phenomena arise such as epistasis, that is, the **context dependency** of the functional effect of mutations. Epistasis was first defined at the genetic level but also has strong molecular implications, since the structural organisation of proteins largely determines how mutations might interfere with one another (Starr and Thornton 2016). A mutation that appears neutral at a certain time may have consequences on the subsequent mutations that can be tolerated by the protein, and as such, molecular epistasis

can either constrain the evolution of proteins by barring subsequent mutations or make new evolutionary paths accessible through permissive mutations. Ancestral protein reconstruction enables to study the relationship between sequence, structure, dynamics and function of proteins (Johansson and Lindorff-Larsen 2018). In particular, it brings insights into the epistatic process and its capacity to drive changes in ligand binding specificity, but also to becloud the mechanisms by which proteins evolved (Siddiq, Hochberg et al. 2017).

1.2.1.2 Homology relationships

Through their common ancestry, proteins share homology relationships with each other. Two homologs within the same species are described as **paralogs** (from gene duplication) whereas homologs derived from a single ancestral gene in the last common ancestor of two different species are termed **orthologs** (from speciation or horizontal transfer) (Koonin 2005). Up to a certain point, homology relationships can be detected at the protein sequence level, the probability that two sequences share a high sequence identity only by chance being very slim. Very distant homologs are more difficult to identify as sequences are usually too diverged to detect similarity. Efficient algorithms relying on profile-profile sequence search exist to reliably detect homology (Steinegger, Meier et al. 2019) and can come in handy to identify suitable templates for 3D structural prediction (see section 1.3.1.1, page 42 and section 1.3.3.1, page 54). In that sense, when sequence identity is low (typically under 50%), orthologs are considered more reliable than paralogs since they are more likely to have a **similar biological function** – in fact, an initial definition of orthology was linked to function conservation and fitness (Koonin 2005) – and similar binding patterns in the case of complexes (Faure, Andreani et al. 2012). In contrast, paralogs are believed to often **differ in function** because they result from duplicated genes that both evolved independently with less evolutionary pressure to conserve function (as there are two copies). Paralogs might show different binding modes in the case of complexes and are commonly used to study function innovation.

IDPs or **IDRs** have less constraints on protein fold conservation than globular proteins or domains and as such, are less amenable to homology search, but they display **conserved**

molecular features such as length (e.g. when linking two globular domains), complexity (in terms of sequence motifs and repetition), amino acid composition or net charge. They might also present conserved key functional residues, important for posttranslational modifications or interactions with other proteins (Zarin, Strome et al. 2019).

1.2.1.3 Introduction to co-evolution

Protein function is dependent on protein structure, which in turn depends on protein sequence. In light of this, as protein function contributes to fitness, conservation of function implies structural constraints, which are directly observable at the protein sequence level through conserved or co-changing residues across homologs. Variations in amino acid types have to be correlated over time in order to maintain the same fold, so that if a mutation occurs, it can be **compensated** by complementary mutation(s) at different position(s) in the sequence. This phenomenon is referred to as **co-evolution** (Figure 1-4).

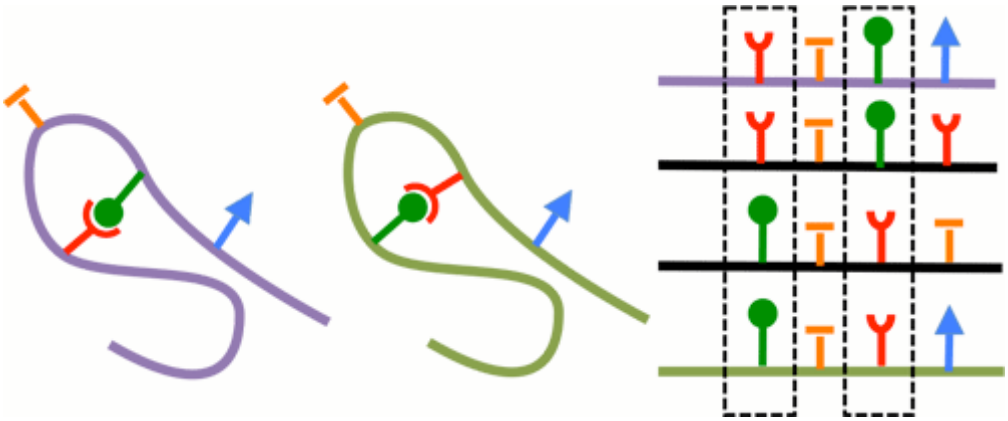


Figure 1-4: Schematic representation of protein sequence covariation. *Purple and green loops on the left represent a same protein fragment structure in two different homologous proteins, both present in the MSA on the left. The various shapes represent residues with different physical-chemical properties. In order to conserve the interaction between green and red residues, when one of them mutates, the other has to follow suit. This behaviour can be directly observed in the MSA. Thus, correlating or co-varying positions in the MSA point towards possible contacts in the 3D structure. Picture taken from GREMLIN's FAQ page (http://gremlin.bakerlab.org/gremlin_faq.php).*

The idea of detecting covariation to predict structural proximity between pairs of amino acid positions emerged thirty years ago and was successfully used at the time for RNA. Mutual information was long used as the primary method, but it suffers from large amounts of statistical and phylogenetic noise. Only within the last decade did significant breakthroughs occur following seminal work (Weigt, White et al. 2009, Marks, Colwell et al. 2011, Morcos,

Pagnani et al. 2011), thanks to methods derived from statistical physics as reviewed in (de Juan, Pazos et al. 2013, Cocco, Feinauer et al. 2018). Among the first successful methods, **statistical coupling analysis** (SCA) detects functionally related networks of residues, using principal component analysis to identify eigenvectors of the covariance matrix reweighted by site-specific conservation factors (Socolich, Lockless et al. 2005). **Direct coupling analysis** (DCA) generally designates a class of methods in which direct couplings between pairs of positions are disentangled from transitive correlations by global statistical analysis of an MSA. Many variants exist, based for example on sparse inversion of the covariance matrix (as in the PSICOV method (Jones, Buchan et al. 2012)) or inference using maximum likelihood estimation (Stein, Marks et al. 2015). DCA methods generally work best for large MSAs containing rich statistics about protein families. Covariation-based methods have applications ranging from protein and RNA structure prediction to the prediction of protein-protein interaction partners and the computational design of novel proteins (Simkovic, Ovchinnikov et al. 2017, Cocco, Feinauer et al. 2018). Most recently, meta predictors and integration of DCA-based predictions into deep learning pipelines gave the best results, leading to dramatic improvements in *ab initio* protein structure modelling (Kryshtafovych, Schwede et al. 2019) (section 1.3.1.2, page 43).

1.2.2 Interface evolution

A crucial question is how proteins maintain specific interactions in the crowded environment of the cell and along evolutionary pathways. Protein interfaces are **more conserved** on average than the remainder of the protein surface, due to the pressure to maintain functional interactions (Teichmann 2002, Mintseris and Weng 2005). This naturally leads to the question of how good sequence identity is as a proxy for the conservation of interface structure and function (Andreani and Guerois 2014). A first hierarchy can be outlined depending on interface stability: stable assemblies and core complexes are relatively robust to sequence perturbations, while transient interfaces and peripheral interactions may be more sensitive. Interactions mediated by short linear motifs (SLiMs), often weaker and more transient than interactions between globular domains, can be rapidly rewired in the context of regulatory

interactions. On the scale of the human interactome, domain-domain interactions form strongly connected modules, while interactions between domains and linear motifs are more likely to connect modules with different biological functions (Kim, Lee et al. 2014). **Interface conservation** thus depends not only on **stability** but also on **structural and functional classes** of protein interactions. To study the evolution of protein interfaces and their structure, we need tools to compare them in order to identify what is conserved in their global architecture, as well as more locally in their amino acid composition. Beyond conservation, we also aim to understand how interfaces have diversified. I will now present these different aspects.

1.2.2.1 Tools to assess and score interface similarity

The 3D structures of protein-protein interfaces need to be compared frequently, either to find out about their evolutionary properties or in the context of structural predictions. Interface structural comparison tools may typically provide three levels of information: the **structural alignment** between two protein interfaces; a **similarity score**, most often based on inter-protein distance comparison; and various **properties** that can be **inferred** from such comparisons, for example the prediction of protein-protein interactions based on similar known complexes.

Many programs exist for the structural alignment of interfaces. Among them, MM-Align (Mukherjee and Zhang 2009) and iAlign (Gao and Skolnick 2010) use dynamic programming to iteratively align interfaces based on inter-residue distance comparisons. **iAlign** can be used for more specific detection of homology relationships or to cluster interfaces based on interface geometry. iAlign also provides a useful interface similarity score called IS-score, which combines inter-protein geometric distance comparison with the evaluation of interface contact overlap (Gao and Skolnick 2010). Interface contact comparison is also used by the **FCC** (fraction of common contacts) method, which accelerates clustering of interface structures by circumventing the need for structural alignment (Rodrigues, Trellet et al. 2012). FCC also facilitates clustering of multi-protein assemblies by accounting for symmetries, which are frequent in structures of homo-multimers. This method is especially useful for

clustering many structures from molecular dynamics trajectories, simulations or protein-protein docking.

MM-Align is one of only a few methods that can process multi-protein assemblies with an arbitrary number of subunits. Subsequent multi-subunit comparison methods include TopMatch, which can compare large oligomers and molecular aggregates (Sippl and Wiederstein 2012), and QSalign, which focuses on identifying evolutionarily conserved quaternary structure states as the most biologically relevant (Dey, Ritchie et al. 2018). QSalign builds upon the fast protein structure alignment method Kpax, which performs fragment comparison followed by dynamic programming to build a global alignment providing structural superimposition (Ritchie, Ghoorah et al. 2012). The resulting QSbio database provides annotations of biological assemblies as well as predictions with an associated confidence measure. Most structural comparison approaches are very computationally intensive. To provide frequent updates, the **VAST+** service built on top of the MMDB database extends the simple heuristic strategy of the VAST protein structure comparison method to provide structural neighbourhood information on the level of biological assemblies (Madej, Lanczycki et al. 2014).

Some databases were designed to explore structural and evolutionary properties of interfaces. **ProtCID** contains clusters of similar interfaces between interacting PFAM domains found in different crystal forms (Xu and Dunbrack 2011). The **InterEvol** database was designed to jointly explore and compare the 3D structure and evolutionary properties of protein complexes in order to reveal the molecular details of interface coevolution (Faure, Andreani et al. 2012). In particular, InterEvol contains information about over a thousand pairs of structural interologs, that is, homologous complexes of known 3D structure sharing similar interfaces that can be used to probe how protein interfaces coevolve.

1.2.2.2 Evolution of different interface regions

It has been known for many years that not all mutated positions have comparable effects on the stability and affinity of protein interactions (Kastritis and Bonvin 2013). The structural organisation of interfaces is strongly connected to their evolutionary properties. The amount

of **accessible surface area buried** by an interface residue upon binding is traditionally the main descriptor related to both changes in affinity and rates of evolution (Chen, Sawyer et al. 2013) and **support** and **core** interface regions can be defined that are more **conserved** compared to the peripheral rim of the interface (Levy 2010). However, the non-interacting surface of proteins also plays a role in fine-tuning the binding affinity, mostly through charged and polar chemical properties that are conserved between orthologous complexes (Kastritis, Rodrigues et al. 2014). Regions outside the interface can also be determining for binding specificity, for instance in the small heat-shock protein family where paralogs avoid hetero-oligomerisation through conformational flexibility at non-interfacial regions (Hochberg, Shepherd et al. 2018). Special positions at the interface such as **hotspots** and **anchor** residues, which significantly contribute to the binding free energy, are also more evolutionarily **conserved** (Walker, Bond et al. 1999). Recent studies of human disease mutations show that although interfaces are robust to common sequence variations, a single missense mutation can have large functional impact, either by affecting protein folding or stability or by inducing local structural changes that disrupt interactions; perhaps unexpectedly, the latter mechanism was observed most frequently (Sahni, Yi et al. 2015, Fragoza, Das et al. 2019). These studies also conclude that a large fraction of disease mutations leading to interactome perturbation do so in an "edgetic" manner, that is, they only affect some specific interactions with a generally small subset of the protein's partners.

1.2.2.3 Compensatory mutations in protein interfaces

Similarly to covariation described earlier in protein monomers (see section 1.2.1, page 25 and Figure 1-4, page 27), protein interfaces must also adapt to mutations by coevolving, but not to maintain protein fold but rather to conserve protein function (i.e. the interaction). That is, when a mutation in one protein partner destabilises an interface, a **compensatory mutation** in the other partner can occur in order to **maintain the interaction**. As in conservation, the structural interface hierarchy plays an important role in coevolution. By analysing selection pressures in a large dataset of 896 protein complex structures, a recent study found that interface **core positions** show **higher conservation** and **coevolution** than those in the

rim and that both conservation and coevolution increase when residues are involved in **increasing numbers of interactions**, as these interactions jointly exert stronger selection pressures (Teppa, Zea et al. 2017). Systematic comparison of interface contacts in over a thousand pairs of homologous complex structures from the InterEvol database highlighted surprising plasticity, especially in polar contacts, while apolar patches and anchor residues display higher contact conservation, even in transient interfaces (Andreani, Faure et al. 2012).

1.2.2.4 Insertions and deletions in protein interfaces

Such detailed investigations about the mechanisms by which protein interactions evolve are crucial to understand how protein interactions can acquire the functionally required specificity. Binding selectivity is especially puzzling since the number of binding mode geometries seems to be limited: when interface regions were directly aligned using iAlign on the basis of backbone geometry and interface contact patterns, only around 1,000 distinct interface architectures could be identified (Gao and Skolnick 2010). As a step towards explaining this apparent dilemma, a 2010 study identified relatively **small insertions and deletions** (mostly up to 8 residues) that differentiate between monomers and homodimers, can **modulate oligomerisation** and most likely determine **interface specificity** (Hashimoto and Panchenko 2010). More recently, a similar conclusion was drawn for heteromeric complexes, which can achieve **evolutionary diversification** and **functional specificity** and avoid promiscuous interactions thanks to **interface "add-ons"**, typically 10-20 residues containing a high proportion of interface hotspots (Plach, Semmelmann et al. 2017). Most likely, insertion/deletion of these add-ons entail evolutionary routes going through promiscuous intermediates. Strikingly, systematic protein-protein interaction profiling for a large number of human alternatively spliced transcripts showed that alternative splicing is another major source of interactome expansion through the insertion/deletion of regions containing either globular domains or SLiMs able to mediate interactions (Yang, Coulombe-Huntington et al. 2016). Protein isoforms can thus display widely different interaction profiles. In the scope of modelling protein assemblies, being aware of the potential existence of these structural add-

ons in interface evolution may help interpret the existence of partly conserved inserted regions in the alignments.

1.2.2.5 Probing evolutionary paths of interface structures

Complementary to large-scale statistical investigations of how the structural organisation of interfaces relates to evolution, a number of detailed case studies have **experimentally probed** the mechanisms by which sequence variations can be accommodated at the interface of protein assemblies.

1.2.2.5.1 DMS of protein complexes

The use and analysis of **DMS** of protein complexes is one way to **disentangle** the complexity of interface **coevolution events**. DMS provides a systematic way of quantifying the effects of mutations through high-throughput assays coupled with next-generation sequencing (see section 1.1.3.2, page 12). Studies of DMS on interfaces can give insights into their robustness to mutagenesis and the evolutionary pathways used to rewire and expand specificity. They also highlight the mechanisms of interface coevolution, as over time mutations most likely occur one at a time and therefore intermediate states must be considered in which interface complementarity or specificity might not be optimal. Combined with structural biology, the DMS approach may provide a powerful means to **understand** the molecular bases underlying **epistatic phenomena** at complex interfaces. For example, one of the first deep mutational scans on interfaces was performed on a PDZ domain model system (McLaughlin, Poelwijk et al. 2012). Single and exhaustive mutagenesis of every position in PDZ distinguished positions tolerant to mutations from those functionally sensitive to substitutions, located around the ligand binding site (McLaughlin, Poelwijk et al. 2012). DMS studies coupled with impressive structural characterisation of PDZ variants enabled to identify a class-bridging but "conditionally neutral" mutation that was found to trigger epistasis by enabling conformational plasticity through a local structural change at the binding site.

1.2.2.5.2 Interface assembly pathways and symmetry

With the rise of protein complex structure determination, especially through recent developments in cryo-EM (section 1.1.3.1, page 9), the amount of information about multi-protein assemblies is increasing (Marsh and Teichmann 2015). A growing number of studies **investigate** the **dynamics** of such multi-protein assemblies and their **assembly pathways**. Recently, a study proposed a classification of protein complexes by number of subunit types and number of repeats and confirmed previous findings that many protein complexes assemble through **ordered pathways**, often evolutionarily **conserved** and reflecting evolutionary pathways (Ahnert, Marsh et al. 2015). Symmetry has also a special role in protein assemblies (Marsh and Teichmann 2015) as most homomeric complexes and many heteromeric complexes exhibit **symmetry**. A remarkable evolutionary feature in symmetric homomeric interfaces is the **multiplicative effect** of a single mutation through symmetry. Although often highlighted as an evolutionary advantage, it can also trigger uncontrolled self-assembly by amplifying the tendency of protein surfaces to interact by chance (Garcia-Seisdedos, Empereur-Mot et al. 2017), thereby creating a new interface only through a single mutation. The corresponding change in sequence might be conserved in only a few related species and the evolutionary trace might thus be difficult to recognise.

1.2.2.5.3 Multi-valence in large assemblies

Although multi-protein complexes involving multiple interfaces between different subunits are often conserved in evolution, the **binding affinity** of individual interfaces may **vary** largely between different species or different paralogous complexes in the same species. **Multi-valence** may globally **buffer** the loss or weakening of an elementary interface in a complex assembly. Such tolerance of interfaces to mutations may vary from one species to another, leading to altered evolutionary rates. Two examples of such evolutionarily-resilient but dynamic, regulatory complexes are the mismatch-repair related MutLβ complex between yMLH1 and yMLH2 with conserved meiotic helicase, yMER3 (Duroc, Kumar et al. 2017) (Figure 1-5A) and the mixed lineage leukaemia (MLL) family of histone methyltransferases interacting with four conserved factors (WDR5, RBBP5, ASH2L and DPY30) (Li, Han et al. 2016)

(Figure 1-5B). In the first example, the interaction between various components of these complexes have compensating strengths in mouse and in yeast. In the second example, paralogs of MLL show very different binding affinities although there are only two significant sequence substitutions between the two.

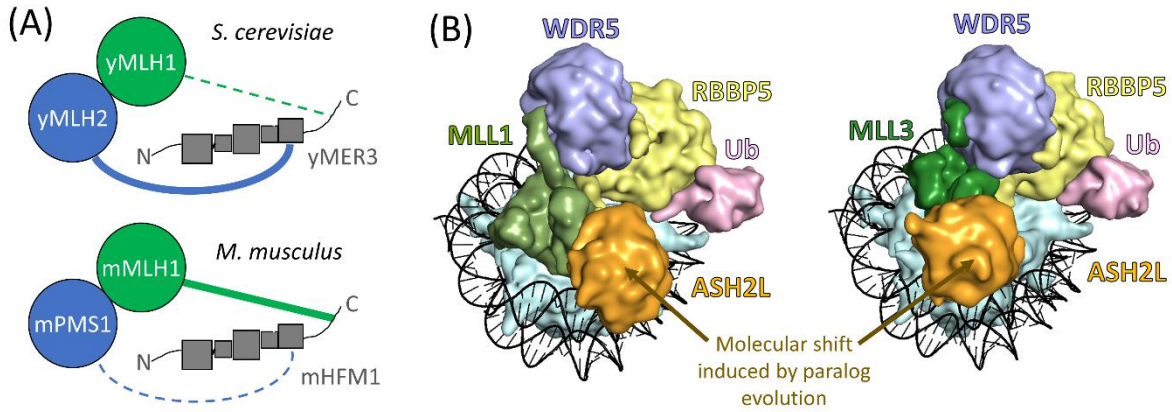


Figure 1-5: Examples of binding compensations through multivalence. *Schematic representation of the interaction networks between the yMLH1-yMLH2 heterodimer and yMER3 in yeast S. cerevisiae and between their mouse orthologs, the mMLH1-mPMS1 heterodimer and mHFM1. yMER3 and mHFM1 are composed of five globular domains represented by squares surrounded by disordered N-terminal and C-terminal extensions (indicated by "N" and "C" labels). The width of the links between each pair of proteins is indicative of the experimentally observed relative interaction strength. (B) Compared architecture of the MLL complexes involving either MLL1 (left, reference PDB structure: 6KIU) or MLL3 (right, reference PDB structure: 6KIW). WDR5 subunit is coloured purple, ASH2L is orange, MLL1 and MLL3 are two different shades of dark green, histone octamer is cyan, RBBP5 is yellow, ubiquitin is pink and DNA is black. Top views of the two complexes (with the nucleosome at the bottom) are provided where the nucleosomes and the RBBP5 subunits are exactly in the same orientation. Due to differences between MLL1 and MLL3, the relative positions of WDR5 and even more ASH2L are quite different between the two complexes even though the same overall architecture is maintained, providing a likely explanation for the large difference in binding affinity for RBBP5-ASH2L between MLL1 and MLL3. These differences in the details of the assembly reflect a different functional role for MLL1 compared to MLL3 and other MLLs.*

This highlights that evolutionary **conservation** can be used as a **guide** for **structural prediction** of protein assemblies but should not be strictly enforced, since variations that make one interface weaker can be counterbalanced by variations in interfaces between other pairs of subunits.

1.2.3 PPI network evolution

On a larger scale, protein-protein interactions form networks called **interactomes** (see section 1.1.7, page 22), which also change in the course of evolution. Many edges (interactions) are rewired, that is, some are gained and others are lost (Ghadie, Coulombe-Huntington et

al. 2018). This can happen either through modification of the interface or through gene loss or gene gain as a result of horizontal transfer, de novo emergence or duplication events followed by divergence leading to the expansion of the protein repertoire (Kolodny, Pereyaslavets et al. 2013) (see section 1.2.1, page 25). Protein interaction and protein function usually go hand in hand. This implies that PPIs undergo evolutionary constraints in order to conserve function. As mentioned in section 1.1.7 (page 20), our current view of protein interactomes is incomplete and biased. On the other hand, protein homology inferring methods that enable the mapping of orthologs and paralogs inter- and intra-species are not perfectly accurate leading to false homology assumptions. These technical details make the study between and within interactomes extremely challenging (Ratmann, Wiuf et al. 2009).

The retention or not of new protein copies and/or new interactions after gene duplication, may be more or less strongly influenced by selective pressure. Proteins that are born through duplication necessarily inherit at least part of their parent's interactions. These proteins have several fates as illustrated in Figure 1-6, the most common one being **pseudogenisation** (i.e. **loss of the copy** through too many detrimental mutations causing it to become a silent pseudo-gene). When the duplicate is kept, it might keep the same functions as the original copy, it might acquire a new function or both copies could grow dependent on each other to ensure the ancestral function. Studies have shown that the evolutionary rates of duplicated genes are accelerated in the period following duplication in yeast and was coupled with an apparent decrease in shared PPIs between the new paralogs (Ratmann, Wiuf et al. 2009).

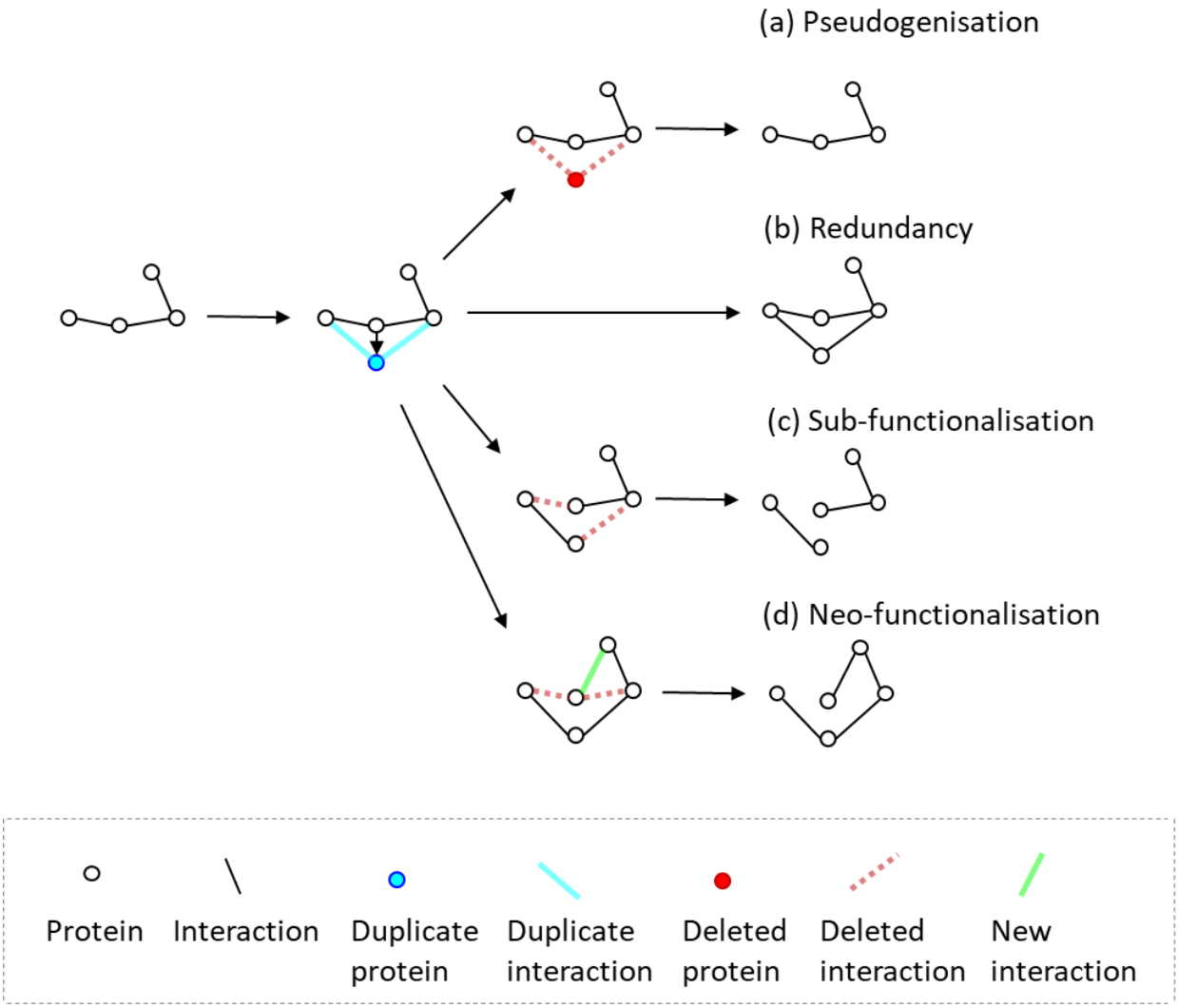


Figure 1-6: Evolutionary fates of a duplicated gene pair in a protein interaction network. *The duplicated gene might be lost through pseudogenisation (a) (the most common fate), or kept. In this case, if both copies are conserved, double-dosage might present a selective advantage (b). If both copies evolved individually and accumulated complementary deleterious mutations affecting different sub-functions, expression of both might be necessary to ensure ancestral function in a collaborative manner (c). Finally, less evolutionary pressure over each copy because of initial redundancy might be beneficial for the exploration and acquisition of a new function in one of the copies (d). Illustration adapted from (Ratmann, Wiuf et al. 2009).*

Another interesting and controversial topic is the link between network topology and network evolution, and more specifically, the relationship between evolutionary rate and **protein centrality** in the network. The conclusions seem to vary according to the data and methods used in various studies. Some studies showed that **protein degree** is **negatively correlated** to **evolutionary rate**; others argue that the observed correlation is an artefact of protein abundance differences. Based on the analysis of PPI structures involving hubs in yeast, (Kim, Lu et al. 2006) defined party hubs as having many interfaces, enabling to bind many partners at once, and date hubs as having only few interfaces, where PPIs would be

able to and would have to share the same interaction surfaces. Independent from gene expression level, they found that **party hubs** had **slower evolutionary rates** than date hubs. Intuitively, because hub proteins interact with more partners, they should proportionally have more surface dedicated to binding than less connected proteins, thus one could assume that they are subjected to higher evolutionary pressure. In that sense, we also have to make the distinction between date and party hubs. A study by Alvarez-Ponce *et al.* on high quality and close-to-complete human PPI networks concluded that network centrality had a significant effect on protein evolutionary rate, with a contribution comparable to that of gene expression (Alvarez-Ponce, Feyertag et al. 2017). They found, however, that **closeness** (i.e. one over the average distance between a protein and all other proteins in the network) was one of the highest contributors and that node degree had low or nearly no correlation with evolutionary rate after correcting for confounding factors. Thus, they hypothesise that **evolutionary rates** are **affected** by the **global position** of proteins in PPI networks rather than by surface constraints imposed by PPIs (Alvarez-Ponce, Feyertag et al. 2017).

An interesting concept is what Ghadie *et al.* call the "dispensable part of the interactome", characterising PPIs that are neutral to disruption in a PPI network. Based on homology-based three-dimensional structural models for PPIs in the human interactome and computational and experimental determination of mutation effects on these structures, they estimate that up to **~20%** of the overall human interactome is **completely dispensable** (Ghadie and Xia 2019). More information on PPI network topology evolution can be found in (Ratmann, Wiuf et al. 2009) or (Ghadie, Coulombe-Huntington et al. 2018).

In this section, I discussed a number of situations illustrating why evolutionary signals may be tricky to recognise in the context of protein assemblies and to exploit for their structural prediction. Depending on a variety of factors such as the local structural framework, the number of paralogs or the number of partners in an assembly, the consistency of a structural model with the evolutionary history of the interface may be difficult to establish. Next, I will present a state-of-the-art report of the successful methods for the prediction of individual protein structures and for the modelling of their assemblies. We will see in the following

sections that combining different computational approaches can help in getting the best from conservation and coevolutionary information.

1.3 COMPUTATIONAL STRUCTURAL PREDICTION

When a direct physical interaction exists between two proteins, more detailed knowledge of the specific interface structure is extremely valuable in order to modulate this interaction and understand its functional role, for example by suggesting positions on each partner that can be mutated to specifically disrupt and restore the interaction. If the complex structure has not been experimentally determined, modelling techniques can be used to obtain predictions for the assembly. Computational techniques for modelling protein complex structures are traditionally divided into **template-based** interface modelling and **template-free** docking methods (Soni and Madhusudhan 2017, Porter, Desta et al. 2019). Template-based approaches are the more accurate of the two but are only possible when a sufficiently close homologous structure exists (Figure 1-7, label 1). In many cases, a homologous complex structure cannot be identified, then one can resort to template-free docking. Docking requires the individual monomeric structures or models of these structures. Fortunately, experimentally resolved monomeric structures are more abundant than experimental structures of complexes and individual protein structures can often be modelled from structures of individual protein homologs if not available (i.e. by **homology** or **comparative modelling**, Figure 1-7, label 2). When monomeric template structures are not available, ***ab initio* modelling** can be performed (also Figure 1-7, label 2). Either of the methods mentioned above can be **guided** by additional data such as **biochemical data**, **conservation** or predicted **binding zones** (see Figure 1-7, label 3 and 4, 6 and 7). Here, I will first present the main principle and examples of tools for monomeric structure prediction. I will then describe briefly binding surface prediction. Finally, I will introduce in greater detail the issue of protein complex prediction central to my thesis, as well as its specific existing evaluation metrics.

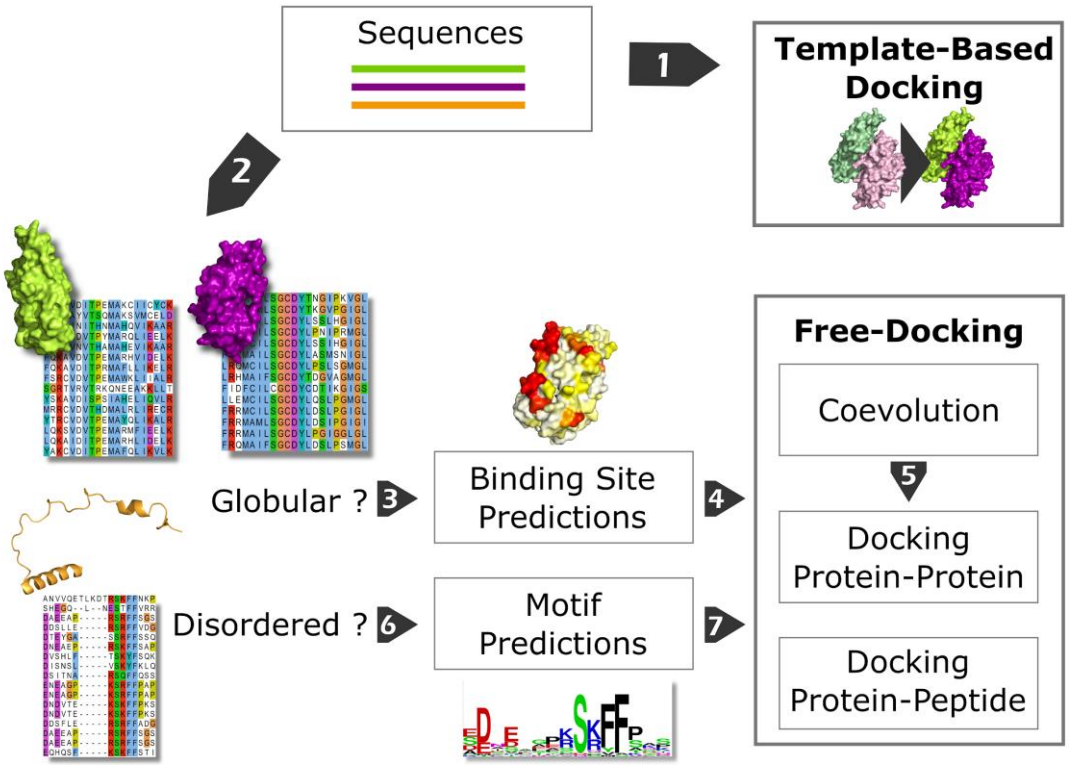


Figure 1-7: Structure modelling flowchart when using evolutionary information. *Flowchart of the protocols and tools described in the review to carry out structural modelling of protein interactions taking into account evolutionary information. When starting from the sequences of interacting proteins, structural modelling of their assembly can follow two strategies, both relying on evolutionary relationships. The first one (1), generally more accurate but restricted to a limited number of cases, uses homology relationships and template-based docking methods to generate structures of assemblies, which are reviewed in two subsections of this review for globular and disordered regions, respectively. The second strategy (2) relies on a combination of approaches involving structural modelling of the partners when possible, evolutionary analysis of the disordered regions and use of evolutionary information to identify binding patches at the surface of globular domains (3, 6). Combined with coevolution analyses, free docking methods can incorporate all these levels of information to produce models of assemblies (4, 5, and 7). These methods are reviewed for both globular and disordered systems.*

1.3.1 Structural prediction of monomers

As mentioned earlier, the monomeric structures of interacting proteins have to be known in order to proceed with docking. When experimental structures are not available, one can resort to monomeric structure prediction, whether through template-based approaches when homologous structures are available or by *ab initio* methods. These methods are constantly evaluated in blind-test scenarios such as regular Critical Assessment of protein Structure Prediction (CASP) sessions or through Continuous Automated Model EvaluatiOn (CAMEO) (see section 1.3.1.3 below).

1.3.1.1 Homology modelling of individual protein structures

Homology modelling is usually the **most accurate** modelling method. It makes use of evolutionary structure conservation and performs especially well when reliable templates are available. Templates should have good structural resolution and sequence identity typically above 30% with the query structure(s), in which case only local adjustments should be applied. The sequence identity threshold can sometimes be pushed down to as far as 15% thanks to effective homolog search algorithms. However, when more **remote templates** are available, **larger perturbations** are needed to make good models as modelling relies on a less reliable alignment due to low sequence identities, which might also contain more insertions and deletions. Template search can be performed efficiently and accurately using the profile-profile alignment toolkit hh-suite (Steinegger, Meier et al. 2019) against the PDB for example.

The main homology modelling pipeline used in my PhD project was **RosettaCM** (Song, DiMaio et al. 2013) which, given one or more templates and the corresponding query-template pairwise alignments, hands out one or more possible models. The full RosettaCM protocol is carried out in three main steps, an initial threading step where amino acids are simply replaced in the template structures by the corresponding query residues, a second step where missing regions and loops are completed, and a third side-chain and backbone optimisation step using a fast relax protocol. RosettaCM is also used by the **RaptorX-TBM** server (Xu and Wang 2019), one of the top-performing servers in recent CASP sessions, after an initial template search and alignment generation step using their DeepThreader algorithm.

Alternative examples of homology modelling tools include SWISS-MODEL (Waterhouse, Bertoni et al. 2018), MODELLER (Webb and Sali 2016) or I-TASSER (Yang, Yan et al. 2015). **SWISS-MODEL** (Waterhouse, Bertoni et al. 2018) is a widely-used and user-friendly homology modelling pipeline with various degrees of user intervention according to the chosen modelling mode. Templates are searched for using BLAST and HHblits and selected based on an estimated quality measurement. Models are then generated using an in-built

modelling engine and are given a quality score (QMEAN) reflecting how they compare to experimental structures of the same size, in other words, how realistic the models are. **MODELLER** (Webb and Sali 2016) uses its own alignment builders. Its modelling is performed iteratively and is guided by spatial restraints deduced empirically based on the identified or given template structure(s) and CHARMM force field terms. MODELLER is often used in other programmes for when modelling of monomeric structures is needed (Mirabello and Wallner 2017, Zimmermann, Stephens et al. 2018). **I-TASSER** is within the top-performing template-based servers in recent CASP evaluations. I-TASSER searches for suitable templates using a profile-profile sequence search, and then generates its models through template threading and free modelling of non-matching regions followed by a reassembling step with Monte Carlo sampling.

Recent advances in the field observed since the beginning of CASP are linked to the development of more accurate sequence-template alignment generation methods, the use of multiple templates, better modelling of non-template-covered regions, better final model refinement and better quality estimation to select the final output models (Kryshtafovych, Schwede et al. 2019).

However, homology modelling is not always possible as it relies on having available suitable templates. Indeed, Interactome3D (Mosca, Céol et al. 2013) lists more than 62,000 proteins involved in interactions, of which about 44% have no experimental monomeric structure and no readily identifiable template structure. Fortunately, for these cases, we can still resort to *ab initio* modelling of individual subunits.

1.3.1.2 *Ab initio* modelling of individual structures

Ab initio modelling, i.e. predicting a protein structure from its sequence only, is much more **challenging**. Decades of effort have been dedicated to methods trying to achieve protein structural prediction from physical principles. Molecular Dynamics (MD), a computer simulation technique widely used to study protein dynamics and conformational changes, could theoretically be used to fold proteins (Lindorff-Larsen, Piana et al. 2011), although its computational cost and imperfect force fields currently limit this application (Geng, Chen et al.

2019). Threading (Xu, Jiao et al. 2008) and fragment-based methods, although not strictly *ab initio*, draw on the knowledge of protein structures and the limited number of folds that proteins seem to adopt (Zhang and Skolnick 2005).

Recent advances enabled the generation of high-accuracy models by drawing on (co-)evolutionary information. A first performance boost was observed thanks to the introduction of **predicted contacts** in the modelling pipeline. The second boost came about by implementing **deep neural network methods** coupled with the prediction of inter-residue **distance** and backbone torsion **angle distributions** and is the secret behind the incredible success of RaptorX (Xu and Wang 2019) and AlphaFold (Senior, Evans et al. 2020) in recent CASP rounds (details about CASP are given in section 1.3.1.3 below). In response to AlphaFold's success in CASP13, trRosetta (Yang, Anishchenko et al. 2020) recently made its appearance, implemented in an AlphaFold-inspired fashion but additionally integrating inter-residue orientation predictions (i.e. dihedral angle predictions between non-covalently bound atoms). Although they did not participate in CASP13, a posteriori results on the CASP13 dataset and on CAMEO show that this extra feature additionally increases the success of structural prediction.

RaptorX, **AlphaFold** and **trRosetta** are able to completely bypass template structures thanks to a deep and convolutional residual neural network architecture (**ResNet**) and through integration of **evolutionary information** taken from MSAs. They all predict distance distributions (baptised "distograms" in AlphaFold for distance histograms) as well as backbone torsion angle distributions. RaptorX additionally predicts secondary structures and trRosetta additionally predicts torsion angles between residues ("anglegrams"). Distograms and all angle predictions can then be used for 3D model reconstruction after translating them into **structural constraints**. Both AlphaFold and trRosetta convert their predicted contact information into smoothed restraints that are used in gradient descent and Rosetta protocols, respectively, and RaptorX uses Crystallography and NMR System (CNS) to predict protein folds.

The input and output of these contact prediction methods might be similar to the DCA approach but they use different workflows. Thanks to convolution layers, the networks used in AlphaFold, RaptorX and trRosetta are able to take in the **global context** of the contact map and go beyond classical pairwise relationships that are mainly detected in DCA, thus enabling them to capture possible **structural motifs**. Additionally, ResNets provide a highly **non-linear model** as opposed to DCA methods, which are mainly linear and they also need **less sequences** in the MSAs to detect useable information. However, ResNets suffer from the **"black box" effect**, like any neural network, meaning that prediction performances might generally be improved using these methods but our understanding of how proteins are successfully folded and what information is used and learnt by these algorithms remains partial.

1.3.1.3 Evaluation of structure prediction methods for individual proteins

CASP is an international **blind-test challenge** for protein fold and protein contact map prediction that occurs every two years since its creation in 1994 and that assesses the state-of-the-art methods in that field (Kryshtafovych, Schwede et al. 2019). In CASP, groups are able to test their methods on targets that have not yet been published at the time of the challenge. Over the years, CASP has accumulated a large set of targets with their associated predictions proposed by various participating groups, which has now become invaluable for method developments and assessments. Targets are separated into several categories according to the availability of templates or biological data (e.g. X-link, NMR or SAXS). Participants can predict contacts or suggest structural models or can restrict themselves to estimating the quality of models generated by other groups.

Unlike CASP, **CAMEO** (Haas, Barbato et al. 2018) provides a **continuous** and fully automated assessment dataset based on weekly pre-releases of sequences in the PDB, meaning that structures are not available at the moment of the prediction. CAMEO offers a maximum of 20 targets per week that are cautiously selected in order to remove any proteins that are too close in sequence to already existing structures in the PDB. As participants can only compare performances if they happen to be predicting CAMEO targets simultaneously, CAMEO also

continuously runs base-line prediction and assessment tools that can be seen as "null models" for easier comparison between sessions.

Benchmarking datasets such as **CulledPDB** (Wang and Dunbrack 2005) or **ProteinNet** (AlQuraishi 2019) complete the assessment landscape. CulledPDB is based on the PDB whereas ProteinNet draws on CASP. ProteinNet was designed with a special focus on emerging machine learning techniques. It provides an additional effective validation set distinct from the official CASP sets along with evolutionary profile information as well as file formats directly compatible with machine learning approaches.

Assessment measurements for protein models usually include the **GDT-TS** (Global Distance Test – Total Score) which measures the similarity between the model and the experimental structure upon superposition of both. GDT-TS can be assimilated to RMSD but is less sensitive to outliers (e.g. poorly modelled loops). GDT-TS calculates the largest set of Cα atoms falling within a defined distance threshold with the reference structure, thus, the higher the score, the better the performance. In CASP, the regularly used metric is an average over GDT-TS results for 1, 2, 4 and 8 Å distance cut-offs (Kryshtafovych, Schwede et al. 2019). Another metric is the local distance difference test (**lDDT**), a superposition-independent score based on inter-atomic distance deviations in the model compared to the reference structure. In CAMEO, an average lDDT is used over four different deviation thresholds (0.5, 1, 2 and 4 Å) and higher scores represent better agreement with the reference model (Haas, Barbato et al. 2018).

1.3.2 Binding surface prediction

A first step towards studying PPIs or predicting their structure might lie in the prediction of their respective binding regions. This information can be used directly by biologists to feed more specific assays or can be integrated into docking methods in order to guide the prediction of protein complex structures as illustrated in Figure 1-7, steps 3 and 4 on page 41. The vast **majority** of binding surface predictors **include evolutionary information** taken from MSAs or homologous structures. Indeed, due to the evolutionary pressures mentioned

above, conservation of amino-acid positions in MSAs provides key guidelines to identify functionally important residues.

A large number of available predictors were extensively described and compared in recent reviews (Aumentado-Armstrong, Istrate et al. 2015, Maheshwari and Brylinski 2015, Esmaielbeiki, Krawczyk et al. 2016). A global ranking of available tools remains difficult due to the variety of training and test datasets and to the metrics used to compare each other. Overall, most recent approaches tend to perform as well or slightly better than standard methods such as the SVM-based **SPPIDER** tool (Porollo and Meller 2007) (using version II with 3D structure) by measures of precision, recall, true positive and false positive rates.

1.3.2.1 Conservation-based predictors

One of the most sensitive tools for spotting out evolutionary constraints is the **Rate4Site** algorithm (Pupko, Bell et al. 2002, Mayrose, Graur et al. 2004), which can be run from the ConSurf web server (Ashkenazy, Abadi et al. 2016). The relative evolutionary rates at each site are estimated using a probabilistic evolutionary model, which takes into account the stochastic process underlying sequence evolution within protein families and the phylogenetic tree of the proteins in the family. As illustrated in panel A of Figure 1-8, most of the interaction regions on the RBBP5 subunit in the multi-protein MLL1 complex associated with the nucleosome are characterised by slower evolutionary rates as measured by Rate4Site. Other conserved regions may correspond to alternative interactions or functional constraints.

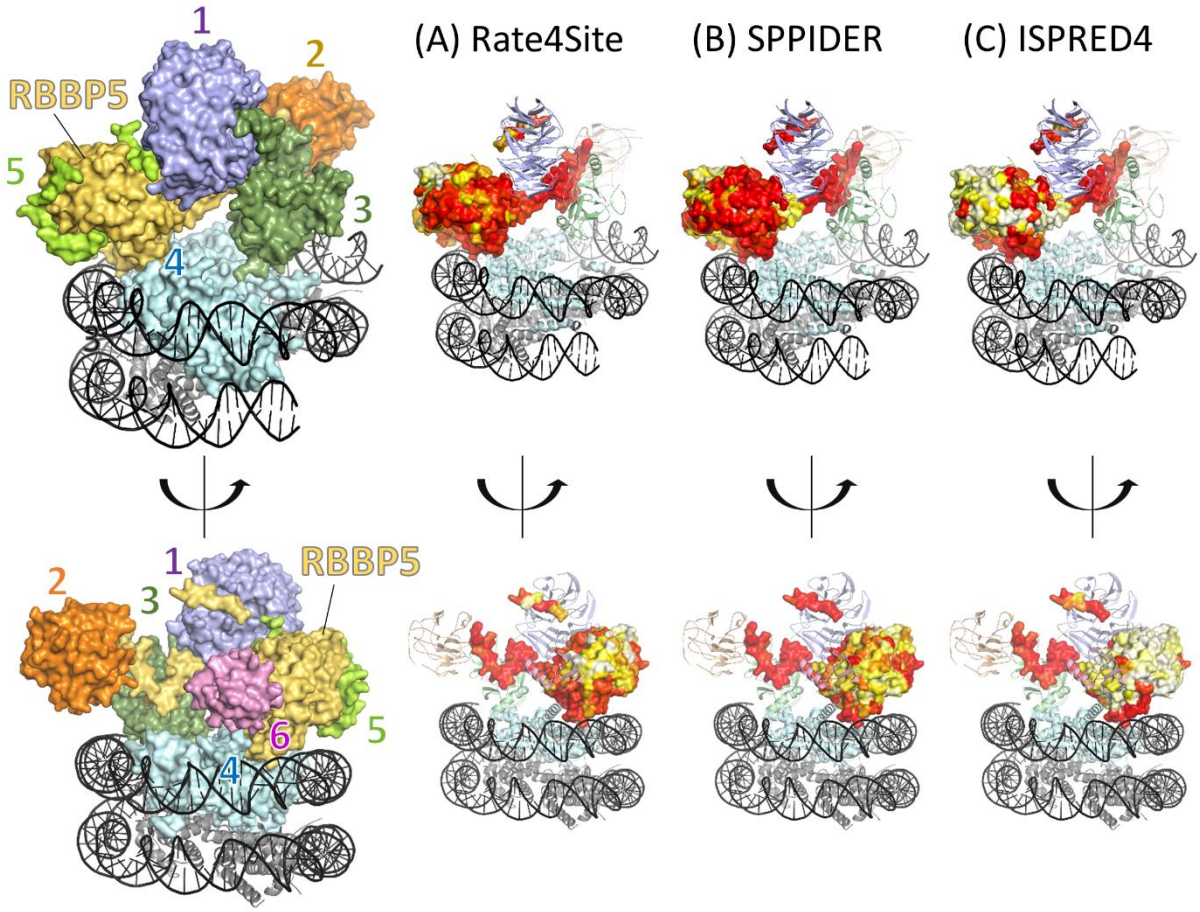


Figure 1-8: Interface residue prediction of the RBBP5 protein using different programs. *(A) Rate4Site, (B) SPPIDER and (C) ISPRED4. Interface residue predictions are displayed on the surface of RBBP5 and colour-coded from white (predicted as non-interface) to yellow to red (highest predicted probability to be involved in an interface). The RBBP5 subunit is involved in interfaces with six different partners (five proteins and one DNA) in the MLL1 complex associated with the nucleosome (interfaces 1, 2, 3, 4 and 6 in reference PDB structure: 6KIU) and in one intra-molecular interaction (interface 5 in reference PDB structure: 6KM7). These interfaces are mediated either by its globular beta-propeller domain (interfaces 1, 4, 5 and 6) or by its N-terminal intrinsically disordered region (interfaces 1, 2 and 3). ISPRED4 prediction exhibits remarkable sensitivity in the detection of interface residues in RBBP5 for all seven interfaces with almost no false positives. As in Figure 1B, WDR5 subunit is coloured purple (1), ASH2L is orange (2), MLL1 is dark green (3), histones octamer is cyan (4), RBBP5 is yellow and lime (5), ubiquitin is pink (6) and DNA is black.*

Most recently developed protein interface prediction methods take conservation into account (Pei and Grishin 2001, Hou, De Geest et al. 2017, Northey, Barešic et al. 2017, Savojardo, Fariselli et al. 2017, Meyer, Beltran et al. 2018, Dequeker, Laine et al. 2019, Sanchez-Garcia, Sorzano et al. 2019, Wang, Yu et al. 2019).

1.3.2.2 Coevolution-based predictors

Beyond conservation features, coevolution scores are also increasingly used either within a single sequence or considering two binding partners. In the **ISPRED4** method (Savojardo,

Fariselli et al. 2017), conservation and coevolution features of each individual protein partner were integrated among other sequence and structure-based descriptors. The rationale behind using coevolution of positions within a single sequence to account for residues in interaction is that neighbouring residues at the surface may co-vary more strongly due to the presence of a bound partner. The increase in **performance** obtained with the covariation score was actually of **similar magnitude** to the contribution of the conservation term. Overall, the performance of ISPRED4 was found significantly higher than all other methods tested on a standard benchmark and on a blind test set from the CAPRI experiment. The example of the RBBP5 subunit in the MLL1 complex (Figure 1-8) illustrates the quality of predictions that can be obtained. For all six binding interfaces in which RBBP5 is involved, residues involved in the interface were correctly spotted out by ISPRED4 without too many false positives.

Coevolution-based scores are also increasingly integrated in interface residue predictors by considering not only single proteins as for ISPRED4, but also pairs of binding partners. Such a strategy can potentially increase the specificity of predictions as originally shown by the development of the **i-Patch** predictor (Hamer, Luo et al. 2010). The **ECLAIR** method (Meyer, Beltran et al. 2018) was designed to predict interfaces at a genomic scale to feed the Interactome INSIDER browser using different features including conservation and coevolution between specific partners. Interestingly, by using DCA and SCA scores as descriptors to account for correlations between interacting positions, the authors observed that the performance of the classifier was increased even when the MSA contained less than 200 sequences. So far, other methods taking into account pairs of interacting proteins rather than single ones, such as **BIPSPI** (Sanchez-Garcia, Sorzano et al. 2019) or **PAIRpred** (Minhas, Geiss et al. 2014), rather used pairs of PSSMs in their descriptors. Future progress in the field will probably come from further integration of these coevolution signals with machine learning algorithms.

1.3.2.3 Homology-based predictors

In the preceding paragraphs, I mainly described integration of evolutionary properties from protein sequences, but homology information can also be extracted from comparison of structures. In that respect, the **PredUS** server (Zhang, Petrey et al. 2010, Hwang, Petrey et al. 2016) provides a complementary view of how homology can help predict interface residues by identifying structural neighbours of a query protein and mapping the frequency of contacts made by binding partners of these structural neighbours. A related method is **PS-HomPPI**, which predicts interface residues between two query proteins based on their frequency at the interface of homologous complexes with known 3D structures (Xue, Dobbs et al. 2011). Such tools integrating structural homology as features for interface prediction were recently reviewed (Xue, Dobbs et al. 2015). Their success rate is high provided numerous structures exist for a given structural family. These homology-based predictors are different from template-based docking strategies, discussed in the next section, in that most do not account for binding partner specificity to predict binding site location.

For practical applications, using a selection or a combination of the different available predictors should be envisioned, following the concept used in consensus approaches such as **CPORT** (de Vries and Bonvin 2011). The choice of tools and parameters also depends on the type of application considered. To increase the chances of success when selecting a small number of residues that will be experimentally mutated in order to perturb an interface, the precision metric should rather be favoured. In contrast, if interface prediction is used with the aim of generating constraints for subsequent docking (as described in the section 1.3.3.2.1, page 57), a higher recall would be advisable to ensure that none of the potential regions of interaction are omitted in targeted sampling, since further scoring and clustering of the candidate interfaces will be used to refine solutions.

1.3.2.4 Predicting binding modes in disordered regions using evolution

An important class of protein-protein interactions, only briefly mentioned so far, are those mediated by IDPs and by exposed flexible loops within folded domains. Their binding generally involves short stretches of adjacent amino acids forming compact clusters known as

SLiMs or **molecular recognition features** (MoRFs) (Van Roey, Uyar et al. 2014). These sequence motifs play fundamental roles in cell functions such as signalling, transport or protein turnover and are involved in many human diseases (Uyar, Weatheritt et al. 2014, Via, Uyar et al. 2015). They trigger **transient** and **reversible interactions** between partners and are often regulated by post-translational modifications. In vivo, these interactions often act in the context of complex multiprotein assemblies as illustrated in Figure 1-8 between RBBP5 and partners 1, 2 and 3. Moderate binding stability of these motifs together with the complexity of the biological context in which they act may hinder evolutionary traces used to spot them out. As noted earlier, the low complexity of linear binding motifs can give rise to complicated compensatory mechanisms in evolution difficult to decrypt from sequence analyses. Here, I will provide a few guidelines and tools that can help predict binding sites in disordered regions and in their folded partners, illustrated as steps 5 and 6 in Figure 1-7, page 41. Docking methods suitable for incorporating these features for modelling purposes will be discussed in section 1.3.3, page 53.

1.3.2.4.1 Prediction of binding motifs in disordered regions

If the conditions for closely related template-based modelling are not met, a first challenge can be the identification of binding regions in disordered stretches (Figure 1-7, label 6). As a first approach, well-annotated binding motifs can be recognised using databases such as the **Eukaryotic Linear Motif resource** (ELM) (Kumar, Gouw et al. 2019), a repository of manually curated and experimentally validated motifs. In cases where no known binding motifs can be found, more general approaches can be used to search for stretches with a tendency to fold upon binding. For instance, the **IUPred2A** server predicts disordered regions in proteins (Meszaros, Erdos et al. 2018) and uses ANCHOR2 to predict binding stretches within these regions. ANCHOR2 uses a biophysics-based model but it does not incorporate any evolutionary constraints (Meszaros, Simon et al. 2009). In fact, using evolutionary information for the recognition of binding motifs in disordered regions requires to pay particular attention to the quality of the generated multiple sequence alignment (Gibson, Dinkel et al.

2015). Rather than automatic tools, it is advisable to use more interactive approaches allowing to tune sequence divergence and prune those with low quality, in order to increase the contrast between the most conserved positions and the highly variable sequence tracts in which they are generally encompassed. Interactive manipulation can be performed, for example, with the **Jalview** sequence analysis workbench (Waterhouse, Procter et al. 2009), or with the **ProViz** visualisation server to investigate evolutionary features in protein sequences (Jehl, Manguy et al. 2016), which maps pre-calculated MSAs across different clades to useful information sources such as ANCHOR2 and secondary structure predictions. Recent target T134 from the CAPRI 7th edition (Lensink, Nadzirin et al. 2019) typically addressed the question of recognising a local motif inside a large disordered sequence stretch capable of binding the dynein domain (see Chapter 4, page 125).

1.3.2.4.2 Prediction of sites in folded domains binding disordered motifs

On the side of the folded domains involved in the recognition of short binding motifs, identification of binding sites can be performed using tools previously mentioned for interface predictions, which generally include evolutionary information. Interaction sites can also be predicted using the **PEP-SiteFinder server** (Saladin, Rey et al. 2014) which generates 3D de novo conformations of peptides based on their sequence and performs a fast blind rigid docking of these conformations on the complete protein surface to map the most favourable binding sites. A more homology-based strategy is also proposed in the **InterPep** pipeline (Johansson-Akhe, Mirabello et al. 2019) which uses distant protein complex structures as structural templates for the identification of residues likely involved in binding flexible peptides. InterPep includes a conservation score among other features and was shown to outperform alternative approaches on a 502-target benchmark, based only on the Rate4Site conservation score. As when predicting binding motifs in disordered regions, care should be taken in the generation of sequence alignments for folded domains when dealing with protein-peptide interactions. Target T66 from CAPRI 6th edition is illustrative of such potential pitfalls (Yu, Andreani et al. 2017). It involved the disordered C-terminal tail of *B. subtilis* SSB protein in complex with a primase. Even though the interaction is conserved from *B. subtilis*

to *E. coli*, large MSAs sampling homologs around these two distant species did not allow the detection of a conserved evolutionary trace at the surface of the primase. Probably due to a spatial switch in the location of the binding site over long evolutionary times, only alignments restricted to closely related Firmicutes species could help identify the binding site using evolutionary information.

1.3.3 Structural prediction of protein assemblies

Once the monomeric structures of our proteins are available (whether experimentally or through prediction as described in section 1.3.1, page 41), one can proceed with protein assembly prediction (Figure 1-10, page 55 and Figure 1-11, page 57). Similarly to monomeric structures, complex structures can be deduced by homology if a suitable template is available or by free-docking otherwise. Figure 1-9 provides a graphical summary of a selection of user-friendly methods that can be used in order to determine the structure of a protein interaction.

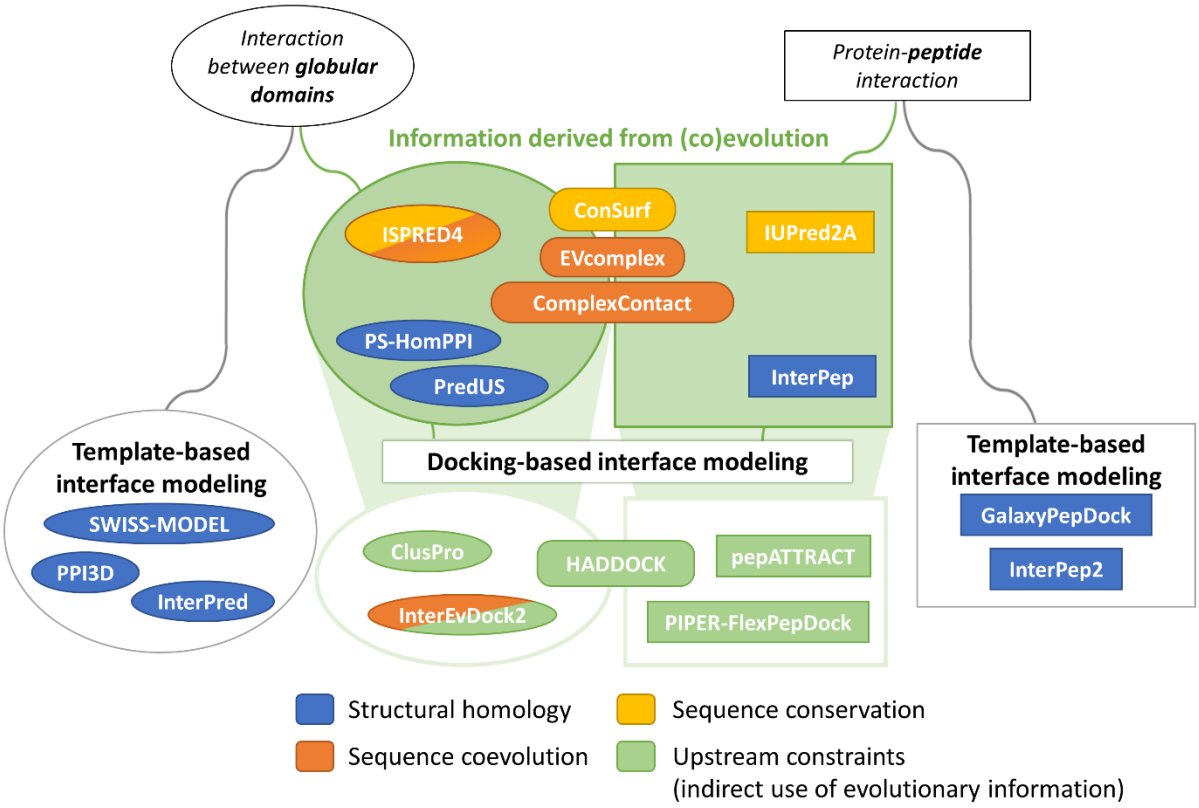


Figure 1-9: Graphical summary of a selection of user-friendly methods used in structural protein interaction prediction. *Those methods are available as web servers, except InterPep and InterPep2 (see Table A-1*

page 167 in the appendix for links and references). Methods for predicting interactions between globular domains are presented on the left in oval shapes, methods for predicting protein-peptide interactions mediated by short motifs are presented on the right in rectangles and methods suitable for both are in the middle in rounded rectangles. The background colour code denotes how structural modelling makes use of evolutionary information: through sequence conservation (light orange), sequence coevolution (dark orange) or structural homology (blue), or indirectly through information provided by upstream methods that predict binding sites, motifs or contacts (green). Template-based modelling, which relies on homologous complexes of known structure, bypasses the use of upstream methods compared to docking-based predictions.

1.3.3.1 Template-based docking

Template-based modelling makes use of homologous structures and is driven by the knowledge that proteins **similar** in **sequence** or **structure**, especially in the interface region, **bind** in a **similar way** (Andreani, Faure et al. 2012). Similarly to template-based modelling of monomeric structures (see section 1.3.1.1, page 42), protein complexes can be modelled using homolog complex structures and results in predictions that are often **more accurate** than free-docking (Figure 1-10) (Lensink, Nadzirin et al. 2019). The higher the sequence identity between the template and the protein complex to be modelled, the more accurate the model. The common cut-off in sequence identity lies within 30% in both partners, a threshold below which we cannot be sure that the homologous proteins interact in the same way as the proteins to be modelled (Aloy, Ceulemans et al. 2003, Faure, Andreani et al. 2012). Additional difficulty lies within the search of suitable for templates. The same procedure can be performed to identify homologous complexes as in monomer template-based modelling but it requires an additional step where the **individual homologs** of each protein partner have to be **matched** (i.e. intersection of homolog lists and removal of homologs that are not in direct contact). Alternatively, several databases exist that map homologous complexes with experimentally determined 3D structure to infer properties about other protein-protein interactions. The **PRISM web server** provides a repository for the prediction and structural modelling of protein interactions using evolutionary conservation of hotspot residues and multi-protein structural alignments to measure interface similarity (Baspinar, Cukuroglu et al. 2014). The **3D-interologs database** infers protein interactions across species by mapping domains to interface structures (Lo, Chen et al. 2010). The **IBIS database** uses the VAST structure comparison method to predict interaction partners and protein binding surfaces by analysing homologous complexes of known structure (Shoemaker, Zhang et al. 2012).

More recently, the **PPI3D web server** was built to search structural data using as query a single sequence or a pair of sequences in order to retrieve 3D structures of protein-protein and domain-domain complexes containing subunits homologous to the query sequence(s) (Dapkunas, Timinskas et al. 2017). PPI3D is especially useful for template-based interface modelling and we regularly rely on it in CAPRI docking challenges (see Chapter 4, page 125).

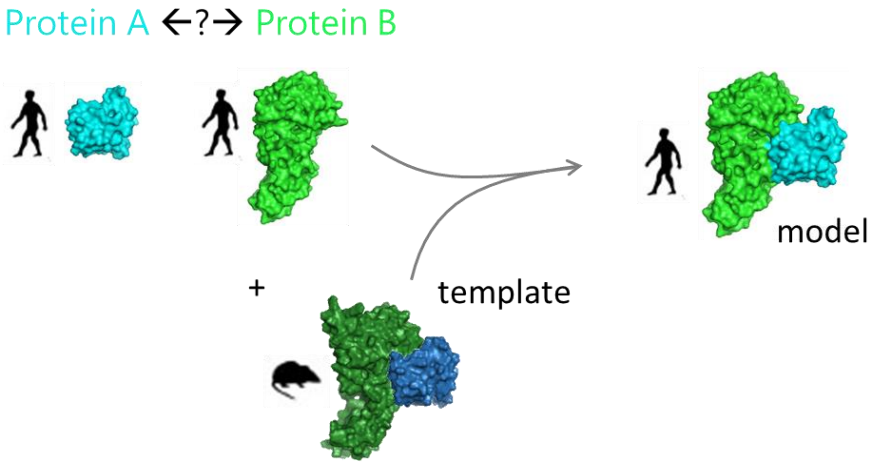


Figure 1-10: Illustration of template-based docking. *When an interologous structure exists for a given protein pair (e.g. in mouse), this interolog can be used as a template in order to deduce a model of the bound structure of our proteins of interest (e.g. in human).*

As for comparative modelling of monomeric structures, the user-friendly server **SWISS-MODE**L can be used for template-based docking as well as the **RosettaCM** package (see section 1.3.1.1, page 42). Another template-based docking and fully automated server is **InterPred** (Mirabello and Wallner 2017), which combines several tools to predict the final complex. First, monomeric structures are modelled using MODELLER if not given and templates are found using a structural alignment algorithm. Monomeric models are then superimposed onto the selected template to give a first set of coarse-grained models. The most likely coarse-grained models are selected using a random forest classifier based on sequence and structural features (e.g. interface size, interface overlap, structural alignment quality and sequence identity with the template). InterPred's final selection consists of the models that changed the least after a last refinement step. **ClusPro**, one of the best performing servers in recent blind-tests (Lensink, Nadzirin et al. 2019), lately integrated template-based docking into its pipeline to expand its ability to make high accuracy interface models (Porter, Padhorny et al. 2019). The **HDOCK** server, which also performed well in these

blind-tests, implements a hybrid strategy involving template-based and template-free docking (Yan, Zhang et al. 2017) (template-free docking will be described in the next section).

For the structural modelling of **protein-peptide** interactions, comparative modelling also remains the most suitable method in cases where closely related structural templates exist. The **GalaxyPepDock** web server (Lee, Heo et al. 2015) provides for that purpose a full pipeline to search for templates and model these categories of complexes in an automated manner. The InterPep2 software uses templates from both protein-peptide complexes and protein-protein interactions (Johansson-Akhe, Mirabello et al. 2020).

Template-based methods are efficient but cannot always be used, as reliable homologous complex structures are not always available. Interactome3D (Mosca, Céol et al. 2013) lists more than 270,000 interactions identified in 18 model species, for which around 88% have no experimental structure and no readily identifiable template structure. They can also lead to wrong assumptions when predicting interfaces in cases where homologs display different binding modes to the query complex, especially for homologs under 30% sequence identity (Faure, Andreani et al. 2012). For these cases, we can resort to template-free protein-protein docking.

1.3.3.2 Template-free docking

Template-free docking performs an **exhaustive search** of the conformational space, starting from two unbound protein structures or models. It is traditionally divided into two steps, illustrated in Figure 1-11. First, several thousands of interface conformations, called decoys, are generated during a **sampling** step. Sampling is then followed by or coupled with **scoring**, during which these decoys are ranked based on their interface properties (Huang 2014, Huang 2015). In an ideal situation, the score should directly reflect how close the decoy is to the true complex (denoted the native complex or bound structure). Many template-free docking programs and pipelines exist, each having their own specialty. Their **performance** is usually **increased** when they are included in **integrative modelling** pipelines, where docking is guided by additional experimental data, evolutionary information (i.e. conservation or coevolution) or predictions of binding areas (Koukos and Bonvin 2019). This data can be

integrated in the scoring step or as constraints during or immediately after the sampling step. Finally, a high-resolution but costly scoring step can be performed on a small selection of promising complex candidates, followed by a final refinement step can help to optimise the structures and remove the remaining imperfections.

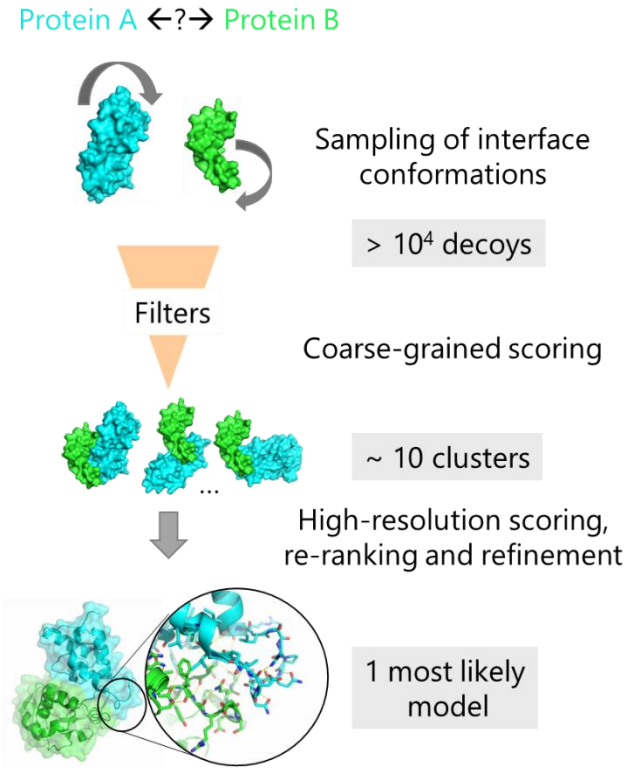


Figure 1-11: Template-free docking pipeline. *The general pipeline for template-free protein-protein docking relies on having the structures of two individual proteins (proteins A and B in blue and green). From top to bottom, a large amount of docking poses are generated, then filtered or ranked according to low-resolution scoring functions resulting in a set of approximately 10 decoys. An additional clustering step can be implemented whether before or after scoring. High-resolution scoring and model refinement can be performed on a small number of decoys in order to generate a model with the highest quality possible.*

1.3.3.2.1 Sampling

There are many different template-free docking tools based on various different criteria. Docking tools perform an extensive or guided search of the structural space for possible solutions and often score these decoys during that step with a simple and fast scoring function (e.g. to remove those that are too unrealistic because they are too clashing). By performing this search, we hope that sampling tools will be able to propose at least one solution that is close to the real complex (denoted near-native complex) and which can later be identified by one or several scoring functions. Traditionally, the bigger of the two proteins is

called the receptor and stays put, whilst the smaller one of the two, the ligand, moves in space and orientation around the receptor. The complex shape of proteins makes it very difficult to exhaustively explore all possible binding conformations, thus bioinformatics tools had to use tricks to minimise the system's enormous number of degrees of freedom. A way to deal with this is to consider both proteins as **rigid units** (i.e. "**rigid-body**" docking), thereby **reducing** the **complexity** to a 6-dimensional search space: typically three translational degrees and three rotational degrees representing the x, y and z axes and the θ, φ and ψ angles in a Cartesian space. This considerably accelerates the conformational search. Some also integrate the concept of protein flexibility during sampling in the form of small conformational readjustments.

Sampling strategies include **Fourier-based** sampling, **local shape matching** and **other global search** methods, with Fourier-based sampling the most commonly used. An objective assessment was made in 2015 of 14 global docking tools on Weng benchmark 4 (Huang 2015) (section 0, page 68) of which some are listed in Table 1-1 along with a few of the top-performing servers in CAPRI 7th edition (Lensink, Nadzirin et al. 2019).

Table 1-1: List of popular docking tools and their properties. *This list was adapted from (Huang 2015) and additionally contains the top-ranking servers in CAPRI 7th edition.*

Program	Scoring function	Assembly search	
ZDOCK 3.0.2	Shape complementarity, electrostatics and knowledge-based pair potentials	FFT-based	(Pierce, Wiehe et al. 2014)
MDockPP	Shape complementarity, electrostatics and knowledge-based pair potentials	FFT-based	(Duan, Qiu et al. 2020)
HDOCK	Shape complementarity and knowledge-based pair potentials	FFT-based	(Yan, Zhang et al. 2017)
PIPER	Shape complementarity, electrostatic interactions and knowledge-based pair potentials	FFT-based	(Kozakov, Brenke et al. 2006)
ClusPro	Shape complementarity, electrostatic interactions, knowledge-based pair potentials, cluster size	FFT-based	(Kozakov, Hall et al. 2017)
GRAMM-X	Shape complementarity, hydrophobic match	FFT-based	(Tovchigrechko and Vakser 2006)
MolFit	Geometric complementarity, hydrophobic complementarity and electrostatic interactions	FFT-based	(Kowalsman and Eisenstein 2007)
SDOCK	van der Waals attractive potential, geometric collision, electrostatic potential and desolvation energy	FFT-based	(Zhang and Lai 2011)

FRODOCK 3.12	Van der Waals, electrostatics, desolvation and knowledge-based pair potentials	SFT-based	(Ramírez-Aportela, López-Blanco et al. 2016)
HEX	Surface complementarity and electrostatics interactions	SFT-based	(Macindoe, Mavridis et al. 2010)
PatchDock	Geometric shape complementarity	Local shape matching	(Schneidman-Duhovny, Inbar et al. 2005)
ATTRACT	Lennard-Jones type effective potentials and electrostatics interactions	Global search	(de Vries, Schindler et al. 2015)
HADDOCK	Physical potentials and experimental or computational constraints	Global search	(van Zundert, Rodrigues et al. 2016)
LZERD	Shape complementarity and clash penalty	Global search	(Christoffer, Terashi et al. 2020)

Fast-Fourier transform-based sampling (FFT)

In the traditional Fourier correlation approach, protein topology is reduced to a simple Cartesian **grid model**, which naturally favours close contacts and penalises steric clashes. Cubes in the grids (typically 1-1.5 Å in size) are labelled according to their position in the protein (inside, outside or on the surface). The grids are then efficiently matched through successive translational increments and **Fast-Fourier transform-based calculations** (FFT) which can be applied to calculate correlating surface shapes but also other properties such as compatible hydrophobicity or electrostatic and van der Waals force fields (Ritchie 2000). **ZDOCK** (Pierce, Wiehe et al. 2014) or **GRAMM-X** (Tovchigrechko and Vakser 2006) are examples of FFT-based methods (see Table 1-1). At the time of Huang's review, ZDOCK version 3.02 outperformed all other evaluated docking programmes with a 30.7% success rate (Huang 2015) (see section 1.3.3.4.4, page 72 for more details on evaluation metrics). The MDockPP server is a GPU-adapted reimplementation of ZDOCK3.02 and was ranked within the best-performing servers in CAPRI 7th edition (Lensink, Nadzirin et al. 2019, Duan, Qiu et al. 2020).

Spherical-Fourier transform-based sampling (SFT)

As a new FFT must be calculated for each rotational increment, calculations can be extremely slowed down when docking large molecules, especially when a small grid size is used. In

answer to this, Spherical-Fourier transforms (SFT) have emerged using **spherical harmonic functions** to represent protein surface shapes in which the 6 degrees of freedom in the docking problem become 5 Euler rotation angles and an intermolecular distance (Figure 1-12). Docking programmes such as **FRODOCK** (Ramírez-Aportela, López-Blanco et al. 2016), which we routinely use, or **HEX** (Macindoe, Mavridis et al. 2010) are based on that principle (see Table 1-1).

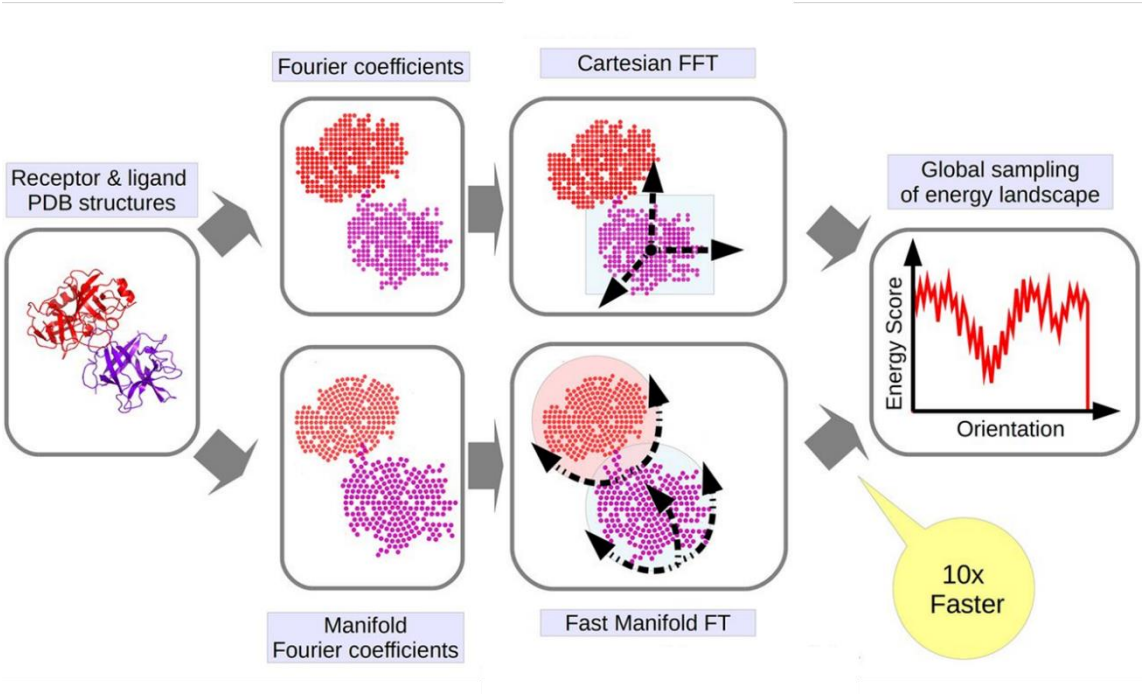


Figure 1-12: Schematic representation of FFT and SFT. *The top path represents FFT sampling in Cartesian space implying three translational and three rotational degrees of freedom around the receptor. The lower path describes SFT sampling in rotational space implying five rotational (two per protein and one defining the angle between the two) and only one translational degree of freedom (vector between the centres of mass of both proteins). This property considerably accelerates the sampling. This picture was taken from (Padhorny, Kazennov et al. 2016).*

FRODOCK2.1 (Fast ROtational DOCKing) ranks its decoys with a linear combination of four different terms: van der Waals, electrostatic and desolvation potentials as well as a knowledge-based term, Tobi (Dong, Fan et al. 2013). FRODOCK typically includes an additional clustering step after sampling to remove redundancy between decoys.

Other sampling strategies

Local shape matching, as in **PatchDock** (Schneidman-Duhovny, Inbar et al. 2005), reduces the sampling space by directly focusing on regions with possible complementary between

the two proteins. In this case, regions of interest are usually protein surface areas with specific and distinct geometric features such as cavities or local knobs and holes. Possible ligand orientations are then rapidly explored through local shape complementarity by reducing receptor and ligand proteins to negative and positive images.

Other global search strategies include **HADDOCK**, **LZERD** or **MDOCKPP**, which were found in the top-ranking docking servers in the latest CAPRI round (Lensink, Nadzirin et al. 2019). **LZERD** (Christoffer, Terashi et al. 2020) is a geometric shape-based docking programme that uses rotation-independent 3D Zernike descriptors to represent protein surfaces and can handle a certain degree of protein flexibility by adjusting the resolution of its descriptors. **ATTRACT** (de Vries, Schindler et al. 2015) has a random search strategy which is combined with physics-based scoring terms. ATTRACT performs efficient minimisation of individual docking poses by reducing protein residues to a group of three pseudo-atoms. It also integrates a minimal side-chain rotamer sampling during minimisation.

Data-driven docking

In most free docking programs and web servers designed to predict interactions between globular domains, restraints can be used to enrich docking solutions by filtering out decoy interfaces that do not involve some residues or residue pairs. For instance, most docking servers include a field where the user can input interface restraints. Examples include **ClusPro** (Kozakov, Hall et al. 2017), **GRAMM-X** (Tovchigrechko and Vakser 2006), **PatchDock** (Schneidman-Duhovny, Inbar et al. 2005), **SwarmDock** (Torchala, Moal et al. 2013), **pyDock-WEB** (Jimenez-Garcia, Pons et al. 2013), **GalaxyTongDock** (Park, Baek et al. 2019), **HDOCK** (Yan, Zhang et al. 2017) and **InterEvDock2** (Quignot, Rey et al. 2018). This feature is especially useful in cases where experimental data are available such as X-link or NMR data (Xue, Dobbs et al. 2015). Interface residue predictions, especially those mentioned above that use conservation or coevolution (Figure 1-7, label 3, page 41), can also be used as restraints in the docking process (Figure 1-7, label 4, page 41).

The user-friendly **HADDOCK** server (van Zundert, Rodrigues et al. 2016) has the additional and interesting ability to set ambiguous constraints. HADDOCK distinguishes between residues that can be confidently assigned as involved in the interaction ("active" residues) and their solvent-accessible neighbours ("passive" residues). HADDOCK was typically designed to integrate experimental data but when no such data is available, interface predictions from software such as CPORT (de Vries and Bonvin 2011) can be used to guide the docking process. HADDOCK consists of three main steps, starting with rigid-body energy minimisation followed by simulated annealing and refinement in explicit solvent. Decoys are scored with a linear combination of physical potentials and experimental data distance constraints.

Decoy clustering

Other than removing possible redundancies to alleviate possible scoring steps that follow (e.g. **FRODOCK**), an additional clustering step after the initial sampling can help to better select near-native configurations. This is based on the assumption that the free energy landscape exhibits a broader and deeper well around the native structure than around non-native structures. One can therefore assume that within the sampling population, near-native regions should be more enriched in decoys.

ClusPro (Kozakov, Hall et al. 2017), a top-performer in the CAPRI challenge, selects the top 1,000 FFT-generated decoys by PIPER and clusters them based on RMSD. Since **near-native** structures are more likely to **cluster together** into big low-energy groups, ClusPro uses cluster size as a selection criterion to return its most probable solutions.

1.3.3.2.2 Scoring functions

According to physical chemistry, the structure that is closest to the native structure should be the one with the **lowest binding free energy**. However, predicting binding free energy is a difficult task as it involves the calculation of entropic contributions and solvent effects. Additionally, accurate selection of near-native poses within predictions relies on a complete sampling of the conformational space. Good **proxies** are **scoring functions**, which implies, however, the loss of the quantitative aspect of scoring as we are reduced to relative decoy

ranking. Scoring functions can be based on different properties such as physics-based (electrostatic or van der Waals interactions, hydrogen bonding, desolvation) or knowledge-based. Shape-complementarity is the most basic one and is integrated into sampling. Sophisticated scoring functions are often used in a rescoring step after sampling.

We particularly use SOAP-PP, InterEvScore and Rosetta's score in our team, thus I will describe all three in more detail below.

SOAP-PP is an example of knowledge-based scoring function and stands for Statistically Optimised Atomic Potential (Dong, Fan et al. 2013). It is an atomic statistical-based score integrating distance-dependent potentials learnt on a set of real complex structures and normalised on a set of incorrect PatchDock decoys generated from the Weng benchmark. This enables it to better differentiate wrong models from near-native ones on three different decoy sets and two different benchmarks (Dong, Fan et al. 2013, Quignot, Rey et al. 2018).

Another scoring function that I used during my PhD was the **Rosetta interface score** (ISC) (Gray, Moughon et al. 2003, Chaudhury, Berrondo et al. 2011). ISC is made of a linear combination of non-bonded atom-pair interaction energies and empirical and statistical potentials among other terms and is calculated by subtracting the total energy of both monomeric structures from the total energy of the complex structure.

InterEvScore (Andreani, Faure et al. 2013) is a scoring function, which couples a coarse-grained two- and three-body statistical potential with coevolutionary information extracted from coMSAs built jointly for the two protein partners. InterEvScore goes beyond conservation and makes use of coevolving pairs (or groups) of positions across the interface. The goal is thus to favour decoys containing contacts that are compatible with the coevolutionary history of the interaction. In InterEvScore, interface contacts are computed for each docking decoy and scored for each species in the coMSAs. InterEvScore can make use of coevolutionary information from coMSAs containing as few as 10 sequences. It was integrated in the recently updated free docking server InterEvDock2 (Quignot, Rey et al. 2018), which I will describe in more detail in the following chapter (Chapter 2, page 77).

Similarly to constraints in the docking, predicted interface residues may also be used in the scoring step. This is the case, for instance, in **DockRank** (Xue, Jordan et al. 2014), where interface residues are predicted based on the 3D structures of interologs and decoys are scored according to how many predicted residues belong to their interface. DockRank gives good results compared to scoring functions of the reference docking programme ZDOCK (Vreven, Hwang et al. 2011), partly owing to the partner-specific trait of their interface residue predictor. Another scoring function driven by conservation is **GraphRank**, integrated in iScore (Geng, Jung et al. 2019). In GraphRank, interfaces are not represented as a set of individual contacts but as labelled graphs in which the nodes represent interface residues, each annotated with its PSSM, and edges encode residue contacts. GraphRank classifies interfaces as native or non-native by comparing them with a reference set of positive and negative examples. The complex graph comparison problem is solved using random graph walking. The resulting similarities with the reference set are given as input to an SVM classifier to estimate how close each decoy is to a native structure. Combined with intermolecular energetic terms in iScore, GraphRank manages to outperform HADDOCK (Dominguez, Boelens et al. 2003) and state-of-the-art docking programme ZDOCK.

1.3.3.2.3 High-resolution scoring and structural refinement

In order to compensate for the imprecision enforced by rigid-body sampling, one can perform re-sampling around already generated and carefully selected docking poses or integrate small minimisations and optimisations in the hope of reaching higher quality models. Additionally, a last minimisation step can be performed in order to return models of improved quality by taking steric clashes, proper repacking or correct H-binding into account for example (see Figure 1-11, page 57). We commonly use the Rosetta package to do so (Lyskov and Gray 2008, Fleishman, Leaver-Fay et al. 2011).

1.3.3.2.4 Docking methods for structural modelling of protein-peptide complexes

From the prediction of binding motifs and of binding sites in the folded domain partner, it is possible to generate structural models using docking tools that were developed and benchmarked for the specific purpose of docking flexible peptides onto folded receptors

(Figure 1-7, labels 4 and 7, page 41). These docking methods were recently reviewed (Ciemny, Kurcinski et al. 2018) and a number of methods have been described in a collection of protocols (Schueler-Furman and London 2017). Based on some interaction details obtained from evolutionary constraints, the sampling complexity can be restricted either by driving **local docking** around specific anchor residues or by **post-filtering** docking poses after global docking. Among available strategies, a recently developed protein-peptide docking protocol (Alam, Goldstein et al. 2017) reached remarkable accuracy using a combination of **PIPER** (Kozakov, Beglov et al. 2013) for exhaustive fragment-receptor rigid-body docking and **Rosetta FlexPepDock** (Raveh, London et al. 2010) for flexible full-atom refinement of the best rigid-body poses. The method also generated top performing models in CAPRI 7th edition (see section 1.3.3.4.2, page 70 and Chapter 4, page 125 for more details on CAPRI and its targets) on challenging targets such as T134-T135 and T121 (Khramushin, Marcu et al. 2019). Interestingly, combining PIPER-FlexPepDock with the **InterPep2** template-based method improves prediction performance over the use of each individual method on a test dataset of 27 non-redundant protein-peptide complexes for which the unbound structure of the protein is also available (Johansson-Akhe, Mirabello et al. 2020).

Exhaustive sampling of the peptide conformations can also be obtained using **CABS-Dock** (Kurcinski, Jamroz et al. 2015) which randomly docks a peptide with restrictions on a binding region and subsequently refines conformations using replica exchange Monte Carlo dynamics (Kurcinski, Badaczewska-Dawid et al. 2020). Other methods such as **HADDOCK** (Trellet, Melquiond et al. 2013) or **pepATTRACT** (de Vries, Rey et al. 2017) use three conformations for the input peptide (extended, helix and polyproline). In HADDOCK, the search can be targeted by defining explicit spatial constraints as specific or ambiguous distance restraints, while in pepATTRACT restriction to a region of interest should be done after global prediction.

1.3.3.3 Covariation-based interface structure prediction

Coevolutionary constraints can be integrated in scoring functions for template-free protein docking or derived from covariation-based methods to guide molecular simulations in order to provide more likely interface models (Figure 1-7, label 5).

Bacterial two-component signalling (TCS) systems involve specific interactions between proteins from large families of paralogs. The two protein partners most often belong to the same operon and their proximity within genomes facilitates the construction of large coupled MSAs associating specifically interacting protein pairs. This makes the TCS system ideally suited for statistical covariation analysis and as such, TCS was the object of the first studies showing that DCA was predictive of residue pairs in contact across the interface (Weigt, White et al. 2009). Coupling these predictions with molecular dynamics simulations enabled the high-resolution modelling of a TCS pair interface (Schug, Weigt et al. 2009).

The idea of sequence matching using genomic proximity and operon structures was extended from the TCS system to several dozens of bacterial complexes in two 2014 studies, where predicted **DCA contacts** were used as **distance restraints** in molecular docking for interface modelling with either PatchDock coupled with Rosetta or HADDOCK (Hopf, Scharfe et al. 2014, Ovchinnikov, Kamisetty et al. 2014). The **EVcomplex** web server provides an interface for users to predict interacting residues in a complex of interest from two input protein sequences (Hopf, Scharfe et al. 2014). Alignments can be built either by using the original genomic proximity method or by pairing best hits, that is, sequences with highest sequence identity to the query, in each genome.

Recently, HADDOCK was also used to predict homodimeric complex structures using DCA restraints in a large-scale study of almost 2,000 protein families (Uguzzoni, John Lovis et al. 2017). This is a special case for the use of DCA-derived restraints since the homodimeric interaction signal is entangled with intra-protein couplings in predictions based on homologous sequences of a single protein that homodimerises (dos Santos, Morcos et al. 2015).

As an alternative to docking or molecular dynamics simulations with restraints, **Monte Carlo simulations** based on a coarse-grained potential energy specifically validated on low-affinity protein complexes were used to **exploit DCA predictions** for the molecular modelling of the eukaryotic Hsp70/Hsp40 and homologous bacterial DnaK/DnaJ interfaces (Malinverni, Jost Lopez et al. 2017). Due to the lack of operon organisation in this system, the authors used random paralog matching to build concatenated MSAs of Hsp40 and Hsp70 family proteins. This is close in spirit to recent work on paralog matching algorithms, which try to predict simultaneously and iteratively pairs of specifically interacting proteins and inter-protein contacts (Bitbol, Dwyer et al. 2016, Gueudre, Baldassi et al. 2016, Marmier, Weigt et al. 2019). So far, that work focused on pairs of proteins for which homologous sequences can be found within operons and further generalisation to any pair of interacting proteins should provide interesting insights into other bacterial and eukaryotic biological processes. Extension to eukaryotic complexes should also benefit from the recent finding that inter-protein contacts identified by DCA-like methods in bacterial complexes are well conserved in homologous eukaryotic protein complex structures (Rodriguez-Rivas, Marsili et al. 2016).

Recent large-scale and blind assessments showed that DCA-type predictions were most efficient for single protein structure prediction when integrated into deep learning pipelines (see section 1.3.1.2, page 43). This idea was generalised to inter-protein contact prediction in the **ComplexContact** web server (Zeng, Wang et al. 2018). ComplexContact first builds two concatenated MSAs for pairs of proteins: one using a genomic context method as discussed above and another relying on a matching method based on phylogenetic species tree ordering. Then, a **deep learning model** trained on single chain proteins predicts inter-protein contacts from these two MSAs. Inter-protein contact prediction results suggest that deep learning greatly enhances DCA performance. Most recently, large-scale interface modelling was performed using protein-protein docking guided by distance constraints between residue pairs that were predicted as coevolving by algorithms of the DCA family, with the goal of predicting protein interaction networks in two bacterial species (Cong, Anishchenko et al. 2019). In apparent contrast to the results obtained with ComplexContact, the authors

found that a deep learning method successfully developed for single protein contact prediction (Jones and Kandathil 2018) did not improve interface model discrimination. This may be because in ComplexContact, only sequences are used and the deep learning layer increases accuracy because it strengthens contacts compatible with the implicitly predicted 3D structures of unbound subunits. In the protein docking study (Cong, Anishchenko et al. 2019), the set of DCA constraints satisfied by a docking model has to be consistent with the explicit monomeric 3D structures of the binding partners, which may explain why deep learning did not bring additional discrimination. Further progress might be obtained by coupling molecular docking to deep learning contact prediction methods specifically trained on protein-protein interfaces.

1.3.3.4 Evaluation

Protein-protein docking approaches can be assessed using datasets with known unbound and complex structures or during blind tests through CAPRI (Critical Assessment of PRediction of Interactions), a challenge similar to CASP for protein folding. Decoy quality assessment and general performance metrics are also summarised below.

1.3.3.4.1 Benchmarking - Testing performance on known cases

Prediction quality of protein interfaces can be assessed using a benchmark of protein **pairs** with **known** experimental **bound** and their corresponding **unbound structures**. Using unbound structures is essential in benchmarking in order to avoid any shape complementarity bias in the prediction and reproduce a scenario as close as possible to real cases. **Weng's Benchmark** is widely used in the docking community (Hwang, Vreven et al. 2010). The latest version 5 (Vreven, Moal et al. 2015) contains 230 complexes, 190 of which are non-antigen-antibody complexes. Complexes are classified into three difficulty categories depending on how much the structures change between bound and unbound states – the RMSD between the native and its superimposed unbound proteins being the indicator of this change. **DOCKGROUND Docking X-ray Unbound Benchmark 4** (Kundrotas, Anishchenko et al. 2018) is another benchmark containing 396 unbound/bound crystal structures of which 39

are not shared with Weng Benchmark 5 (TM-score < 0.6 and sequence identity < 23% with all of Weng's complexes).

Unfortunately, one of the biggest limitations of experimental benchmarks is that their size is highly dependent on the availability of experimental bound and their corresponding unbound structures in the PDB. A way to overcome this limitation is to enrich the benchmark with complexes for which at least one of the unbound structures is unavailable by modelling the unbound state. Both DOCKGROUND and PPI4DOCK (Yu and Guerois 2016) followed that logic.

DOCKGROUND currently contains subsets of complexes for which at least one unbound structure was simulated using Langevin Dynamics simulations in CHARMM (1918 complexes) or modelled using I-TASSER (165 complexes, see section 1.3.1.1, page 42) or Phyre2 (963+171 complexes). In either of these subsets, the "unbound quality" of the generated models was assessed according to their RMSD with the bound complex (in theory, the more different to the native complex, the more reliably unbound). 100 well-selected GRAMM-generated decoys are available on their website for the 165 I-TASSER-modelled complexes.

PPI4DOCK (Yu and Guerois 2016) is a large benchmark developed in our team made of 1417 binary docking targets where unbound structures were modelled by homology and of which the "unbound quality" is guaranteed by the use of unbound-assured templates. PPI4DOCK was constructed starting from an initial batch of 3157 non-redundant, high-resolution heterodimers from InterEvol (Faure, Andreani et al. 2012). Homologs were searched for each partner individually using the HH-suite package (Steinegger, Meier et al. 2019) and filtered out to only have good quality homologs that were not co-crystallised with any homolog of the opposite partner (i.e. unbound templates). Unbound models were then generated with these identified templates and the homology modelling RosettaCM protocol (Song, DiMaio et al. 2013). PPI4DOCK has been used to benchmark two different sampling programmes and four different scoring functions with comparable performances to when using the Weng Benchmark 4 (see Chapter 2). Thus, in the case of PPI4DOCK, using modelled unbound struc-

tures only seems to mildly affect the docking with a decreasing impact for increasing sequence identity with the chosen template. PPI4DOCK is split into five difficulty categories going from "very easy" to "super hard" depending on the divergence between bound and unbound states and how many clashes are generated when superimposing the unbound models onto the native complex. The whole PPI4DOCK benchmark can be downloaded from http://biodev.cea.fr/interevol/ppi4dock/ and contains pre-generated co-MSAs for each case.

1.3.3.4.2 The CAPRI initiative - Testing performance in real-case scenarios

CAPRI is a community-wide experiment for the comparative evaluation of structural protein assembly prediction methods (Janin, Henrick et al. 2003) and was initially created as a satellite of CASP (section 1.3.1.3, page 45). Regular rounds of **blind prediction** and scoring provide challenging, unpublished protein complex targets for all method developers to test and improve their docking programs and pipelines. The CAPRI targets are therefore complementary to more traditional docking benchmarks. In most cases, only the sequences of the interacting macromolecules are provided to predictors, who must submit an ordered list of interface models within a few weeks. Predicted models are then assessed by comparing them to the experimental structure. A CAPRI evaluation meeting, held every three years, gives a fantastic opportunity for state-of-the-art assessment and discussion of the best methods and the remaining challenges. The most recent CAPRI meeting (7th edition) featured prediction rounds held between 2016 and 2019, in which more than 50 research groups participated to predict challenging and diverse targets involving protein-protein, protein-peptide and protein-oligosaccharide complexes (see Chapter 4). Analysis of these latest results by the CAPRI assessors showed overall progress due to slightly improved methods and better integration of template-based interface modelling techniques with docking, rescoring and refinement (Lensink, Nadzirin et al. 2019).

1.3.3.4.3 Assessing decoy quality - CAPRI and DockQ criteria

Decoy **quality** can be assessed using different criteria. Previous studies, such as (Chen and Weng 2003), used to base themselves on the **interface RMSD** (I-RMSD) between Cα atoms of the decoy and the experimental native structure – a decoy being considered as a near-

native (i.e. correct) if its value was under a certain threshold (e.g. 2.5 Å). However, this gives an incomplete picture of decoy quality. The nowadays commonly accepted evaluation is based on the rules established by the CAPRI community (Mendez, Leplae et al. 2003) in which decoys are classified in **three** near-native **categories**: High, Medium or Acceptable in decreasing order of quality according to their closeness to the native complex, and are Incorrect otherwise. This quality is defined using **three** different **measurements**, namely the fraction of native contacts (**Fnat**), the ligand backbone RMSD (**L-RMSD**, backbone being C, N, O and Cα atoms) and the backbone **I-RMSD** with the native complex. The thresholds for these measurements are illustrated in Figure 1-13. Contacts are defined by a distance of 5 Å or less between heavy atoms from two different residues on opposite chains. The decoy needs to satisfy the Fnat threshold and at least one of the two RMSD metric thresholds of the same category in order to be labelled as such.

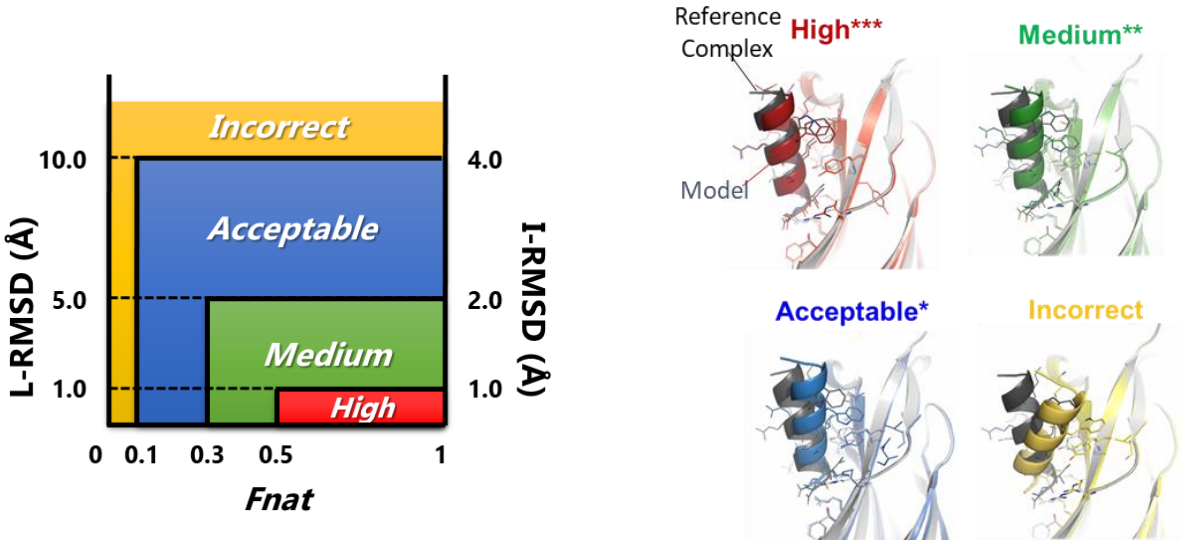


Figure 1-13: CAPRI thresholds. *Left: Illustration of all three CAPRI thresholds defining each of the three decoy quality categories Acceptable, Medium and High. The Fnat metric and at least one of the other two metrics has to be satisfied in order to give a decoy the corresponding label e.g. a decoy with Fnat > 0.5 and L-RMSD < 1 Å and/or I-RMSD < 1 Å would be classified as High. Right: Illustration of all three CAPRI categories on a protein-peptide complex (reference structure in grey, decoy of various quality in colour).*

In today's machine learning era, it is becoming increasingly important to have a **continuous** evaluation **metric** rather than a discrete one for model fitting and to avoid threshold effects. More recently, a continuous metric was published, **DockQ** (Basu and Wallner 2016), that closely reproduces the CAPRI criteria on a test set of 15,328 CAPRI-submitted decoys (Figure

1-14). DockQ is effectively the average value between the three previously mentioned metrics, Fnat, L-RMSD and I-RMSD, in which both RMSDs are first subjected to inverse square scaling in order to bring them, similarly to the Fnat, in the [0, 1] range (the scaled L-RMSD and I-RMSD being equal to 0.5 for 8.5 Å and 1.5 Å respectively). The inverse square method also has the benefit of giving less weight to high RMSD values since above a certain threshold, decoys are wrong, no matter how high the RMSD goes.

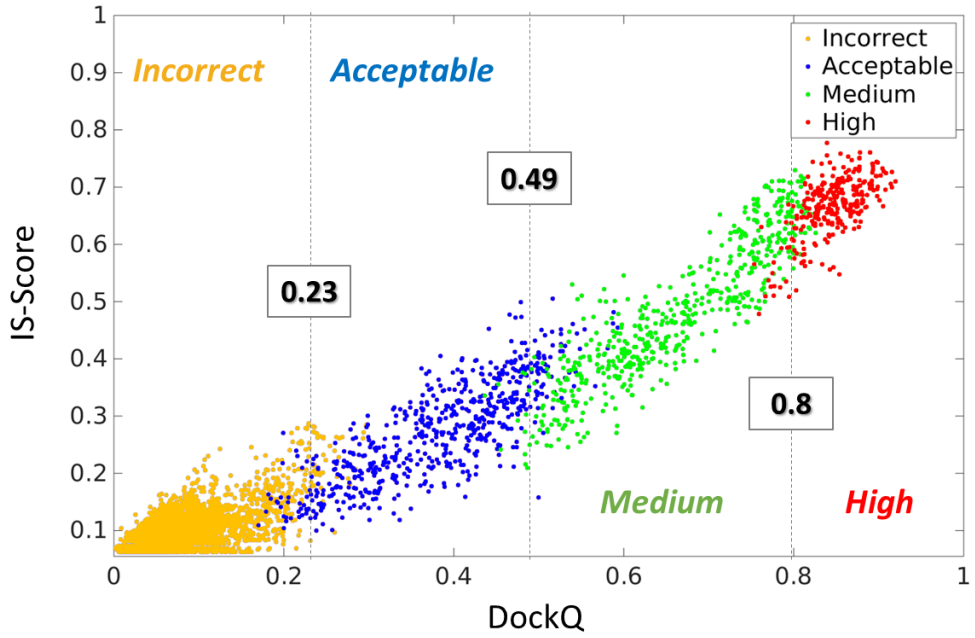


Figure 1-14: Scatter plot of the novel DockQ decoy evaluation criteria against IS-Score. *This figure was adapted from (Basu and Wallner 2016) and illustrates how well the DockQ score reflects the commonly-accepted CAPRI criteria for decoy quality evaluation on a CAPRI-set of 15,328 decoys. Models are coloured according to CAPRI classification as Incorrect (yellow), Acceptable (blue), Medium (green) or High (red). DockQ thresholds that best reproduce the CAPRI categories are 0.23, 0.49 and 0.8 i.e. decoys with a score bellow 0.23, between 0.23 and 0.49, between 0.49 and 0.8 or above 0.8, could relatively safely (±0.02) be classified as Incorrect, Acceptable, Medium or High, respectively.*

The interest in DockQ can be expected to increase in the following years and its use was discussed in the most recent CAPRI meeting. However, performances are mainly measured using the original CAPRI criteria throughout this work in order to better compare our methods with others in the literature.

1.3.3.4.4 Docking performance measurements

The general performance of various docking and/or scoring methods on a set of cases can be evaluated using several different metrics. The top N **success rate** (SR) is the most

common performance measurement and consists of the fraction of cases in a benchmark that have at least one near-native decoy in the top N ranked decoys. N is usually equal to 10, but starting with recent CAPRI challenge round 47, only the top 5 submissions are evaluated. The top N hit rate (HR), also called hit count in (Chen and Weng 2003), corresponds to the overall proportion of hits within the top N ranked decoys and gives a better view of how enriched the top N ranking decoys are in near-natives.

As DockQ only recently appeared in the field, the community is still establishing the best way to integrate this continuous score into a general evaluation metric. Bonvin's team recently used the **discounted cumulative gain** (DCG) (Geng, Jung et al. 2019). The DCG for each case can roughly be assimilated to a weighted average and is calculated as follows:

$$DCG = \sum_{rank=1}^{N} \frac{2^{(DockQ_{rank})} - 1}{rank}$$

where rank is the rank of the decoy, $DockQ_{rank}$ is the DockQ score of the decoy with that rank and N is the top N decoys that are taken into account for this measurement. The 1/rank factor gives more importance to the quality of the top scoring decoys. In order to better compare cases with different numbers of decoys, the DCG is normalised by an ideal DCG (iDCG), which is calculated by reordering all decoys by decreasing DockQ score. The final normalised value (nDCG) for each case can be extrapolated into a single value by calculating the average nDCG over all cases in the benchmark.

1.4 OVERVIEW OF THIS MANUSCRIPT

Proteins are of fundamental importance in cells and knowledge of their 3D structure can help study their function in the cellular context with possible applications in therapeutic field (e.g. inhibitor of PPIs, drug design etc.). Proteins evolve under the constraint of maintaining functional interactions. This constraint is reflected in the evolutionary history of protein partners shown in coupled MSAs. As a consequence, these alignments provide valuable information for the purpose of interface computational prediction. The use of this information in docking is the central theme of this PhD project.

Apart from the introduction and the conclusions and perspectives chapters (Chapter 1 and Chapter 5), this manuscript is split into three other chapters. Chapter 2 and Chapter 4 correspond to already published articles and Chapter 3 to a paper in the process of being submitted.

In Chapter 2, I present our team's molecular docking server, InterEvDock2. I participated in major developments during my first year of PhD to make it more automated and user-friendly. InterEvDock2 predicts 10 most probable complex models from a pair of input sequences or oligomeric or monomeric structures using the FRODOCK sampling programme and a unique consensus scoring approach between three highly complementary scoring functions: the physics-based FRODOCK score, atomic-statistical potentials from SOAP-PP and the evolutionary-guided InterEvScore. Thanks to a completely automated modelling pipeline using the RosettaCM protocol, users are able to dock their proteins, even if the monomeric structures are unknown. Using strategic breakpoints throughout the pipeline, the user also has a say in the template's choice if wanted. Of particular interest to biologists, constraints can be added in order to filter out any irrelevant docking poses. Finally, I validated InterEvDock2's performance on a large set of 812 cases from our PPI4DOCK dataset.

In order to further develop our discrimination capacity between wrong and correct predictions, I decided to pursue the integration of evolutionary information at a much finer level of detail into scoring in Chapter 3. In the evolutionary-based InterEvScore, evolutionary information is given at the residue level in coMSAs and thus can only be easily mixed with a

residue-scale potential. The high complementarity between InterEvScore and the atomic statistical potentials from SOAP-PP encouraged us to derive evolutionary information at the atomic level using homology modelling. Coupled with a more efficient scoring implementation, average scores over a query protein pair and its homologs can be easily calculated for each decoy. This methodology showed promising results on several scores and we are currently preparing its integration in our InterEvDock2 server.

Finally, I present in Chapter 4, our key strategies and latest performances in the famous CAPRI challenge (Critical Assessment of PRediction of Interactions). CAPRI is an international blind-test challenge, where groups are invited to test their complex structural prediction pipelines on regularly dispatched targets over two years. The structures of these targets are only publically available once the challenge is over, thereby providing real-life test scenarios to defy and improve our docking methods. Throughout my PhD project, I have had the chance of participating in 10 of the challenges in CAPRI 7th edition as well as three prediction rounds from CAPRI 8th edition that is currently underway. I was also able to attend the CAPRI 7th evaluation meeting in April 2019, gathering all participating groups. As official results for the recent prediction rounds are not yet released, I will only focus on the targets in CAPRI 7th in this final chapter.

CHAPTER 2
InterEvDock2

Acquiring the 3D structure of protein interfaces is of high use for structural biologists to study their protein of interest and understand its functions in the cellular context. As experimental techniques are sometimes too time consuming, expensive or impossible, there is a high demand for structural prediction tools of protein complexes. Computational biologists are therefore encouraged to provide easy access to their general prediction pipelines and to make them as user-friendly and automated as possible to suite the majority of the scientific community. In light of this, our team developed the molecular docking server, InterEvDock. I participated in the implementation of major developments into the server (now InterEvDock2) during my first year of PhD and will present them in the following chapter. Three main features were added including the possibility of specifying constraints and the automated monomer homology pipeline in which I took part. I also took over the majority of the benchmarking. This chapter is based on our published paper (Quignot, Rey et al. 2018).

As introduced above, computational modelling of protein assemblies provides crucial insights for the functional characterisation of macromolecular interactions occurring in the crowded cellular environment. Predictions of protein-protein interfaces can be used to design experiments to investigate the role of important interactions and possibly interfere with them, typically using mutagenesis. Models of macromolecular complexes are also useful to complement integrative structural biology (Ward, Sali et al. 2013) and to deepen our understanding of disease-associated mutations (Gress, Ramensky et al. 2017) and protein interaction networks (Vakser 2014).

A number of servers have been developed for protein-protein docking, which can be separated into template-based modelling servers, which aim to identify suitable structural templates for the protein complex, and template-free docking servers. Recent resources to find templates for interface modelling starting from the sequences of two protein partners include KBDOCK (Ghoorah, Devignes et al. 2016), focused on domain-domain interactions, and PPI3D (Dapkunas, Timinskas et al. 2017). Recently released servers taking protein sequences as input for homology-based interface modelling include SnapDock (Estrin and Wolfson 2017), HOMCOS (Kawabata 2016) and SWISS-MODEL Quaternary Structure Prediction (Bertoni, Kiefer et al. 2017). Many template-free docking servers implement a rigid-body docking approach, sometimes followed by rescoring: PatchDock (Schneidman-Duhovny, Inbar et al. 2005), FireDock (Mashiach, Schneidman-Duhovny et al. 2008), HexServer (Macindoe, Mavridis et al. 2010), ZDOCK (Pierce, Wiehe et al. 2014), FRODOCK 2.0 (Ramírez-Aportela, López-Blanco et al. 2016), pyDockWEB (Jimenez-Garcia, Pons et al. 2013), ClusPro (Kozakov, Hall et al. 2017), GRAMM-X (Tovchigrechko and Vakser 2006), InterEvDock (Yu, Vavrusa et al. 2016). A hybrid approach combining template-based and template-free docking was recently proposed in the HDOCK server (Yan, Zhang et al. 2017). Some free docking servers include specific features such as symmetric docking (SymmDock (Schneidman-Duhovny, Inbar et al. 2005), ZDOCK); local docking around an initial guess (RosettaDock (Lyskov and Gray 2008), recently moved to the ROSIE server (Moretti, Lyskov et al. 2018)); docking with more than two proteins (ClusPro, GRAMM-X); and docking including degrees

of flexibility (SwarmDock (Torchala, Moal et al. 2013), ATTRACT (de Vries, Schindler et al. 2015), HADDOCK (de Vries, van Dijk et al. 2010)).

Attempts to address the limitations of computational docking have led to placing increasing focus on data-driven docking (Rodrigues and Bonvin 2014) and many servers now allow the user to specify interface residues and/or distance restraints, including ZDOCK, FRODOCK 2.0 (for refinement), pyDockWEB, ClusPro, GRAMM-X, SwarmDock, HADDOCK and ATTRACT. Some servers such as ClusPro and pyDockSAXS (Jimenez-Garcia, Bernado et al. 2020) can specifically use experimental SAXS data.

One of the main features differentiating existing docking servers is the nature of the scoring function used to discriminate correct from incorrect docking models. Most scoring strategies use either physics-based or statistical potentials. Understanding how binding partners co-evolved can provide essential clues to improve the structural prediction of protein interfaces. Several servers enable the prediction of inter-molecular contacts such as EVcomplex (Hopf, Scharfe et al. 2014), GREMLIN (Ovchinnikov, Kamisetty et al. 2014) and I-COMS (Iserte, Simonetti et al. 2015); however, such methods still have limited applicability due to the difficulty in building large enough joint multiple sequence alignments (MSAs) for the two protein partners. We developed the InterEvScore scoring function incorporating co-evolutionary information into the docking process, which improves predictions for as few as 10 sequences in the joint MSAs (Andreani, Faure et al. 2013). We integrated this scoring function into the InterEvDock pipeline (Yu, Vavrusa et al. 2016). InterEvDock is based on rigid-body sampling by FRODOCK (Garzon, Lopez-Blanco et al. 2009) followed by re-scoring using the SOAP-PP atomic statistical potential (Dong, Fan et al. 2013) and InterEvScore (Andreani, Faure et al. 2013) and consensus model selection. To date, the InterEvDock web server is the only free docking server allowing to directly predict the structure of protein-protein interactions using co-evolutionary information. We successfully used the InterEvDock strategy to guide our predictions in recent Critical Assessment of Predicted Interactions (CAPRI) rounds: for CAPRI rounds 28–35, our group ranked first by making correct predictions for 10 out of 18 targets (Yu, Andreani et al. 2017).

Very often, the individual structures of the exact proteins involved in a complex of biological interest are not known. On the other hand, structural models can be obtained for a large fraction of proteins in interaction networks thanks to homology modelling (Mosca, Céol et al. 2013), making them amenable to protein-protein docking. To date, most free docking servers, except the HDOCK server, allow users to provide only input structures but no input sequences for the protein partners.

Based on the user-oriented considerations mentioned above, here we introduce the InterEvDock2 server which represents a major evolution over the original InterEvDock. Protein sequences can now be provided as input, and not only 3D structures. To handle sequence inputs, we have added a module that performs comparative modelling prior to docking based on an automatic template search protocol. In case the user has biological input such as a position that is known to be involved in the interface between the two protein partners or a pair of residues known to be in contact, restraints with a tunable distance threshold can be specified for use in the docking procedure. This is crucial to ensure that all available biologically relevant information is used for InterEvDock2 predictions. In addition, InterEvDock2 implements the possibility to submit structures of oligomers as input to the pairwise free docking. Such an option is generally complicated in co-evolution analyses since the joint MSAs have to be generated for every chain of an oligomer. This process is now fully automated in InterEvDock2, allowing users to submit inputs such as homodimers or more complex structures as that of the nucleosome made of ten subunits. InterEvDock2 also benefits from improved accuracy by integrating the most recent FRODOCK 2.1 algorithm for rigid-body docking and scoring (Ramírez-Aportela, López-Blanco et al. 2016) and implementing an improved consensus selection and from a speed-up in the generation of joint MSAs for the two protein partners. The InterEvDock2 pipeline was benchmarked on 812 complexes from the PPI4DOCK database (Yu, Vavrusa et al. 2016) designed to ensure unbiased evaluation of the performance of free docking from unbound homology models. 29% of those 812 cases have an acceptable or better solution among the top 10 consensus models returned by InterEvDock2. As InterEvDock, InterEvDock2 also outputs a list of the 10 residues most

likely involved in the interface and at least one residue was correctly predicted in 91% of the 812 benchmark cases.

2.1 THE INTEREVDOCK2 SERVER

2.1.1 Web interface

Users are expected to provide for each protein partner either an input sequence or an input structure (Figure 2-1). Input structures can be uploaded or retrieved automatically from the Protein Data Bank (PDB) by typing in the PDB code and optionally one or more chain identifier(s). More options are available through the "advanced options" menu (see Figure 2-2). Optional breakpoints can be selected, either after template search to choose among up to 20 suggested templates prior to modelling (Figure 2-2A), or after modelling for interactive visualisation of the models prior to docking (Figure 2-2B). When input sequences are provided, users can specify which template to use for homology modelling; as for structure inputs, the template can be uploaded or directly retrieved from the PDB. If providing a template, users can also optionally enforce the query-template alignment for modelling. It is also possible to provide only a query-template alignment obtained from a previous server run in which a template search was performed (without modifying the identifiers), in which case the input sequence and the template PDB will be automatically retrieved based on the alignment. Several options are offered to tune the modelling: by default only loops (insertions) shorter than 14 residues are rebuilt during the modelling and N-terminal and C-terminal extensions are not modelled, but maximal lengths for modelling of loops, N-terminal and C-terminal extensions can be defined by the user (Figure 2-2E). Additionally, for input structures or sequences, users may define constraints that will be used to filter docking solutions; these constraints can be a single interface residue or a pair of residues in contact. Users can optionally specify the distance that will be used for each constraint (Figure 2-2C). An InterEvDock2 session identifier can also be provided in order to re-use docking results from a previous run and test different constraints (Figure 2-2D). As in InterEvDock, users may input the joint MSAs used for co-evolution-based scoring; otherwise the joint MSAs will be built by the server through an automated procedure. In case an oligomeric structure is submitted as one of the two docking partners, the joint MSAs will also be automatically calculated and processed by the server for every chain of the oligomer. A demonstration case

using sequences as input to the docking (and optionally a constraint) is available from the InterEvDock2 submission page.

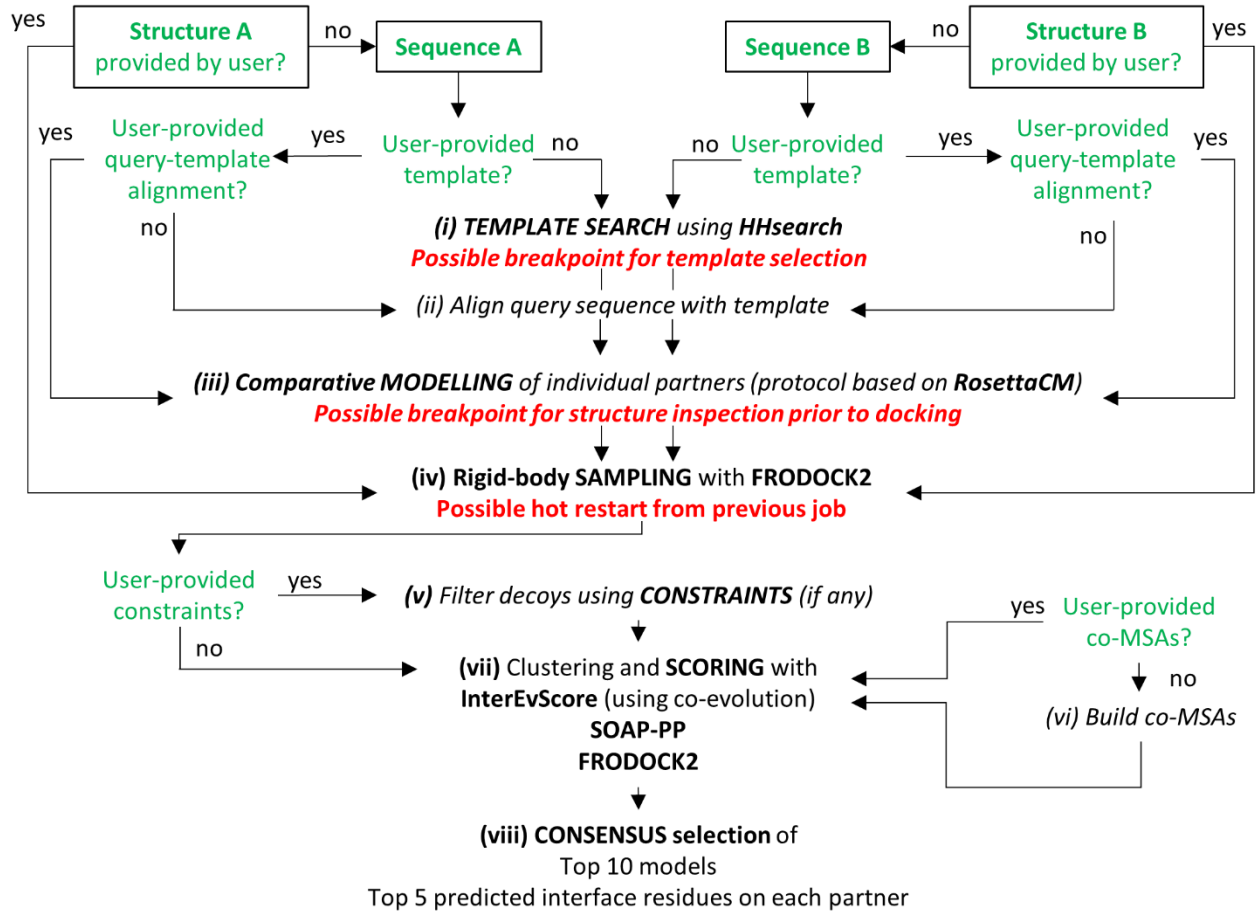
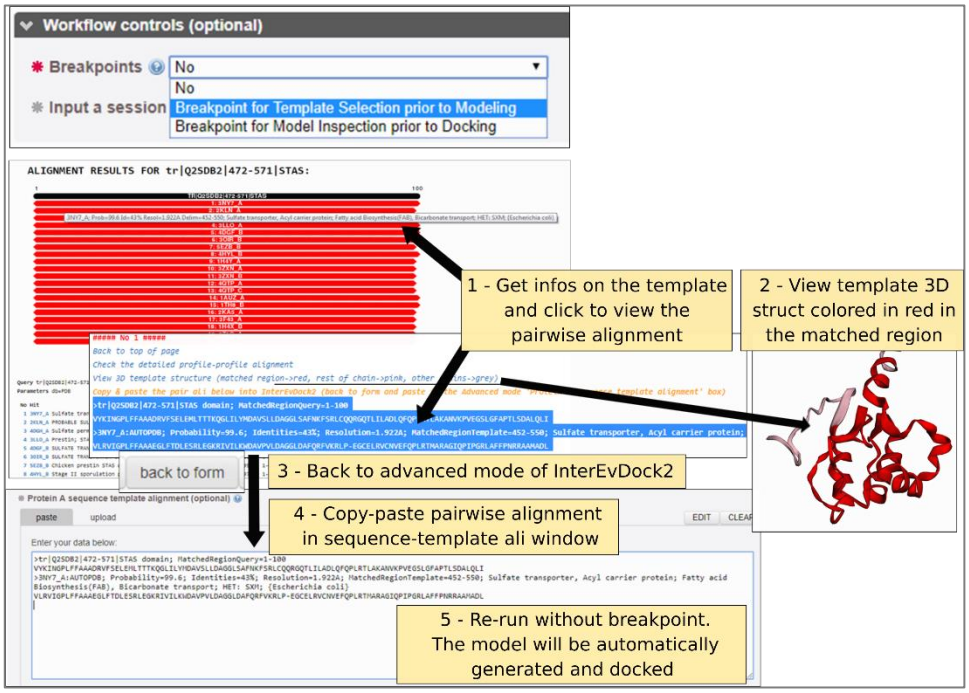
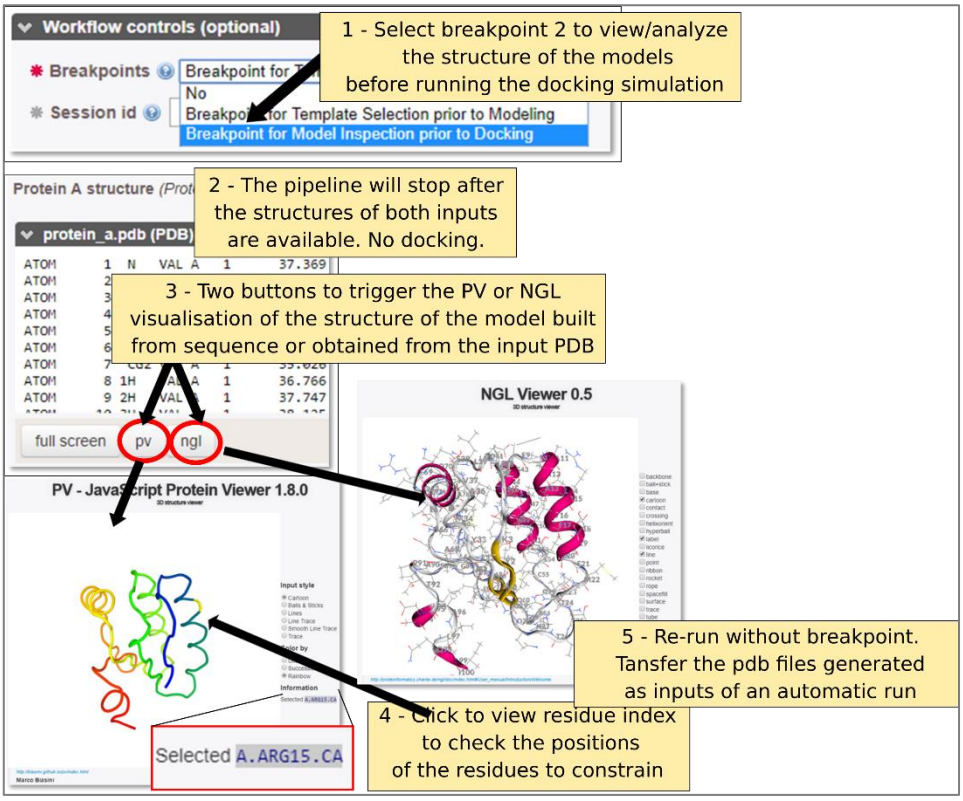


Figure 2-1: InterEvDock2 pipeline. *Eight steps can be performed in InterEvDock2 depending on the user input, out of which three mandatory steps (iv), (vii) and (viii) are always performed as they were in the original InterEvDock pipeline (Yu, Vavrusa et al. 2016), except that the FRODOCK algorithm was updated to version 2.1 (Ramírez-Aportela, López-Blanco et al. 2016) and the consensus calculation was slightly modified to save time and improve results. New features are available allowing the user to provide only an input sequence for one or both partners: (i) if the user does not provide a template, search for a suitable template using HHsearch (Soding 2005, Remmert, Biegert et al. 2011); (ii) if the user provides a template but no query-template alignment, alignment of query sequence with template sequence using MAFFT (Katoh and Standley 2013); (iii) once a template and a query-template alignment are available for each partner with no user-provided structure, comparative modelling using a RosettaScripts (Fleishman, Leaver-Fay et al. 2011) protocol based on RosettaCM (Song, DiMaio et al. 2013) to build a 3D model for (at least part of) the input sequence. Once a 3D structure or a structural model is available for each partner: (iv) exhaustive sampling using the rigid-body method FRODOCK 2.1; (v) (new feature) if the user provides information on residues (or pairs of residues) involved in the interface, applying constraints to filter sampled solutions; (vi) if the user does not provide a joint MSA for the two protein partners, co-MSA generation; (vii) clustering and scoring by three scores, FRODOCK 2.1, SOAP-PP (Dong, Fan et al. 2013) and InterEvScore (Andreani, Faure et al. 2013); (viii) clustering and selection of the InterEvDock2 consensus top 10 decoys. Green text indicates user input. Red text indicates possible breakpoints and hot restart. Italics indicate the optional steps in the pipeline, depending on the input provided by the user. Details for each step are provided in the Supplementary Methods (Appendix B. page 170).*

A


Workflow controls (optional)
Breakpoints
No
Input a session
No
Breakpoint for Template Selection prior to Modeling
Breakpoint for Model Inspection prior to Docking
ALIGNMENT RESULTS FOR tr|Q2SDB2|472-571|STAS:
1 - Get infos on the template and click to view the pairwise alignment
2 - View template 3D struct colored in red in the matched region
back to form
3 - Back to advanced mode of InterEvDock2
Protein A sequence template alignment (optional)
paste
upload
EDIT
Enter your data below:
4 - Copy-paste pairwise alignment in sequence-template ali window
5 - Re-run without breakpoint. The model will be automatically generated and docked

B


Workflow controls (optional)
Breakpoints
Session id
No
Breakpoint for Template Selection prior to Modeling
Breakpoint for Model Inspection prior to Docking
1 - Select breakpoint 2 to view/analyze the structure of the models before running the docking simulation
Protein A structure
protein_a.pdb (PDB)
2 - The pipeline will stop after the structures of both inputs are available. No docking.
3 - Two buttons to trigger the PV or NGL visualisation of the structure of the model built from sequence or obtained from the input PDB
full screen
pv
ngl
NGL Viewer 0.5
PV - JavaScript Protein Viewer 1.8.0
5 - Re-run without breakpoint. Tansfer the pdb files generated as inputs of an automatic run
4 - Click to view residue index to check the positions of the residues to constrain
Selected A.ARG15.CA

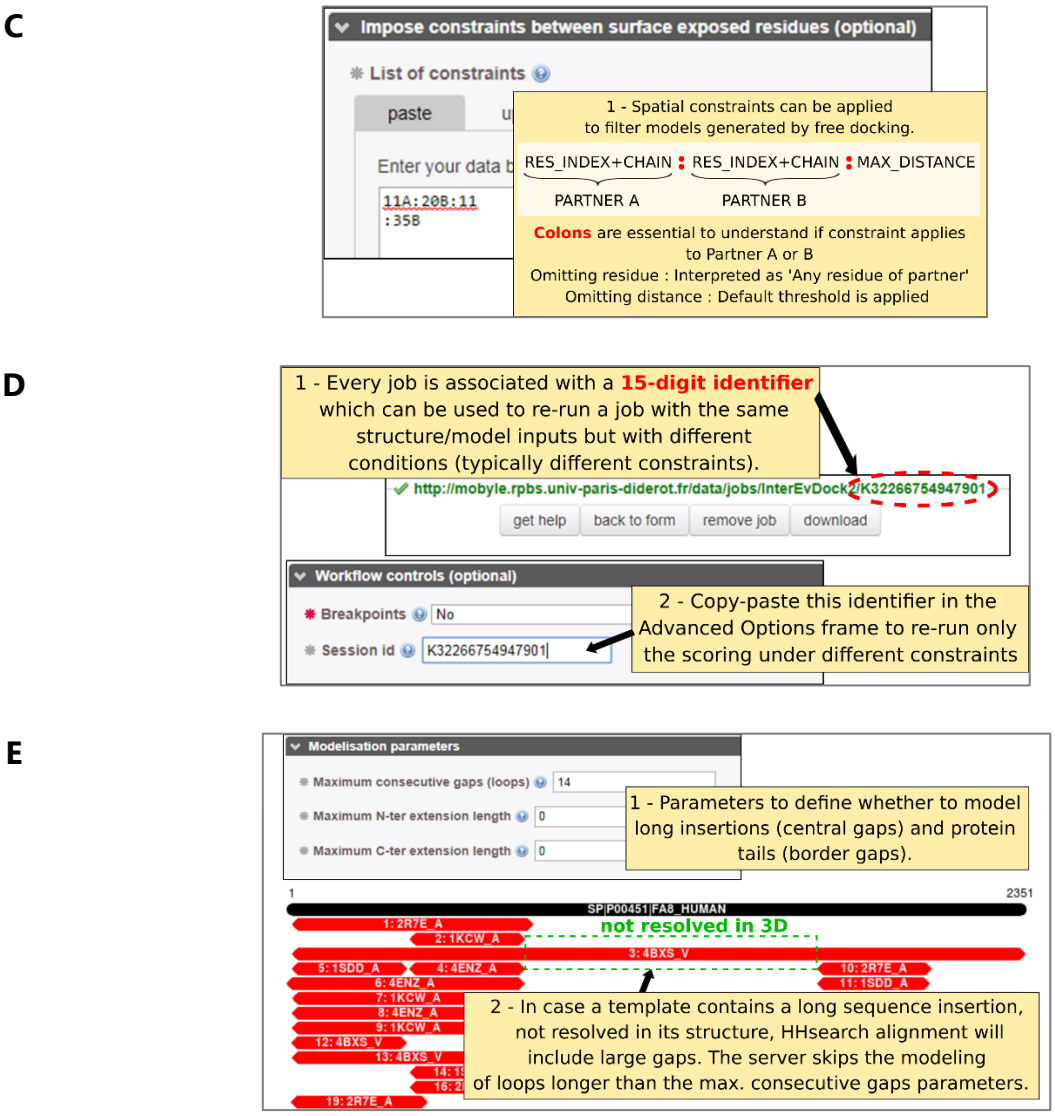


Figure 2-2: Explanation of InterEvDock2 advanced options. *(A) Interruption of runs (breakpoint) for users to select templates for modelling instead of automatic selection. (B) Interruption of runs (breakpoint) for users to inspect their structural input prior to docking. (C) Imposing distance constraints to include biological information. (D) Re-running scoring only with different constraints without having to re-run the docking part. (E) Controlling the size of loops and extensions to model.*

The web page resulting from an InterEvDock2 submission contains information about the best-ranked decoys, which can be explored interactively thanks to the PV WebGL applet (M. Biasini, https://dx.doi.org/10.5281/zenodo.12620). Detailed results are available in a downloadable archive, also containing a script for easy loading and offline visualisation of the best docking solutions with PyMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC).

The InterEvDock2 server benefits from parallelised implementation in the dedicated infrastructure built at RPBS and from data privacy ensured in the Mobyle framework.

2.1.2 Molecular docking procedure

Figure 2-1 presents the InterEvDock2 pipeline which consists of eight steps (for more details about each step, see section A.). The three core docking steps – sampling with FRODOCK2.1 (iv), clustering with FRODOCK2.1 and scoring with InterEvScore and SOAP-PP (vii) and consensus calculation (viii) – are always performed. Step (vi) consists in automatically generating the joint MSAs used by InterEvScore to account for co-evolution in the scoring process, unless the joint MSAs are provided by the user. Steps (iv), (vi) and (vii) are unchanged compared to the original InterEvDock pipeline (Yu, Vavrusa et al. 2016), except that the FRODOCK algorithm was updated to version 2.1 (Ramírez-Aportela, López-Blanco et al. 2016). In the final step (viii) a consensus list of 10 most likely models is calculated. Since decoys well ranked by at least two different scoring methods (out of the three methods used in InterEvDock2) have higher chances of being correct, the 3*top 10 models for each score are re-ranked according to the number of similar decoys (defined as ligand RMSD ≤ 10 Å) within the top 50 models of the other two scores (down to a minimum of two similar decoys). In case of a tie, priority is given to InterEvScore top 10 models, then SOAP-PP, then FRODOCK. If necessary, the consensus list is then filled up to 10 models by selecting the best models from each score (4 from InterEvScore, 3 from SOAP-PP and 3 from FRODOCK). When building the consensus, models that are structurally redundant (i.e. ligand RMSD ≤ 10 Å) with previously selected models are excluded, so that the final list contains 10 structurally non-redundant models.

2.1.3 Docking from input sequences

If the user provides only an input sequence for one or both partners, steps (i) to (iii) can be applied. (i) If the user does not provide a template, the profile-profile comparison tool HHsearch is used to search for templates (Soding 2005, Remmert, Biegert et al. 2011); only templates with HHsearch probability higher than 95% are selected. The web server returns a list of up to 20 templates selected according to HHsearch probability, query-template sequence identity and structural resolution (see details in section A.). By setting the breakpoint after template search, the user can choose to start modelling from any of these templates by copy-pasting the query-template alignment to the server submission form; otherwise the

best template found by the automatic procedure is used. If no suitable template is identified, no modelling is performed. (ii) If the user provides a template but no query-template alignment, the query sequence is aligned with the template sequence using MAFFT (Katoh and Standley 2013). (iii) Once a template and a query-template alignment are available for each protein with no user-provided structure, comparative modelling using a RosettaScripts (Fleishman, Leaver-Fay et al. 2011) protocol based on RosettaCM (Song, DiMaio et al. 2013) is performed to build a 3D model for (at least part of) the input sequence. Due to runtime considerations, compared to the procedure used to build the PPI4DOCK database (Yu, Vavrusa et al. 2016), the comparative modelling protocol implemented in the InterEvDock2 web server involves fewer optimisation cycles (see protocol details in the Supplementary Methods, Appendix B. page 170). This protocol is quite robust for templates with relatively high homology but it can lead to loss of precision for more remote templates (typically when both templates have less than 50% sequence identity with the query proteins). By default, to avoid spending time reconstructing regions that are not present in the template, only loops (insertions) shorter than 14 residues are rebuilt during the modelling and N-terminal and C-terminal extensions are not modelled, but maximal lengths for modelling of loops, N-terminal and C-terminal extensions can be tuned by the user.

2.1.4 User-defined constraints

Step (v) applies if the user provides information on residues (or pairs of residues) involved in the interface: restraints are applied to filter sampled solutions. The distance used to enforce restraints can be modulated which offers the possibility to integrate data from various sources. The default distance was set to 8 Å for constraints on single positions and 11 Å for pair constraints (see section d. for a detailed justification of these thresholds). When constraints are provided by the user, the output returned by the server will provide information about whether or not each constraint was used during docking (e.g. constraints on residues not exposed on the surface of the protein are excluded).

2.1.5 Runtime

The core docking steps (iv), (vii) and (viii) take altogether around 30 min for proteins of size 200 residues and 1 hour for proteins of size 400-500 residues. Template search and query-template alignment steps (i) and (ii) take only a few minutes, whatever the size of the proteins. The comparative modelling step (iii) was optimised for speed as reported above and typically takes 5 to 20 minutes depending on the size of the proteins and the query-template sequence identities. Compared to InterEvDock, InterEvDock2 benefits from a large speed-up in step (vi) for the generation of joint MSAs for two protein partners which was a key bottleneck. This step now typically lasts ~3 min for proteins of 200 residues and ~15 min for proteins of 400-500 residues.

2.2 RESULTS

2.2.1 Benchmarking on PPI4DOCK

To assess the predictive power of the InterEvDock2 server on 3D models, we have set up the most extensive benchmark to date, using unbound models as input of the docking simulations. The PPI4DOCK database (Yu and Guerois 2016) was designed to ensure unbiased evaluation of free docking performance and contains 1417 non-redundant heterodimeric docking targets based on unbound homology models. The InterEvDock2 pipeline was tested on the subset of 812 protein complexes from PPI4DOCK for which pairs of joint MSAs with more than 10 sequences could be obtained (excluding any antibody complex) and FRODOCK 2.1 (Ramírez-Aportela, López-Blanco et al. 2016) was able to generate at least one acceptable or better decoy among the top 10,000 decoys. The list of the 812 complexes used for benchmarking and detailed results are provided in https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/table.html. This benchmark dataset is roughly an order of magnitude larger than other typical docking benchmarks, among which the widely used Weng benchmark (Hwang, Vreven et al. 2010). For each of the 812 targets, PPI4DOCK provides unbound homology models of the two protein partners as well as the joint MSAs used for docking and scoring in the InterEvDock2 pipeline. As on the web server, the predictions for each case consist in the top 10 consensus interface models and the top 10 interface residues, which are used to assess the InterEvDock2 performance. A solution is defined as acceptable or better according to the criteria defined by the CAPRI consortium (Mendez, Leplae et al. 2003).

The prediction performance of InterEvDock2 is reported in Table 2-1. Among the 812 targets, 29% (239) have at least one model of acceptable or better quality in the top 10 consensus obtained from the InterEvDock2 pipeline, which represents a significant improvement over the top 10 success rates of the three individual scores used to build the consensus (see Figure 2-3). The 812 complexes belong to four difficulty levels (PPI4DOCK categories) based on the quality of the superimposed interface model (two unbound models superimposed

on the bound structure): "very easy" (174 complexes), "easy" (498 complexes), "hard" (118 complexes) and "very hard" (22 complexes). Other "very hard" and all "super hard" PPI4DOCK targets do not satisfy the condition that FRODOCK 2.1 was able to generate at least one acceptable or better decoy among the top 50,000 decoys, since they may require flexibility in the docking process (Yu and Guerois 2016), and are therefore not included in the present benchmark. As expected, the InterEvDock2 top 10 consensus success rate decreases with increasing difficulty of the test cases, from 43% for the "very easy" PPI4DOCK category to 30% for the "easy" category, 11% for the "hard" category and 5% for the "very hard" category. Analysis of InterEvDock2 performance depending on the minimum sequence identity between the target and template shows a moderate drop in success rate for models built with remote templates (< 30% sequence identity) and an increased success rate for models built with very close templates (>=95% sequence identity), compared to the overall InterEvDock2 success rate (see Supplementary Materials, Appendix 177 Table B-2).

Table 2-1: InterEvDock2 performance on PPI4DOCK. *Prediction performance of the InterEvDock2 server on 812 complexes of the PPI4DOCK benchmark, split into four levels of difficulty: very easy, easy, hard and very hard. The benchmark is made of the 812 targets of the PPI4DOCK benchmark (1417 cases) (Yu and Guerois 2016) for which pairs of co-evolved MSAs with more than 10 sequences could be obtained and FRODOCK 2.1 (Ramírez-Aportela, López-Blanco et al. 2016) was able to generate at least one acceptable or better decoy (Mendez, Leplae et al. 2003) among the top 50,000 decoys. In the upper part of the table, top 10 success rates are reported as the number of cases (and percentage between brackets) for which at least one model out of 10 is an acceptable or better solution. Assessed methods are InterEvScore (Andreani, Faure et al. 2013), SOAP-PP (Dong, Fan et al. 2013), FRODOCK 2.1 (Ramírez-Aportela, López-Blanco et al. 2016), InterEvDock2 consensus (this work and (Yu, Vavrusa et al. 2016)) and Zdock3.0.2 (Pierce, Hourai et al. 2011). In the lower part of the table, the number (and percentage) of cases for which at least one residue out of the top 10 or top 2 residues was correctly predicted as present in the complex interface is assessed for InterEvDock2 and Zdock3.0.2 (see*

calculation details in Supplementary Materials, Appendix B. page 174). The best results for each category are highlighted in bold.

		All	Very easy	Easy	Hard	Very hard
Number of cases		812	174	498	118	22
Top 10 success rate	**InterEvScore**	171 (21%)	44 (25%)	115 (23%)	11 (9%)	1 (5%)
	SOAP_PP	194 (24%)	55 (32%)	126 (25%)	12 (10%)	1 (5%)
	FRODOCK 2.1	164 (20%)	55 (32%)	102 (20%)	5 (4%)	**2 (9%)**
	InterEvDock2 consensus	**239 (29%)**	**75 (43%)**	**150 (30%)**	**13 (11%)**	1 (5%)
	Zdock 3.0.2	126 (15%)	33 (19%)	83 (17%)	9 (8%)	1 (5%)
Residue interface prediction (≥1 correct in top 5 receptor OR top 5 ligand)	**InterEvDock2**	**735 (91%)**	**160 (92%)**	**450 (90%)**	**103 (87%)**	**22 (100%)**
	Zdock3.0.2	680 (84%)	145 (83%)	427 (86%)	91 (77%)	17 (79%)
Residue interface prediction (≥1 correct in top 5 receptor AND top 5 ligand)	**InterEvDock2**	**414 (51%)**	**103 (59%)**	**263 (53%)**	**39 (33%)**	**9 (41%)**
	Zdock3.0.2	345 (43%)	76 (44%)	228 (46%)	33 (28%)	8 (34%)
Residue interface prediction (≥1 correct in top 1 receptor OR top 1 ligand)	**InterEvDock2**	**613 (75%)**	**140 (80%)**	**385 (77%)**	**71 (60%)**	**17 (77%)**
	Zdock3.0.2	532 (66%)	111 (64%)	344 (69%)	64 (54%)	13 (58%)
Residue interface prediction (≥1 correct in top 1 receptor AND top 1 ligand)	**InterEvDock2**	**278 (34%)**	**75 (43%)**	**184 (37%)**	**17 (14%)**	2 (9%)
	Zdock3.0.2	195 (24%)	44 (25%)	133 (27%)	15 (12%)	**3 (14%)**

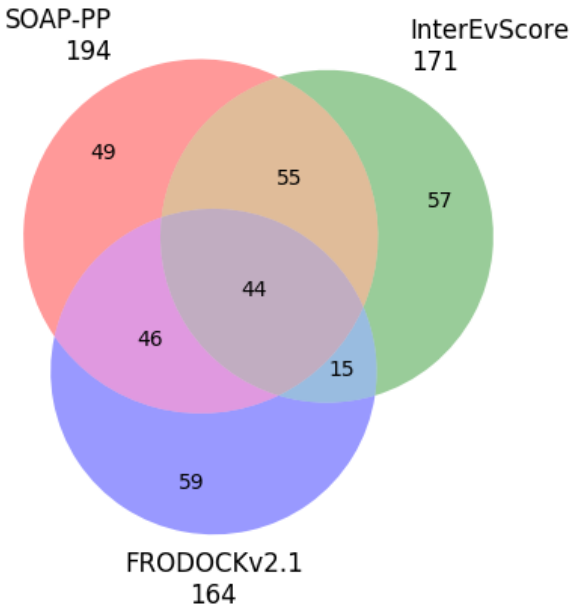


Figure 2-3: Venn diagram of prediction performances for the three scoring components in InterEvDock2. *Out of the 812 cases in the PPI4DOCK set used to benchmark InterEvDock2, 171, 194 and 164 cases have at least one decoy of acceptable or better quality in the top 10 decoys scored by InterEvScore, SOAP-PP and FRODOCK2.1 respectively. However, as illustrated in the Venn diagram below, the three scores are quite complementary, as 57, 49 and 59 cases were detected by InterEvScore, SOAP-PP or FRODOCK2.1 alone respectively, thereby highlighting the interest of using a consensus between the three scores.*

Direct comparisons with previous benchmarks are difficult because the benchmark dataset used here is much larger than others datasets typically used to assess docking and scoring performance. Comparison with previously reported success rates on the Weng benchmark (Hwang, Vreven et al. 2010, Yu, Vavrusa et al. 2016) are details in the Supplementary materials (see Appendix B. page 177 and Table B-3 and Table B-4). An interesting feature of the Weng benchmark compared to PPI4DOCK is that it contains targets where one partner is multimeric. Out of the 85 cases from the Weng benchmark that can be used for InterEvDock2 benchmarking, 16 contain a multimeric partner. The InterEvDock2 top 10 consensus contains an acceptable or better solution for 7 out of these 16 cases (44%). This success rate is comparable to the overall success rate of InterEvDock2 on the much larger PPI4DOCK benchmark (29%) and on the 85 cases of the Weng benchmark (32%). Additionally, docking using multimeric partners has the advantage that potentially "sticky" interface regions involved in multimeric interactions of one partner are buried in the multimeric interface and therefore masked for the docking process.

In Table 2-1, the InterEvDock2 performance is also compared to the performance of the widely-used rigid-body docking programme Zdock3.0.2, assessed on the same 812 complexes from the PPI4DOCK benchmark. For each case, 54,000 decoys are generated and ranked by Zdock3.0.2. In 126 out of 812 cases (15%), an acceptable or better solution is found among the top 10 decoys. Altogether, these benchmarking results highlight the added value of the InterEvDock2 processing pipeline, in particular the clustering and consensus scoring steps.

Of key interest for experimental biologists, the InterEvDock2 output offers a list of 10 residues most likely involved in the complex interface (5 predicted residues on each partner) that can be targeted for mutagenesis. For these residue predictions, we reach 91% success rate, with 735 of the 812 benchmark cases having at least one of the 10 predicted residues involved in the actual interface (Table 2-1). As was found for the 85 cases from the Weng benchmark used to assess the original InterEvDock performance (Yu, Vavrusa et al. 2016), this success rate is remarkably stable with increasing difficulty: from 92% for very easy cases to 90% for easy cases to 87% for hard cases. Predictions of the InterEvDock2 server can also be used as a prior to constrain more thorough docking simulations including flexibility. In that perspective, in 51% of the cases, at least one correct residue is predicted on both sides of the interface (59% for very easy targets, 53% for easy targets and 33% for hard targets). Results are also presented in Table 2-1 and Supplementary Materials Figure B-1 for only the top 2 predicted residues (one on each partner): at least one of the two predicted residues is correct in 75% of the cases and both are correct in 34% of the cases, highlighting the practical value of InterEvDock2 residue prediction. All those results are significantly higher than a reference interval given by random selection of residues on the surface of the protein (for calculation details see Appendix B. page 174 and Figure B-1.

2.2.2 Predictions of CAPRI targets

The InterEvDock2 pipeline was challenged through our participation in all CAPRI rounds since 2013. Focusing on heteromeric targets evaluated at the sixth CAPRI evaluation meeting (rounds 28-35), our group ranked first with 10 correctly predicted targets out of 18. Among

those 10 targets, our best prediction among ten submitted models was of high quality in 1 case, medium quality in 7 cases and acceptable quality in 2 cases (Yu, Andreani et al. 2017). In 15 of the total 18 targets, evolutionary information was available in the form of either co-evolution or conservation, providing key constraints to guide docking towards the correct solution. Although the InterEvDock2 pipeline was not specifically designed to handle homo-oligomeric docking, we were also among the highest ranking groups in the two joint CASP-CAPRI experiments involving mostly predictions of homodimers (Lensink, Velankar et al. 2016, Lensink, Velankar et al. 2017). Of note, for most CAPRI targets since 2013, only sequence information was provided to the participants. Figure 2-4 illustrates an InterEvDock2 run for CAPRI target T95 (round 31) involving docking between the nucleosome (a decameric structure) and the PRC1 ubiquitin ligase (a trimeric structure).

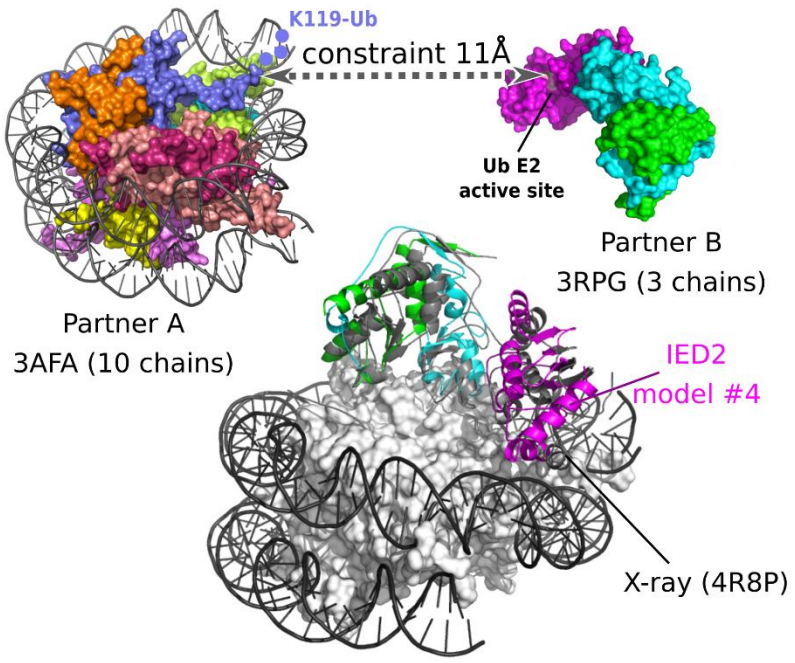


Figure 2-4: Successful CAPRI target T95 prediction with InterEvDock2. *Successful example of docking from multimeric inputs in a CAPRI target. Prediction for CAPRI target T95 involving docking between two multimeric inputs: the nucleosome and the PRC1 ubiquitin ligase. These multimeric inputs were directly used as inputs in the InterEvDock2 server (PDB identifiers 3afa for the nucleosome and 3rpg for the ubiquitin ligase). A constraint is additionally used between a residue close to the ubiquitinated lysine K119 and the active site of the ubiquitin ligase (constraint between residues 117C and 85A at distance 11 Å). The first acceptable solution (ranked #4 in the Top 10 InterEvDock2 consensus) is superimposed on the reference crystal structure (PDB: 4r8p).*

2.2.3 Description of docking case studies from input sequences and using constraints

To illustrate the biological relevance of InterEvDock2 predictions, we consider two docking case studies derived from the PPI4DOCK benchmark (Figure 2-5). The first case is a complex between the Rho-family GTPase Cdc42 and the conserved, catalytic domains of exchange factor intersectin. Details of this interaction (PDB identifier: 1ki1) and structure-based mutagenesis revealed key features of the activation of Cdc42 by intersectin (Snyder, Worthylake et al. 2002). This case was tested on the InterEvDock2 server by providing input sequences of the interacting regions in the two partners. Unbound templates were imposed for both proteins as in the PPI4DOCK benchmark; otherwise the automatic template search might have found the bound partners belonging to PDB 1ki1 or other bound templates. The unbound templates (4f38A and 3odoA) have sequence identities of 54% and 25% with the modelled regions of Cdc42 and intersectin, respectively. Among the top 10 consensus models returned by InterEvDock2, one acceptable solution is found as top 2 (Figure 2-5A).

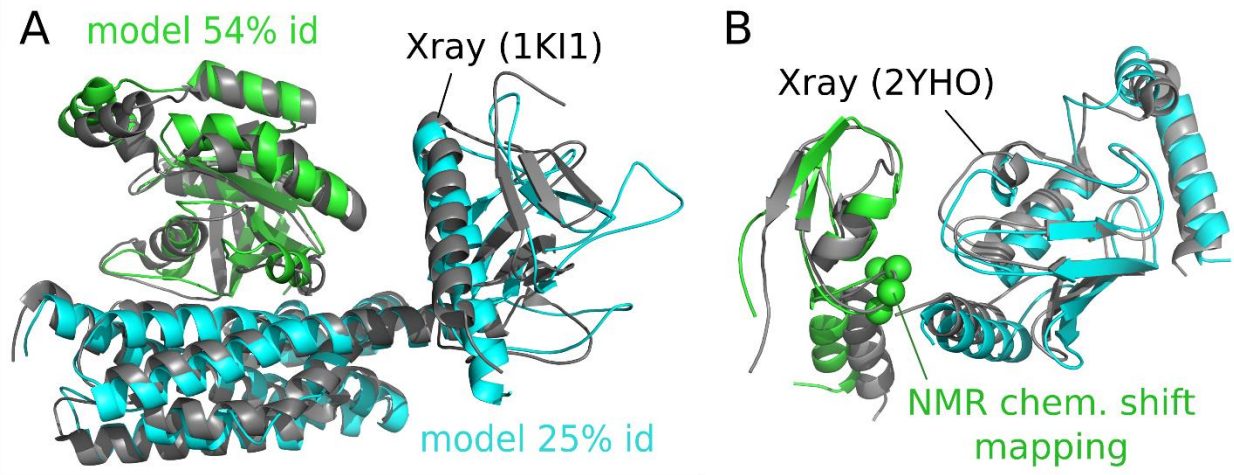


Figure 2-5: Successful prediction of a case in PPI4DOCK using InterEvDock2. *Successful examples from the PPI4DOCK database. (A) Top 2 consensus model found by InterEvDock2 for docking between unbound homology models of Cdc42 (green, modeled using an unbound template at 54% sequence identity) and the conserved, catalytic domains of intersectin (cyan, modeled using an unbound template at 25% sequence identity). The model is superimposed on the reference crystal structure (PDB identifier: 1ki1) (gray). It is acceptable with interface RMSD 4.03 Å. (B) Best model found in the InterEvDock2 top 10 consensus for docking between PPI4DOCK unbound homology models of the RING domain of IDOL (green) and UBE2D (cyan) when four residues experimentally known to be important for the interaction are used as constraints (with default distance 8 Å). The model is superimposed on the reference crystal structure (PDB identifier: 2yho) (gray). The model is acceptable with interface RMSD 2.29 Å and is ranked first of the top 10 consensus. The four residues used as constraints from chemical shift mapping are shown as green spheres (M388, V389, C390 and C391).*

The second case illustrates the interest of docking with user-defined restraints. We consider a complex between the RING domain of E3 ubiquitin protein ligase IDOL and ubiquitin-conjugating enzyme E2 UBE2D (PDB identifier: 2yho) (Zhang, Fairall et al. 2011). This interaction is involved in the regulation of cholesterol uptake. Nuclear magnetic resonance (NMR) chemical shift mapping was used to confirm the interacting region prior to crystallographic studies. This NMR analysis showed four residues (M388, V389, C390, and C391) in the RING domain of IDOL to have particularly high chemical shift variation upon binding of UBE2D. The PPI4DOCK models of the interacting regions of IDOL and UBE2D (built by homology modelling using unbound templates for the two proteins, respectively 2yhnA and 3bzhA with sequence identities of 100% and 61%) were submitted to InterEvDock2. Two runs were performed, one without constraints and one using the four residues identified by NMR as interface constraints. Among the top 10 consensus models returned by InterEvDock2, the highest-ranked acceptable solution (medium quality according to the CAPRI criteria) was ranked number 6 in the run without constraints. When using the constraints derived from experimental NMR data, there were two acceptable or better solutions in the InterEvDock2 top 10 consensus: one was ranked first (Figure 2-5B) and the second ranked number 6.

2.3 CONCLUDING REMARKS

InterEvDock2 represents a major, user-oriented evolution of InterEvDock. InterEvDock2 is still the only free docking server taking into account co-evolutionary information, relying on a combination of complementary scoring functions to identify the most likely interface models. The previous InterEvDock version was limited by its requirement of only dealing with monomeric inputs. InterEvDock2 greatly expands the range of applications to homo- and hetero-oligomers by handling multimeric chains in the two input proteins used for pairwise docking and the automated processing of their joint MSAs. Benchmarking results on PPI4DOCK emphasize the usefulness of InterEvDock2 in generating interface models of good quality in the scope of integrative structural biology. The InterEvDock2 server returns docking results within typical runtimes of 30 minutes (for proteins of around 100 residues) to 2 hours (for proteins of around 500 residues) even when starting from input sequences, while performing well on our benchmark of 812 cases docked from unbound homology models. The server also benefits from a user-friendly submission and visualisation interface, including breakpoints after template search and homology modelling, and options for offline in-depth analysis with PyMOL. InterEvDock2 is thus designed as a useful tool for biologists who can very easily submit docking runs starting from simple input sequences and specify constraints to make use of any previously acquired experimental knowledge. InterEvDock2 results can assist biologists in designing hypotheses about molecular interaction mechanisms and interface mutations to investigate the functional role of an interaction.

CHAPTER 3
Reconciling evolutionary information and atomic detail in scoring

In computational structural biology, we are constantly trying to improve the performance of our prediction methods. As we have seen previously, there is a lot to learn from a protein's evolutionary history. Proteins evolve under the constraint of maintaining functional interactions and this constraint is reflected in coupled MSAs. As a consequence, these alignments provide valuable information for the purpose of interface computational prediction. This chapter is dedicated to the exploratory concept of extrapolating evolutionary information to the atomic level of detail. Its use in scoring interface predictions combined with atomic-resolution scoring functions has shown promising results. The results of this work are in the process of being submitted for publication (pre-print deposited in HAL: and BioRxiv (Quignot, Granger et al. 2020)) and we are currently implementing this methodology in another update of our docking server InterEvDock2.

As described in the previous chapter, evolutionary information can be especially useful to guide molecular docking (Geng, Jung et al. 2019). The benchmarking of InterEvDock2 showed us that InterEvScore presents a high complementarity with SOAP-PP (Quignot, Rey et al. 2018). As both scores are based on statistical potentials but SOAP-PP has an atomic level of detail, we hypothesised that a score integrating evolutionary information at an atomic scale might pick up on finer properties to better distinguish near-natives from the rest of the decoys.

In InterEvScore, evolutionary information is given implicitly at residue-level through coMSAs and combined with a coarse-grained statistical potential. A major challenge in deriving evolutionary information to an atomic level of detail is finding a suitable way of representing residue-scale information from coMSAs at an atomic level. Here, we present a novel strategy to couple evolutionary information with atomic scores in order to improve decoy discrimination. We reconstruct an equivalent and hypothetical interfacial atomic contact network for each interface decoy and for each pair of homologs present in the coMSAs, by using a threading-like strategy to generate explicit backbone and side-chain coordinates. These models can, in turn, be scored with non-evolutionary atomic-resolution scoring functions such as SOAP-PP (Dong, Fan et al. 2013) or Rosetta interface score (ISC) (Gray, Moughon et al. 2003, Chaudhury, Berrondo et al. 2011).

Here, we show that including explicit evolutionary information improves the top 10 success rate of SOAP-PP and ISC by 6 and 13 percentage points respectively, on a large benchmark of 752 docking cases for which evolutionary information can be used (Yu and Guerois 2016). It also improves the top 10 success rate of the residue-level statistical potential from InterEvScore by 6.5 percentage points. We then use a consensus approach to take advantage of the complementarity between different scores. The top 10 success rate of a consensus integrating FRODOCK2.1 with InterEvScore and SOAP-PP increases from 32% to 36% when including the homology-enriched score variants. A more time-consuming consensus combining all scores with an explicit homolog representation reaches 40% top 10 success rate.

3.1 METHODS

3.1.1 Docking benchmark

As for InterEvDock2, we performed evaluation of docking methods on cases from the large docking benchmark our team developed to ensure unbiased evaluation, PPI4DOCK (Yu and Guerois 2016). Each case in PPI4DOCK is associated to a coMSA, i.e. a pair of joint MSAs for the two docking partners. We excluded antigen-antibody interactions and cases with less than 10 sequences in their coMSAs, in order to focus on cases with enough co-evolutionary information. Sampling was performed using FRODOCK2.1 (see detailed parameters in supplementary methods appendix C. A. page 180) and only the top 10,000 decoys ranked by FRODOCK2.1 were kept. Near-native decoys were defined as being of Acceptable or better quality in accordance with the CAPRI criteria (Mendez, Leplae et al. 2003). To focus the study on scoring performance, cases that did not have a near-native within the top 10,000 FRODOCK2.1 decoys were excluded from the benchmark. This resulted in a final benchmark of 752 cases (supplementary Table C-).

Performance was measured by top N success rate. We especially focus on the top 10 success rate traditionally used as a docking metric, and the top 50 success rate since consensus computation typically involves the top 50 decoys of each score (see section 3.1.2.1). Additional metrics are available in the supplementary information (supplementary methods appendix C. page 180).

3.1.2 Scoring functions

In addition to FRODOCK2.1's integrated score (Ramírez-Aportela, López-Blanco et al. 2016), we rescored decoys and their threaded homologs with InterEvScore, SOAP-PP, and Rosetta interface score (ISC).

InterEvScore (Andreani, Faure et al. 2013) was re-implemented with the great help of master's student Pierre Granger to accelerate the scoring step (see supplementary methods appendix C. page 180). We also use a faster implementation of SOAP-PP (Dong, Fan et al. 2013)

developed in collaboration with Pablo Chacón (see supplementary methods appendix C. page 180).

Rosetta interface score (ISC) includes a linear combination of non-bonded atom-pair interaction energies and empirical and statistical potentials among other terms (Gray, Moughon et al. 2003, Chaudhury, Berrondo et al. 2011). This score is calculated by subtracting the total energy of both monomeric structures from the total energy of the complex structure. Since Rosetta ISC is sensitive to small variations and clashes at the interface, we included high-resolution interface side-chain optimisation as a scoring option (see supplementary methods appendix C. page 180). Decoys for which Rosetta scoring did not converge after 10 iterations were assigned the worst score for that case. As Rosetta ISC scoring can take up to a couple of minutes per structure, we scored only the top 1,000 FRODOCK2.1 decoys (noted later 1k) per case rather than 10,000 (noted 10k).

3.1.2.1 Consensus scores

Consensus calculations were performed similarly to InterEvDock2 (see Chapter 2) to obtain a set of 10 most likely decoys depending on the agreement between several scoring functions. Here, we apply consensus scoring to combinations of 3 to 5 different scoring functions. For a given set of scoring functions, ordered according to their individual performances from best to worst performing, the top 10 decoys of each scoring function receive a convergence count based on the number of similar decoys (defined as L-RMSD ≤ 10 Å) that are found in the top 50 decoys of each other scoring function. The final 10 consensus decoys are selected iteratively by decreasing convergence count (if > 1). In the case of a tie, decoys are selected according to the ranking order of their respective scoring functions. Note that decoys are added to the top 10 consensus only if they are not structurally redundant with the already selected ones (L-RMSD > 10 Å). If necessary, the consensus list is completed up to 10 decoys by selecting the top 4, 3, 3 decoys for a consensus between three scoring functions (or the top 3, 3, 2, 2 or top 2, 2, 2, 2, 2 decoys for a consensus between four or five scoring functions, respectively).

3.1.3 Homology-enriched docking pipeline

For a pair of query proteins A and B for which we are trying to predict the 3D structure of the complex, the homology-enriched docking pipeline consists of four steps outlined in Figure 3-1. We dock proteins A and B using FRODOCK2.1 (Ramírez-Aportela, López-Blanco et al. 2016), thereby sampling a maximum of 10,000 decoys that can be reconstructed from the input query proteins using rotation and translation coordinates (Figure 3-1A). In parallel, we construct coMSAs and subsample them to a subset of M pairs of homologs (proteins A_1 and B_1, A_2 and B_2, ..., A_M and B_M, homologs of query proteins A and B respectively) (see section 3.1.3.1, page 109). We model the unbound structures of this subset of M pairs of homologs, using the threading function from RosettaCM's pipeline (Song, DiMaio et al. 2013) and the unbound query protein structures as templates (see Figure 3-1B and section 3.1.3.2, page 110). We then generate homolog equivalents to each query decoy by applying the translation and rotation coordinates generated during the query docking to each pair of homologs. Figure 3-1C illustrates this reconstruction for the first pair of homologs (proteins A_1 and B_1). To obtain the final score of each decoy, we average scores over the query decoy itself and its equivalent homolog decoys (Figure 3-1D). Note that for one case, we have to compute (M+1) x N scores to obtain the final ranking of N decoys. The scoring functions we used are described in section 3.1.2, page 106. All steps of the pipeline are easily parallelisable to reduce end-user runtime, whether through MPI (sampling step) or by splitting over decoys (scoring steps).

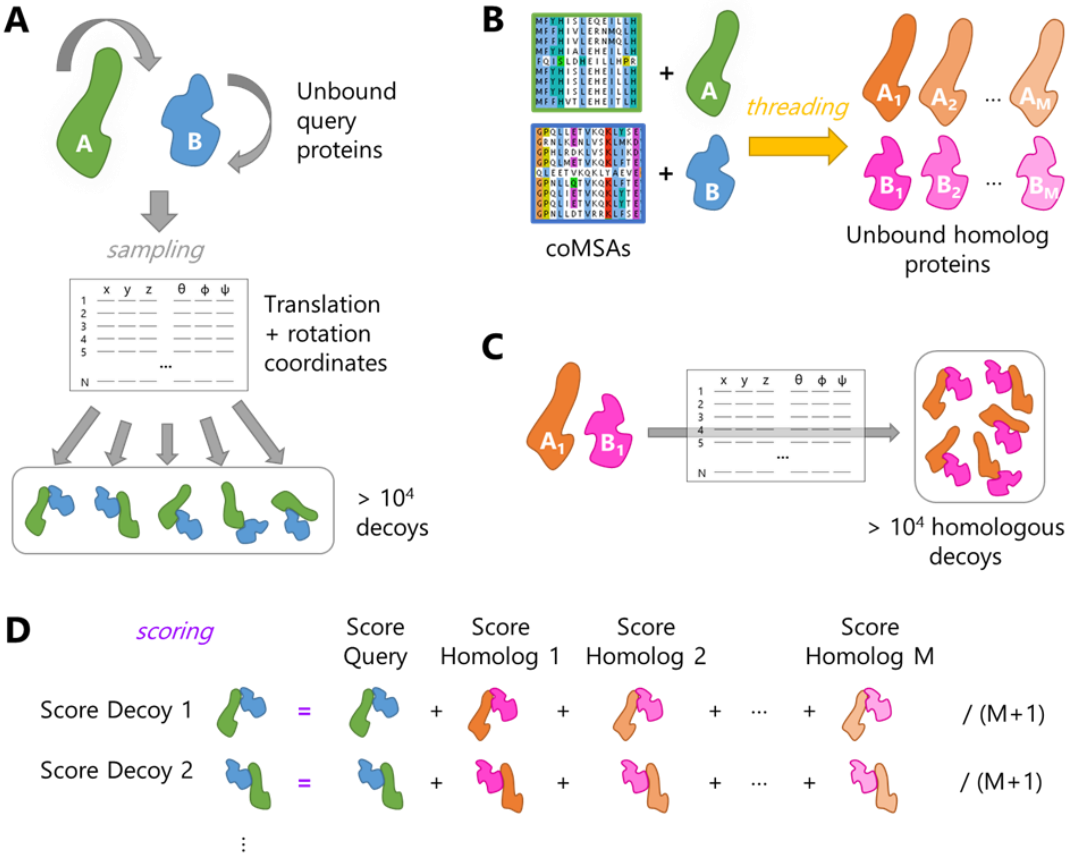


Figure 3-1: Docking pipeline with explicit modelling of decoy homologs. *A. Upon docking of query unbound structures (proteins A and B in green and blue), FRODOCK2.1 outputs a rotation and translation matrix to reconstruct the corresponding decoys. B. In order to generate their homologous counterparts, the unbound structures of each homolog (proteins A_1 and B_1, A_2 and B_2, ..., A_M and B_M, in various shades of orange and magenta) are threaded based on the query unbound structures (proteins A and B) and the homologous sequence alignments in the coMSAs of the query proteins. C. For each homolog pair (such as homolog 1 illustrated here), decoys can be reconstructed using the same rotation and translation matrix as for the query. D. The final score of each decoy (left column) corresponds to the average score over itself and its M homolog equivalents for a given scoring function.*

3.1.3.1 Subsampling homologs in the coMSAs

Homologous sequences used in scoring were taken from the coMSAs provided with the PPI4DOCK benchmark, reduced to maximum M=40 and then to M=10 sequences (plus the query sequence) to limit computational time. Indeed, it was already seen with InterEvScore that co-evolutionary information can be extracted from alignments with as few as 10 sequences (Andreani, Faure et al. 2013). The sequences in the coMSAs are ordered by decreasing average sequence identity with the query sequences. This is taken into account when sub-selecting sequences in order to keep a representative subset of sequences in both re-

duced coMSAs. Sequence selection was performed in three steps. First, the number of sequences was cut at 100, as in the InterEvDock2 pipeline. Then the alignment was filtered with hhfilter 3.0.3 (Remmert, Biegert et al. 2011) from the hh-suite package. hhfilter was applied with the "-diff X" option on the concatenated coMSAs and the value of X was adjusted for each case in order to return a reduced alignment with no more than 41 sequences (i.e. the query + 40 homologs). At this stage, we obtain a first set of reduced coMSAs with maximum 40 sequences, which we call coMSA^{40}, and that are representative of the full diversity of the initial coMSAs. Finally, 11 equally distributed sequences (i.e. the query + 10 homologs) were uniformly selected within coMSA^{40} in order to preserve sequence diversity compared to the initial coMSAs (see supplementary methods appendix C. page 181). The final set of reduced coMSAs is called coMSA^{10}.

3.1.3.2 Threading models

Pairwise alignments between the template structure and the homolog sequence to be modelled were directly extracted from the reduced coMSAs. The templates used for threading were the unbound template structures provided in the PPI4DOCK benchmark (Yu and Guerois 2016) (see supplementary methods appendix C. page 181).

Rosetta's threading programme, the first step in the RosettaCM pipeline (Song, DiMaio et al. 2013), was used to thread the homologous sequences onto the template structure. We used Rosetta 3.8 (version 2017.08.59291). No insertion, N- or C-terminus were modelled. This resulted in gapped and mainly structurally conserved threaded models of the homologs, where backbone coordinates remained unchanged and side-chain rotamers were different from the template's side-chains only if the residue type changed between the template and the homologous sequence (Figure 3-1B).

3.2 RESULTS

3.2.1 Consensus approach with implicit homology scoring

In previous work (see Chapter 2), we integrated evolutionary information implicitly at the coarse-grained level by scoring decoys with residue-based InterEvScore (noted IES) (Andreani, Faure et al. 2013). In IES, for each decoy, we enumerate all residue-level interface contacts. We then use a residue-level statistical potential to score decoys by considering all sequences in a coMSA and assuming the same contacts were conserved in all homologous interfaces.

We also combined InterEvScore with complementary scores FRODOCK2.1 and SOAP-PP (supplementary Figure C-2A) in a three-way consensus score, denoted Cons^3, which preferentially selects decoys supported by several scores (section 3.1.2.1). Compared to individual scores, we observed a notable boost of about 8 points in top 10 success rate using Cons^3, which captures a near-native in the top 10 decoys in 32% of the cases (Table 3-1 and Figure 3-2A).

Table 3-1: Performance of consensus scores including InterEvScore implicit homology scoring. *Scores used in three-way consensus score Cons^3 were SOAP-PP on the top 10,000 FRODOCK2.1 decoys (SPP/10k), InterEvScore on full coMSAs and on the top 10,000 FRODOCK2.1 decoys (IES/10k) and FRODOCK2.1 (FD2.1). Performances of individual scores used in the consensus are reported in terms of top 10 and top 50 success rates, since consensus calculation relies on the top 50 decoys ranked by each component score.*

Score	Top 10 success rate	Top 50 success rate
FD2.1	164 (21.8%)	292 (38.8%)
IES/10k	182 (24.2%)	287 (38.2%)
SPP/10k	183 (24.3%)	328 (43.6%)
Cons^3	**241 (32.0%)**	/

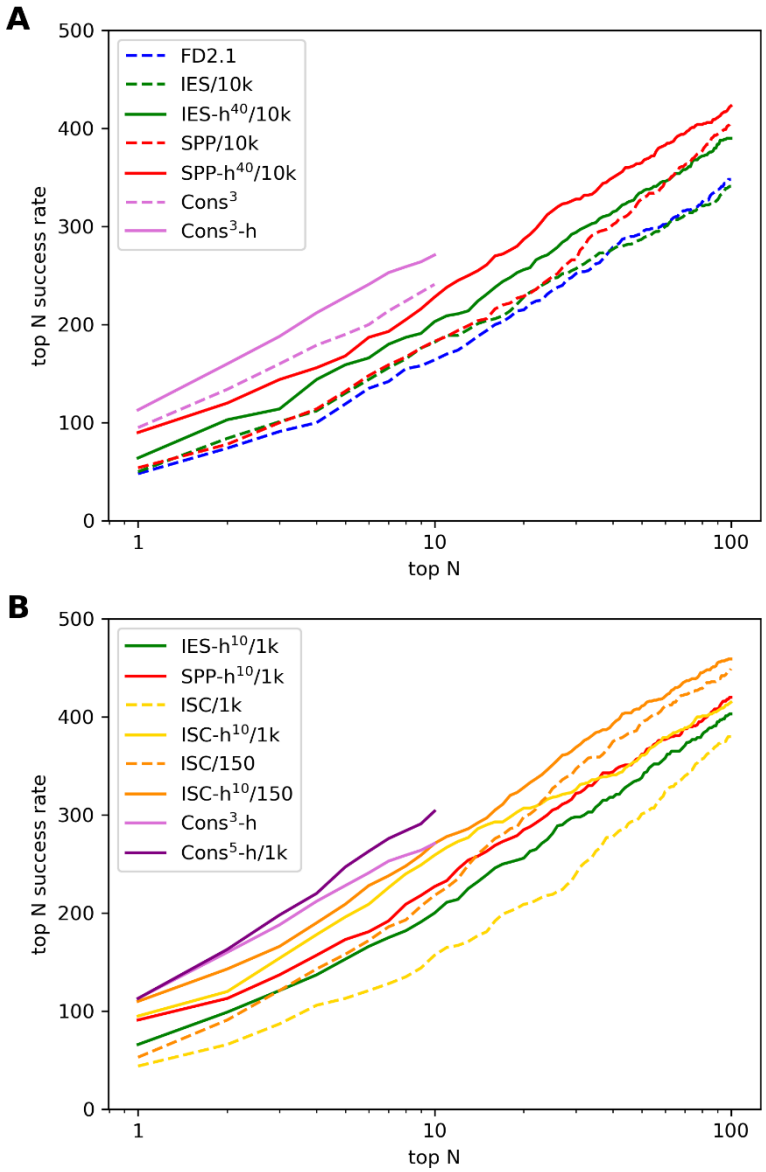


Figure 3-2: Success rate as a function of the number of selected decoys for individual and consensus scores. *Illustration of the success rate on an increasing number of top N decoys with N going from 1 to 100. (A) FRODOCK2.1 (FD2.1), SOAP-PP (SPP) and InterEvScore (IES) individual and consensus scores (dashed lines) and their homology-enriched variants on coMSA^{40} and 10,000 decoys (10k) (solid lines). (B) Rosetta ISC scores (dashed lines) together with homology-enriched variants of individual scores on coMSA^{10} and 1,000 decoys (1k) and selected homology-enriched consensus scores (solid lines). Performances were measured on 752 benchmark cases. Note that consensus scores produce only a selection of 10 decoys, hence they stop at N=10.*

This complementarity between the examined scores, in particular SOAP-PP and InterEvScore, (supplementary Figure C-2A) prompted us to attempt a more explicit integration of evolutionary information into the various scores. An initial attempt at deriving InterEvScore to the atomic level will be described first in section 3.2.2. Then, following the pipeline described in methods section 3.1.3, page 108 (Figure 3-1, page 109), in the following sections, we include

evolutionary information into individual scores InterEvScore and SOAP-PP through explicit atomic-level models of homologous decoys.

3.2.2 First steps towards an atomic version of InterEvScore

The first natural step towards an evolutionary score at atomic scale was to try to derive our own score, InterEvScore, to this scale. InterEvScore makes use of coarse-grained statistical potentials for each residue contact type. It has different scoring components depending on the use of 2- and/or 3-way contacts (i.e. "2-body" or "2/3-body"), on the restriction to only the best contact per interface residue (i.e. "best") or all contact contributions and on the use or not of evolutionary information (i.e. "evol"). Co-evolutionary information is deduced from the coMSAs by assuming that contacts observed in the decoy are maintained in the homologs of both proteins. The interface propensity of these assumed contacts contribute to the final score if they are a part of an apolar patch in the decoy. The term that works best in InterEvScore is the 2/3-body best-evol ($2/3B_{evol}^{best}$) which is composed of the best score of each interface residue within its 2- and 3-body contacts and their equivalent contacts in the various species of the coMSAs. Figure 3-3 illustrates the contribution and calculation of the $2B_{evol}$ term for one contact in the protein complex.

For the atomic version of InterEvScore (IES^{at}), atomic potentials were calculated similarly to the original residue-level InterEvScore for 158 different atom types (the same types as in SOAP-PP (Dong, Fan et al. 2013)) instead of 20 residue types, on a set of 1,050 structurally non-redundant and non-obligate interfaces from the InterEvol database (Faure, Andreani et al. 2012). Because of the high number of different atom types compared to residue types, we limited ourselves to 2-body potentials only, bearing in mind from previous studies that this moderately affected the performance at residue-scale (Andreani, Faure et al. 2013).

3.2.2.1 Atomic scoring without co-evolution

Scoring the query complex with a basic IES^{at} excluding coevolutionary information is straightforward since atomic contacts are directly defined in the decoy structure. This version of IES^{at} worked best when summing the propensities of all "best" 2-body atomic contacts

per residue contact at the interface of the query structure. Performances in terms of top 10 success rates on 54,000 ZDOCK3.0.2 decoys (Pierce, Hourai et al. 2011) for 54 cases from the Weng benchmark v4 of both the residue-scale InterEvScore and IES^{at} without co-evolution are illustrated in Figure 3-3 as green and yellow striped bars respectively. For comparison purposes, the "best" equivalent 2-body version of the original residue-level InterEvScore was used here. The performances show the additional value of considering atomic contacts (28%) rather than a coarse-grained representation of the interface (22%) and are consistent with our initial observations in InterEvDock2 stating that a finer potential might capture more properties that are unique to near-native decoys than a coarse-grained one.

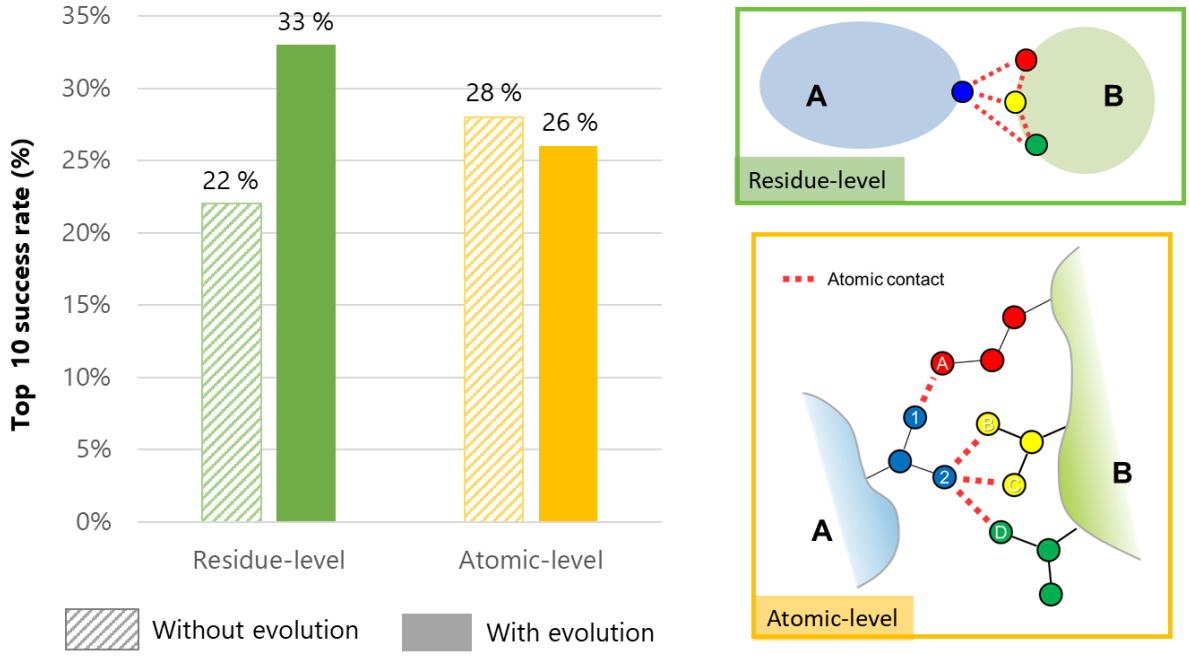


Figure 3-3: Performance of an atomic InterEvScore. *Performances were measured in terms of top 10 success rate on ZDOCK decoys of 54 cases from the Weng benchmark. They are shown for the residue-level InterEvScore (in green) and the best atomic-scale InterEvScore, IES^{at}, (in gold) with and without taking evolution into account (i.e. with or without using the MSAs, respectively; full colour and stripy motif respectively). Here, the residue-level InterEvScore (illustrated in the green box) consisted in summing over the potentials of the best 2-body residue contact (depicted with red dotted lines) per interface residue (blue, red, yellow and green elements). The best performing IES^{at} (gold box) consisted in summing over the potentials of the best atomic contact per residue pair at the interface. In both versions, co-evolutionary information is taken into account implicitly by assuming that homolog contacts remain as in the query proteins.*

3.2.2.2 Adding co-evolution to the atomic InterEvScore

As co-evolutionary information is only given at a residue level in the coMSAs, implementations of IES^{at} with co-evolution relied on various approximations. When

including homology information, we would ideally need to derive equivalent atomic contacts if different residue types are involved in the coMSAs, which cannot be easily inferred from coMSAs alone. In the best-performing IES^{at} including coevolutionary information, I approximated this by taking the atomic pair with the highest propensity for each equivalent residue contact in the homologs, as well as summing over all 2-body contacts in the query structure. Performances for these two versions are also illustrated in Figure 3-3 as green and yellow fully-coloured bars. Unfortunately, as opposed to what was expected, the benefit of adding evolutionary information seen in the residue-level InterEvScore (33% against 22% top 10 success rate) was not observed for the atomic version (26% against 28%, Figure 3-3). This is probably still due to the assumptions made in IES^{at} where residue contacts are just carried across in the homologs and are represented by their best-scoring atomic contact. We thus need a more detailed and explicit representation of the side-chains of the various homologs in order to score the corresponding atomic contacts. In the rest of this chapter, we therefore drop the atomic version of InterEvScore (IES^{at}) and come back to residue-level IES, but we introduce atomic-level information through explicit modelling of the coMSA homologs.

3.2.3 InterEvScore with explicitly modelled homologs

For efficiency, we represent homologs at atomic resolution by threading their sequences onto the query structure (section 3.1.3.2, page 110). As a first step to validate this new representation of evolutionary information, we test the performance of InterEvScore on these threaded models and compare it with the original InterEvScore. With the threaded models, contacts are re-defined in each homolog at an explicit level, rather than implicitly deduced from the coMSAs as in the original InterEvScore. In practice, we calculate the threaded homolog version of InterEvScore (denoted IES-h) by scoring query decoys and their threaded homolog equivalents with the InterEvScore residue-level statistical potentials (section 3.1.2). The final score of each query decoy corresponds to the average score over the query decoy itself and its homologs.

Table 3-2 and Figure 3-2A show the performance of IES-h^{40}, i.e. IES-h computed using threaded homologs from the set of reduced coMSAs with maximum 40 sequences (coMSA^{40}, see section 3.1.3.1, page 109). Results for the original InterEvScore with complete coMSAs (IES) and coMSAs^{40} (IES^{40}) are also shown for comparison. Reducing the number of sequences to maximum 40 does not strongly affect performance in terms of top 10 and top 50 success rates. However, the top 10 success rate increases from 23.8% to 27.0% when using explicit threaded models (IES-h^{40}) instead of only implicit coMSA information (IES^{40}). Of note, a variant of InterEvScore without evolutionary information, where only the query decoy gets scored by the statistical potential has a much lower top 10 success rate of 20.5% (supplementary Table C-6).

Table 3-2: Performance of InterEvScore using coMSAs without or with threaded models. *Top 10 and top 50 success rates of InterEvScore on complete coMSAs (IES, reported in section 3.2.1 and Table 3-1) and coMSA^{40} (IES^{40}) compared to InterEvScore using explicit threaded models of homologs in coMSA^{40} (IES-h^{40}) on 10,000 decoys (/10k). Performances were measured on 752 benchmark cases.*

	Top 10 success rate	**Top 50 success rate**
IES/10k	182 (24.2%)	287 (38.2%)
IES^{40}/10k	179 (23.8%)	284 (37.8%)
IES-h^{40}/10k	**203 (27.0%)**	**335 (44.5%)**

The difference in performance between IES^{40}/10k and IES-h^{40}/10k can be explained by the fact that, in IES-h^{40}, contacts are not extrapolated from the query interface network anymore but are redefined in each homolog based on their modelled interface structure.

3.2.4 Homology-enriched SOAP-PP

Having explicit structures at atomic resolution corresponding to each homolog enables us to score them directly using an atomic potential such as SOAP-PP (Dong, Fan et al. 2013), which might be able to better exploit the atomic detail of homologs for the final ranking of query decoys. As for the threaded version of InterEvScore, homology-enriched SOAP-PP (SPP-h^{40}) consists in the average SOAP-PP score over all homologs including the query decoy itself.

SPP-h^{40} performs better than SOAP-PP on the query decoys alone (Table 3-3 and Figure 3-2A). Using threaded homology models in this way gives a large performance boost to SOAP-PP (+6 percentage points on the top 10 success rate). SPP-h^{40} also outperforms InterEvScore and IES-h^{40} (section 0) as well as the FRODOCK2.1 score (section 3.2.1).

Table 3-3: Performance of SOAP-PP against SPP-h^{40}. *Top 10 and top 50 success rates of SOAP-PP (SPP) compared to its homology-enriched version SPP-h^{40} over sequences in coMSA^{40} on 10,000 decoys (/10k). Performances were measured on 752 benchmark cases.*

	Top 10 success rate	**Top 50 success rate**
SPP/10k	183 (24.3%)	328 (43.6%)
SPP-h^{40}/10k	**228 (30.3%)**	**365 (48.5%)**

3.2.5 Homology-enriched Rosetta interface score (ISC)

Since we build atomic-level homolog models of decoys, we can score them explicitly using a physics-based score such as Rosetta ISC. As Rosetta scoring is much more computationally expensive (about 750 times slower) than SOAP-PP and InterEvScore, to compute homology-enriched ISC, the number of decoys was reduced to 1,000 (as ranked by FRODOCK2.1) and the number of homologs to 10 (coMSAs^{10}, section 3.1.3.1, page 109).

As above, homology-enriched ISC consisted in the average score of the query and its homologous decoys (ISC-h^{10}). For easier comparison, homology-enriched InterEvScore and SOAP-PP were evaluated in the same conditions (*i.e.* 1,000 decoys and coMSAs^{10}) (Table 3-4 and Figure 3-2B). Their success rates are very similar to those with 10,000 decoys and coMSAs^{40} (supplementary Table C-7). Even though ISC on query decoys performs worse than SPP-h and IES-h, ISC-h^{10} largely outperforms the best-performing individual score, SPP-h^{10}, with 34.4% top 10 success rate (259 cases) compared to 30.2% (227). With only 165 successful cases in common, SPP-h^{10} and ISC-h^{10} remain very complementary (supplementary Figure C-2B).

Note that for scores calculated on the top 1,000 FRODOCK2.1 decoys, success rates are technically capped to 77.1%, as only 580 cases out of the 752 in our benchmark have a near-

native within this subset of decoys. In light of this, the ISC-h^{10}/1k performance is all the more remarkable.

Table 3-4: Scoring performance of Rosetta homology-enriched ISC. *Scoring performance of ISC on query decoys only and using the threaded homology models (ISC-h^{10}) on top 1,000 FRODOCK2.1 decoys (1k) and coMSA^{10} as well as the performance of SPP-h^{10} and IES-h^{10} on 1,000 FRODOCK2.1 decoys with coMSAs^{10} for easier comparison. Performances were measured as the top 10 ant top 50 success rates on 752 benchmark cases.*

	Top 10 success rate	**Top 50 success rate**
IES-h^{10}/1k	200 (26.6%)	338 (44.9%)
SPP-h^{10}/1k	227 (30.2%)	**362 (48.1%)**
ISC/1k	157 (20.9%)	301 (40.0%)
ISC-h^{10}/1k	**259 (34.4%)**	361 (48.0%)

3.2.5.1 Using ISC to re-score homology-enriched decoys

ISC-h^{10} showed the highest top 10 success rate from all scores tested above, but scoring 1,000 x 11 decoys with Rosetta ISC is too time consuming in a generalised docking context as it takes approximatively 137 CPU hours per case (supplementary Table C-8). One way to alleviate the total scoring time is to score only a pre-selected amount of decoys using Rosetta ISC as a second step in the scoring pipeline.

In Cons^{3}, we pre-selected the top 50 decoys of FRODOCK2.1, InterEvScore and SOAP-PP. Similarly, here we use the top 50 decoys of the top-performing homology-enriched score variants tested above, namely SPP-h^{40}/10k and IES-h^{40}/10k, as well as FRODOCK2.1. These scores have a high complementarity in terms of top 10 success rate with only 67 cases found in common between all three (supplementary Figure C-2C). Using this subset of 150 pre-selected decoys for ISC scoring (referred to with /150h) reduced scoring times approximately by a factor 7. We enrich near-natives in this set of 150 decoys since they were pre-selected by three already well-performing scores, but only 476 out of 752 cases in our benchmark possess a near-native in this subset.

In terms of top 10 success rate, both ISC-h^{10} and ISC perform better on 150 than 1,000 decoys with 36.0% and 29.0% top 10 success rate instead of 34.4% and 20.9%, respectively (Table 3-5 and Figure 3-2B). Here again, the addition of evolutionary information to ISC

through the threaded homolog models remarkably increases its performance. ISC-h^{10}/150h has the best performance of all tested scores so far, for a much lower computational cost than ISC-h^{10}/1k.

Table 3-5: Performance of ISC and ISC-h^{10} on 150 pre-selected decoys. *Below are summarised the top 10 success rates of ISC and ISC-h^{10}. Top 10 success rates of ISC/150h and ISC-h^{10}/150h were calculated after a pre-selection of maximum 150 decoys taken from the 3 x top 50 decoys of IES-h^{40}/10k, SPP-h^{40}/10k, and FRODOCK2.1. Scoring was performed on all 752 benchmark cases.*

Score	Top 10 success rate	Top 50 success rate
ISC/150h	218 (29.0%)	394 (52.4%)
ISC-h^{10}/150h	**271 (36.0%)**	**411 (54.7%)**

3.2.6 Homology-enriched consensus scoring

As a first step, we calculate Cons^{3}-h, the homology-enriched variant of the Cons3 base consensus score presented in section 3.2.1. Calculating a three-way consensus using higher-performing homology-enriched variants (Cons^{3}-h) instead of their original counterparts (Cons^{3}) increases the top 10 success rate from 32% to 36% (Table 3-6 and Figure 3-2A). Consensus Cons^{3}-h performs as well as ISC-h^{10}/150h, while calculated on the same top 150 decoys, and computation is about 20 times faster for Cons^{3}-h than for ISC-h^{10}/150h.

Out of the 271 successful cases for Cons^{3}-h and ISC-h^{10}/150h, only 199 cases are common. Moreover, ISC and ISC-h^{10} remain complementary to SPP-h^{40}/10k, IES-h^{40}/10k and FRODOCK2.1 (supplementary Figure C-2D and Figure C-2E). This led us to test four- and five-way consensus approaches to combine ISC optimally with other homology-enriched scores. We tested two four-way consensuses that integrate ISC without homology on 1,000 or 150 decoys (Cons^{4}-h/1k and Cons^{4}-h/150h respectively) and two five-way consensuses that integrate ISC both with and without homology on 1,000 or 150 decoys (Cons^{5}-h/1k and Cons^{5}-h/150h respectively). Performances are reported in Figure 3-2B and Table 3-6, together with time estimates when parallelising the whole pipeline on 4 CPUs.

Table 3-6: Performance of homology-enriched consensus scores. *Performance of three-, four- and five-way consensus scores in terms of top 10 success rates on 752 benchmark cases and approximate timescales for the whole pipeline (including sampling with FRODOCK2.1, homology model generation, scoring steps and consensus calculation). Scores used in Cons^{3} were SOAP-PP/10k, InterEvScore/10k and FRODOCK2.1. Scores used in all homology-based consensuses (Cons^{X}-h) were FRODOCK2.1, SPP-h^{40}/10k, IES-h^{40}/10k, ISC and ISC-*

h^{10}. The three-way consensus included the first three scores, four-way consensuses included all scores up to ISC and five-way consensuses included all of them. $Cons^{X}$-h/150h included ISC scores over 150 decoys only and $Cons^{X}$-h/1k included ISC scores over 1k decoys.

Consensus	Top 10 success rate	Whole pipeline time estimates on 4 CPU*
$Cons^{3}$	241 (32.0%)	**15 min**
$Cons^{3}$-h	271 (36.0%)	**15 min**
$Cons^{4}$-h/150h	276 (36.7%)	45 min
$Cons^{4}$-h/1k	282 (37.5%)	3 h 15
$Cons^{5}$-h/150h	289 (38.4%)	5 h 30
$Cons^{5}$-h/1k	**304 (40.4%)**	34 h 30

** all steps are parallelisable using MPI (sampling) or over the decoys (scoring)*

With five-way consensus $Cons^{5}$-h/1k, top 10 success rate rises to 304 cases (40.4%). Unfortunately, computation time strongly increases, since we have to compute ISC-h^{10} on 1,000 decoys. The most time-effective consensus, $Cons^{3}$-h, has 36.0% top 10 success rate and the same top 1 success rate as $Cons^{5}$-h/1k (Figure 3-2B and supplementary Table C-9).

3.3 DISCUSSION

In InterEvScore (Andreani, Faure et al. 2013), evolutionary information improved protein-protein scoring performance when given implicitly through coMSAs and coupled with a coarse-grained, residue-level statistical potential. Combining InterEvScore with complementary scoring functions FRODOCK2.1 and SOAP-PP by computing a consensus (see Chapter 2) improved over the individual scores, reaching 32% top 10 success rate (see Table 3-1). However, this strategy did not take full advantage of the three scores' complementarity and we thus decided to combine directly evolutionary information from coMSAs with atomic scores such as SOAP-PP. To this aim, we threaded coMSA homologs of docked query proteins and scored homologous decoys together with each query decoy.

With this new explicit implementation of evolutionary information, we tested a variant of InterEvScore where we scored decoys and their modelled homologs with a residue-level statistical potential. This modified version (named IES-h) had a slightly improved success rate compared to the implicit homology version (see Table 3-2). The explicit representation of homologous decoys enabled us to build homology-enriched versions of atomic scores SOAP-PP (SPP-h) and Rosetta ISC (ISC-h). For both, adding homology drastically improved top 10 success rates (see Table 3-3 and Table 3-4) even when coMSAs were down-sampled to a maximum of 10 homologous sequences. The Rosetta homology-enriched version, ISC-h^{10}, had outstanding performances, but it also was the most time-consuming score, about 750 times slower than SOAP-PP or InterEvScore. A first compromise between computation time and performance was to run ISC-h^{10} on a pre-selection of 150 decoys defined by the top 50 decoys of SPP-h^{40}/10k, IES-h^{40}/10k and FRODOCK2.1 (see Table 3-5). This score had the same top 10 success rate (36%) as a much faster consensus score involving the same top 150 decoys. Taking further advantage of this complementarity, different four- and five-way consensus calculations managed top 10 success rates from 36.7% to 40.4% at runtimes ranging from 45 minutes to 34.5 hours on four CPUs (Table 3-6).

Our homology enriched scoring scheme is robust to change in the definition of near-natives (supplementary Table C-12) and in evaluation metrics (supplementary Table C-13). Using a

more stringent definition of near-natives (as being of at least Medium quality according to CAPRI criteria) still allows homology enrichment to boost predictive performance of scoring functions. However, consensus scores become less efficient than the best individual scoring functions, probably because when grouping decoys with a relatively loose similarity criterion (see methods section 3.1.2.1, page 107), we do not manage to selectively uprank Medium quality decoys (supplementary Table C-12).

We further tried to understand the origin of the large performance improvements obtained through homology enrichment. Scoring performance improves when near-natives are recognised better (positive selection) or when wrong decoys are down-ranked (negative selection). In the homology-enriched scores described in this work, correct decoys could be upweighted by conserved interfaces in the homologous decoys and, at the same time, incorrect decoys could be discredited by statistically incompatible, clashing, or incomplete homologous decoys (since insertions in reference to the query structures were not modelled). We decided to first explore the simplest explanations, namely, deletions and/or clashes at the interface of homologs that would pull down the average score of the incorrect decoys. However, this does not seem to be the main driving force of ISC-h^{10}'s success over ISC, as the number of gaps or the number of clashes (defined as heteroatom contacts under 1.5 Å) at the interface of homologous decoys do not strongly correlate with the given scores. Additionally, ranking using only the repulsive van der Waals component of the Rosetta score (fa_rep) performs very poorly in comparison to other scoring schemes with at most 34 out of 752 cases with correctly identified near-natives in the top 10 (supplementary Table C-10). Finally, IES-h, SPP-h or ISC-h variants where only the worst homologous decoys are taken into account when scoring each query decoy showed systematically worse performance than using the full range of homologous decoys for each query decoy (supplementary Table C-10). This means that the performance of the homology-enriched scores is positively driven by recognition of correct decoys rather than exclusion of incorrect decoys through the presence of clashes or gaps.

Improvement of the SOAP-PP and Rosetta ISC scoring functions by homology enrichment is significant (supplementary Figure C-3) and consistent over difficulty categories (supplementary Table C-11). When splitting results over PPI4DOCK difficulty categories, we observe that the strongest relative gain for the SPP-h and ISC-h homology-enriched scores compared to their versions without homology occurs on "very_easy" cases, followed by "easy" cases (supplementary Table C-11). A few cases are gained in the "hard" category, but the "very_hard" category remains largely inaccessible to the tested scores, even though our benchmark is limited to cases where at least one near-native decoy was sampled in the top 10,000 FRODOCK2.1 decoys (there are only 16 such "very_hard" cases). Consensus scoring also consistently improves results over the "very_easy", "easy" and "hard" categories, in order of decreasing improvement. We hypothesise that correct ranking of very_easy and easy decoys is mainly dependent on the ability to score positively native-like models while more difficult categories would also require integration of flexibility, an ongoing challenge of protein docking (Torchala, Moal et al. 2013, Desta, Porter et al. 2020).

In this work, we developed a strategy to enrich scoring functions with evolutionary information by including atomic-level models for as few as ten homologs. This strategy improves performance of several scores with different properties: InterEvScore (supplementary Table C-14), SOAP-PP and Rosetta ISC. This means that homology enrichment can in principle be applied to any scoring function with at most a ten-fold increase in runtime. This enrichment works with a very small number of sequences compared e.g. to the large MSAs needed by covariation methods to pick up coevolutionary signal, highlighting complementarity between the two approaches, which may be exploited by using additional DCA-derived constraints, e.g. in intermediate cases with a few hundred homologous sequences (Simkovic, Ovchinnikov et al. 2017, Cong, Anishchenko et al. 2019). The increase in docking success rate also opens interesting perspectives regarding the large-scale application of structural prediction to interaction networks. Finally, with the rise of machine learning techniques in computational biology, one can expect interesting future developments using these approaches to further enhance the extraction of (co)evolutionary signal from coMSAs

CHAPTER 4
The CAPRI challenge

CAPRI consists in the ultimate blind-test scenario, where docking teams can put their methods to the test by predicting the structures of newly resolved and yet unpublished protein-protein interactions. Throughout my PhD, I was able to participate in 10 such docking challenges. Each case brought on its own difficulties, meaning that we had to always adapt our proceedings to the target at hand. However, there were still general guidelines that we followed to tackle these cases.

Resolving CAPRI challenges was always a team effort with regular discussions, adapted strategies and consensus ranking of the models, enriching my knowledge on how to solve a structure from A to Z. This chapter summarises our proceedings in resolving targets T131 to T136 from the 7th edition of CAPRI and is partly based on our published article (Nadaradjane, Quignot et al. 2019). I had the luck of taking part in the concluding international CAPRI meeting in April 2019 that takes place every three years. Three new target prediction rounds have been launched since, and a fourth is ongoing in autumn 2020 dedicated to the current coronavirus situation. Since the official results are not yet available, I will not describe these new rounds in this chapter.

The CAPRI experiment (Janin, Henrick et al. 2003, Lensink, Mendez et al. 2007) is a unique opportunity for methods developers to assess their computational tools and strategies in a wide range of applications, often away from benchmark cases used for methods development and assessment. In CAPRI rounds 38 to 45, a wide variety of targets and challenges was proposed from 2016 to 2018, including diverse classes of conserved prokaryotic assemblies, metazoan cytokine-receptor complexes or host-pathogen interactions. The originality of the challenges also arose from cases of complexes involving polysaccharides and a redesigned interface. In this chapter, I will especially focus on CAPRI rounds 42 to 45, in which I was able to actively participate from 2017 to 2018, and for which official evaluation results are available.

When considering CAPRI targets, a distinction should be made as to whether a homologous template for the complex can be detected or not. The success of CAPRI participants (including our group) for these two categories is quite reflective of the difference in difficulty represented by the two classes of challenges. On average, in the case when a homologous interface template exists, about 20 groups manage to propose successful models among the top 5 models while usually fewer than 10 groups get correct models when no such template is available. In case a template is not available, reaching an Acceptable solution is already a significant challenge, which generally assesses whether the relative orientation between binding partners has been correctly predicted. When a template assembly can be used, the challenge moves toward the quality of the detailed modelling and refinement strategy rather than the docking protocol itself. In the 7th CAPRI edition, most challenges could be addressed using some constraints from a comparative modelling strategy (Table 4-1) and only 4 targets were tackled relying on free docking protocols.

Table 4-1: Summary of CAPRI targets in rounds 42-45. *The table also summarises our group's strategy for addressing each target.*

Round	Target	Short partner ids	Category	Target specificity	Provided info†	Our strategy	Ref complex PDB code
42	T131	HopQI / CEACAM1	protein - protein	pathogen / host		Free docking + biological information	6GBG (Moonens, Hamway et al. 2018), 6AW2 (Bonsor, Zhao et al. 2018)
	T132	HopQII / CEACAM1	protein - protein	pathogen / host		Template-based docking based on T131 solutions	6GBH (Moonens, Hamway et al. 2018)
43	T133	E^{des3} / Im^{-des3}	protein - protein	redesigned interface	wild-type PDB code, affinities	Docking perturbations including rigid-body moves, loops and side-chain refinement	6ERE (Netzer, Listov et al. 2018)
44	T134	DLC8 / MAG(57-aa)	protein - peptide	binding segment prediction	DLC8 PDB code	Evolution-driven motif recognition + template-based docking + constrained refinement	6GZJ (Myllykoski, Eichel et al. 2018)
	T135	DLC8 / MAG(12-aa)	protein - peptide		DLC8 PDB code + 12-aa peptide	Template-based docking + constrained refinement	6GZL (Myllykoski, Eichel et al. 2018)
45	T136	LdcA decamer	protein - protein	homodecamer	clues about homologous structures	Template-based docking + rigid-body perturbations	6Q6I

† Information provided by CAPRI organisers in addition to the identity and sequence of the target partners and the stoichiometry (which were systematically provided when relevant)

Over the past 10 years, our group focused on the integration of evolutionary information in the rigid-body docking toolbox (Andreani and Guerois 2014, Quignot, Rey et al. 2018). From an extensive survey of protein complex structures and sequences conserved in evolution contained in the InterEvol database (Faure, Andreani et al. 2012), we extracted rules and methods to recognise models of interfaces that have most likely co-evolved with the development of InterEvScore (Andreani, Faure et al. 2012, Andreani, Faure et al. 2013). More recently, we combined this evolutionary information with the FRODOCK rigid-body programme (Ramírez-Aportela, López-Blanco et al. 2016) and with alternative scoring approaches such as SOAP-PP (Dong, Fan et al. 2013) to release the InterEvDock2 web service

ADDIN EN.CITE (. Consistent with many challenges proposed in CAPRI, inputs can be submitted as multi-subunit structures but also as sequences since a module for structural modelling of individual partners by homology has been incorporated based on RosettaCM protocols (Song, DiMaio et al. 2013). With this framework, we previously participated in several rounds of CAPRI with a significant number of correct predictions which allowed us to rank among the top performing groups (Lensink, Velankar et al. 2016, Yu, Andreani et al. 2017). In parallel to the development of InterEvScore, the importance of covariation analysis in the field of structural biology has been emphasized by the successful implementation of direct evolutionary coupling analyses methods such as DCA (Morcos, Pagnani et al. 2011), EVFOLD (Marks, Colwell et al. 2011), PSICOV (Jones, Buchan et al. 2012) or CCMpred (Seemayer, Gruber et al. 2014) which have fostered structure prediction of monomeric proteins (Xu 2019) and enabled large scale prediction of structural assemblies (Cong, Anishchenko et al. 2019) for proteins with sufficiently large numbers of homologs in sequence databases. In contrast, InterEvScore can run with a limited set of sequences, ranging from 10 to 100, and thus provides a complementary way to integrate evolutionary information.

In CAPRI rounds 42 to 45, we explored for each target the extent to which evolutionary information could be used. First, we systematically assessed whether a template-based modelling approach could be used, looking for close and remote homology relationships with complexes of known structures. In case only remotely related homologs were detected, we not only considered global homology relationships but also focused on anchoring clusters of residues conserved in evolution. We found that for one group of targets, such a strategy, focusing on recurrent anchoring patterns conserved in evolution, provided key constraints to improve the quality of our models. A third way of exploiting evolutionary information was in the generation of subunit structures prior to docking. There was eventually no target involving rigid-body docking between partners conserved in evolution for which we could use InterEvScore itself and this CAPRI session rather opened onto alternative strategies that could be used to exploit evolutionary information for docking applications. Altogether, the strategy adopted was rarely exactly the same from one target to the next, reflecting the wide variety of macromolecular assembly modes, either through folding upon binding processes,

through multivalent contact points emerging from symmetric arrangements or relying on subtle loop conformation to ensure specific and tight recognition. In this report, we attempt to account for that variety providing hints that might be used depending on the nature of the targets. We also discuss how these observations echo with our large experience in modelling protein complexes.

4.1 METHODS

In this section, we present the pipeline used to prepare all targets and the general strategies followed for the two types of challenges: protein-protein and protein-peptide docking.

4.1.1 Target preparation

All CAPRI targets in rounds 42-45 were provided as sequences, sometimes with additional information about stoichiometry and possible templates. HHsearch (Soding 2005) was systematically used to search for homologous structures in the Protein Data Bank (PDB). When a suitable template was available for individual partners, we generally used homology modelling with a RosettaCM-based protocol (Song, DiMaio et al. 2013) relying on the HHsearch alignment. We evaluated evolutionary conservation for individual protein partners using the Rate4Site algorithm (Pupko, Bell et al. 2002) or the Consurf web server (Ashkenazy, Abadi et al. 2016).

For all protein-protein and protein-peptide targets (since the peptides in targets T134-T135 were actually protein fragments), the PPI3D (Dapkunas, Timinskas et al. 2017) and HHpred (Zimmermann, Stephens et al. 2018) web servers were queried to search for available structures of homologous complexes.

4.1.2 Protein-protein docking challenge (T131-T132, T133, T136)

When structures of homologous complexes were available (T133, T136), our protein-protein docking strategy always started with comparative interface modelling using a RosettaCM-based protocol (Song, DiMaio et al. 2013). The available interface templates were close in sequence identity for T133 (wild-type complex at 80% sequence identity with redesigned interface) and more remote for T136 (2 templates with 40% overall sequence identity but only 28% and 20% N-terminal domain sequence identity).

When no homologous complex structure was available (T131-T132), free docking was used instead. Our standard docking pipeline (Quignot, Rey et al. 2018) at that time was based on

rigid-body interface sampling followed by clustering and consensus rescoring using three scores: the SOAP-PP atomic-level statistical potential (Dong, Fan et al. 2013), our InterEvScore residue-level statistical potential including coevolutionary information (Andreani, Faure et al. 2013) and the FRODOCK scoring function (Ramírez-Aportela, López-Blanco et al. 2016). In these CAPRI rounds 42-45, the free docking targets involved host-pathogen complexes (T131-T132) for which no co-alignment could be built with joint sequences in multiple species for the two partners, therefore our usual InterEvScore-based strategy was not applied. We derived T132 models from T131 free docking models by a template-based strategy rather than *ab initio* docking, and then re-ranked them after refinement and interface analysis.

We performed final refinement of all docked interfaces using Rosetta-based protocols. For targets T131-T132, T133, T136, docking perturbations using RosettaDock were performed, with symmetry constraints for T136. For all targets, we used RosettaRelax protocols (Tyka, Keedy et al. 2011, Nivon, Moretti et al. 2013) for final refinement (under symmetry constraints for T136). For target T136, which involved a multi-domain homodimer, Rosetta kinematic loop modelling (Mandell, Coutsias et al. 2009) was used to rebuild domain linkers after perturbations and refinement.

4.1.3 Protein-peptide docking challenge (T134-135)

For target T134, a first step was to scan the long fragment of the MAG protein in order to identify the most likely 12-residue binding stretch. The strategy for this step is further described in the results section. For targets T134-T135, the corresponding binding motif was anchored using homologous interfaces containing the canonical TQT binding motif as templates, then the interface was refined by extending the motif at the N- and/or C-terminal tail with the Rosetta FloppyTail protocol (Kleiger, Saha et al. 2009) and finally by using RosettaRelax (Tyka, Keedy et al. 2011, Nivon, Moretti et al. 2013) to relieve the strong clashes induced by the template-target superimposition.

4.2 RESULTS

Prediction results for all CAPRI targets from rounds 42-45 (Table 4-2 and Table 4-3) are discussed below. We split targets into three categories according to our docking strategy: *ab initio* free docking, straightforward template-based docking, and finally targets for which available structures of homologous or similar interfaces were used together with evolutionary information to identify recurrent conserved interaction motifs and to guide interface modelling accordingly.

Table 4-2: Results for CAPRI targets in rounds 42-45. *Results are provided for the top 5 and top 20 models submitted by our group vs. all other groups by indicating the quality of the best model in this range: - for incorrect, * for Acceptable, ** for Medium and *** for High.*

Round	Target	Short partner ids	Category	Top 5 our group†	Top 5 other groups†	Top 20 our group†	Top 20 other groups†
42	T131	HopQI / CEACAM1	protein - protein	-	**	**	**
	T132	HopQII / CEACAM1	protein - protein	-	**	*	**
43	T133	E^{des3} / Im^{des3}	protein - protein	**	**	**	**
44	T134	DLC8 / MAG(57-aa)	protein - peptide	***	***	***	***
	T135	DLC8 / MAG(12-aa)	protein - peptide	***	***	***	***
45	T136	LdcA decamer	protein - protein	** / ** / *	** / ** / **	** / ** / **	** / ** / **

† For T136, multiple interfaces were assessed that are denoted by multiple results separated by a / sign.

Table 4-3: Assessment summary for our best submitted CAPRI targets. *This table includes the following assessment metrics (as provided by the CAPRI assessment team on the CAPRI website https://www.ebi.ac.uk/msd-srv/capri/): fraction of native contacts (fnat), ligand RMSD (L_rmsd), interface RMSD on backbone atoms (I_rmsdbb), and individual RMSDs of the two partners (M_rmsd_1 and M_rmsd_2). For each target (and each interface whenever relevant), this information is provided for our best submitted model*

among the top 5. If that model is incorrect and if we submitted a better model (in terms of I_rmsdbb) within the top 20, the metrics are additionally provided for that model (in yellow in the table below).

Model id	capriround_target(.interface)	fnat	L_rmsd (Å)	I_rmsdbb (Å)	M_rmsd_1 (Å)	M_rmsd_2 (Å)	classification
T131_P07.M03	capri42_T131	0.055	44.163	11.364	0.686	1.907	incorrect
T131_P07.M12	capri42_T131	0.52	3.247	1.7	0.53	1.22	medium
T132_P07.M03	capri42_T132	0.06	45.543	9.156	0.787	2.973	incorrect
T132_P07.M14	capri42_T132	0.209	11.578	3.905	0.611	2.555	acceptable
T133_P05.M02	capri43_T133	0.66	3.433	1.577	1.694	0.749	medium
T134_P19.M01	capri44_T134	0.895	1.21	0.356	1.098	0.364	high
T135_P19.M04	capri44_T135	0.895	1.812	0.459	1.513	0.366	high
T136_P03.M04	capri45_T136.1	0.687	2.813	1.709	2.002	2.001	medium

4.2.1 Protein-protein docking using *ab initio* free docking strategy (targets T131-T132)

This category included two host-pathogen protein complexes, T131 and T132 (round 42).

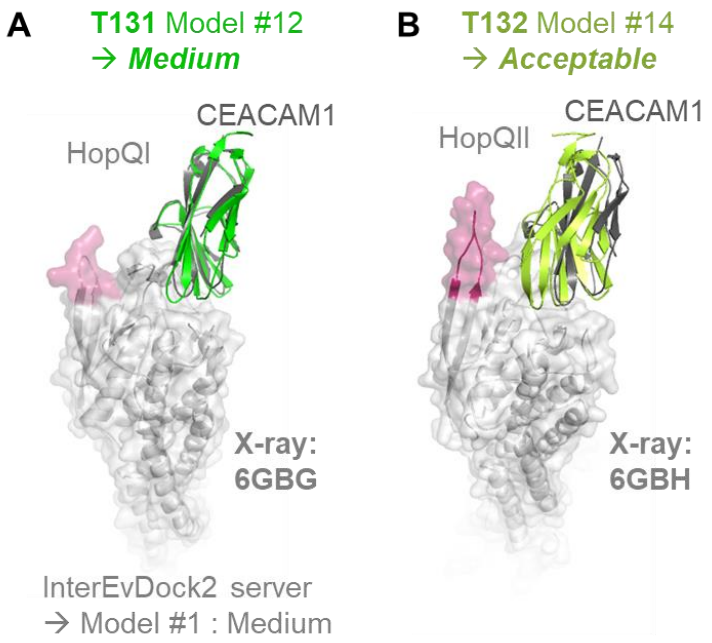


Figure 4-1: Representation of models predicted by free docking. *Comparisons between the structure of the target complexes (coloured as light and dark grey cartoons for the receptor and ligand, respectively) and the best predicted model (coloured as green cartoon) with the index of the model indicated after the hash symbol for (A) HopQI-CEACAM1 (T131), (B) HopQII-CEACAM1 (T132). The hairpin coloured in red was previously published as involved in the interaction although this turned out to be incorrect. It biased the ranking of our models, although for T131, the InterEvDock2 server ranked as best model a Medium quality prediction.*

4.2.1.1 Targets 131 & 132: Success of the InterEvDock server undermined by misleading biological information.

T131 and T132 involved host-pathogen protein complexes of the N-terminal domain of human cell adhesion protein CEACAM1 with, respectively, the cell adhesion proteins HopQ type I and HopQ type II from *Helicobacter pylori*. Both CEACAM1 and HopQ type I structures were solved experimentally (PDB codes: 4WHD for CEACAM1 and 5LP2 for HopQ type I) while the structure of HopQ type II could be obtained by comparative modelling (see Methods) using the structure of type I as a template (sequence identity 56%). This target was well suited for free docking and we used the InterEvDock2 web server to generate a set of most likely solutions. After docking HopQ type I with the CEACAM1 N-terminal domain, model #1 returned by the server is of Medium quality compared to the released structures of the complex (I-RMSD 1.45 Å with respect to PDB code 6AW2 and 2.94 Å with respect to PDB code 6GBG) (Figure 4-1B). However, it was suggested in previously published literature (Javaheri, Kruse et al. 2016) that a specific region of HopQ, called the insertion domain, was important for binding CEACAM1, as its deletion reduced the affinity of HopQ to CEACAM1 and a peptide derived from the HopQ insertion domain could inhibit the infectious phenotype triggered by H. pylori. For that reason, upon model submission we downgraded the free docking models that did not involve the insertion domain shown in red on Figure 4-1A. In the end, our submitted models #12 for HopQ type I (T131) and #14 for HopQ type II (T132) (Figure 4-1B) were assessed of Medium and Acceptable quality, respectively. Without this erroneous information we would not have downgraded them beyond the top 10 threshold; the published structures (PDB codes: 6AW2, 6GBG, 6GBH) eventually showed that CEACAM1 does not interact with the HopQ insertion domain. Nevertheless, this example highlights the performance of the InterEvDock2 server for free docking applications even when no coevolutionary information can be used.

4.2.2 Taking evolution into account in template-based docking strategies (targets T133, T136)

This category included protein-protein targets T133 (round 43) and T136 (round 45).

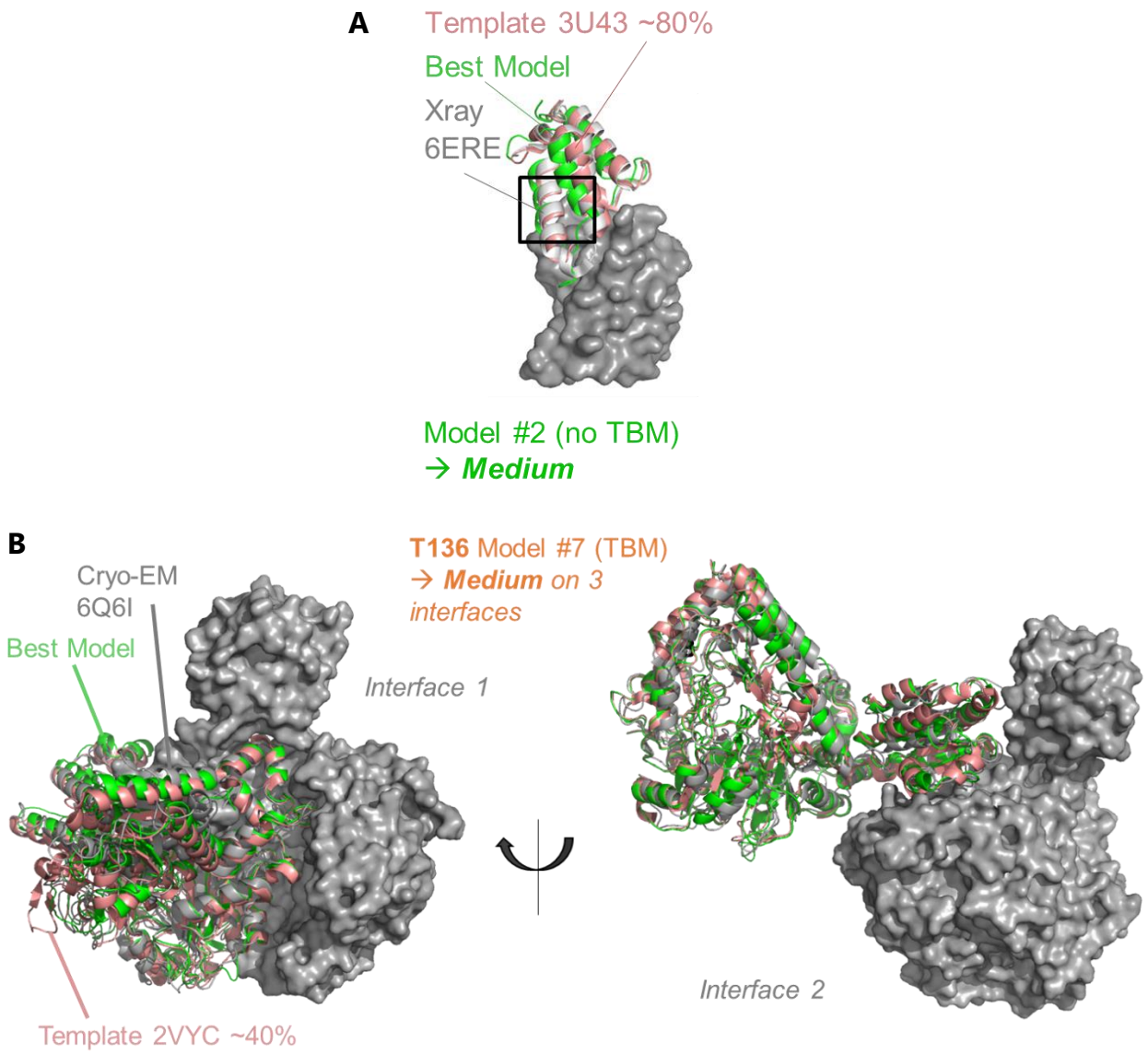


Figure 4-2: Template-based prediction of CAPRI targets T133 and T136. *Representation of the best model (coloured as green cartoon) compared to the experimental structure (grey cartoon and surface for the ligand and receptor subunits, respectively) and the template (light red cartoon) which could be used for template-based modelling (TBM). (A) Structures for target T133 (redesigned Edes3 / Imdes3 interface). (B) Structures for target T136 (homodecamer LdcA) for the two main interfaces out of the three created through oligomerisation. Our model #7 was of Medium quality for all three interfaces.*

4.2.2.1 Target 133: Optimisation of an interface locally but drastically remodelled by design.

Target 133 is an interesting target to test refinement strategies. It consists of a redesigned orthogonal version of the wild-type Colicin E2 DNase-Im2 complex (PDB code: 3U43) which shares 80% sequence identity with the wild-type. At this level of sequence identity, our survey of complex interologs (Faure, Andreani et al. 2012) revealed that changes in rigid-body orientations are very unlikely. However, CAPRI organisers mentioned that the complex displayed a different binding mode (including rigid body orientation) with respect to the wild-type complex, which prompted us to combine not only loop and side-chain refinement strategies but also rigid-body docking perturbations. A posteriori, the released structure (PDB code: 6ERE) only had a L-RMSD of 0.7 Å with respect to the original (Figure 4-2A) indicating that no rigid-body motions were required to reach a High quality model. Because we had to consider simultaneously the three degrees of freedom listed above, we did not resample deeply enough the conformations in the mutated regions. In the end, our submitted model #2 was a Medium quality model that corresponded to the least rigid-body perturbed model. As all CAPRI participants, we did not manage to optimise the interface so as to reach a High quality model although it could have been expected given the high level of identity with the template. This target confirmed that at high sequence identity, assemblies usually do not dramatically change their relative orientation and that it is more efficient to optimise the structure locally rather than to include rigid-body perturbations.

4.2.2.2 Target 136: Combining multi-domain and multi-subunit template-based modelling in a symmetric homomultimer.

Target 136 was the lysine decarboxylase LdcA from *Pseudomonas aeruginosa*, a challenging large complex assembling as a homodecamer of subunits themselves composed of three domains. Two templates were available sharing overall 40% identity with the target sequence (PDB codes: 2VYC and 3N75). However, the N-terminal domain exhibited more divergence, sharing 28% sequence identity with 2VYC and only 20% with 3N75. The domains move

slightly with respect to each other in the two template structures and alter the interfaces. For that reason, we explored two strategies for this target, following either a conservative template-based modelling protocol (see Methods) or a modelling protocol involving rigid-body perturbations between domains, maintaining the D5 symmetry and rebuilding the domain linkers after refinement. This complex combinatorial strategy was implemented as a RosettaScripts protocol (Fleishman, Leaver-Fay et al. 2011) (see supplementary materials, appendix D. page 192). The best overall model we generated (model #7) was assessed as Medium for all three interfaces and was derived from 2VYC, the template that shared the highest sequence identity in the N-terminal domain (Figure 4-2B), with a L-RMSD of 2.33 Å with the now-revealed cryo-EM structure (PDB code: 6Q6I). Somewhat disappointingly, none of the resampling protocols using rigid-body perturbations of the domains improved the quality of the models significantly, suggesting once again that above 30% sequence identity, no major change in orientation might be expected while they might be much more pronounced at lower identity.

4.2.3 Evolutionarily conserved and recurrent structural motifs as guide for docking (targets T134-T135).

This category included protein-peptide targets T134-T135 (round 44).

When a homologous complex is available to perform template-based docking, it is often possible to reach an Acceptable solution even at low sequence identities without extensive conformational sampling or resampling. Reaching Medium quality models is often more challenging and may require extensive refinement comprising simultaneous conformational resampling of loops and small rigid-body perturbations. As discussed in the previous section, it is difficult to decipher which region should be resampled and whether some regions may be considered rather invariant or pivotal in the evolution of interfaces. CAPRI 7th targets provided several examples in which, beyond global homology relationships, we could rely on local recurrent structural motifs that can be seen as rather invariant features in the evo-

lution of interfaces and useful to guide docking. When spotted out, these conserved recurrent motifs may be used as useful constraints to drive the resampling around an initial template-based model and reach higher quality models.

4.2.3.1 Targets 134 & 135: Evolution-driven recognition of Small Linear interaction Motifs in non-trivial cases.

For targets 134 and 135, we had to predict the structure of a complex of Dynein Light Chain subunit 8, DLC8 (a dimer), with a peptide of myelin associated glycoprotein, MAG. The challenge consisted in a two-step prediction process in which we were first asked to identify and model the small 12-residue interaction motif out of a longer 50 amino-acid segment. Then, in a second step, the prediction had to be repeated with the knowledge of the exact 12-residue ligand sequence. The first step was not trivial because none of the reported sequence motifs known to bind DLC8 could be identified (Figure 4-3). In particular, a glutamine observed as a central DLC8-binding residue in many structures of ligand peptides bound to DLC8 could not be identified in the MAG segment (Figure 4-3A). This challenge of identifying a motif in a long disordered stretch of a protein provided an original and interesting test case, closely matching issues we often encounter during collaborations with experimental biologists. To address this challenge, we first gathered all the structures and ligands available for DLC8 and its homologs and created multiple sequence alignments for all known binding partners. Concatenation of these alignments centred in the region of the binding motif led to the definition of an enriched sequence profile (Figure 4-3B) that was applied as a sliding window to score the likelihood of a motif over the sequence of the MAG protein segment. We further enriched the motif taking advantage of a publication proposing a similar strategy for dynein binders (Erdos, Szaniszlo et al. 2017). Figure 4-3C illustrates that a single short motif in the sequence led to a positive score. Our model #1 derived from that analysis reached a High quality assessment score for target T134. The moves that we further incorporated in subsequent optimisations of the N-terminal and C-terminal tails of T135 peptide degraded the assessment for model #1 to a Medium quality model but T135 model #2 was

rated High. Generalisation of recurrent structural anchoring motifs using known experimental structures expanded by evolutionary information from divergent homologs can be used as a powerful means to increase the sensitivity of Small Linear Motif (SLiM) recognition. SLiMs can be organised as relatively independent multivalent anchoring points, some able to provide strong affinity gains as the conserved glutamine in DLC8 ligands, others providing more moderate and diffuse affinity gains. The T134-T135 example highlights that combining moderate anchors can compensate for the absence of a strong anchoring point, providing guidelines for the identification of binding motifs when analysing complex interactomes.

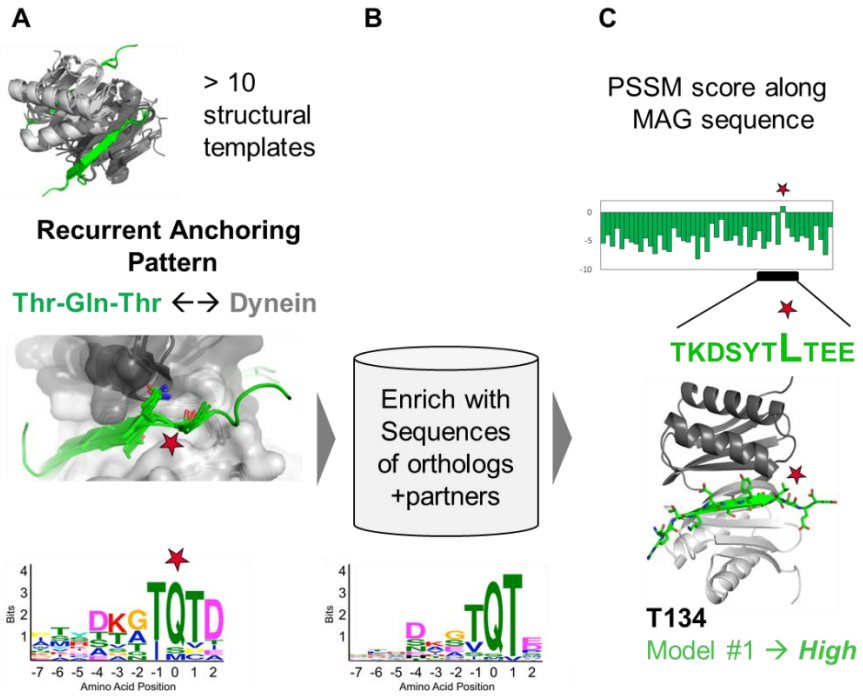


Figure 4-3: Examples of recurrent anchoring patterns used to constrain docking models between the dynein light chain and its binding partner MAG (CAPRI target T134). *(A) Most available structural templates (dynein in grey, binding partners in green) emphasize the importance of the T-Q-T motif in the binding; a profile (or Position-Specific Scoring Matrix, PSSM) based on templates is shown with the central Q highlighted by a red star. (B) Using external information from homologs of all dynein binding partners, the definition of the motif could be enlarged and translated into an enriched profile. (C) By scanning the profile along the MAG sequence, we identified the region most likely bound by dynein. The red star highlights the single position for which a positive score was obtained and the central L which plays the role of the Q in other dynein-bound ligands.*

4.3 DISCUSSION

Rounds 42-45 of the 7th session of CAPRI contained a total of six targets introducing several types of challenges for docking, such as challenging multimeric complexes (T136), a redesigned interface (T133) and a target where predictors were prompted to identify a short binding motif in a longer protein segment (T134).

None of the targets in this edition of CAPRI were relevant for the use of our InterEvScore-based free docking protocol, which requires building a joint multiple sequence alignment for two protein partners that reflects how the interface coevolves. However, we still made use of evolutionary information in one form or another for all targets (except for T131, which was entirely based on free docking from known structures of the interacting monomers). Most often this was done through the use of homologous interfaces for template-based docking or to derive recurrent interface features, and these strategies helped us to successfully model interfaces for targets T133, T136. In T134, we identified the correct MAG binding motif based on enriched sequence profiles combining structural and evolutionary information.

Altogether, we were able to generate Acceptable or better models in the top 5 for four out of six targets, including two with Medium models and two with High quality models. We missed T131-T132 where we downgraded Medium and Acceptable models below the top 5 due to misleading biological information from the literature, while our InterEvDock2 server can generate a top 1 Medium quality model for T131. For T133, even though we submitted a correct model in the top 5, we think we may have reached higher quality models by following a slightly different strategy. No group submitted a High quality model even though this might have been expected given the presence of a template at 80% sequence identity (the wild-type complex) that was already Medium with respect to the redesigned interface. In our case, we did not resample deeply enough the fine details of the mutated regions because we explored larger moves, while we should have trusted that the redesigned interface would maintain the same global binding mode as the wild-type complex and differ only (but significantly) in the local arrangement of interface features.

Overall, this CAPRI session revealed interesting ways to include evolutionary information beyond our usual docking pipeline. This includes of course classical template-based docking, for which the CAPRI targets in rounds 42-45 further reinforced previous observations that models derived from templates above 30% sequence identity should be optimised only locally, while templates below 30% sequence identity should always be considered but should be perturbed more extensively, including rigid-body moves. Finally, two targets in CAPRI rounds 42-45 highlighted structural interface motifs recurrently found among similar and homologous interfaces and conserved in evolution, confirming the importance of such anchors and stressing the need for improved ways to identify and encode them.

CHAPTER 5
Conclusions and perspectives

My thesis focused on improving the general prediction power of docking and scoring methods, in particular by drawing on co-evolutionary information. Apart from improving their performance, part of our mission as structural predictors is also to make our work accessible to the scientific community. In light of this, I participated in major developments of our docking server, InterEvDock2, described in Chapter 2. Based on input proteins, InterEvDock2 suggests 10 most plausible interface models selected by combining physics-based scoring terms, statistical potentials and co-evolutionary information. InterEvDock2 now also accepts oligomeric structure inputs or sequence inputs, for which it can automatically model monomer structures for docking. The user can also integrate previous knowledge about the interaction if available in the form of single or pairwise constraints between residues to filter out any non-relevant solutions. The complete pipeline can be run fully automatically or in a more user-controlled manner, using strategic breakpoints throughout the process and/or self-tuned parameters. I validated the performance of InterEvDock2 on a large benchmark of 812 heterodimeric docking cases with homology modelled unbound structures. InterEvDock2 was capable of finding a correct complex structure in as much as 32 % of these cases (Table 2-1, page 93). Of particular value to biologists is also its high performance in predicting interface residues with a 75% probability of having at least one correct prediction out of two predicted residues (one on each partner, Table 2-1, page 93).

My work then focused on finding a more efficient and higher-resolution way of integrating evolutionary information to discriminate near-native structures from wrong complexes in scoring (Chapter 3, page 102). I managed to derive the implicit evolutionary information present in the sequence alignments to an atomic level of detail, using modelled homologous interfaces. This explicit representation is directly compatible with atomic-scale scoring and yields a significant increase from 32% to 40% success in predictive performance on a large benchmark (Table 3-6, page 119) by applying the same consensus approach between scores as in InterEvDock2. This strategy of atomic integration of evolutionary information is directly compatible with our InterEvDock2 pipeline as it relies on efficient scoring and will be integrated in the server during its next update.

Finally, CAPRI consists in the ultimate blind-test scenario, where docking teams can put their methods to the test by predicting the structures of newly resolved and yet unpublished protein-protein interactions. Chapter 4 describes the strategies applied by our team, which enabled us to rank first in number and precision of correct predictions in the latest CAPRI round (2016-2019, Table 4-2 shows our performance for the rounds in which I participated). I was able to participate in 10 such docking challenges throughout my PhD. Resolving these challenges required a lot of team work and organisation as each target called for a different search strategy. According to the latest CAPRI results, template-based approaches generally tend to be the most accurate, when good templates can be found. Use of evolutionary information in addition to complementary scoring functions also enabled more efficient selection of near-native models. Currently, InterEvDock2 is not well-suited to CAPRI's server round since it does not integrate template-based docking. I am currently working on a third update of InterEvDock in collaboration with Pierre Tufféry's team at RPBS, which will make use of the already optimised and automated multimeric template search in Proteo3Dnet (Postic, Marcoux et al. 2020). Together with the integration of our atomic-derived evolutionary information and an extra clash removal step at the end of the pipeline, InterEvDock3 will be equipped for participation in the next server rounds of CAPRI.

Some of the most promising recent developments in the structural prediction of protein interactions rely on coevolution to provide specific constraints for assembly modelling. The DCA class of methods was recently showcased as a breakthrough for individual protein structure prediction, especially when integrated into deep learning pipelines (Xu and Wang 2019, Senior, Evans et al. 2020, Yang, Anishchenko et al. 2020) (section 1.3.1.2, page 43). DCA-like approaches were also applied to interface structural modelling with some success, but challenges remain, especially to obtain enough statistical information by building large coupled MSA pairing up interacting homologs. Further integration of DCA-like methods and other coevolution-based methods with machine learning and deep learning algorithms will likely prompt future progress and expand the range of applications.

DMS is also a promising direction for future research, providing a more systematic way of quantifying the effects of mutations through high-throughput assays coupled with next-generation sequencing (section 1.1.3.2, page 12 and section 1.2.2.5.1, page 33). In close relation to how covariation-based methods make use of natural sequences to infer 3D contacts, DMS was very recently used as a source of synthetic sequence information to predict the 3D structure of a few protein domains of limited size (up to 56 amino acids), of one ribozyme RNA and one protein-protein interface composed of two interacting helices in leucine-zipper domains of a transcription factor complex as well as the structural prediction of RNA (Rollins, Brock et al. 2019, Schmiedel and Lehner 2019, Zhang, Xiong et al. 2020). Remaining technological and computational challenges explain the limited amount of DMS data so far usable for such applications. In particular, applying this deep mutagenesis technique to larger single proteins and interactions between standard-size proteins remains an issue that may be alleviated by future developments of DNA synthesis and sequencing technologies. In addition to protein-protein interface structural modelling and the prediction of interacting protein pairs, coevolutionary constraints, especially those derived from DCA-like methods, can be used to study interface dynamics and interaction specificity, to shed light on protein-ligand and protein-nucleic acid interactions and to help in designing new interacting proteins (Morcos and Onuchic 2019). The development of binding affinity predictors is also closely related to that of docking scoring functions (Gromiha, Yugandhar et al. 2017, Geng, Xue et al. 2019). Recent work has shown that binding affinity prediction for interactions between peptide-binding domains and disordered motifs benefits from transfer between binding domain families and from the design of custom machine learning algorithms (Cunningham, Koytiger et al. 2020). Future advances in binding affinity prediction for globular and disordered systems should also take advantage of a more systematic use of evolutionary information, whether it be conservation, coevolution, or more innovative DMS data.

Integrating flexibility into docking is still an ongoing challenge. In the traditional docking protocol, flexibility is taken into account during a second step after rigid-body sampling on a small number of carefully selected models (e.g. HADDOCK (van Zundert, Rodrigues et al.

2016), RosettaDock (Lyskov and Gray 2008), FiberDock (Mashiach, Nussinov et al. 2010), FireDock (Mashiach, Schneidman-Duhovny et al. 2008)). For many proteins, however, poses that are close to the native complex might already be overlooked during sampling, even with a powerful scoring function, when backbone and side chain conformations are too drastic. Other ways around this problem is the integration of flexibility before the sampling stage on the unbound monomers, followed by rigid-body cross-docking of structural ensembles, as demonstrated by (Krol, Chaleil et al. 2007). This method, however, extremely increases the number of outputted conformations to be scored. Programmes integrating elastic network model simplified representations of proteins coupled with a normal mode-based flexibility description might be better suited to a more modulated sampling and was implemented in the coarse-grained ATTRACT docking programme to approximately account for global conformational changes during the first stages of sampling (de Vries, Schindler et al. 2015). The RMSD calculation algorithm RapidRMSD also uses normal modes as well as linear collective motions to efficiently calculate structural changes between flexible docking poses (Neveu, Popov et al. 2018). In the end, even though flexibility integration is more of a sampling issue, scoring functions can also contribute in a rigid-docking context by better identifying the near-native pose. Integration of flexibility also tremendously increases the number of models to score, thus, it is important to develop efficient and discriminative scoring functions.

Of high interest in biology would be to be able to reconstruct a complete structural protein interactome. This would help understand cellular dynamics, especially in the therapeutics field. For instance, structural insights in a protein interaction network would allow easier development of target-specific drugs that would only minimally disrupt the rest of the interactome. The structural prediction of the whole network, not just the structure of protein pairs that were confirmed to interact beforehand, is an ongoing challenge, especially because of noisy data or even lack thereof. Cross-docking studies constitute a step towards that goal by trying to predict what two proteins interact within a set of proteins (Lopes, Sacquin-Mora et al. 2013). Experimentally acquired data centralised in protein interaction databases could also be used in a similar fashion to Ghadie and Xia (2019) (Ghadie and Xia

2019). In their study, the authors predict the effect of key disease-associated mutations based on the structural reconstruction of the human interactome by combining different levels of PPI information found in databases and reconstructing the structures through homology modelling (Ghadie and Xia 2019).

Machine learning techniques are getting increasing attention in the field of structural biology. The use of deep learning techniques in particular enabled major advances in protein fold prediction, as can be seen in the results on recent CASP targets (Xu and Wang 2019, Senior, Evans et al. 2020, Yang, Anishchenko et al. 2020) (section 1.3.1.2, page 43). Coupled with the use of co-evolutionary information and the innovative idea of predicting distance distributions rather than binary contacts, these latest techniques were able to predict accurate enough distance maps to predict close-to-correct models. One could imagine that protein-protein docking might benefit from these recent advances in order to improve performance and precision. For instance, machine learning could be used to predict protein interface regions by returning a local score for each residue and/or by attributing a general score for a given structure (in that case, the recent and continuous DockQ score is quite convenient), such as applied in (Pittala and Bailey-Kellogg 2020), or by predicting if continuous surface patches are at the interface, as in (Gainza, Sverrisson et al. 2020). One of the major challenges in machine learning techniques, however, lies within the careful preparation of the training, validation and test sets; this goes in hand with the problem of data availability for such practices. Although there are large enough amounts of non-redundant protein structures to train and test protein folding models, structures of protein complexes are less abundant. Coupled with redundancy filtering, there would be very few complexes left to avoid overfitting the model. Similarly to knowledge-based scoring functions, machine learning models can be trained on experimental structures and/or on decoys. For optimal training, datasets should cover the whole range of different structure qualities in a balanced fashion, meaning that negative (false interfaces) as well as positive (correct interfaces) inputs should be equally well represented. Therein lies another challenge as there are many ways of sam-

pling an incorrect decoy but only a few to generate near-native decoys without being redundant. In that sense, the generation of homology models of decoys, similar to the strategy that was applied in Chapter 3 but for a larger number of homologs, might be a way around that problem. We could also consider the use of transfer learning from protein fold prediction models.

Other than the datasets, thought also has to be given to what model is best suited as well as what information (features) should be inputted to the model and the best way of encoding it. Inspiration can be found in well-established protein fold quality assessors, such as Ornate (Pages, Charmettant et al. 2019), KORP (Lopez-Blanco and Chacon 2019) or GraphQA (Baldassarre, Hurtado et al. 2020), fold predictor trRosetta (Yang, Anishchenko et al. 2020) or antibody-antigen interface predictor, PECAN (Pittala and Bailey-Kellogg 2020) for example. In Ornate, proteins are encoded as 3D grids of fixed size and centred on each residue in the protein to exclude orientation-dependency. Each cube in the grid has a smoothed atom-occupation probability and could be given a set of additional features. One could also consider a simpler representation of residues, as is the case in KORP, where they are represented by only three backbone atoms each to avoid side-chain orientation-dependency. One could also consider representing interfaces as multi-level distance and angle maps as outputted by the trRosetta framework. Another intuitive way of representing proteins are graphs, as in the message-passing algorithm GraphQA or in the graph convolution network PECAN. In the graph, nodes (residues) and edges (contacts) can both be given specific features, such as residue type, conservation and co-evolution profiles, surface accessibility or secondary structure for nodes, and distance distributions for edges. In an interface prediction context, one could additionally add the chain number as a feature or encode receptor and ligand as two separate but communicating graphs as in PECAN.

Finally, although deep learning methods are difficult to interpret, efforts are being made to better understand what is effectively learnt by these methods. For instance, (Pittala and Bailey-Kellogg 2020) used attention layers in the context of epitope and paratope prediction

to visualise the regions in antigens and antibodies that are given the most attention in the network.

I will finish on protein interface design, a hot topic in light of the current Covid-19 situation to design suitable artificial antibodies against the virus. Protein interface design is a field very much related to assembly prediction and also relies on the understanding of the key factors that are important for the interaction between macromolecules. In that sense, protein design can learn from docking and vice versa and both can benefit from methods integrating complementary information taken from evolutionary analysis or DMS for example, particularly in understanding the importance of anchor residues that govern interfaces and how they coordinate to result in a stable interface. Protein design usually involves many cycles of computational prediction and experimental selection. The more traditional approach consists in tweaking an already existing structure to engineer new ones that will bind a particular target. As natural proteins are sometimes difficult to modify without disrupting their overall structure, some structural biologists turn towards *de novo* approaches, in which proteins are created from scratch (Netzer and Fleishman 2016). A difficulty in this field is not only to be able to predict protein shape from the sequence alone but also to make sure that they carry out their assumed function (e.g. binding). A fragment-based method was recently successfully applied to develop antibodies against the respiratory syncytial virus fusion protein (RSVF). The author's method, TopoBuilder, shapes a new stable protein by assembling fragments around an already-existing continuous or discontinuous epitope (Sesterhenn, Yang et al. 2020). Recent developments in SARS-CoV-2 research include the design of miniproteins to inhibit binding of the virus' spike protein to the human angiotensin-converting enzyme 2 (ACE2) receptor (Cao, Goreshnik et al. 2020). The study used two approaches, one similar to TopoBuilder based on fragment reconstruction around the binding domain of ACE2, and another more systematic approach to find new binding sites with the virus. Results are promising with affinities beyond the nanomolar range and cryo-EM structures confirming the computational models.

References

Ahnert, S. E., J. A. Marsh*, et al.* (2015). "Principles of assembly reveal a periodic table of protein complexes." *Science* **350**(6266): aaa2245.

Ako-Adjei, D., W. Fu*, et al.* (2015). "HIV-1, human interaction database: current status and new features." *Nucleic Acids Res* **43**(Database issue): D566-570.

Alam, N., O. Goldstein*, et al.* (2017). "High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock." *PLoS Comput Biol* **13**(12): e1005905.

Aloy, P., H. Ceulemans*, et al.* (2003). "The relationship between sequence and interaction divergence in proteins." *J Mol Biol* **332**(5): 989-998.

AlQuraishi, M. (2019). "ProteinNet: a standardized data set for machine learning of protein structure." *BMC Bioinformatics* **20**(1): 311.

Alvarez-Ponce, D., F. Feyertag and S. Chakraborty (2017). "Position Matters: Network Centrality Considerably Impacts Rates of Protein Evolution in the Human Protein-Protein Interaction Network." *Genome Biol Evol* **9**(6): 1742-1756.

Andreani, J., G. Faure and R. Guerois (2012). "Versatility and invariance in the evolution of homologous heteromeric interfaces." *PLoS Comput Biol* **8**(8): e1002677.

Andreani, J., G. Faure and R. Guerois (2013). "InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution." *Bioinformatics* **29**(14): 1742-1749.

Andreani, J. and R. Guerois (2014). "Evolution of protein interactions: from interactomes to interfaces." *Arch Biochem Biophys* **554**: 65-75.

Andreani, J., C. Quignot and R. Guerois (2020). "Structural prediction of protein interactions and docking using conservation and coevolution." *Wiley Interdisciplinary Reviews-Computational Molecular Science*.

Andreeva, A., E. Kulesha*, et al.* (2020). "The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures." *Nucleic Acids Res* **48**(D1): D376-D382.

Ashkenazy, H., S. Abadi*, et al.* (2016). "ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules." *Nucleic Acids Res* **44**(W1): W344-350.

Aumentado-Armstrong, T. T., B. Istrate and R. A. Murgita (2015). "Algorithmic approaches to protein-protein interaction site prediction." *Algorithms Mol Biol* **10**: 7.

Bai, X. C., G. McMullan and S. H. Scheres (2015). "How cryo-EM is revolutionizing structural biology." *Trends Biochem Sci* **40**(1): 49-57.

Bajpai, A. K., S. Davuluri*, et al.* (2020). "Systematic comparison of the protein-protein interaction databases from a user's perspective." *J Biomed Inform* **103**: 103380.

Baldassarre, F., D. M. Hurtado*, et al.* (2020). "GraphQA: Protein Model Quality Assessment using Graph Convolutional Networks." *Bioinformatics*.

Barradas-Bautista, D., M. Rosell*, et al.* (2018). "Structural Prediction of Protein-Protein Interactions by Docking: Application to Biomedical Problems." *Adv Protein Chem Struct Biol* **110**: 203-249.

Baspinar, A., E. Cukuroglu*, et al.* (2014). "PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes." *Nucleic Acids Res* **42**(Web Server issue): W285-289.

Basu, S. and B. Wallner (2016). "DockQ: A Quality Measure for Protein-Protein Docking Models." *PLoS One*

11(8): e0161879.

Bertoni, M., F. Kiefer*, et al.* (2017). "Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology." *Sci Rep* **7**(1): 10480.

Bitbol, A. F., R. S. Dwyer*, et al.* (2016). "Inferring interaction partners from protein sequences." *Proc Natl Acad Sci U S A* **113**(43): 12180-12185.

Bonsor, D. A., Q. Zhao*, et al.* (2018). "The Helicobacter pylori adhesin protein HopQ exploits the dimer interface of human CEACAMs to facilitate translocation of the oncoprotein CagA." *EMBO J* **37**(13).

Bruckner, A., C. Polge*, et al.* (2009). "Yeast two-hybrid, a powerful tool for systems biology." *Int J Mol Sci* **10**(6): 2763-2788.

Cao, L., I. Goreshnik*, et al.* (2020). "De novo design of picomolar SARS-CoV-2 miniprotein inhibitors." *Science*.

Chaudhury, S., M. Berrondo*, et al.* (2011). "Benchmarking and analysis of protein docking performance in Rosetta v3.2." *PLoS One* **6**(8): e22477.

Chen, J., N. Sawyer and L. Regan (2013). "Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area." *Protein Sci* **22**(4): 510-515.

Chen, R. and Z. Weng (2003). "A novel shape complementarity scoring function for protein-protein docking." *Proteins* **51**(3): 397-408.

Christoffer, C., G. Terashi*, et al.* (2020). "Performance and enhancement of the LZerD protein assembly pipeline in CAPRI 38-46." *Proteins* **88**(8): 948-961.

Ciemny, M., M. Kurcinski*, et al.* (2018). "Protein-peptide docking: opportunities and challenges." *Drug Discov Today* **23**(8): 1530-1537.

Cocco, S., C. Feinauer*, et al.* (2018). "Inverse statistical physics of protein sequences: a key issues review." *Rep Prog Phys* **81**(3): 032601.

Collavin, L., A. Lunardi and G. Del Sal (2010). "p53-family proteins and their regulators: hubs and spokes in tumor suppression." *Cell Death Differ* **17**(6): 901-911.

Cong, Q., I. Anishchenko*, et al.* (2019). "Protein interaction networks revealed by proteome coevolution." *Science* **365**(6449): 185-189.

Csermely, P., R. Palotai and R. Nussinov (2010). "Induced fit, conformational selection and independent dynamic segments: an extended view of binding events." *Nat Prec*.

Cunningham, J. M., G. Koytiger*, et al.* (2020). "Biophysical prediction of protein-peptide interactions and signaling networks using machine learning." *Nat Methods*.

Dapkunas, J., A. Timinskas*, et al.* (2017). "The PPI3D web server for searching, analyzing and modeling protein-protein interactions in the context of 3D structures." *Bioinformatics* **33**(6): 935-937.

de Juan, D., F. Pazos and A. Valencia (2013). "Emerging methods in protein co-evolution." *Nat Rev Genet* **14**(4): 249-261.

de Vries, S. J. and A. M. Bonvin (2011). "CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK." *PLoS One* **6**(3): e17695.

de Vries, S. J., J. Rey*, et al.* (2017). "The pepATTRACT web server for blind, large-scale peptide-protein docking." *Nucleic Acids Res* **45**(W1): W361-W364.

de Vries, S. J., C. E. Schindler*, et al.* (2015). "A web interface for easy flexible protein-protein docking with ATTRACT." *Biophys J* **108**(3): 462-465.

de Vries, S. J., M. van Dijk and A. M. Bonvin (2010). "The HADDOCK web server for data-driven biomolecular docking." *Nat Protoc* **5**(5): 883-897.

Dequeker, C., E. Laine and A. Carbone (2019). "Decrypting protein surfaces by combining evolution, geometry, and molecular docking." *Proteins* **87**(11): 952-965.

Desta, I. T., K. A. Porter*, et al.* (2020). "Performance and Its Limits in Rigid Body Protein-Protein Docking." *Structure* **28**(9): 1071-1081 e1073.

Dey, S., D. W. Ritchie and E. D. Levy (2018). "PDB-wide identification of biological assemblies from conserved quaternary structure geometry." *Nat Methods* **15**(1): 67-72.

Dominguez, C., R. Boelens and A. M. Bonvin (2003). "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information." *J Am Chem Soc* **125**(7): 1731-1737.

Dong, G. Q., H. Fan*, et al.* (2013). "Optimized atomic statistical potentials: assessment of protein interfaces and loops." *Bioinformatics* **29**(24): 3158-3166.

dos Santos, R. N., F. Morcos*, et al.* (2015). "Dimeric interactions and complex formation using direct coevolutionary couplings." *Sci Rep* **5**: 13652.

Duan, R., L. Qiu*, et al.* (2020). "Performance of human and server prediction in CAPRI rounds 38-45." *Proteins* **88**(8): 1110-1120.

Duroc, Y., R. Kumar*, et al.* (2017). "Concerted action of the MutLβ heterodimer and Mer3 helicase regulates the global extent of meiotic gene conversion." *Elife* **6**.

Eames, M. and T. Kortemme (2007). "Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance." *Structure* **15**(11): 1442-1451.

El-Gebali, S., J. Mistry*, et al.* (2019). "The Pfam protein families database in 2019." *Nucleic Acids Res* **47**(D1): D427-D432.

Erdos, G., T. Szaniszlo*, et al.* (2017). "Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway." *PLoS Comput Biol* **13**(12): e1005885.

Esmaielbeiki, R., K. Krawczyk*, et al.* (2016). "Progress and challenges in predicting protein interfaces." *Brief Bioinform* **17**(1): 117-131.

Estrin, M. and H. J. Wolfson (2017). "SnapDock-template-based docking by Geometric Hashing." *Bioinformatics* **33**(14): i30-i36.

Faure, G., J. Andreani and R. Guerois (2012). "InterEvol database: exploring the structure and evolution of protein complex interfaces." *Nucleic Acids Res* **40**(Database issue): D847-856.

Fleishman, S. J., A. Leaver-Fay*, et al.* (2011). "RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite." *PLoS One* **6**(6): e20161.

Fowler, D. M. and S. Fields (2014). "Deep mutational scanning: a new style of protein science." *Nat Methods* **11**(8): 801-807.

Fragoza, R., J. Das*, et al.* (2019). "Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations." *Nat Commun* **10**(1): 4141.

Gainza, P., F. Sverrisson*, et al.* (2020). "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning." *Nat Methods* **17**(2): 184-192.

Gao, M. and J. Skolnick (2010). "iAlign: a method for the structural comparison of protein-protein interfaces." *Bioinformatics* **26**(18): 2259-2265.

Gao, M. and J. Skolnick (2010). "Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected." *Proc Natl Acad Sci U S A* **107**(52): 22517-22522.

Garcia-Seisdedos, H., C. Empereur-Mot*, et al.* (2017). "Proteins evolve on the edge of supramolecular self-assembly." *Nature* **548**(7666): 244-247.

Garzon, J. I., J. R. Lopez-Blanco*, et al.* (2009). "FRODOCK: a new approach for fast rotational protein-protein docking." *Bioinformatics* **25**(19): 2544-2551.

Geng, C., Y. Jung*, et al.* (2019). "iScore: A novel graph kernel-based function for scoring protein-protein docking models." *Bioinformatics*.

Geng, C., L. C. Xue*, et al.* (2019). "Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it?" *Wiley Interdisciplinary Reviews: Computational Molecular Science* **9**(5): e1410.

Geng, H., F. Chen*, et al.* (2019). "Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins." *Comput Struct Biotechnol J* **17**: 1162-1170.

Ghadie, M. and Y. Xia (2019). "Estimating dispensable content in the human interactome." *Nat Commun* **10**(1): 3205.

Ghadie, M. A., J. Coulombe-Huntington and Y. Xia (2018). "Interactome evolution: insights from genome-wide analyses of protein-protein interactions." *Curr Opin Struct Biol* **50**: 42-48.

Ghoorah, A. W., M. D. Devignes*, et al.* (2016). "Classification and Exploration of 3D Protein Domain Interactions Using Kbdock." *Methods Mol Biol* **1415**: 91-105.

Gibson, T. J., H. Dinkel*, et al.* (2015). "Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad." *Cell Commun Signal* **13**: 42.

Gray, J. J., S. Moughon*, et al.* (2003). "Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations." *Journal of Molecular Biology* **331**(1): 281-299.

Gress, A., V. Ramensky and O. V. Kalinina (2017). "Spatial distribution of disease-associated variants in three-dimensional structures of protein complexes." *Oncogenesis* **6**(9): e380.

Gromiha, M. M., K. Yugandhar and S. Jemimah (2017). "Protein-protein interactions: scoring schemes and binding affinity." *Curr Opin Struct Biol* **44**: 31-38.

Gueudre, T., C. Baldassi*, et al.* (2016). "Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis." *Proc Natl Acad Sci U S A* **113**(43): 12186-12191.

Guo, J. T., D. Xu*, et al.* (2003). "Improving the performance of DomainParser for structural domain partition using neural network." *Nucleic Acids Res* **31**(3): 944-952.

Haas, J., A. Barbato*, et al.* (2018). "Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12." *Proteins: Structure, Function, and Bioinformatics* **86**: 387-398.

Hamer, R., Q. Luo*, et al.* (2010). "i-Patch: interprotein contact prediction using local network information." *Proteins* **78**(13): 2781-2797.

Hashimoto, K. and A. R. Panchenko (2010). "Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states." *Proc Natl Acad Sci U S A* **107**(47): 20352-20357.

Hochberg, G. K. A., D. A. Shepherd*, et al.* (2018). "Structural principles that enable oligomeric small heat-shock protein paralogs to evolve distinct functions." *Science* **359**(6378): 930-935.

Hong, S. H., K. Joo and J. Lee (2019). "ConDo: protein domain boundary prediction using coevolutionary information." *Bioinformatics* **35**(14): 2411-2417.

Hopf, T. A., C. P. Scharfe*, et al.* (2014). "Sequence co-evolution gives 3D contacts and structures of protein complexes." *Elife* **3**.

Hou, Q. Z., P. F. G. De Geest*, et al.* (2017). "Seeing the trees through the forest: sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest." *Bioinformatics* **33**(10):

1479-1487.

Huang, S. Y. (2014). "Search strategies and evaluation in protein-protein docking: principles, advances and challenges." *Drug Discov Today* **19**(8): 1081-1096.

Huang, S. Y. (2015). "Exploring the potential of global protein-protein docking: an overview and critical assessment of current programs for automatic ab initio docking." *Drug Discov Today* **20**(8): 969-977.

Hwang, H., D. Petrey and B. Honig (2016). "A hybrid method for protein-protein interface prediction." *Protein Sci* **25**(1): 159-165.

Hwang, H., T. Vreven*, et al.* (2010). "Protein-protein docking benchmark version 4.0." *Proteins* **78**(15): 3111-3114.

Iserte, J., F. L. Simonetti*, et al.* (2015). "I-COMS: Interprotein-COrrelated Mutations Server." *Nucleic Acids Res* **43**(W1): W320-325.

Janin, J., K. Henrick*, et al.* (2003). "CAPRI: a Critical Assessment of PRedicted Interactions." *Proteins* **52**(1): 2-9.

Javaheri, A., T. Kruse*, et al.* (2016). "Helicobacter pylori adhesin HopQ engages in a virulence-enhancing interaction with human CEACAMs." *Nat Microbiol* **2**: 16189.

Jehl, P., J. Manguy*, et al.* (2016). "ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences." *Nucleic Acids Res* **44**(W1): W11-15.

Jiang, Y. and C. G. Kalodimos (2017). "NMR Studies of Large Proteins." *J Mol Biol* **429**(17): 2667-2676.

Jimenez-Garcia, B., P. Bernado and J. Fernandez-Recio (2020). "Structural Characterization of Protein-Protein Interactions with pyDockSAXS." *Methods Mol Biol* **2112**: 131-144.

Jimenez-Garcia, B., C. Pons and J. Fernandez-Recio (2013). "pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring." *Bioinformatics* **29**(13): 1698-1699.

Johansson-Akhe, I., C. Mirabello and B. Wallner (2019). "Predicting protein-peptide interaction sites using distant protein complexes as structural templates." *Sci Rep* **9**(1): 4267.

Johansson-Akhe, I., C. Mirabello and B. Wallner (2020). "InterPep2: Global Peptide-Protein Docking using Interaction Surface Templates." *Bioinformatics*.

Johansson, K. E. and K. Lindorff-Larsen (2018). "Structural heterogeneity and dynamics in protein evolution and design." *Curr Opin Struct Biol* **48**: 157-163.

Jones, D. T., D. W. Buchan*, et al.* (2012). "PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments." *Bioinformatics* **28**(2): 184-190.

Jones, D. T. and S. M. Kandathil (2018). "High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features." *Bioinformatics* **34**(19): 3308-3315.

Kabsch, W. and C. Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers* **22**(12): 2577-2637.

Kastritis, P. L. and A. M. Bonvin (2013). "On the binding affinity of macromolecular interactions: daring to ask why proteins interact." *J R Soc Interface* **10**(79): 20120835.

Kastritis, P. L., J. P. Rodrigues*, et al.* (2014). "Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface." *J Mol Biol* **426**(14): 2632-2652.

Katoh, K. and D. M. Standley (2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability." *Mol Biol Evol* **30**(4): 772-780.

Kawabata, T. (2016). "HOMCOS: an updated server to search and model complex 3D structures." *J Struct Funct Genomics* **17**(4): 83-99.

Keshava Prasad, T. S., R. Goel, *et al.* (2009). "Human Protein Reference Database--2009 update." *Nucleic Acids Res* **37**(Database issue): D767-772.

Khramushin, A., O. Marcu, *et al.* (2019). "Modeling beta-sheet peptide-protein interactions: Rosetta FlexPepDock in CAPRI rounds 38-45." *Proteins*.

Kim, I., H. Lee, *et al.* (2014). "Linear motif-mediated interactions have contributed to the evolution of modularity in complex protein interaction networks." *PLoS Comput Biol* **10**(10): e1003881.

Kim, P. M., L. J. Lu, *et al.* (2006). "Relating three-dimensional structures to protein networks provides evolutionary insights." *Science* **314**(5807): 1938-1941.

Kleiger, G., A. Saha, *et al.* (2009). "Rapid E2-E3 assembly and disassembly enable processive ubiquitylation of cullin-RING ubiquitin ligase substrates." *Cell* **139**(5): 957-968.

Kolodny, R., L. Pereyaslavets, *et al.* (2013). "On the universe of protein folds." *Annu Rev Biophys* **42**: 559-582.

Koonin, E. V. (2005). "Orthologs, paralogs, and evolutionary genomics." *Annu Rev Genet* **39**: 309-338.

Koukos, P. I. and A. Bonvin (2019). "Integrative modelling of biomolecular complexes." *J Mol Biol*.

Kowalsman, N. and M. Eisenstein (2007). "Inherent limitations in protein-protein docking procedures." *Bioinformatics* **23**(4): 421-426.

Kozakov, D., D. Beglov, *et al.* (2013). "How good is automated protein docking?" *Proteins* **81**(12): 2159-2166.

Kozakov, D., R. Brenke, *et al.* (2006). "PIPER: an FFT-based protein docking program with pairwise potentials." *Proteins* **65**(2): 392-406.

Kozakov, D., D. R. Hall, *et al.* (2017). "The ClusPro web server for protein-protein docking." *Nat Protoc* **12**(2): 255-278.

Krol, M., R. A. Chaleil, *et al.* (2007). "Implicit flexibility in protein docking: cross-docking and local refinement." *Proteins* **69**(4): 750-757.

Kryshtafovych, A., T. Schwede, *et al.* (2019). "Critical assessment of methods of protein structure prediction (CASP)-Round XIII." *Proteins* **87**(12): 1011-1020.

Krystkowiak, I., J. Manguy and N. E. Davey (2018). "PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants." *Nucleic Acids Res* **46**(W1): W235-W241.

Kumar, M., M. Gouw, *et al.* (2019). "ELM-the eukaryotic linear motif resource in 2020." *Nucleic Acids Res*.

Kundrotas, P. J., I. Anishchenko, *et al.* (2018). "Dockground: A comprehensive data resource for modeling of protein complexes." *Protein Sci* **27**(1): 172-181.

Kurcinski, M., A. Badaczewska-Dawid, *et al.* (2020). "Flexible docking of peptides to proteins using CABS-dock." *Protein Sci* **29**(1): 211-222.

Kurcinski, M., M. Jamroz, *et al.* (2015). "CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site." *Nucleic Acids Res* **43**(W1): W419-424.

Larkin, M. A., G. Blackshields, *et al.* (2007). "Clustal W and Clustal X version 2.0." *Bioinformatics* **23**(21): 2947-2948.

Lee, H., L. Heo, *et al.* (2015). "GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization." *Nucleic Acids Res* **43**(W1): W431-435.

Lensink, M. F., R. Mendez and S. J. Wodak (2007). "Docking and scoring protein complexes: CAPRI 3rd Edition." *Proteins* **69**(4): 704-718.

Lensink, M. F., N. Nadzirin, *et al.* (2019). "Modeling protein-protein, protein-peptide and protein-

oligosaccharide complexes: CAPRI 7(th) edition." *Proteins*.

Lensink, M. F., S. Velankar*, et al.* (2017). "The challenge of modeling protein assemblies: The CASP12-CAPRI experiment." *Proteins*.

Lensink, M. F., S. Velankar*, et al.* (2016). "Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment." *Proteins* **84 Suppl 1**: 323-348.

Levy, E. D. (2010). "A simple definition of structural regions in proteins and its use in analyzing interface evolution." *J Mol Biol* **403**(4): 660-670.

Li, Y., J. Han*, et al.* (2016). "Structural basis for activity regulation of MLL family methyltransferases." *Nature* **530**(7591): 447-452.

Lindorff-Larsen, K., S. Piana*, et al.* (2011). "How fast-folding proteins fold." *Science* **334**(6055): 517-520.

Lo, Y. S., Y. C. Chen and J. M. Yang (2010). "3D-interologs: an evolution database of physical protein- protein interactions across multiple genomes." *BMC Genomics* **11 Suppl 3**: S7.

Lopes, A., S. Sacquin-Mora*, et al.* (2013). "Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information." *PLoS Comput Biol* **9**(12): e1003369.

Lopez-Blanco, J. R. and P. Chacon (2019). "KORP: knowledge-based 6D potential for fast protein and loop modeling." *Bioinformatics* **35**(17): 3013-3019.

Lyskov, S. and J. J. Gray (2008). "The RosettaDock server for local protein-protein docking." *Nucleic Acids Res* **36**(Web Server issue): W233-238.

Macindoe, G., L. Mavridis*, et al.* (2010). "HexServer: an FFT-based protein docking server powered by graphics processors." *Nucleic Acids Res* **38**(Web Server issue): W445-449.

Madej, T., C. J. Lanczycki*, et al.* (2014). "MMDB and VAST+: tracking structural similarities between macromolecular complexes." *Nucleic Acids Res* **42**(Database issue): D297-303.

Maheshwari, S. and M. Brylinski (2015). "Predicting protein interface residues using easily accessible on-line resources." *Brief Bioinform* **16**(6): 1025-1034.

Malinverni, D., A. Jost Lopez*, et al.* (2017). "Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and coevolutionary sequence analysis." *Elife* **6**.

Mandell, D. J., E. A. Coutsias and T. Kortemme (2009). "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling." *Nat Methods* **6**(8): 551-552.

Marks, D. S., L. J. Colwell*, et al.* (2011). "Protein 3D structure computed from evolutionary sequence variation." *PLoS One* **6**(12): e28766.

Marmier, G., M. Weigt and A. F. Bitbol (2019). "Phylogenetic correlations can suffice to infer protein partners from sequences." *PLoS Comput Biol* **15**(10): e1007179.

Marsh, J. A. and S. A. Teichmann (2015). "Structure, dynamics, assembly, and evolution of protein complexes." *Annu Rev Biochem* **84**: 551-575.

Marsh, J. A., S. A. Teichmann and J. D. Forman-Kay (2012). "Probing the diverse landscape of protein flexibility and binding." *Curr Opin Struct Biol* **22**(5): 643-650.

Mashiach, E., R. Nussinov and H. J. Wolfson (2010). "FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking." *Nucleic Acids Res* **38**(Web Server issue): W457-461.

Mashiach, E., D. Schneidman-Duhovny*, et al.* (2008). "FireDock: a web server for fast interaction refinement in molecular docking." *Nucleic Acids Res* **36**(Web Server issue): W229-232.

Mayrose, I., D. Graur*, et al.* (2004). "Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior." *Mol Biol Evol* **21**(9): 1781-1791.

McKusick-Nathans Institute of Genetic Medicine, J. H. U. B., MD) Online Mendelian Inheritance in Man, OMIM®.

McLaughlin, R. N., Jr., F. J. Poelwijk*, et al.* (2012). "The spatial architecture of protein function and adaptation." *Nature* **491**(7422): 138-142.

Mendez, R., R. Leplae*, et al.* (2003). "Assessment of blind predictions of protein-protein interactions: current status of docking methods." *Proteins* **52**(1): 51-67.

Meszaros, B., G. Erdos and Z. Dosztanyi (2018). "IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding." *Nucleic Acids Res* **46**(W1): W329-W337.

Meszaros, B., I. Simon and Z. Dosztanyi (2009). "Prediction of protein binding regions in disordered proteins." *PLoS Comput Biol* **5**(5): e1000376.

Meyer, M. J., J. F. Beltran*, et al.* (2018). "Interactome INSIDER: a structural interactome browser for genomic studies." *Nat Methods* **15**(2): 107-114.

Minhas, F., B. J. Geiss and A. Ben-Hur (2014). "PAIRpred: partner-specific prediction of interacting residues from sequence and structure." *Proteins* **82**(7): 1142-1155.

Mintseris, J. and Z. Weng (2005). "Structure, function, and evolution of transient and obligate protein-protein interactions." *Proc Natl Acad Sci U S A* **102**(31): 10930-10935.

Mirabello, C. and B. Wallner (2017). "InterPred: A pipeline to identify and model protein-protein interactions." *Proteins* **85**(6): 1159-1170.

Miryala, S. K., A. Anbarasu and S. Ramaiah (2018). "Discerning molecular interactions: A comprehensive review on biomolecular interaction databases and network analysis tools." *Gene* **642**: 84-94.

Moonens, K., Y. Hamway*, et al.* (2018). "Helicobacter pylori adhesin HopQ disrupts trans dimerization in human CEACAMs." *EMBO J* **37**(13).

Morcos, F. and J. N. Onuchic (2019). "The role of coevolutionary signatures in protein interaction dynamics, complex inference, molecular recognition, and mutational landscapes." *Curr Opin Struct Biol* **56**: 179-186.

Morcos, F., A. Pagnani*, et al.* (2011). "Direct-coupling analysis of residue coevolution captures native contacts across many protein families." *Proc Natl Acad Sci U S A* **108**(49): E1293-1301.

Moretti, R., S. Lyskov*, et al.* (2018). "Web-accessible molecular modeling with Rosetta: The Rosetta Online Server that Includes Everyone (ROSIE)." *Protein Sci* **27**(1): 259-268.

Morrow, J. K. and S. Zhang (2012). "Computational prediction of protein hot spot residues." *Curr Pharm Des* **18**(9): 1255-1265.

Mosca, R., A. Céol and P. Aloy (2013). "Interactome3D: adding structural details to protein networks." *Nat Methods* **10**(1): 47-53.

Mukherjee, S. and Y. Zhang (2009). "MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming." *Nucleic Acids Res* **37**(11): e83.

Myllykoski, M., M. A. Eichel*, et al.* (2018). "High-affinity heterotetramer formation between the large myelin-associated glycoprotein and the dynein light chain DYNLL1." *J Neurochem* **147**(6): 764-783.

Nadalin, F. and A. Carbone (2018). "Protein-protein interaction specificity is captured by contact preferences and interface composition." *Bioinformatics* **34**(3): 459-468.

Nadaradjane, A. A., C. Quignot*, et al.* (2019). "Docking Proteins and Peptides Under Evolutionary Constraints in CAPRI rounds 38-45." *Proteins*.

Nakane, T., A. Kotecha*, et al.* (2020). "Single-particle cryo-EM at atomic resolution." *Nature*.

Ncbi Resource Coordinators (2018). "Database resources of the National Center for Biotechnology

Information." *Nucleic Acids Res* **46**(D1): D8-D13.

Netzer, R. and S. J. Fleishman (2016). "PROTEIN DESIGN. Inspired by nature." *Science* **352**(6286): 657-658.

Netzer, R., D. Listov*, et al.* (2018). "Ultrahigh specificity in a network of computationally designed protein-interaction pairs." *Nat Commun* **9**(1): 5286.

Neveu, E., P. Popov*, et al.* (2018). "RapidRMSD: rapid determination of RMSDs corresponding to motions of flexible molecules." *Bioinformatics* **34**(16): 2757-2765.

Nivon, L. G., R. Moretti and D. Baker (2013). "A Pareto-optimal refinement method for protein design scaffolds." *PLoS One* **8**(4): e59004.

Northey, T., A. Barešic and A. C. R. Martin (2017). "IntPred: a structure-based predictor of protein-protein interaction sites." *Bioinformatics*.

Orban-Nemeth, Z., R. Beveridge*, et al.* (2018). "Structural prediction of protein models using distance restraints derived from cross-linking mass spectrometry data." *Nat Protoc* **13**(3): 478-494.

Orchard, S., M. Ammari*, et al.* (2014). "The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases." *Nucleic Acids Res* **42**(Database issue): D358-363.

Oughtred, R., C. Stark*, et al.* (2019). "The BioGRID interaction database: 2019 update." *Nucleic Acids Res* **47**(D1): D529-D541.

Ovchinnikov, S., H. Kamisetty and D. Baker (2014). "Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information." *Elife* **3**: e02030.

Padhorny, D., A. Kazennov*, et al.* (2016). "Protein-protein docking by fast generalized Fourier transforms on 5D rotational manifolds." *Proc Natl Acad Sci U S A* **113**(30): E4286-4293.

Pages, G., B. Charmettant and S. Grudinin (2019). "Protein model quality assessment using 3D oriented convolutional neural networks." *Bioinformatics* **35**(18): 3313-3319.

Park, T., M. Baek*, et al.* (2019). "GalaxyTongDock: Symmetric and asymmetric ab initio protein-protein docking web server with improved energy parameters." *J Comput Chem* **40**(27): 2413-2417.

Pauling, L., R. B. Corey and H. R. Branson (1951). "The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain." *Proc Natl Acad Sci U S A* **37**(4): 205-211.

Pei, J. and N. V. Grishin (2001). "AL2CO: calculation of positional conservation in a protein sequence alignment." *Bioinformatics* **17**(8): 700-712.

Pierce, B. G., Y. Hourai and Z. Weng (2011). "Accelerating protein docking in ZDOCK using an advanced 3D convolution library." *PLoS One* **6**(9): e24657.

Pierce, B. G., K. Wiehe*, et al.* (2014). "ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers." *Bioinformatics* **30**(12): 1771-1773.

Pittala, S. and C. Bailey-Kellogg (2020). "Learning context-aware structural representations to predict antigen and antibody binding interfaces." *Bioinformatics* **36**(13): 3996-4003.

Plach, M. G., F. Semmelmann*, et al.* (2017). "Evolutionary diversification of protein-protein interactions by interface add-ons." *Proc Natl Acad Sci U S A* **114**(40): E8333-E8342.

Porollo, A. and J. Meller (2007). "Prediction-based fingerprints of protein-protein interactions." *Proteins* **66**(3): 630-645.

Porter, K. A., I. Desta*, et al.* (2019). "What method to use for protein-protein docking?" *Curr Opin Struct Biol* **55**: 1-7.

Porter, K. A., D. Padhorny*, et al.* (2019). "Template-based modeling by ClusPro in CASP13 and the potential for using co-evolutionary information in docking." *Proteins* **87**(12): 1241-1248.

Postic, G., Y. Ghouzam*, et al.* (2017). "An ambiguity principle for assigning protein structural domains." *Sci Adv* **3**(1): e1600552.

Postic, G., J. Marcoux*, et al.* (2020). "Probing Protein Interaction Networks by Combining MS-Based Proteomics and Structural Data Integration." *J Proteome Res* **19**(7): 2807-2820.

Pupko, T., R. E. Bell*, et al.* (2002). "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues." *Bioinformatics* **18 Suppl 1**: S71-77.

Quignot, C., P. Granger*, et al.* (2020). "Atomic-level evolutionary information improves protein-protein interface scoring." *bioRxiv*.

Quignot, C., J. Rey*, et al.* (2018). "InterEvDock2: an expanded server for protein docking using evolutionary and biological information from homology models and multimeric inputs." *Nucleic Acids Res* **46**(W1): W408-W416.

Ramírez-Aportela, E., J. R. López-Blanco and P. Chacón (2016). "FRODOCK 2.0: Fast Protein-Protein docking server." *Bioinformatics*: btw141.

Ratmann, O., C. Wiuf and J. W. Pinney (2009). "From evidence to inference: probing the evolution of protein interaction networks." *HFSP J* **3**(5): 290-306.

Raveh, B., N. London and O. Schueler-Furman (2010). "Sub-angstrom modeling of complexes between flexible peptides and globular proteins." *Proteins* **78**(9): 2029-2040.

Remmert, M., A. Biegert*, et al.* (2011). "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment." *Nat Methods* **9**(2): 173-175.

Ritchie, D. W., A. W. Ghoorah*, et al.* (2012). "Fast protein structure alignment using Gaussian overlap scoring of backbone peptide fragment similarity." *Bioinformatics* **28**(24): 3274-3281.

Ritchie, D. W. K., G. J. L. (2000). "Protein docking using spherical polar Fourier correlations." *Proteins* **39**(2): 178-194.

Rodier, F., R. P. Bahadur*, et al.* (2005). "Hydration of protein-protein interfaces." *Proteins* **60**(1): 36-45.

Rodrigues, J. P. and A. M. Bonvin (2014). "Integrative computational modeling of protein interactions." *FEBS J* **281**(8): 1988-2003.

Rodrigues, J. P., M. Trellet*, et al.* (2012). "Clustering biomolecular complexes by residue contacts similarity." *Proteins* **80**(7): 1810-1817.

Rodriguez-Rivas, J., S. Marsili*, et al.* (2016). "Conservation of coevolving protein interfaces bridges prokaryote-eukaryote homologies in the twilight zone." *Proc Natl Acad Sci U S A* **113**(52): 15018-15023.

Rollins, N. J., K. P. Brock*, et al.* (2019). "Inferring protein 3D structure from deep mutation scans." *Nat Genet* **51**(7): 1170-1176.

Sahni, N., S. Yi*, et al.* (2015). "Widespread macromolecular interaction perturbations in human genetic disorders." *Cell* **161**(3): 647-660.

Saladin, A., J. Rey*, et al.* (2014). "PEP-SiteFinder: a tool for the blind identification of peptide binding sites on protein surfaces." *Nucleic Acids Res* **42**(Web Server issue): W221-226.

Sanchez-Garcia, R., C. O. S. Sorzano*, et al.* (2019). "BIPSPI: a method for the prediction of partner-specific protein-protein interfaces." *Bioinformatics* **35**(3): 470-477.

Savojardo, C., P. Fariselli*, et al.* (2017). "ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model." *Bioinformatics* **33**(11): 1656-1663.

Schmiedel, J. M. and B. Lehner (2019). "Determining protein structures using deep mutagenesis." *Nat Genet*

51(7): 1177-1186.

Schneidman-Duhovny, D., Y. Inbar*, et al.* (2005). "PatchDock and SymmDock: servers for rigid and symmetric docking." *Nucleic Acids Res* **33**(Web Server issue): W363-367.

Schueler-Furman, O. and N. London (2017). Modeling Peptide-Protein Interactions. Methods and Protocols., Humana Press.

Schug, A., M. Weigt*, et al.* (2009). "High-resolution protein complexes from integrating genomic information with molecular simulation." *Proc Natl Acad Sci U S A* **106**(52): 22124-22129.

Seemayer, S., M. Gruber and J. Soding (2014). "CCMpred--fast and precise prediction of protein residue-residue contacts from correlated mutations." *Bioinformatics* **30**(21): 3128-3130.

Senior, A. W., R. Evans*, et al.* (2020). "Improved protein structure prediction using potentials from deep learning." *Nature* **577**(7792): 706-710.

Sesterhenn, F., C. Yang*, et al.* (2020). "De novo protein design enables the precise induction of RSV-neutralizing antibodies." *Science* **368**(6492).

Shannon, P., A. Markiel*, et al.* (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome Res* **13**(11): 2498-2504.

Shi, Q., W. Chen*, et al.* (2019). "DNN-Dom: predicting protein domain boundary from sequence alone by deep neural network." *Bioinformatics* **35**(24): 5128-5136.

Shoemaker, B. A., D. Zhang*, et al.* (2012). "IBIS (Inferred Biomolecular Interaction Server) reports, predicts and integrates multiple types of conserved interactions for proteins." *Nucleic Acids Res* **40**(Database issue): D834-840.

Shoemaker, S. C. and N. Ando (2018). "X-rays in the Cryo-Electron Microscopy Era: Structural Biology's Dynamic Future." *Biochemistry* **57**(3): 277-285.

Siddiq, M. A., G. K. Hochberg and J. W. Thornton (2017). "Evolution of protein specificity: insights from ancestral protein reconstruction." *Curr Opin Struct Biol* **47**: 113-122.

Sillitoe, I., N. Dawson*, et al.* (2019). "CATH: expanding the horizons of structure-based functional annotations for genome sequences." *Nucleic Acids Res* **47**(D1): D280-D284.

Simkovic, F., S. Ovchinnikov*, et al.* (2017). "Applications of contact predictions to structural biology." *IUCrJ* **4**(Pt 3): 291-300.

Sippl, M. J. and M. Wiederstein (2012). "Detection of spatial correlations in protein structures and molecular complexes." *Structure* **20**(4): 718-728.

Snyder, J. T., D. K. Worthylake*, et al.* (2002). "Structural basis for the selective activation of Rho GTPases by Dbl exchange factors." *Nat Struct Biol* **9**(6): 468-475.

Socolich, M., S. W. Lockless*, et al.* (2005). "Evolutionary information for specifying a protein fold." *Nature* **437**(7058): 512-518.

Soding, J. (2005). "Protein homology detection by HMM-HMM comparison." *Bioinformatics* **21**(7): 951-960.

Song, Y., F. DiMaio*, et al.* (2013). "High-resolution comparative modeling with RosettaCM." *Structure* **21**(10): 1735-1742.

Soni, N. and M. S. Madhusudhan (2017). "Computational modeling of protein assemblies." *Curr Opin Struct Biol* **44**: 179-189.

Starr, T. N. and J. W. Thornton (2016). "Epistasis in protein evolution." *Protein Sci* **25**(7): 1204-1218.

Stein, R. R., D. S. Marks and C. Sander (2015). "Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models." *PLoS Comput Biol* **11**(7): e1004182.

Steinegger, M., M. Meier*, et al.* (2019). "HH-suite3 for fast remote homology detection and deep protein annotation." *BMC Bioinformatics* **20**(1): 473.

Szklarczyk, D., J. H. Morris*, et al.* (2017). "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible." *Nucleic Acids Res* **45**(D1): D362-D368.

Teichmann, S. A. (2002). "The constraints protein-protein interactions place on sequence divergence." *J Mol Biol* **324**(3): 399-407.

Teppa, E., D. J. Zea and C. Marino-Buslje (2017). "Protein-protein interactions leave evolutionary footprints: High molecular coevolution at the core of interfaces." *Protein Sci* **26**(12): 2438-2444.

Thul, P. J., L. Akesson*, et al.* (2017). "A subcellular map of the human proteome." *Science* **356**(6340).

Torchala, M., I. H. Moal*, et al.* (2013). "SwarmDock: a server for flexible protein-protein docking." *Bioinformatics* **29**(6): 807-809.

Tovchigrechko, A. and I. A. Vakser (2006). "GRAMM-X public web server for protein-protein docking." *Nucleic Acids Res* **34**(Web Server issue): W310-314.

Trellet, M., A. S. Melquiond and A. M. Bonvin (2013). "A unified conformational selection and induced fit approach to protein-peptide docking." *PLoS One* **8**(3): e58769.

Tripathi, A. and V. A. Bankaitis (2017). "Molecular Docking: From Lock and Key to Combination Lock." *J Mol Med Clin Appl* **2**(1).

Tyka, M. D., D. A. Keedy*, et al.* (2011). "Alternate states of proteins revealed by detailed energy landscape mapping." *J Mol Biol* **405**(2): 607-618.

Uguzzoni, G., S. John Lovis*, et al.* (2017). "Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis." *Proc Natl Acad Sci U S A* **114**(13): E2662-E2671.

Uhlen, M., C. Zhang*, et al.* (2017). "A pathology atlas of the human cancer transcriptome." *Science* **357**(6352).

UniProtConsortium (2019). "UniProt: a worldwide hub of protein knowledge." *Nucleic Acids Res* **47**(D1): D506-D515.

Uyar, B., R. J. Weatheritt*, et al.* (2014). "Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer?" *Mol Biosyst* **10**(10): 2626-2642.

Vakser, I. A. (2014). "Protein-protein docking: from interaction to interactome." *Biophys J* **107**(8): 1785-1793.

Van Roey, K., B. Uyar*, et al.* (2014). "Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation." *Chem Rev* **114**(13): 6733-6778.

van Zundert, G. C. P., J. Rodrigues*, et al.* (2016). "The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes." *J Mol Biol* **428**(4): 720-725.

Via, A., B. Uyar*, et al.* (2015). "How pathogens use linear motifs to perturb host cell networks." *Trends Biochem Sci* **40**(1): 36-48.

Vreven, T., H. Hwang and Z. Weng (2011). "Integrating atom-based and residue-based scoring functions for protein-protein docking." *Protein Sci* **20**(9): 1576-1586.

Vreven, T., I. H. Moal*, et al.* (2015). "Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2." *J Mol Biol* **427**(19): 3031-3041.

Walker, D. R., J. P. Bond*, et al.* (1999). "Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features." *Oncogene* **18**(1): 211-218.

Wang, G. and R. L. Dunbrack, Jr. (2005). "PISCES: recent improvements to a PDB sequence culling server." *Nucleic Acids Res* **33**(Web Server issue): W94-98.

Wang, X., B. Yu*, et al.* (2019). "Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique." *Bioinformatics* **35**(14): 2395-2402.

Ward, A. B., A. Sali and I. A. Wilson (2013). "Biochemistry. Integrative structural biology." *Science* **339**(6122): 913-915.

Waterhouse, A., M. Bertoni*, et al.* (2018). "SWISS-MODEL: homology modelling of protein structures and complexes." *Nucleic Acids Res* **46**(W1): W296-W303.

Waterhouse, A. M., J. B. Procter*, et al.* (2009). "Jalview Version 2--a multiple sequence alignment editor and analysis workbench." *Bioinformatics* **25**(9): 1189-1191.

Webb, B. and A. Sali (2016). "Comparative Protein Structure Modeling Using MODELLER." *Curr Protoc Protein Sci* **86**: 2 9 1-2 9 37.

Weigt, M., R. A. White*, et al.* (2009). "Identification of direct residue contacts in protein-protein interaction by message passing." *Proc Natl Acad Sci U S A* **106**(1): 67-72.

Woolhouse, M. E., J. P. Webster*, et al.* (2002). "Biological and biomedical implications of the co-evolution of pathogens and their hosts." *Nat Genet* **32**(4): 569-577.

wwPDBconsortium (2019). "Protein Data Bank: the single global archive for 3D macromolecular structure data." *Nucleic Acids Res* **47**(D1): D520-D528.

Xia, B., A. Mamonov*, et al.* (2015). "Accounting for observed small angle X-ray scattering profile in the protein-protein docking server ClusPro." *J Comput Chem* **36**(20): 1568-1572.

Xu, D. (2012). "Protein databases on the internet." *Curr Protoc Protein Sci* **Chapter 2**: Unit2 6.

Xu, J. (2019). "Distance-based protein folding powered by deep learning." *Proc Natl Acad Sci U S A* **116**(34): 16856-16865.

Xu, J., F. Jiao and L. Yu (2008). "Protein structure prediction using threading." *Methods Mol Biol* **413**: 91-121.

Xu, J. and S. Wang (2019). "Analysis of distance-based protein structure prediction by deep learning in CASP13." *Proteins* **87**(12): 1069-1081.

Xu, K., I. Bezakova*, et al.* (2011). "Path lengths in protein-protein interaction networks and biological complexity." *Proteomics* **11**(10): 1857-1867.

Xu, Q. and R. L. Dunbrack (2011). "The protein common interface database (ProtCID)--a comprehensive database of interactions of homologous proteins in multiple crystal forms." *Nucleic Acids Res* **39**(Database issue): D761-770.

Xue, L. C., D. Dobbs*, et al.* (2015). "Computational prediction of protein interfaces: A review of data driven methods." *FEBS Lett* **589**(23): 3516-3526.

Xue, L. C., D. Dobbs and V. Honavar (2011). "HomPPI: a class of sequence homology based protein-protein interface prediction methods." *BMC Bioinformatics* **12**: 244.

Xue, L. C., R. A. Jordan*, et al.* (2014). "DockRank: ranking docked conformations using partner-specific sequence homology-based protein interface prediction." *Proteins* **82**(2): 250-267.

Yamada, T. and P. Bork (2009). "Evolution of biomolecular networks: lessons from metabolic and protein interactions." *Nat Rev Mol Cell Biol* **10**(11): 791-803.

Yan, Y., D. Zhang*, et al.* (2017). "HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy." *Nucleic Acids Res* **45**(W1): W365-W373.

Yang, J., I. Anishchenko*, et al.* (2020). "Improved protein structure prediction using predicted interresidue orientations." *Proc Natl Acad Sci U S A* **117**(3): 1496-1503.

Yang, J., R. Yan*, et al.* (2015). "The I-TASSER Suite: protein structure and function prediction." *Nat Methods*

12(1): 7-8.

Yang, X., J. Coulombe-Huntington*, et al.* (2016). "Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing." *Cell* **164**(4): 805-817.

Yip, K. M., N. Fischer*, et al.* (2020). "Atomic-resolution protein structure determination by cryo-EM." *Nature*.

Yu, J., J. Andreani*, et al.* (2017). "Lessons from (co-)evolution in the docking of proteins and peptides for CAPRI Rounds 28-35." *Proteins* **85**(3): 378-390.

Yu, J. and R. Guerois (2016). "PPI4DOCK: large scale assessment of the use of homology models in free docking over more than 1000 realistic targets." *Bioinformatics* **32**(24): 3760-3767.

Yu, J., M. Vavrusa*, et al.* (2016). "InterEvDock: a docking server to predict the structure of protein-protein interactions using evolutionary information." *Nucleic Acids Res* **44**(W1): W542-549.

Yu, J., M. Vavrusa*, et al.* (2016). "InterEvDock: A docking server to predict the structure of protein-protein interactions using evolutionary information." *submitted*.

Zarin, T., B. Strome*, et al.* (2019). "Proteome-wide signatures of function in highly diverged intrinsically disordered regions." *Elife* **8**.

Zeng, H., S. Wang*, et al.* (2018). "ComplexContact: a web server for inter-protein contact prediction using deep learning." *Nucleic Acids Res* **46**(W1): W432-W437.

Zhang, C. and L. Lai (2011). "SDOCK: a global protein-protein docking program using stepwise force-field potentials." *J Comput Chem* **32**(12): 2598-2612.

Zhang, L., L. Fairall*, et al.* (2011). "The IDOL-UBE2D complex mediates sterol-dependent degradation of the LDL receptor." *Genes Dev* **25**(12): 1262-1274.

Zhang, Q. C., D. Petrey*, et al.* (2010). "Protein interface conservation across structure space." *Proc Natl Acad Sci U S A* **107**(24): 10896-10901.

Zhang, Y. and J. Skolnick (2005). "The protein structure prediction problem could be solved using the current PDB library." *Proc Natl Acad Sci U S A* **102**(4): 1029-1034.

Zhang, Z., P. Xiong*, et al.* (2020). "Accurate inference of the full base-pairing structure of RNA by deep mutational scanning and covariation-induced deviation of activity." *Nucleic Acids Res* **48**(3): 1451-1465.

Zheng, W., X. Zhou*, et al.* (2020). "FUpred: detecting protein domains through deep-learning-based contact map prediction." *Bioinformatics* **36**(12): 3749-3757.

Zhou, H., B. Xue and Y. Zhou (2007). "DDOMAIN: Dividing structures into domains using a normalized domain-domain interaction profile." *Protein Sci* **16**(5): 947-955.

Zimmermann, L., A. Stephens*, et al.* (2018). "A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core." *J Mol Biol* **430**(15): 2237-2243.

A. Supplementary materials for Chapter 1

Table A-1: Links to web resources *to explore the evolution of interface structures, predict binding sites and model protein-protein and protein-peptide complex structures using evolutionary information.*

Category	Name	Type	Link to web service	Ref.
Interface evolution comparison tools	MM-Align	software	https://zhanglab.ccmb.med.umich.edu/MM-align/	(Mukherjee and Zhang 2009)
	iAlign	software	http://pwp.gatech.edu/cssb/ialign/	(Gao and Skolnick 2010)
	FCC	software	https://github.com/haddocking/fcc	(Rodrigues, Trellet et al. 2012)
	TopMatch	server	https://topmatch.services.came.sbg.ac.at/	(Sippl and Wiederstein 2012)
	QSalign	software	https://github.com/elevywis/QSalign	(Dey, Ritchie et al. 2018)
	VAST+	server	https://www.ncbi.nlm.nih.gov/Structure/vastplus/vastplus.cgi	(Madej, Lanczycki et al. 2014)
Interface structure and evolution databases	QSbio	database	http://www.qsbio.org/	(Dey, Ritchie et al. 2018)
	PRISM	database	http://cosbi.ku.edu.tr/prism/	(Baspinar, Cukuroglu et al. 2014)
	3D-interologs	database	http://3d-interologs.life.nctu.edu.tw/	(Lo, Chen et al. 2010)
	IBIS	database	https://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi	(Shoemaker, Zhang et al. 2012)
	ProtCID	database	http://dunbrack2.fccc.edu/ProtCiD/	(Xu and Dunbrack 2011)
	InterEvol	database	http://biodev.cea.fr/interevol/	(Faure, Andreani et al. 2012)
	Periodic table of protein complexes	database	http://www.periodicproteincomplexes.org/	(Ahnert, Marsh et al. 2015)
	Interactome INSIDER	database	http://interactomeinsider.yulab.org/	(Meyer, Beltran et al. 2018)
	Interactome3D	database	https://interactome3d.irbbarcelona.org/	(Mosca, Céol et al. 2013)
Protein evolution tools	Jalview	software	https://www.jalview.org/	(Waterhouse, Procter et al. 2009)
	ProViz	server	http://slim.icr.ac.uk/proviz/	(Jehl, Manguy et al. 2016)

Binding site prediction using evolution	ConSurf	server	https://consurf.tau.ac.il/	(Ashkenazy, Abadi et al. 2016)
	Rate4Site	software	https://www.tau.ac.il/~itay-may/cp/rate4site.html	(Mayrose, Graur et al. 2004)
	SPPIDER	server	http://sppider.cchmc.org/	(Porollo and Meller 2007)
	IntPred	server & software	http://www.bioinf.org.uk/intpred/	(Northey, Barešic et al. 2017)
	EL-SMURF	software	http://github.com/QUST-AIBBDRC/EL-SMURF/	(Wang, Yu et al. 2019)
	DynJet2	software	http://www.lcqb.upmc.fr/dynJET2/	(Dequeker, Laine et al. 2019)
	ISPRED4	server	https://ispred4.biocomp.unibo.it/	(Savojardo, Fariselli et al. 2017)
	PredUS	server	http://honig.c2b2.columbia.edu/predus	(Hwang, Petrey et al. 2016)
	PS-HomPPI	server	http://ailab-projects2.ist.psu.edu/PSHOMP-PIv2	(Xue, Dobbs et al. 2011)
	CPORT	server	http://milou.science.uu.nl/services/CPORT/	(de Vries and Bonvin 2011)
Template-based protein-protein docking	PPI3D	server	http://bioinformatics.ibt.lt/ppi3d/	(Dapkunas, Timinskas et al. 2017)
	SWISS-MODEL	server	https://swissmodel.expasy.org/	(Waterhouse, Bertoni et al. 2018)
	InterPred	server	http://bioinfo.ifm.liu.se/inter/interpred/	(Mirabello and Wallner 2017)
	HDOCK	server	http://hdock.phys.hust.edu.cn/	(Yan, Zhang et al. 2017)
Free and guided docking servers	ClusPro	server	https://cluspro.bu.edu/	(Kozakov, Hall et al. 2017)
	GRAMM-X	server	http://vakser.compbio.ku.edu/resources/gramm/grammx/	(Tovchigrechko and Vakser 2006)
	PatchDock	server	https://bioinfo3d.cs.tau.ac.il/PatchDock/	(Schneidman-Duhovny, Inbar et al. 2005)
	SwarmDock	server	https://bmm.crick.ac.uk/~svc-bmm-swarmdock/	(Torchala, Moal et al. 2013)
	InterEvDock2	server	https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/	(Quignot, Rey et al. 2018)
	GalaxyTong-Dock	server & software	http://galaxy.seoklab.org/tongdock	(Park, Baek et al. 2019)
	pyDock	server & software	https://life.bsc.es/pid/pydock/	(Jimenez-Garcia, Pons et al. 2013)

	HADDOCK	server & software	https://haddock.science.uu.nl/	(van Zundert, Rodrigues et al. 2016)
Docking scoring functions including evolutionary information	InterEvScore	software	http://biodev.cea.fr/interevol/interevscore/	(Andreani, Faure et al. 2013)
	DockRank	software	http://ailab-projects2.ist.psu.edu/Dock-Rank/	(Xue, Jordan et al. 2014)
	iScore	software	https://github.com/DeepRank/iScore	(Geng, Jung et al. 2019)
Covariation-based prediction of interface contacts	EVcomplex	server	https://evcouplings.org/complex	(Hopf, Scharfe et al. 2014)
	ComplexContact	server	http://raptorx2.uchicago.edu/ComplexContact/	(Zeng, Wang et al. 2018)
Binding motif databases and prediction tools	Eukaryotic Linear Motif resource	database	http://elm.eu.org/	(Kumar, Gouw et al. 2019)
	IUPred2A	server	https://iupred2a.elte.hu/	(Meszaros, Erdos et al. 2018)
	PSSMsearch	server	http://slim.icr.ac.uk/pssmsearch/	(Krystkowiak, Manguy et al. 2018)
Peptide binding site prediction	PEP-SiteFinder	server	https://bioserv.rpbs.univ-paris-diderot.fr/services/PEP-SiteFinder/	(Saladin, Rey et al. 2014)
	InterPep	software	http://wallnerlab.org/InterPep/	(Johansson-Akhe, Mirabello et al. 2019)
Protein-peptide docking servers	Galaxy-PepDock	server	http://galaxy.seoklab.org/pepdock	(Lee, Heo et al. 2015)
	PIPER-FlexPepDock	server	http://piperfpd.furmanlab.cs.huji.ac.il/	(Alam, Goldstein et al. 2017)
	CABS-Dock	server	http://biocomp.chem.uw.edu.pl/CABSdock	(Kurcinski, Jamroz et al. 2015)
	pepATTRACT	server	https://bioserv.rpbs.univ-paris-diderot.fr/services/pepATTRACT/	(de Vries, Rey et al. 2017)
	InterPep2	software	http://wallnerlab.org/InterPep2	(Johansson-Akhe, Mirabello et al. 2020)

B. Supplementary materials for Chapter 2

a. InterEvDock2 pipeline

If the user provided only an input sequence or a query-template alignment for one or both partners, preparatory steps are performed:

Step (i) Template search:

If the user did not provide a template, a profile is built for the query sequence using HHblits (Remmert, Biegert et al. 2011) against the Uniprot20 database and used by HHsearch (Soding 2005) to query the PDB70 database to find a suitable template. If no template with HHsearch probability over 95% is found, then the run is stopped.

HHsearch templates are re-ordered with the following rules: all templates with HHsearch probability equal to the maximum probability are re-ordered by decreasing sequence identity; in case of equal HHsearch probability and sequence identity, HHsearch E-value and template PDB resolution are used as sorting criteria. In case there is an HHsearch match with sequence identity ≥ 70% with the query and which covers at least 50 residues, which is not among the highest HHsearch probability hits due to profile divergence, this match is extracted and set as the first hit among the re-ranked HHsearch matches. All templates with resolution worse than 7 Å are excluded. Only template regions with a DSSP (Kabsch and Sander 1983) assignment are kept for modelling.

The top 20 templates for each query sequence are provided to the user once the run is over. The run is stopped after this step if the user selected a breakpoint after template selection; the user can then choose a template and use the corresponding query-template alignment to restart the run at the modelling step (iii). Otherwise, the first template hit is used for modelling.

Step (ii) Query-template alignment:

If the user provided a template but no query-template alignment, the query sequence is (re)aligned with the template sequence using MAFFT (Katoh and Standley 2013) with the E-INS-i algorithm.

Note: if the user did not provide an input sequence nor a template but provided a query-template alignment (following a template search breakpoint) where the template header starts with ">PDBID_chain:AUTOPDB" (e.g. ">1ki1_D:AUTOPDB") the template PDB coordinates and the input sequence will be automatically retrieved following the information in the alignment.

Step (iii) Modelling:

Once a template and a query-template alignment are available for each partner with no user-provided structure, comparative modelling is performed using a RosettaScripts (Fleishman, Leaver-Fay et al. 2011) protocol based on RosettaCM (Song, DiMaio et al. 2013) to build a 3D model for (at least part of) the input sequence. By default, N-terminal and C-terminal regions of the query protein that are not aligned with the template sequence are not modelled and insertions (loops) longer than 14 residues are not rebuilt. This can be adjusted by the user through 3 tunable parameters: the maximum length of loops to be rebuilt, the maximum length of the N-terminal extension and the maximum length of the C-terminal extension to model.

The RosettaScripts protocol consists in a single "hybridize" mover step where the first 2 (centroid) stages are sped-up by setting options stage1_increase_cycles and stage2_increase_cycles to 0.1 (instead of default value 1.0). This speed-up was specifically introduced for InterEvDock2 and was not present in the protocol for building models in the PPI4DOCK database. It is robust for relatively high homology levels but can lead to loss of precision in the docking results for models built from more remote templates (typically when both templates are below 50% sequence identity with the query).

Once a 3D structure or a structural model is available for each partner, the molecular docking steps are performed:

Step (iv) Sampling:

Exhaustive rigid-body sampling with FRODOCK 2.1 (frodock) (Ramírez-Aportela, López-Blanco et al. 2016).

Step (v) Constraints:

This is an optional step performed only if the user provided information on residues (or pairs of residues) involved in the interface: after checking the user-provided constraints to remove any constraints involving residues not present or buried in the structure/model, apply constraints with FRODOCK 2.1 (frodockonstraints) to filter sampled solutions.

Step (vi) Joint multiple sequence alignments:

If the user did not provide a joint MSA for the two protein partners, a joint MSA is generated automatically by the server. Each query sequence is used as input to a single blastp search against the Uniprot-KB database, with threshold sequence identity > 30%, coverage > 75% and E-value < 10–4. Only one sequence per species is kept (the sequence with the highest sequence identity, and highest coverage if sequence identities are identical). Pairs of sequences belonging to the same species are collected. Redundant paired sequences with sequence identity higher than 90% are removed. The sequences are re-aligned by MAFFT. In the end, a set of two MSAs containing exactly the same number of sequences in the same species order. When fewer than 10 sequences are retrieved, a warning message in the server progress log indicates that models selected by InterEvScore may be less reliable. In case more refinement is needed in the construction of the joint MSAs, users may use the InterEvolAlign server (Faure, Andreani et al. 2012).

Step (vii) Clustering and scoring:

All decoys (or only decoys remaining after filtering if constraints were provided) are clustered by FRODOCK 2.1 (frodockcluster) at ligandRMSD 4.0 Å. The best 10 000 FRODOCK2 cluster representatives are rescored using InterEvScore (Andreani, Faure et al. 2013) and SOAP-PP (Dong, Fan et al. 2013).

Step (viii) Consensus calculation:

The consensus calculation returns a list of 10 models accounting for the fact that decoys well ranked by at least two different scoring methods have higher chances of being correct. The 3*top 10 models for each score (FRODOCK2.1, InterEvScore, SOAP-PP) are re-ranked according to the number of similar decoys (defined as ligand RMSD ≤ 10 Å) within the top 50 models of the other two scores (down to a minimum of two similar decoys). In case of a tie, priority is given to InterEvScore top 10 models, then SOAP-PP, then FRODOCK. If necessary, the consensus list is then filled up to ten models by selecting the best models from each score (4 from InterEvScore, 3 from SOAP-PP and 3 from FRODOCK). When building the consensus, models that are structurally redundant (i.e. minimum ligand RMSD of 10 Å) with previously selected models are excluded, so that the final list contains 10 structurally non-redundant models.

The top 5 residues for each partner (ranked starting with the residue most likely to be at the interface) are chosen as the five most frequently occurring residues at the interface of the top 10 models of each score (FRODOCK2.1, InterEvScore, SOAP-PP). In case of a tie, priority is given to residues with a higher frequency in the top 10 models of only InterEvScore, then SOAP-PP, then FRODOCK. A residue is considered at the interface if any of its non-hydrogen atoms is within a 5 Å radius of any non-hydrogen atom on the opposite partner protein, as defined in CAPRI (Mendez, Leplae et al. 2003).

b. PPI4DOCK benchmark (Yu and Guerois 2016)

The list of the 812 complexes used for benchmarking, as well as results of the InterEvDock2 pipeline, is provided in http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/table.html. The full benchmark data (including for each target models for the two unbound partners, co-alignments for both protein partners and the reference protein complex for evaluation) can be downloaded from http://biodev.cea.fr/interevol/ppi4dock/PPI4DOCK.zip.

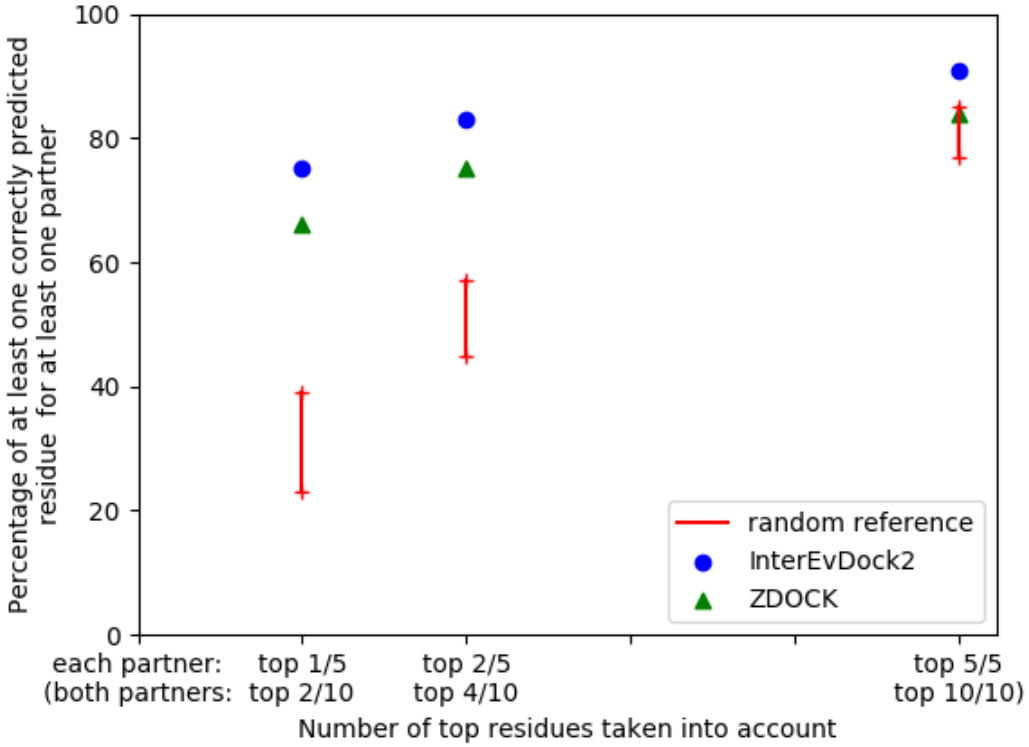
c. Success rates for interface residue predictions

The number and percentage of cases for which at least one residue out of 10 could be predicted correctly as present in the complex interface is assessed. Contacts are defined as in CAPRI (Mendez, Leplae et al. 2003), i.e. two residues are assumed to be in contact if any non-hydrogen atom in the first residue is within 5 Å of any atom in the second residue.

For Zdock3.0.2 (Pierce, Hourai et al. 2011) interface predictions, we take the 5 residues on each partner (10 predicted residues in total) that occur most often in the interfaces of the top 10 Zdock3.0.2 models. In case of a tie, we draw at random among residues having the same frequency of occurrence. For instance, if 2 receptor residues occur in all 10 interfaces, those are selected to be among the top 5 predicted residues; if the following 6 residues occur in 9 out of 10 interfaces, we draw 3 at random among those 6 to obtain a total of 5 predicted residues. This type of ties occurs relatively frequently for Zdock3.0.2 predictions (contrary to InterEvDock2 predictions, where residues are discriminated better thanks to the consensus approach and the fact that InterEvScore predictions are prioritised over SOAP_PP and FRODOCK predictions). Therefore, for Zdock3.0.2 predictions, we repeat the procedure 1000 times and report the average success rate, which provides a robust way to evaluate those predictions.

We also calculate a random reference for interface residue predictions. For this purpose, we randomly draw residues from the surface of each partner and assess whether these residues are located at the complex interface. Surface residues are defined as residues with at least 5% relative accessible solvent area as in (Pierce, Wiehe et al. 2014). This procedure is repeated 10,000 times when drawing 1, 2 or 5 residues per partner (i.e. 2, 4 or 10 residues in total). 99.9% confidence intervals (shown as red bars in Figure B-1) are extracted from the sorted list of success rates as the 5th and 9995th values.

A



B

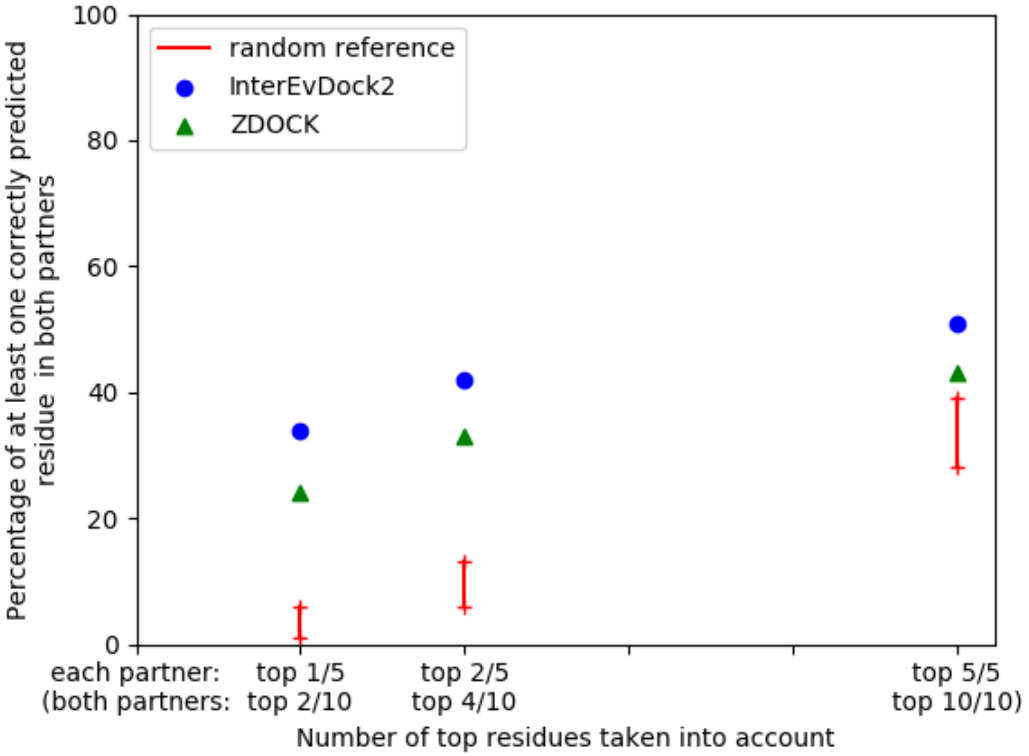


Figure B-1: Illustration of the residue prediction success rate of InterEvDock2, ZDOCK3.0.2 and a random reference. *Success rate (expressed as a percentage) for the interface residue prediction when taking 1, 2 or 5 predicted residues in each partner (top 2, top 4 or top 10 predicted residues in total) for InterEvDock2 and ZDOCK3.0.2 compared to a random reference. The plots show the results for when at least one (A) or both (B)*

of the two protein partners have at least one correct residue in the top 1, 2 and 5 of each partner (i.e. at least one correct residue in the top 2, 4 and 10 total predicted residues). In other words, (A) corresponds to the predictions marked as "≥1 correct in top x receptor OR top x ligand" (x = 1, 2, 5) in Table 2-1 and the (B) corresponds to the predictions marked as "≥1 correct in top x receptor AND top x ligand" (x = 1, 2, 5). As seen in Table 2-1, InterEvDock2 has a slightly higher number of correctly predicted residues than ZDOCK3.0.2 and both scores are clearly above the random reference, especially when looking at only the top 1 predicted residue per partner (top 2 predicted residues in total).

d. Default constraint thresholds

The default distance thresholds for single constraints and constraint pairs were set to 8 Å and 11 Å respectively. In order to determine the ideal default distance for a single constraint, the interface residues in the native interface (residues containing a non-hydrogen atom at a distance of less than 5 Å from any non-hydrogen atom of the opposite chain) of 812 cases used to benchmark InterEvDock2 were identified. We then calculated the minimum distance of these residues to the opposite chain in all decoys of acceptable or better quality and in an equivalent number of incorrect decoys (considering separately incorrect decoys with a fraction of native interface residues above or below 10%) within the top 10,000 decoys ranked by FRODOCK2.1 after the FRODOCK2.1 clustering step. The default distance of 8 Å was chosen so as to recover at least 80% of the acceptable or better decoys with at least 80% of the native interface residues having a minimum distance to the opposite chain under this threshold. As a comparison, we found that only 45% and 8% of the incorrect decoys with a fraction of native interface residues above or below 10% respectively were retained under these same conditions (55% and 92%, respectively, were filtered out).

The optimal default distance of 11 Å for constraint pairs was determined using the same reasoning except that we used the minimum distance between the residue pairs present at the real interface (two residues on opposite chains are considered a pair when any non-hydrogen atom of the first residue is within 5 Å of any non-hydrogen atom of the second residue) and observed their distribution in the three different types of decoys. The default distance of 11 Å was chosen so that at least 80% of the acceptable or better decoys had at least 80% of the native interface residue pairs within this distance threshold. As a comparison, we found that only 23% and 1% of the incorrect decoys with a fraction of native interface

residues above or below 10% respectively were retained under these same conditions (77% and 99%, respectively, were filtered out).

e. Performance according to sequence identity with PPI4DOCK template

Table B-2: InterEvDock2 performance according to target-template sequence identity in PPI4DOCK. *This table summarises the prediction performance of the InterEvDock2 consensus on the 812 PPI4DOCK cases, split by sequence identity between the target and the template used to model the unbound structures in the benchmark (the smaller of the target-template identities for the two protein partners is used). It shows only a moderate drop in success rate for models built with remote templates (< 30% sequence identity) and an increase in success rate for models built with very close templates (>=95% sequence identity).*

	All cases	0-30% id	30-60% id	60-95% id	95-100% id
Number of cases	812	227	239	154	192
Top 10 success rate	239 (29%)	57 (25.1%)	70 (29.3%)	47 (30.5%)	65 (33.9%)

f. Performance comparison with the Weng benchmark

All cases from the Weng benchmark presented in Table B-3 and Table B-4 were benchmarked using input structures and co-alignments as in the original InterEvDock paper (Yu, Vavrusa et al. 2016). All input files are provided in http://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock2/table-weng.html.

Table B-3: InterEvDock2 performance on 47 cases in common between PPI4DOCK and Weng benchmarks. *(A) Table summarising the prediction performance of the 3 scoring components of InterEvDock2 (InterEvScore, SOAP-PP and FRODOCK 2.1) and the InterEvDock2 consensus itself (this work) on 47 cases that are in common between the PPI4DOCK and the Weng benchmark (Hwang, Vreven et al. 2010). Note that the reference complexes are not exactly the same since a different representative might be chosen in PPI4DOCK compared to the Weng benchmark cases. PPI4DOCK also uses homology models for docking while the Weng benchmark uses X-ray structures. Those are however cases where homologs with a known unbound structure exist with very high sequence identity to the query sequences. The highest success rates for each category are highlighted in bold. (B) 95% confidence intervals were calculated for results in (A) with a bootstrap analysis over 10,000 iterations where the performance was repeatedly calculated on a random set of 47 cases chosen from the original list of 47 (drawing with replacement). All performance values were ordered from smallest to largest and the 250th and the 9750th values correspond to the lowest and highest values of the 95% confidence intervals. This analysis shows that results in table (A) are variable due to the small number of cases.*

A

			PPI4DOCK category				Weng benchmark difficulty		
		All	**Very-easy**	**Easy**	**Hard**	**Very-hard**	**Rigid-body**	**Medium**	**Difficult**
Number of cases		47	11	35	0	1	27	12	8
Top 10 success rate (PPI4DOCK)	**InterEvScore**	14	**4**	10	0	0	10	2	**2**
	SOAP-PP	11	2	9	0	0	7	4	0
	Frodock v2.1	8	1	7	0	0	4	3	1
	InterEvDock2 Consensus	**19**	**4**	**15**	0	0	**12**	**5**	**2**
Top 10 success rate on the Weng benchmark	**InterEvScore**	11	4	7	0	0	9	2	0
	SOAP-PP	**15**	**6**	**9**	0	0	**12**	**3**	0
	Frodock v2.1	7	2	5	0	0	6	1	0
	InterEvDock2 Consensus	13	4	**9**	0	0	10	**3**	0

B

			PPI4DOCK category				Weng benchmark difficulty		
		All	Very-easy	Easy	Hard	Very-hard	Rigid-body	Medium	Difficult
Number of cases		47	11	35	0	1	27	12	8
Top 10 success rate (PPI4DOCK)	**InterEvScore**	[8, 20]	[1, 8]	[5, 16]	[0, 0]	[0, 0]	[5, 16]	[0, 5]	[0, 5]
	SOAP-PP	[5, 17]	[0, 5]	[4, 14]	[0, 0]	[0, 0]	[3, 12]	[1, 8]	[0, 0]
	Frodock v2.1	[3, 13]	[0, 3]	[3, 12]	[0, 0]	[0, 0]	[1, 8]	[0, 7]	[0, 3]
	InterEvDock2 Consensus	[13, 26]	[1, 8]	[9, 21]	[0, 0]	[0, 0]	[6, 18]	[1, 10]	[0, 5]
Top 10 success rate on previous paper's benchmark (Weng)	**InterEvScore**	[6, 17]	[1, 8]	[3, 12]	[0, 0]	[0, 0]	[4, 14]	[0, 5]	[0, 0]
	SOAP-PP	[9, 21]	[2, 11]	[4, 15]	[0, 0]	[0, 0]	[6, 18]	[0, 7]	[0, 0]
	Frodock v2.1	[3, 12]	[0, 5]	[1, 9]	[0, 0]	[0, 0]	[2, 11]	[0, 3]	[0, 0]
	InterEvDock2 Consensus	[7, 19]	[1, 8]	[4, 15]	[0, 0]	[0, 0]	[5, 16]	[0, 7]	[0, 0]

Table B-4: InterEvDock2 performance on 85 cases from the Weng benchmark. *For comparison purposes, the prediction performance using the new pipeline on the same 85 cases from the Weng benchmark as the previous InterEvDock paper (Yu, Vavrusa et al. 2016) are reported below for individual scoring components of InterEvDock2 (InterEvScore, SOAP-PP and FRODOCK 2.1) and the InterEvDock2 consensus (this work) as well*

as performances from SwarmDock and Zdock3.0.2 taken as is from the previous InterEvDock paper (Yu, Vavrusa et al. 2016).

	All	**Rigid-body**	**Medium**	**Difficult**
	85	43	23	19
InterEvScore	20 (24%)	14 (33%)	4 (17%)	2 (11%)
SOAP-PP	22 (26%)	17 (40%)	4 (17%)	1 (5%)
Frodock v2.1	20 (24%)	14 (33%)	5 (22%)	1 (5%)
InterEvDock2 Consensus	27 (32%)	19 (44%)	7 (30%)	1 (5%)
SwarmDock server 2013	25 (29%)	18 (42%)	6 (26%)	1 (5%)
Zdock 3.0.2	17 (20%)	12 (28%)	3 (13%)	2 (11%)

C. Supplementary materials for Chapter 3

a. Supplementary methods

1. Docking parameters

In the docking pipeline based on FRODOCK2.1, all parameters were set to default except for the following. Docking with the frodock executable used the "-O" option for "other" complexes (not enzyme and not antibody-antigen). Clustering with frodockcluster was run with the –d 4 option, i.e. setting a LRMSD threshold of 4 Å for clustering.

2. Scoring functions

We employed an in house implementation of SOAP-PP that enables much more efficient scoring since decoy coordinates do not need to be explicitly generated. Note that only a slight reduction in performance on the 752 benchmark cases compared to the original SOAP-PP implementation has been observed (supplementary Table C-15).

We also re-implemented InterEvScore for efficiency reasons. We introduced two variations compared to the best original InterEvScore (Andreani, Faure et al. 2013): we defined interface contacts through distance thresholds, instead of tessellation ("distance mode") and we took evolutionary information into account for all interface residues instead of apolar patches only (so-called "standard mode" in the original implementation). InterEvScore outputs several scoring variants; here, we used the $2/3\mathrm{B}_{\mathrm{evol}}^{\mathrm{best}}$ and the $2\mathrm{B}^{\mathrm{best}}$ (Andreani, Faure et al. 2013). In $2/3\mathrm{B}_{\mathrm{evol}}^{\mathrm{best}}$, each interface residue contributes to the final score through the potential of its best 2- or 3-body contact and the potential of its equivalents in the homolog sequences. $2/3\mathrm{B}_{\mathrm{evol}}^{\mathrm{best}}$ was found to perform best when scoring with homolog sequences (InterEvScore with implicit homology) (Andreani, Faure et al. 2013) and thus was used in this context. $2\mathrm{B}^{\mathrm{best}}$ was used when scoring explicitly modelled side-chain models of our homologs (InterEvScore with explicit homology, IES-h). Indeed, we found that 3-body

potentials are less discriminative than 2-body potentials in the context of explicitly modelled decoys (supplementary Table C-16).

We use Rosetta 3.8 (version 2017.08.59291) and the beta_nov15 Rosetta score. Before scoring with Rosetta ISC, we perform high-resolution interface side-chain optimisation by using 'use_input_sc' and 'docking_local_refine' options of Rosetta's docking_protocol executable. We also tried adding the 'dock_min' option (for even more conservative modelling and shorter scoring runtimes) but scoring results were degraded.

3. Details on coMSA calculation

Compared to the original PPI4DOCK database (Yu and Guerois 2016), coMSAs were slightly adjusted by realigning the first sequence (query) with all other sequences (considered as a block) using MAFFT (Katoh and Standley 2013).

When building reduced coMSA^{40} from the readjusted PPI4DOCK coMSAs, coMSAs that already had under 40 sequences before the hhfilter step were not filtered.

The 10 sequences in coMSA^{10} were selected from coMSA^{40} as follows: Euclidian division was performed of the number of sequences in the coMSAs^{40} (including the query) over 10 with q and r, the quotient and remainder of this division. Starting from the first sequence, the next sequence is selected every q+1 for the first r steps, then every q until the end, including the last sequence resulting in 11 sequences with the first being the query and other 10, the homolog sequences.

4. Threading models

The PPI4DOCK benchmark contains docking targets based on unbound homology models of pairs of binding partners for which an experimental complex structure is available. The use of homology modelling for unbound partners enables to expand the benchmark, by alleviating the need to identify complexes for which experimental structures of the interface and the exact two binding partners have been solved. This makes the benchmark larger, but

as a counterpart, in PPI4DOCK the unbound structures used for docking are themselves homology models.

In a docking context where we know the structures of the unbound partners, we would build homology models for all sequences in the coMSA by using the two query structures as modelling templates. However, since in PPI4DOCK the unbound query structures are themselves homology models, this would mean building a model by using a homology model as a template, and we felt this succession of modelling steps would lead to a loss in model precision. Therefore, the templates used for threading coMSA sequences were the unbound templates used to build the PPI4DOCK unbound models.

Template protein sequences were directly extracted from their structures and aligned onto the coMSAs using MAFFT (sequence-profile alignment) (Katoh and Standley 2013) from which the pairwise homolog-template alignments were directly extracted. coMSAs were stripped down to positions that were covered by the query sequence. In order to ensure that the template structure exactly matched the template sequences in the stripped pairwise alignments, both template sequences were re-aligned using clustalw (Larkin, Blackshields et al. 2007) and identified irrelevant residues in the template structure were removed.

Threading implies that the side-chains of our homologs are mapped very conservatively onto the query template structure.

b. Supplementary results

1. Supplementary tables

Table C-5: List of the 752 docking cases used as a benchmark set in this study. *This subset of the 1417 cases in PPI4DOCK contains all cases with at least 10 sequences in the coMSAs and at least one acceptable decoy in the top 10,000 FRODOCK2.1 decoys.*

1a2y_AB	1azs_FD	1c1y_AB	1dkf_AB	1em8_AB	1flt_BC	1g3n_AB	1gl4_AB	1hx1_AB
1a4y_CD	1b4u_AD	1c4z_AD	1dl7_AB	1euv_AB	1fm0_AB	1g3n_AC	1gla_DH	1hyr_BC
1a9n_CB	1b6c_AB	1cg5_BC	1dlf_AB	1ewy_AB	1fo0_ED	1g8k_AB	1got_AB	1i1q_BD
1agr_AB	1blx_AB	1cgi_AB	1dvf_BD	1ezv_TS	1fq1_AB	1gaq_AB	1gpw_AB	1i2m_AB
1aro_AB	1bqh_AE	1cmx_AB	1e50_AB	1f45_AB	1fqj_AB	1gcq_AC	1gxd_AB	1i4d_AC
1ava_AB	1bqq_AB	1co7_AB	1e96_AB	1f6f_AB	1fr2_AB	1gcq_BC	1h1v_AB	1i85_BD
1awc_CD	1buh_AB	1d4v_BF	1eaw_AB	1f6m_AC	1fvu_CB	1gcv_CB	1hcf_BC	1i8k_AB
1axi_BD	1bzx_AB	1de4_AC	1ebd_BC	1fle_AB	1fx0_CD	1ggp_AB	1he8_AB	1i8l_AB

1iar_AB
1ib1_BD
1ikn_AB
1ikn_CB
1iod_CB
1ixs_AB
1j05_AB
1j2j_AB
1j7d_AB
1jb0_AE
1jb0_CE
1jk0_AB
1jql_AB
1jr3_CD
1jtd_AB
1jwy_AB
1jzd_BC
1k5d_AC
1k9o_AB
1ka9_AB
1kb5_AB
1kcg_AC
1kgy_AC
1ki1_AB
1ksh_AB
1ktz_BD
1kxq_AB
1kz7_AB
1l0o_AC
1l9b_BD
1lb1_AB
1m2o_CD
1m2t_AB
1m2v_AB
1ma9_AB
1mbx_AB
1mfa_AB
1mox_BD
1mqk_AB
1n4x_AB
1nb5_AC
1nbf_AB
1npe_AB
1nql_AB
1nvv_AC
1nvv_BC
1oaq_AB
1oc0_AB
1oey_AB
1of5_AB
1ofu_AB
1oph_AB
1out_BC
1p2j_AB
1p4l_BH
1p4l_CG
1p8v_AF
1pk1_AB
1ppf_AB
1pvh_AB
1q5q_GN
1qa9_AB
1qdl_BD
1qo3_CB
1qop_BD
1r0r_AB
1r8s_AB
1rbl_AH
1rjc_AB
1rv6_BC
1s1q_AB
1sg1_AC
1sg1_BC
1shw_AB
1shy_AB
1spg_BC
1spp_AB
1sq0_AB
1stf_AB
1sv0_AB
1t0p_AB
1t8o_AB
1ta3_AB
1taw_BD
1tco_AC
1tdq_AB
1te1_AB
1tfx_AB
1tgs_AB
1tgz_AB
1to2_AB
1tue_AB
1tx4_AB
1u0s_AB
1u2g_BC
1u75_AB
1u7f_AB
1uac_AB
1uad_AB
1uea_AB
1uex_CB
1us7_AB
1usu_AB
1uw4_AB
1uzx_AB
1v4x_AD
1v7p_AC
1vg0_AB
1w98_AB
1wdw_AB
1wmh_AB
1wmu_BC
1wpx_AB
1wq1_AB
1wqj_AB
1wr6_AB
1wrd_AB
1wt5_BD
1x75_BD
1x86_AB
1x9f_EF
1xcg_AB
1xd3_AB
1xg2_AB
1xqs_AB
1y8x_AB
1yc0_AB
1ycs_AB
1yvb_AB
1z3e_AB
1z5x_AB
1z5y_AB
1z7k_AB
1z7m_BG
1z7x_AB
1zc3_AB
1ze3_AB
1zhh_AB
1zjd_AB
1zr0_AB
2a19_AB
2a1j_AB
2a40_AB
2a5d_AB
2a9m_AB
2ast_CB
2atp_AC
2aw2_AB
2b4s_CD
2b5i_AC
2ba0_CH
2bcg_AB
2bcj_AD
2bcn_AB
2bex_AB
2bkk_AB
2bkr_AB
2bky_AC
2blf_AB
2bo9_AB
2bto_BH
2btq_BD
2c2v_AE
2c5l_AB
2cch_AB
2cg5_AB
2cjs_BC
2ckh_AB
2czv_BD
2d5r_AB
2d7t_AB
2de6_AD
2dsq_CB
2dzn_AB
2e27_AB
2e2d_AB
2e3x_AB
2efe_AB
2ejf_AB
2eke_AB
2ey4_AB
2f5z_BC
2f8x_CD
2fd6_AD
2fep_BD
2fju_AB
2fnj_CB
2fu5_AB
2g45_AB
2goo_DF
2gtp_AB
2gwf_AB
2gzd_AC
2h62_AD
2h62_BC
2hle_AB
2hrk_AB
2htm_AC
2hue_AB
2hy5_BC
2hy5_FC
2ibg_AB
2ie4_AB
2ih3_DL
2ihb_AB
2inc_BF
2io0_AB
2io5_AB
2iy0_AC
2iy1_AB
2j0s_AB
2j0t_AB
2j3t_AC
2j59_AB
2jb0_DH
2jdi_AD
2jdi_GH
2jgz_AB
2ngr_AB
2nps_AB
2npt_AB
2nqd_AB
2nxx_AB
2nz8_AB
2o25_AB
2o26_BD
2o2v_AB
2o8v_BD
2ocf_BD
2ode_AB
2oi9_CB
2omz_AB
2otu_AB
2oul_AB
2oxg_AB
2oxq_BD
2oz4_AB
2ozb_CB
2p45_AB
2pbd_AB
2pop_CD
2ptt_AB
2pu9_AC
2puk_AB
2pvg_AC
2q5w_BD
2qe7_AD
2qho_AB
2qi9_AE
2qi9_BE
2qkl_AB
2qwo_AB
2r0l_CB
2r25_AB
2r40_AB
2rex_AB
2sic_BD
2uyz_AB
2v1y_AB
2v3b_AB
2v4z_AB
2v5q_AB
2v7q_BE
2v8s_AB
2vje_BD
2vol_BD
2vrw_AB
2vso_AB
2vut_AB
2vxs_FA
2w19_DH
2w83_DB
2wbl_AC
2wdt_AB
2wiu_AB
2wnv_AB
2wnv_AC
2wnv_BC
2wo2_AB
2wo3_AB
2wp8_AC
2wqa_DE
2ws9_32
2wus_AB
2x5i_CB
2xac_BC
2xbb_AB
2xko_BD
2xqr_AB
2xwu_AB
2yc2_AB
2yho_AB
2ynm_DF
2yvj_AB
2z0d_AB
2z35_AB
2z3q_AB
2z5c_AC
2z7f_AB
2za4_AB
3a33_BC
3a4u_AB
3a6p_AC
3a7a_AB
3a8k_AB
3a8y_AB
3ab0_CB
3agj_AB
3aji_AB
3alq_BF
3amj_AD
3bbp_AD
3bdw_AB
3bh7_AB
3bik_AB
3bp6_AB
3bp8_AC
3bpl_AC
3bs5_AB
3bt2_BE
3buk_AC
3bwu_AB
3bx1_AB
3bx7_BD
3by4_AB
3c5w_CB
3cbj_AB
3cji_CB
3cki_AB
3cph_AB
3cpj_AB
3cx8_AB
3d1k_BD
3d2f_AB
3d2u_CB
3d3b_AB
3d65_AB
3d7t_AB
3daw_AB
3dbh_CB
3dge_BC
3dlq_AB
3dur_AB
3dwg_AC
3e1z_AB
3ejb_AB
3eno_AB
3er9_AB
3evs_BC
3f1p_AB
3f1s_AB
3f5c_AB
3f5c_AC
3f7p_AB
3f9k_BC
3fap_AB
3fc6_AB
3ff7_BD
3ff8_AC
3fga_AB
3fmo_AB
3fn1_AB
3fpn_AB
3g33_CD
3g3a_AB
3g9v_AB
3gjx_BC
3gni_AB
3gpr_AC
3gqb_AB
3gqi_AB
3gym_AB
3h11_AB
3h2u_AB
3h9r_AB
3hax_DC
3hct_AB
3hei_AB
3hh2_AB
3hhs_AB
3icq_AC
3ifw_AB
3ima_AB
3imz_CD
3jv4_AB
3jv6_AB
3jw0_AB
3jw0_CB
3k1i_AB
3k2m_AB
3k51_BF
3k9m_AB
3k9o_AB
3kb3_AB
3kbt_AB
3kdj_AB
3kfd_AF
3kld_AB
3kmu_AB
3knb_AB
3ks0_AC
3kse_AB
3kud_AB
3kyc_CB
3kyj_AB
3l1z_AB
3lb8_AB
3lbx_AB
3ldq_AB
3lpe_AB
3lqc_AB
3ltf_CD
3lvj_BD
3lvl_BD
3m0a_CD
3m0d_DC
3m18_AB
3m7f_AB
3m7q_AB
3mca_AB
3mdy_AB
3mhv_BD
3mi9_AB
3mkb_CB
3msx_AB
3n1f_AB
3n3a_BD
3n3k_AB
3n5b_CD
3n9y_AB
3nig_AC
3nmv_AB
3ny7_AB
3o2p_AB
3of6_BD
3oky_BD
3or1_CE
3p5t_AD
3p71_AB
3pb1_AB
3pv6_AB
3q3j_DH
3q66_BA
3q9n_AB
3qb4_AB
3qb7_AB
3qht_AB
3qn1_AB
3qq8_AB
3qt2_AC
3qvg_AB
3qwq_AB
3qwr_AC
3r07_AB
3r1g_AB
3r2c_AB
3rpf_BD
3t62_AB
3tg1_AB
3tjz_AB
3tmp_AB
3tx7_AB
3u7u_AB
3uai_AB
3udw_AB
3uir_AB
3ulq_AB
3ulr_AB
3uou_AB
3v2a_BC
3v2a_BD
3v64_AC
3vmf_AB
3von_AC
3vpb_AF
3vr4_CB
3vti_BD
3vyt_BD
3wxw_CB
3ygs_AB
3zdm_EF
3zhp_AB
3zl7_AB
3zo0_AC
3zu7_AB
43c9_AB
4a49_AB
4a63_AB
4a8x_AC
4ag1_AB
4auq_FE
4b8a_AB
4bfi_AB
4bgd_AB
4bi8_AB
4bmo_BD
4bnr_AB
4bos_AC
4bos_AD
4bsr_AD
4bv4_AC
4bvx_AB
4c4k_BA
4c9r_CD
4ccg_BA
4cdk_AB
4crw_AB
4ct4_AB
4cxa_AB
4cym_AD
4cym_BD
4czx_BD
4d0k_AB
4d0l_AB
4d0n_AB
4dcn_AB
4dfc_AB
4dhi_AB
4djd_BF
4doh_AB
4doh_AC
4doh_CB
4dri_AB
4ds8_AB
4dss_BC
4dxe_BD
4e4d_CE
4eb5_AD
4ekd_AB
4emj_AB
4es4_BD
4etw_AB
4ext_AC
4ezm_BD
4ffb_CB
4ffy_BC
4fjv_AB
4fou_AB
4fq0_AB
4fqx_AC
4ged_AB
4gh7_AB
4gmj_AB
4goj_AB
4gok_AB
4grw_DB
4grw_EA
4gs7_AC
4gs7_AD
4gsl_AD
4h2w_AD
4h3k_AB
4h5s_AB
4hdo_AB
4hgm_BA
4hr6_CB
4hr7_AB
4hrl_AB
4hrn_DC
4i18_AC
4i2l_CD
4i2l_CF
4i5l_AB
4i6l_AB
4i6m_AB
4i6n_AB
4ii2_AB

4ij3_AB	4jgh_CD	4krp_AD	4lnu_CB	4mng_CB	4nqa_AB	4qtt_AB	4u5y_AB	4xh9_AB
4ij3_AC	4jhp_AB	4ksk_AB	4lry_AC	4ms4_AB	4ocm_CB	4qxf_AB	4u65_AC	4xl1_AB
4ilh_AB	4jqw_AB	4kt0_CE	4lw4_AC	4msv_CF	4oic_AB	4rca_AB	4u65_BC	4y8d_AB
4ilw_AB	4jx1_AB	4kt1_AB	4lx0_AB	4n0g_AB	4p1b_FD	4rku_NG	4ui0_AC	4ydy_AB
4imi_AB	4k1r_AB	4kvg_AB	4lxr_AB	4n3y_AC	4p2a_AB	4rr2_AB	4ut7_AB	4yfc_AB
4iop_AB	4k5a_AB	4l0p_AB	4m4r_AB	4n6e_BD	4p5o_BD	4rsu_IJ	4ut9_CB	4yii_AB
4iso_AB	4k71_AB	4l41_AB	4m69_AB	4n6o_AB	4p78_AD	4tu3_AB	4v3l_AD	4yn0_AB
4iyp_AB	4k81_AB	4l41_CB	4mcx_AC	4naw_AB	4pbv_AB	4tvs_AB	4v3l_DB	4ypg_CA
4j4l_AB	4kax_AB	4lcd_AC	4mdk_AB	4ni2_AB	4per_AB	4tx3_AB	4wlr_AC	5aie_AB
4jd2_FH	4kgq_HJ	4ldt_AB	4mjs_AB	4nif_AB	4pky_AB	4txo_AB	4wqo_CD	
4jd2_GH	4kml_AB	4ldt_CA	4mmz_CB	4nik_AB	4qci_AC	4txv_AB	4ww7_AB	
4je4_AB	4kng_AC	4lhu_AC	4mn4_DC	4nkg_AB	4qt8_AB	4u30_AB	4x0l_AC	
4jeg_AB	4kng_EC	4lld_AB	4mn8_AC	4nl9_AB	4qts_AB	4u32_AB	4x0l_CB	

Table C-6: InterEvScore statistical potential. *The IES^{query} score represents only the statistical potential part of InterEvScore ($2B^{best}$) without any evolutionary information, used to re-rank either the top 10,000 (10k) or the top 1,000 (1k) FRODOCK2.1 decoys. These results are shown for comparison with the homology-enriched IES-h variants described in the main results.*

	Top 10 success rate	**Top 50 success rate**
IES^{query}/10k	154 (20.5%)	284 (37.8%)
IES^{query}/1k	165 (21.9%)	297 (39.5%)
IES-h^{40}/10k	**203 (27.0%)**	335 (44.5%)
IES-h^{10}/1k	200 (26.6%)	**338 (44.9%)**

Table C-7: Scoring performance of homology-enriched SCORES. *Scoring performance of ISC on query decoys only and using the threaded homology models (ISC-h^{10}) on top 1,000 FRODOCK2.1 decoys (1k) and $coMSA^{10}$ as well as the performance of SPP-h^{40} and IES-h^{40} on top 10,000 (10k) with $coMSAs^{40}$ and the performance of SPP-h^{10} and IES-h^{10} on 1,000 FRODOCK2.1 decoys with $coMSAs^{10}$ for easier comparison. Performances were measured as the top 10 success rate on 752 benchmark cases. This table is the same as Table 3-4 except that it includes $coMSA^{40}$/10k success rates for comparison purposes.*

	Top 10 success rate		**Top 50 success rate**	
	$coMSA^{40}$/10k	$coMSA^{10}$/1k	$coMSA^{40}$/10k	$coMSA^{10}$/1k
IES-h	203 (27.0%)	200 (26.6%)	335 (44.5%)	338 (44.9%)
SPP-h	228 (30.3%)	227 (30.2%)	**365 (48.5%)**	**362 (48.1%)**
ISC	/	157 (20.9%)	/	301 (40.0%)
ISC-h	/	**259 (34.4%)**	/	**361 (48.0%)**

Table C-8: Numbers and timescales (on one CPU) of various elements and programmes. *Times and numbers correspond to measurements on our 752-case PPI4DOCK benchmark. Decoys and docking mentioned below all refer to FRODOCK2.1 docking. The number of decoys generated per case changes according to the size of the complex, it averages at 9,651 with a maximum threshold of 10,000. Docking and decoy generation times are size-dependent but an average value is shown below.*

Number of cases in our benchmark	**752**
Average number of sequences in our coMSAs	134
Average number of residues per case (receptor + ligand)	389
Maximum number of decoys generated in docking	10,000
Average number of decoys per case	9,651
Docking time with FRODOCK2.1	45 min - 1 h
Structure generation time for 1,000 decoys with FRODOCK2.1	1 min
Threading time with Rosetta per structure	1-2 min
SOAP-PP scoring time for 1,000 decoys	1 min
Original SOAP-PP scoring time for 1,000 decoys	15 min
InterEvScore scoring time for 1,000 decoys	1 min
Rosetta's ISC scoring time for 1,000 decoys	12 h 30
Consensus calculation time per case	20 s (3 scores) – 20 min (5 scores)

Table C-9: Top 1 and top 5 compared to top 10 success rates for consensus scores.

	Top 1 success rate	Top 5 success rate	Top 10 success rate
Cons3	95 (12.6%)	190 (25.3%)	241 (32.0%)
Cons3-h	**113 (15.0%)**	228 (30.3%)	271 (36.0%)
Cons4-h/150h	104 (13.8%)	223 (29.7%)	276 (36.7%)
Cons4-h/1k	111 (14.8%)	230 (30.6%)	282 (37.5%)
Cons5-h/150h	109 (14.5%)	230 (30.6%)	289 (38.4%)
Cons5-h/1k	**113 (15.0%)**	**247 (32.8%)**	**304 (40.4%)**

Table C-10: Performance of the repulsive term in Rosetta's score and ISC-h^{10}/1k on the worst third or worst homologs *Top 10 success rate of the fa_rep van der Waals repulsive terme in Rosetta's scoring without (fa_rep /1k) and with homology through threaded homologs (fa_rep-h^{10}/1k) as well as ISC-h^{10}/1k using only the worst scoring third of homologs selected for each decoy individually (ISC-$h^{10/w3}$/1k) or the worst scoring homolog for each decoy (ISC-$h^{10/w1}$/1k) over 752 cases.*

	Top 10 success rate
fa_rep/1k	9 (1.2%)
fa_rep-h^{10}/1k	34 (4.5%)
ISC/1k	157 (20.9%)
ISC-h^{10}/1k	**259 (34.4%)**
ISC-$h^{10/w3}$/1k	227 (30.2%)
ISC-$h^{10/w1}$/1k	200 (26.6%)
SPP/10k	183 (24.3%)
SPP-h^{40}/10k	228 (30.3%)
SPP-$h^{40/w3}$/10k	207 (27.5%)
SPP-$h^{40/w1}$/10k	188 (25.0%)

Table C-11: Performance over PPI4DOCK difficulty categories. *Top 10 success rates separated over the four difficulty categories in our benchmark for FRODOCK2.1, InterEvScore and its threaded-homology variants, SOAP-PP and ISC and their evolutionary variants and the six consensus scores presented in section 3.2.6. Performances were measured on 752 benchmark cases.*

		total	very_easy	easy	hard	very_hard
		752	169	473	94	16
Individual scores	**FD2.1**	164 (21.8%)	55 (32.5%)	102 (21.6%)	5 (5.3%)	2 (12.5%)
	IES / 10k	182 (24.2%)	55 (32.5%)	118 (24.9%)	8 (8.5%)	1 (6.2%)
	IES^{40} / 10k	179 (23.8%)	52 (30.8%)	118 (24.9%)	8 (8.5%)	1 (6.2%)
	IES-h^{40} / 10k	203 (27.0%)	52 (30.8%)	141 (29.8%)	10 (10.6%)	0 (0.0%)
	IES-h^{10} / 1k	200 (26.6%)	56 (33.1%)	133 (28.1%)	10 (10.6%)	1 (6.2%)
	SPP / 10k	183 (24.3%)	52 (30.8%)	120 (25.4%)	11 (11.7%)	0 (0.0%)
	SPP-h^{40} / 10k	228 (30.3%)	65 (38.5%)	146 (30.9%)	15 (16.0%)	2 (12.5%)
	SPP-h^{10} / 1k	227 (30.2%)	65 (38.5%)	146 (30.9%)	16 (17.0%)	0 (0.0%)
	ISC / 1k	157 (20.9%)	52 (30.8%)	99 (20.9%)	6 (6.4%)	0 (0.0%)
	ISC-h^{10} / 1k	259 (34.4%)	86 (50.9%)	158 (33.4%)	14 (14.9%)	1 (6.2%)
	ISC / 150h	218 (29.0%)	71 (42.0%)	139 (29.4%)	8 (8.5%)	0 (0.0%)

Consensuses	**ISC-h^{10} / 150h**	271 (36.0%)	83 (49.1%)	173 (36.6%)	13 (13.8%)	2 (12.5%)
	Cons^{3}	241 (32.0%)	75 (44.4%)	152 (32.1%)	13 (13.8%)	1 (6.2%)
	Cons^{3}-h	271 (36.0%)	82 (48.5%)	174 (36.8%)	13 (13.8%)	2 (12.5%)
	Cons^{4}-h/150h	276 (36.7%)	84 (49.7%)	180 (38.1%)	11 (11.7%)	1 (6.2%)
	Cons^{4}-h/1k	282 (37.5%)	82 (48.5%)	184 (38.9%)	16 (17.0%)	0 (0.0%)
	Cons^{5}-h/150h	289 (38.4%)	93 (55.0%)	181 (38.3%)	14 (14.9%)	1 (6.2%)
	Cons^{5}-h/1k	304 (40.4%)	94 (55.6%)	191 (40.4%)	18 (19.1%)	1 (6.2%)

Table C-12: Performance with a more stringent near-native definition. *Top 10 success rate with near-natives defined as being of at least Medium quality according to CAPRI criteria.*

	Top 10 success rate	**Top 50 success rate**
FD	61 (8.1%)	103 (13.7%)
IES/10k	49 (6.5%)	84 (11.2%)
IES^{40}/10k	50 (6.6%)	87 (11.6%)
IES-h^{40}/10k	60 (8.0%)	112 (14.9%)
IES-h^{10}/1k	66 (8.8%)	107 (14.2%)
SPP/10k	60 (8.0%)	101 (13.4%)
SPP-h^{40}/10k	87 (11.6%)	145 (19.3%)
SPP-h^{10}/1k	85 (11.3%)	136 (18.1%)
ISC/1k	50 (6.6%)	93 (12.4%)
ISC/150h	70 (9.3%)	138 (18.4%)
ISC-h^{10}/1k	94 (12.5%)	130 (17.3%)
ISC-h^{10}/150h	99 (13.2%)	159 (21.1%)
Cons^{3}	62 (8.2%)	/
Cons^{3}-h	76 (10.1%)	/
Cons^{4}-h/150h	77 (10.2%)	/
Cons^{4}-h/1k	84 (11.2%)	/
Cons^{5}-h/150h	84 (11.2%)	/
Cons^{5}-h/1k	86 (11.4%)	/

Table C-13: Performance in terms of top 150 nDCG. *Average nDCG were calculated and normalised over the top 150 decoys for each individual scores over 752 cases (see section 1.3.3.4.4).*

	Top 150 success rate	**nDCG /150**	**nDCG /150 (excluding cases with nDCG = 0)**
FD	387	0.118	0.147
IES/10k	377	0.135	0.180
IES^{40}/10k	371	0.134	0.180
IES-h^{40}/10k	417	0.157	0.195
IES-h^{10}/1k	431	0.165	0.201
SPP/10k	444	0.138	0.157
SPP-h^{40}/10k	455	0.180	0.207
SPP-h^{10}/1k	458	0.186	0.213
ISC/1k	437	0.115	0.137
ISC/150h	**476**	0.149	0.169
ISC-h10/1k	451	0.182	0.213
ISC-h10/150h	**476**	**0.208**	**0.236**

Table C-14: Performance of consensus scores including InterEvScore implicit homology scoring. *Performance of three- and four-way consensus scores in terms of top 10 success rates on 752 benchmark cases. Scores used in Cons[3] were SOAP-PP on the top 10,000 or top 1,000 FRODOCK2.1 decoys (SPP/10k or SPP/1k), InterEvScore on the top 10,000 or top 1,000 FRODOCK2.1 decoys (IES/10k or IES/1k) and FRODOCK2.1 (FD2.1). Scores used in Cons[4] were SPP/10k, IES/10k, FRODOCK2.1 and Rosetta's interface score on the top 1,000 FRODOCK2.1 decoys (ISC/1k). Performances of individual scores used in the consensuses are reported in terms of top 10 and top 50 success rates, since consensus calculation relies on the top 50 decoys ranked by each component score.*

Score	Top 10 success rate	Top 50 success rate
FD2.1	164 (21.8%)	292 (38.8%)
IES/10k	182 (24.2%)	287 (38.2%)
IES/1k	196 (26.1%)	295 (39.2%)
SPP/10k	183 (24.3%)	**328 (43.6%)**
SPP/1k	187 (24.9%)	295 (39.2%)
Cons[3]	**241 (32.0%)**	/
ISC/1k	157 (20.9%)	301 (40.0%)
Cons[4]	235 (31.2%)	/

We try to improve the baseline consensus performance by incorporating Rosetta's physics-based Interface Score (ISC) (section 3.1.2). As Rosetta scoring is more computationally expensive than the other two scores (about 750 times slower than SOAP-PP and InterEvScore calculations), we score only the top 1,000 decoys (as ranked by FRODOCK2.1) with ISC. This score is denoted ISC/1k as opposed to IES/10k and SPP/10k. As such, ISC is individually less well performing than the other scores in terms of top 10 success rate, even when InterEvScore and SOAP-PP are computed only on the top 1,000 FRODOCK2.1 decoys (supplementary Table C-14). However, the top 50 success rate is higher for ISC/1k than for any other individual score, except for SOAP-PP calculated on 10,000 decoys (supplementary Table C-14). In spite of this, integrating the top 50 decoys ranked by ISC/1k with the top 50 of the other three scores into a four-way consensus, denoted Cons[4], slightly degrades performance compared to Cons[3] (supplementary Table C-14) while strongly increasing computation time.

Table C-15: Performances as reported in the InterEvDock2 paper. *Top 10 success rates of original scores in InterEvDock2 with percentages calculated over the same 752 cases compared with equivalent scores in this article. Original InterEvScore was run on the original PPI4DOCK coMSA and on the realigned coMSAs used*

throughout the present study (see section 3.). Original SOAP-PP was run using the much slower Python implementation from the original publication.

	Top 10 success rate of original scores in InterEvDock2	Top 10 success rate of new scores
FRODOCK2.1	164 (21.8%)	164 (21.8%)
InterEvScore	171 (22.7%) (original coMSAs) 177 (23.5%) (realigned coMSAs)	182 (24.2%)
SOAP-PP	194 (25.8%)	183 (24.3%)
3-way consensus	239 (31.8%)	241 (32.0%)

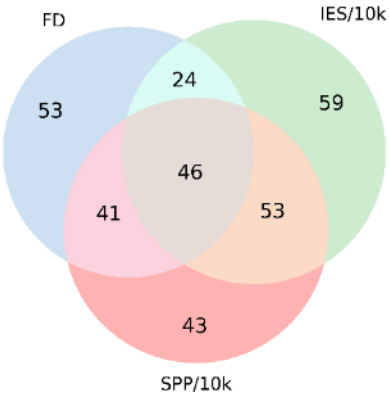
Table C-16: Performances of InterEvScore with 2-body and 2/3-body potentials. *Top 10 success rates of InterEvScore with complete coMSAs (IES) on 10,000 decoys, InterEvScore using homology models (IES-h) on coMSA^{40} and 10,000 decoys and on coMSA^{10} and 1,000 decoys using only 2-body potentials or 2- and 3-body potentials.*

	$2/3B^{best}$	$2B^{best}$
IES/10k	182 (24.2%)	164 (21.8%)
IES/1k	196 (26.1%)	192 (25.5%)
IES^{query}/10k	147 (19.5%)	154 (20.5%)
IES^{query}/1k	172 (22.9%)	165 (21.9%)
IES-h^{40}/10k	161 (21.4%)	**203 (27.0%)**
IES-h^{10}/1k	182 (24.2%)	200 (26.6%)

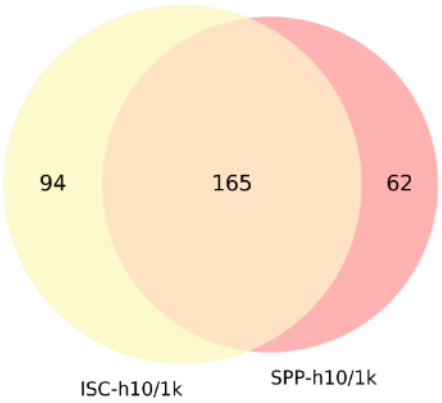
2. Supplementary figures

Figure C-2: Venn diagrams between scores. *Top 10 success rate intersections between scores on 752 cases. FD: FRODOCK2.1, IES: InterEvScore on complete coMSAs, SPP: SOAP-PP and ISC: Rosetta's interface score. /10k and /1k denote that 10,000 and 1,000 decoys were scored. –h10 and –h40 denote homology-enriched scores with 10 or 40 homolog models (coMSA^{10} or coMSA^{40}).*

A



B



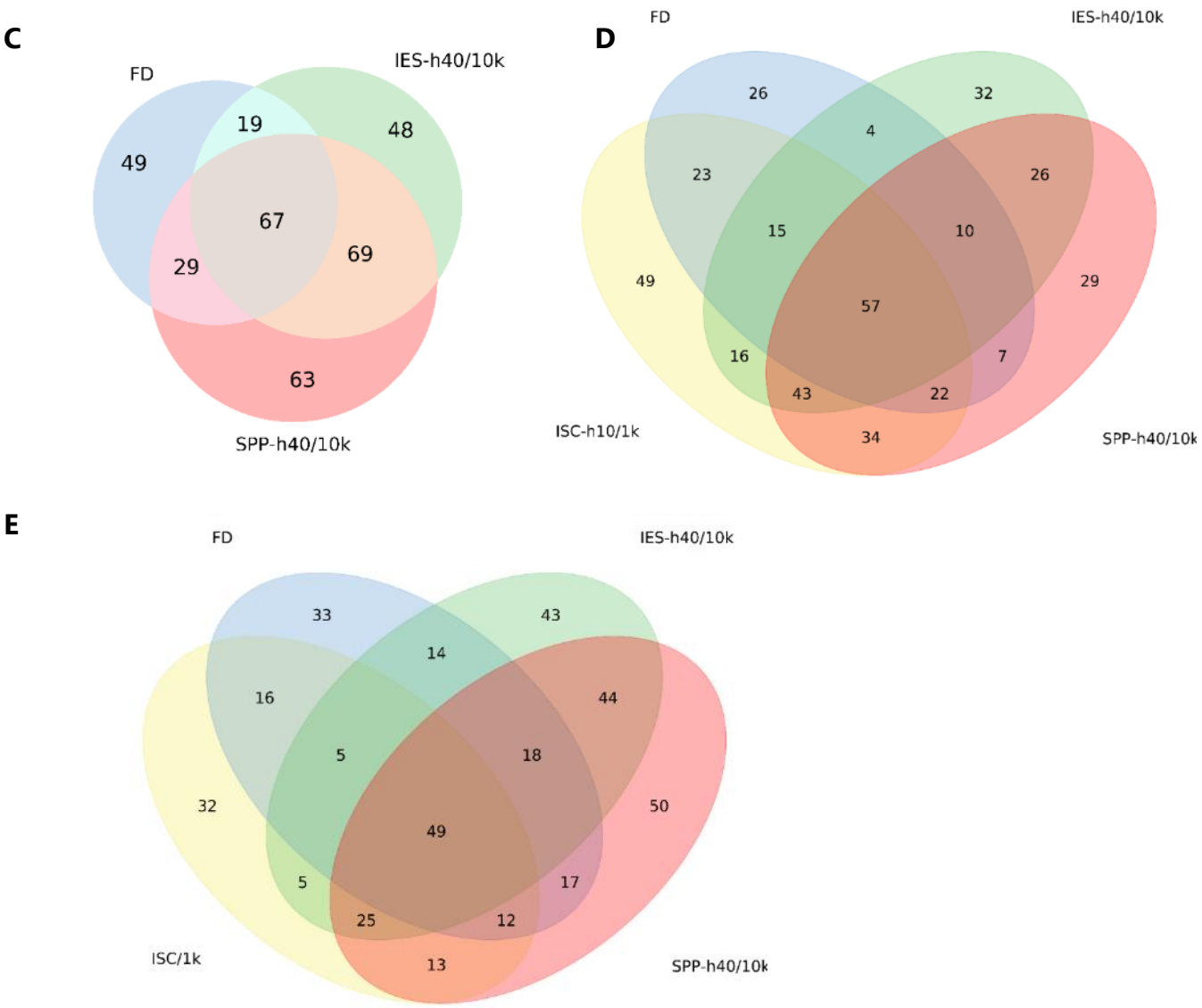
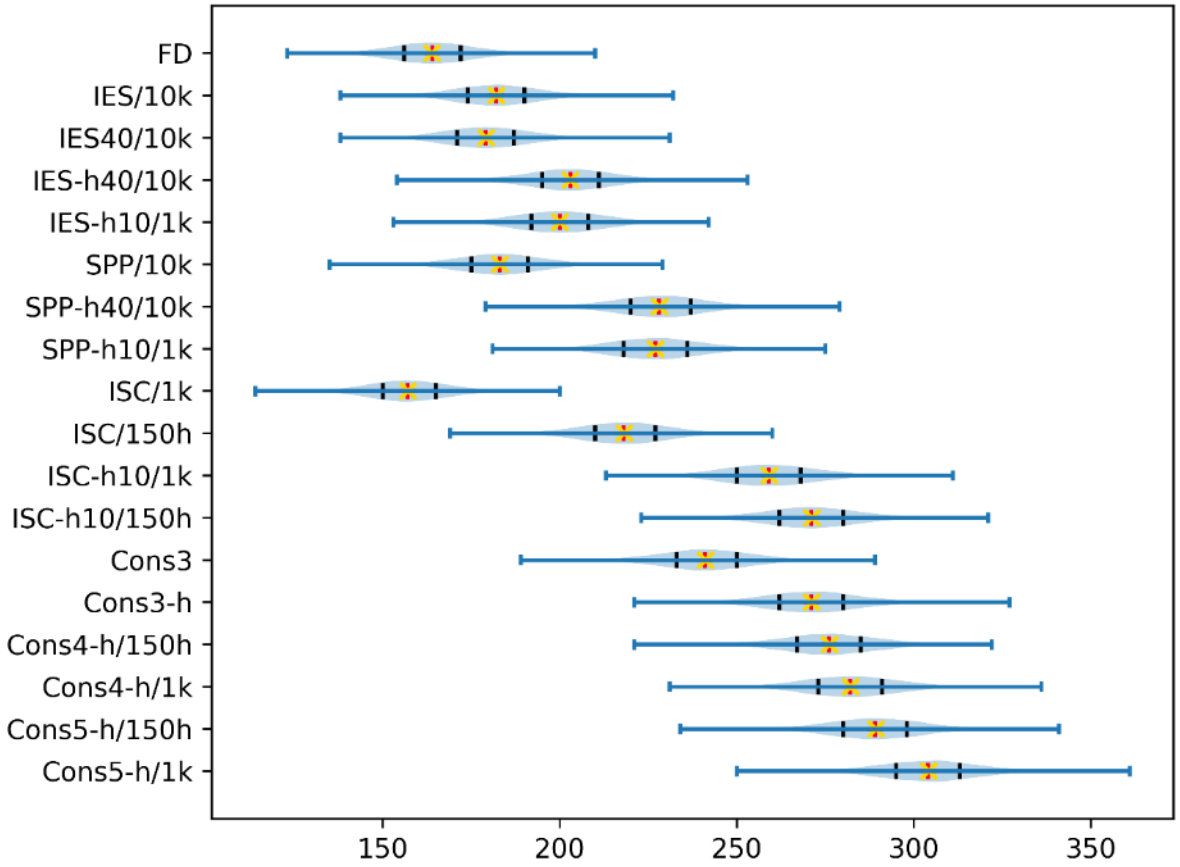


Figure C-3: Bootstrap performance distributions. *Bootstrap top 10 success rate distributions for 10,000 iterations over the 752 cases in our benchmark (blue). Measured top 10 success rates are marked in red and average success rates over all bootstrap iterations are marked as yellow crosses. Black bars indicate 25th and 75th percentiles of the bootstrap distribution. A two-sample t-test with unequal variances (Welch's t-test) on*

all score pairs in this plot systematically outputs p-values < 10^{-10} except for $Cons^{3}$-h against ISC-h^{10}/150h, thus all distribution means are statistically different relative to each other except for these two scores.



D. Supplementary materials for Chapter 4

a. RosettaScript protocol for round CAPRI45

RosettaScript protocol adapted for the round CAPRI45 to combine sampling of domain orientations, loop remodeling under symmetry constrains.

```
<ROSETTASCRIPTS>
      <SCOREFXNS>
            <ScoreFunction name="scorefxn_loopcen" patch="score4L"
weights="cen_std.wts">
                  <Reweight scoretype="atom_pair_constraint" weight="1.0"/>
            </ScoreFunction>
            <ScoreFunction name="scorefxn_loopfa" weights="talaris2013.wts"/>
            <ScoreFunction name="score_docking_low_cst" symmetric="0"
weights="interchain_cen">
                  <Reweight scoretype="atom_pair_constraint" weight="1.0"/>
            </ScoreFunction>
            <ScoreFunction name="ref15sfxn" symmetric="0"
weights="beta_nov15.wts"/>
            <ScoreFunction name="ref15sfxn_symm" symmetric="1"
weights="beta_nov15.wts">
                  <Reweight scoretype="atom_pair_constraint" weight="2.0"/>
            </ScoreFunction>
      </SCOREFXNS>
      <RESIDUE_SELECTORS>
            <Or name="chain_symm">
                  <!-- Used for the selection of 5 out of 6 subunits loaded each
made of two chains (Nter and Cter domain without linker) -->
                  <!--  (i) These subunits are used to define the symmetry file,
                        (ii) Then, deleted before sampling the orientations of
Nter domain vs Cter,
                        (iii) They are rebuilt by symmetry in the end -->
                  <Chain chains="B"/>
                  <Chain chains="C"/>
                  <Chain chains="D"/>
                  <Chain chains="E"/>
                  <Chain chains="F"/>
                  <Chain chains="G"/>
                  <Chain chains="H"/>
                  <Chain chains="I"/>
                  <Chain chains="J"/>
                  <Chain chains="K"/>
            </Or>
            <Or name="chainD">
                  <Chain chains="D"/>
            </Or>
            <Index name="D1Cter" resnums="807P"/>
            <Index name="D1tyr" resnums="784P"/>
            <Index name="D2Nter" resnums="1A"/>
            <Index name="D3pro" resnums="651A"/>
            <Index name="D4lys" resnums="619A"/>
            <Index name="frag2del" resnums="603-661"/>
            <Index name="r146" resnums="146A"/>
            <Index name="r147" resnums="147A"/>
            <Index name="r148" resnums="148A"/>
            <Index name="r149" resnums="149A"/>
```

```
            <Index name="r150" resnums="150A"/>
      </RESIDUE_SELECTORS>
      <TASKOPERATIONS>
            <InitializeFromCommandline name="ifcl"/>
            <RestrictToRepacking name="rtr"/>
            <IncludeCurrent name="keep_curr"/>
      </TASKOPERATIONS>
      <FILTERS>
            <!-- Filters are used to restrict the sampling of the Nter so that
the Nter and Cter can be tethered after sampling -->
            <AtomicDistance atomname1="CA" atomname2="CA" distance="8.0"
name="CNmin" residue1="807P" residue2="1A"/>
            <AtomicDistance atomname1="CA" atomname2="CA" distance="11.0"
name="CNmax" residue1="807P" residue2="1A"/>
            <AtomicDistance atomname1="CA" atomname2="CA" distance="10.0"
name="Cst1min" residue1="807P" residue2="632A"/>
            <AtomicDistance atomname1="CA" atomname2="CA" distance="30.0"
name="Cst1max" residue1="807P" residue2="632A"/>
            <AtomicDistance atomname1="CA" atomname2="CA" distance="5.0"
name="Cst2min" residue1="784P" residue2="619A"/>
            <AtomicDistance atomname1="CA" atomname2="CA" distance="16.0"
name="Cst2max" residue1="784P" residue2="619A"/>
            <AtomicDistance atomname1="CA" atomname2="CA" distance="10.5"
name="Cst3min" residue1="807P" residue2="651A"/>
            <AtomicDistance atomname1="CA" atomname2="CA" distance="12.0"
name="Cst3max" residue1="807P" residue2="651A"/>
            <CompoundStatement invert="false" name="CNthresh">
                  <AND filter_name="CNmax"/>
                  <ANDNOT filter_name="CNmin"/>
                  <AND filter_name="Cst2max"/>
                  <ANDNOT filter_name="Cst2min"/>
                  <AND filter_name="Cst1max"/>
                  <ANDNOT filter_name="Cst1min"/>
                  <AND filter_name="Cst3max"/>
                  <ANDNOT filter_name="Cst3min"/>
            </CompoundStatement>
      </FILTERS>
      <MOVERS>
            <DeleteRegionMover name="del_symm" residue_selector="chain_symm"/>
            <AddConstraints name="add_D1D2_cst">
                  <DistanceConstraintGenerator atom_name1="CA" atom_name2="CA"
function="BOUNDED 7.0 12.0 0.5 0.5 TAG" name="D1D2_cst"
residue_selector1="D1Cter" residue_selector2="D2Nter"/>
            </AddConstraints>
            <AddConstraints name="add_D1D3_cst">
                  <DistanceConstraintGenerator atom_name1="CA" atom_name2="CA"
function="BOUNDED 7.0 12.0 0.5 0.5 TAG" name="D1D3_cst"
residue_selector1="D1Cter" residue_selector2="D3pro"/>
            </AddConstraints>
            <AddConstraints name="add_D1D4_cst">
                  <DistanceConstraintGenerator atom_name1="CA" atom_name2="CA"
function="BOUNDED 5.0 16.0 0.5 0.5 TAG" name="D1D4_cst"
residue_selector1="D1tyr" residue_selector2="D4lys"/>
            </AddConstraints>
            <!-- Defining the connectivity between residues -->
            <!-- File contents: FOLD_TREE  EDGE 1 248 -1  EDGE 248 661 -1  EDGE
248 715 1  EDGE 715 662 -1  EDGE 715 807 -1 -->
            <AtomTree fold_tree_file="constraint_foldtree.cst"
name="def_foldtree"/>

            <!-- Parsed protocol for sampling at low resolution the orientation
of the Nter domain -->
            <!-- CNthresh distance filter is used to constrain the extremities
of the domain, so they can be subsequently tethered by a linker -->
```

```
            <Docking conserve_foldtree="0" design="0" fullatom="0"
ignore_default_docking_task="0" jumps="1" local_refine="0" name="dock_low"
optimize_fold_tree="0" score_high="ref15sfxn" score_low="score_docking_low_cst"
task_operations="ifcl"/>
            <LoopOver drift="false" filter_name="CNthresh" iterations="50"
mover_name="dock_low" ms_whenfail="FAIL_DO_NOT_RETRY" name="repeat_docklow"/>
            <ParsedProtocol name="combine_docklow">
                  <Add mover_name="repeat_docklow"/>
            </ParsedProtocol>

            <!-- Defining the connectivity between residues ==> Getting back
initial configuration in terms of foldtree -->
            <!-- File contents: FOLD_TREE  EDGE 1 661 -1  EDGE 1 662 1  EDGE 662
807 -1 -->
            <AtomTree fold_tree_file="constraint_foldtree2ini.cst"
name="def_foldtree_ini"/>
            <RemoveConstraints constraint_generators="D1D2_cst"
name="rm_D1D2_cst"/>
            <RemoveConstraints constraint_generators="D1D3_cst"
name="rm_D1D3_cst"/>
            <RemoveConstraints constraint_generators="D1D4_cst"
name="rm_D1D4_cst"/>
            <DeleteRegionMover name="del_frag2del" residue_selector="frag2del"/>
            <AddChain file_name="input_chainD.pdb" name="addDsubunit"
new_chain="1" scorefxn="talaris2013"/>
            <DumpPdb fname="dump_preloop.pdb" name="writepose_preloop"
scorefxn="talaris2013" tag_time="0"/>
            <AddConstraints name="add_csts_hel">
                  <!-- Adding constraints to obtain a helix structure -->
                  <FileConstraintGenerator
filename="constraints_helix_centroid.cst" name="loop_hel"/>
            </AddConstraints>
            <LoopModeler config="kic" fast="0" name="l1"
scorefxn_cen="scorefxn_loopcen" scorefxn_fa="scorefxn_loopfa">
                  <Loop cut="710" skip_rate="0.0" start="702" stop="721"/>
                  <Build skip="0"/>
                  <Centroid skip="0"/>
                  <Fullatom skip="1"/>
            </LoopModeler>
            <SwitchResidueTypeSetMover name="switch_repr" set="fa_standard"/>
            <DeleteRegionMover name="del_chainD" residue_selector="chainD"/>
            <SwitchChainOrder chain_order="21" name="switch_chain"/>
            <BridgeChains chain1="1" chain2="2" motif="2HA-3LX" name="bridge"
overlap="2" scorefxn="scorefxn_loopcen"/>
            <MutateResidue name="mutres146" new_res="GLY"
residue_selector="r146"/>
            <MutateResidue name="mutres147" new_res="LEU"
residue_selector="r147"/>
            <MutateResidue name="mutres148" new_res="LEU"
residue_selector="r148"/>
            <MutateResidue name="mutres149" new_res="PRO"
residue_selector="r149"/>
            <MutateResidue name="mutres150" new_res="PRO"
residue_selector="r150"/>
            <ParsedProtocol name="mutate_linker">
                  <Add mover_name="mutres146"/>
                  <Add mover_name="mutres147"/>
                  <Add mover_name="mutres148"/>
                  <Add mover_name="mutres149"/>
                  <Add mover_name="mutres150"/>
            </ParsedProtocol>
            <DumpPdb fname="dump_postlinker.pdb" name="writepose_postlinker"
scorefxn="talaris2013" tag_time="0"/>
            <SetupForSymmetry definition="input_symm_def.symm"
```

```
name="setup_symm"/>
            <!-- Adding constraints to create hydrogen bonds or other type of
interactions such as salt bridges..etc,
                  so that we can avoid buried polar residues or unsatisfied
polar residues -->
            <AddConstraints name="add_csts_relax">
                  <FileConstraintGenerator filename="constraints_relax_fa.cst"
name="cst_final_relax"/>
            </AddConstraints>
            <FastRelax bondangle="false" bondlength="false" cartesian="false"
min_type="dfpmin_armijo_nonmonotone" name="rlx_symm"
ramp_down_constraints="false" repeats="1" scorefxn="ref15sfxn_symm"
task_operations="ifcl,rtr,keep_curr"/>
      </MOVERS>
      <APPLY_TO_POSE/>
      <PROTOCOLS>
            <!-- Load a complex of 6 subunits each composed of two domains (Nter
/ Cter without linker) which were initially used to define symmetry -->
            <!-- Remove 5 subunits, keep only subunit 1 composed of two domains
chain A and chain P -->
            <Add mover="del_symm"/>

            <!-- Define a set of spatial restraints and a fold-tree to generate
restricted and efficient sampling of the Nter domain with respect to Cter-->
            <Add mover="add_D1D2_cst"/>
            <Add mover="add_D1D3_cst"/>
            <Add mover="add_D1D4_cst"/>
            <Add mover="def_foldtree"/>

            <!-- Run the rigid body sampling of Nter orientation -->
            <Add mover="combine_docklow"/>

            <!-- Remove the constraints used for sampling -->
            <Add mover="def_foldtree_ini"/>
            <Add mover="rm_D1D2_cst"/>
            <Add mover="rm_D1D3_cst"/>
            <Add mover="rm_D1D4_cst"/>
            <!-- Remove a segment of subunit 2 which was kept to restrict the
Nter domain moves so that it does not clash subsequently with subunit 2-->
            <Add mover="del_frag2del"/>

            <!-- Recover the subunit 2 coordinates from the original file.
                        The Nter domain of subunit 1 packs against this subunit
2
                        Presence of subunit 2 will prevent that the linker to be
created between Nter-Cter clashes with neighbouring subunits -->
            <Add mover="addDsubunit"/>

            <!-- Dump the pose for checking -->
            <Add mover="writepose_preloop"/>

            <!-- Extend the loop between Nter and Cter domain, only backbone
trace and do not close it (keep two chains)-->
            <Add mover="add_csts_hel"/>
            <Add mover="l1"/>
            <Add mover="switch_repr"/>

            <!-- subunit 2 used to restrict loop conformation can be removed -->
            <Add mover="del_chainD"/>
            <!-- Cter was chain A while Nter was chain P, they have to be
switched before bridging the loop -->
            <Add mover="switch_chain"/>
            <Add mover="bridge"/>
```

```
            <!-- Alter the sequence of the linker to retrieve the correct side-
chains -->
            <Add mover="mutate_linker"/>

            <!-- Dump the pose for checking -->
            <Add mover="writepose_postlinker"/>

            <!-- Regenerate the symmetric subunits and relax -->
            <Add mover="setup_symm"/>
            <Add mover="add_csts_relax"/>
            <Add mover="rlx_symm"/>
      </PROTOCOLS>
      <OUTPUT/>
</ROSETTASCRIPTS>
```

université PARIS-SACLAY

ECOLE DOCTORALE N°569
Innovation thérapeutique : du fondamental à l'appliqué (ITFA)

Titre : Exploration et modélisation structurale d'interactions protéiques guidées par l'information évolutive

Mots clés : Interactions protéine-protéine, modélisation structurale, bioinformatique, information de co-évolution, amarrage protéine-protéine

Résumé : Les protéines sont des acteurs centraux du vivant et agissent rarement seules. La structure 3D de leurs interactions aide à mieux comprendre les mécanismes des processus biologiques dans lesquels elles sont impliquées. L'objectif de cette thèse était d'améliorer la performance des méthodes de prédiction structurale, notamment en utilisant l'information de (co-)évolution. J'ai participé à des évolutions majeures de notre serveur de docking, InterEvDock2, qui, à partir de deux protéines, propose 10 modèles d'interface en croisant des scores aux propriétés différentes. Le serveur accepte désormais aussi en entrée des structures oligomériques ou des séquences protéiques, pour lesquelles il modélise la structure monomérique, ainsi que des contraintes connues a priori sur l'interaction. Sur un large ensemble de 812 cas test, InterEvDock2 prédit une structure de complexe correcte dans 32 % des cas. J'ai ensuite recherché un moyen plus explicite d'intégrer dans les fonctions de score l'information évolutive contenue dans les alignements de séquences. J'ai rendu cette information compatible avec l'utilisation de scores atomiques par la modélisation 3D des interfaces homologues. Ceci améliore la performance prédictive de 32 à 40% sur une large base de test. De plus, durant ma thèse, j'ai pu participer à 10 tests de docking à l'aveugle via CAPRI (Critical Assessment of Predicted Interactions). Les stratégies qui ont permis à notre équipe d'être classée première sur la période 2016-2019 sont décrites dans le dernier chapitre de ce manuscrit. Ce travail vise à aider les biologistes à étudier les protéines ou voies biologiques d'intérêt en utilisant des méthodes de prédiction performantes et constitue un pas en avant dans l'objectif final de la prédiction des interactomes.

Title : Exploration and structural modelling of protein interactions using evolutionary information

Keywords : Protein-protein interactions, structural modelling, bioinformatics, co-evolutionary information, protein-protein docking

Abstract: Protein complexes are of fundamental importance in most biological processes and mainly carry out their function in networks. The structure of their interface can give us crucial information to understand the mechanisms behind these processes. This thesis focuses on the improvement of the performance of structural prediction methods, in particular by exploiting co-evolutionary information.
As part of my PhD project, I participated in major developments in our docking server, InterEvDock2, which suggests 10 interface models for a pair of input proteins using a mix of different scoring properties. InterEvDock2 now also accepts oligomeric structure inputs or sequence inputs, for which it can model monomeric structures, as well as user constraints taken from prior knowledge of the interaction. I validated the performance of InterEvDock2 on a large benchmark of 812 docking cases and found that InterEvDock2 was capable of finding a correct complex structure in as much as 32 % of these cases. My work then focused on finding a more efficient and explicit way of integrating implicitly defined evolutionary information into scoring. I made this information directly compatible with atomic-scale scoring thanks to homologous interface modelling. This strongly increases predictive power, from 32% to 40% on a large benchmark. Moreover, throughout my PhD, I was able to participate in 10 blind-test docking challenges through CAPRI (Critical Assessment of Predicted Interactions). The strategies applied by our team, which enabled us to rank first in the latest CAPRI round for 2016-2019, are described in the last chapter of this manuscript. This work aims at helping biologists study their proteins or biological pathways of interest using well performing prediction methods. It constitutes a step towards the final goal of interactome prediction.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France