Structural Complex Prediction Based on Protein Interface Recognition

by

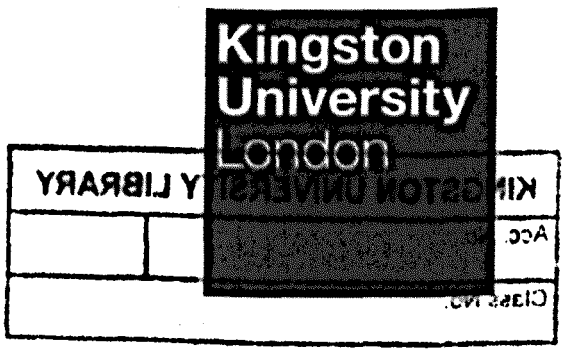
Reyhaneh Esmaielbeiki

Submitted in partial fulfilment of the requirements of

Kingston University for the degree of

Doctor of Philosophy

August, 2013



[Page intentionally left blank]

Supervision:

Dr Jean-Christophe Nebel (Director of Studies) [1]

Dr Dimitrios Makris [2]

Prof. Declan Naughton [3]

[1] Bioinformatics & Genomic Signal Processing Research Group
School of Computing and Information Systems
Faculty of Science, Engineering and Computing Kingston University
Penrhyn Road
Kingston upon Thames
Greater London
KT1 2EE
United Kingdom

[2] School of Computing and Information Systems
Faculty of Science, Engineering and Computing Kingston University
Penrhyn Road
Kingston upon Thames
Greater London
KT1 2EE
United Kingdom

[3] School of Life Sciences
Faculty of Science, Engineering and Computing Kingston University
Penrhyn Road
Kingston upon Thames
Greater London
KT1 2EE
United Kingdom

[Page intentionally left blank]

Abstract

This dissertation contributes to the state of the art in protein interface prediction and detection of native-like docked poses by re-ranking them using protein interface knowledge. We started by investigating binding site patterns among homologues of a target protein in order to create a 3D motif. This structural binding site descriptor enables the re-ranking of docked complexes of the target protein. Although 3D motifs provide biological insight of protein interactions and have usage in real applications, they are not suitable for high through-put analysis. Therefore, we introduced a novel protein interface prediction framework which uses a weighted scoring schema to detect interface residues of the target protein using its homologues. The weights quantify both homology closeness between the target protein and its homologues and the diversity between the interacting partners of these homologues. The main novelty of this predictor is that it takes into account the nature of homologues interacting partners. It was further exploited for the development of a method for re-ranking docked conformations using predicted interface residues. We have evaluated both our interface predictor and re-ranking of docked poses using standard benchmarks. Comparisons to current state-of-the-art methods reveal that the proposed approaches outperform all their competitors. However, similarly to current interface predictors, our framework does not explicitly refer to pairwise residue interactions which leaves ambiguity when assessing quality of complex conformations. In addition, the performance of our interface predictor generally does not outperform the best available homologue interfaces if it was used as prediction. Therefore, we investigated the detection of the best homologue using the 'binding site transitivity' concept: given two query protein chains, interfaces of the first query protein are structurally compared against binding sites of the homologues' partners of the second query chain. This method not only allows detection of the best homologue for a reasonable number of proteins but also produces a docked structure of the two query chains. Finally, experiment suggests a meta interface predictor combining the prediction of our former interface predictor with the latter predictor based on binding site transitivity could further improve interface prediction.

[Page intentionally left blank]

To my wonderful husband
Mohammad

[Page intentionally left blank]

Acknowledgment

First and foremost I would like to thank my first supervisor, Dr Jean-Christophe Nebel, for his continued guidance, contribution of time and ideas all throughout my PhD. His enthusiasm and motivation for his research has always given me inspiration and energy even during the tough times of my PhD. His insightful guidance, suggestions and encouragements have helped me build my skills as an academic. Thank you for giving me the confidence and opportunity to challenge myself by thinking and working outside my comfort zone. Moreover, I deeply appreciate the long hours you spent reading my thesis, your intelligent comments and superior knowledge significantly contributed to development of this thesis. Also I would like to thank my co-supervisor Prof. Declan Naughton for all his comments regarding the biological aspects of the research and his support for publishing my first journal paper. In addition, I would like to thank Dr Dimitrios Makris for giving me the opportunity to extend my teaching experience, by involvement with his postgraduate module.

Moreover, my time at Kingston University was made enjoyable due to many friends and groups which became part of my life. In particular, I would like to thank Jesús Martínez del Rincón and Maria Valera for their friendship and support through the ups and downs of my PhD period.

Furthermore, in a special way I would like to thank my amazing parents and sister for their endless support. Thanks for raising me with a love for science and for believing in me through all the steps of my PhD journey.

Finally, most of all, I would like to dedicate this thesis to my wonderful husband, Mohammad. Without his love, patience, support and encouragement I would have not been able to make it all this way. Thanks for all your dedication and understanding which allowed me to achieve my dreams.

[Page intentionally left blank]

Declarations

I hereby declare that this dissertation describes my solely own research, which was carried out at Kingston University, except where otherwise indicated. Other sources are acknowledged by explicit references. Some of the research presented in this thesis has already been published or is under review for publication. For a complete list of publications, please refer to the next page.

This thesis has not been previously accepted in substance for any degree and is not being concurrently submitted to any other University for examination either in the United Kingdom or overseas.

Reyhaneh Esmaielbeiki

[Page intentionally left blank]

List of Publications

Chapter 3

- **R. Esmaielbeiki**, D. Naughton and J.C. Nebel, "Structure prediction of LDLR-HNP1 complex based on docking enhanced by LDLR binding 3D motif" , Protein & Peptide Letters, 19(4),458-467, 2012.

Chapter 4

- **R. Esmaielbeiki** and J.C. Nebel, "Scoring docking conformations using predicted protein interfaces", BMC Bioinformatics, Accepted subject to corrections, 2013.

- **R. Esmaielbeiki** and J.C. Nebel, "Unbiased Protein Interface Prediction Based on Ligand Diversity Quantification", Open Access Series in Informatics (OASICS), Vol. 26, German Conference on Bioinformatics (GCB) 2012, Jena, Germany, Sep. 19-22.

Chapter 5

- **R. Esmaielbeiki** and J.C. Nebel, "Protein interface prediction based on structural-transitivity of protein binding sites", BMC Bioinformatics, to be submitted in autumn 2013.

[Page intentionally left blank]

Glossary of Terms

3D	Three Dimensional
QP	Query Protein
PDB	Protein Data Bank
PQS	Protein Quaternary Structure
ASA	Accessible Surface Area
RSA	Relative Solvent Accessibility
PPI	Protein-Protein Interactions
CAPRI	Critical Assessment of PRedicted Interactions
PCA	Principle Component Analysis
SVM	Support Vector Machine
NN	Neural Network
RF	Random Forest
MSA	Multiple Sequence Alignment
S-MSA	Structure-based Multiple Sequence Alignment
PSSM	Position Specific Scoring Matrix
NMR	Nuclear Magnetic Resonance
FFT	Fast Fourier Transform
MC	Monte Carlo
GA	Genetic Algorithm
MD	Molecular Dynamic
GH	Geometric Hashing
ACP	Atomic Contact Potential
EM	Energy Minimization
ET	Evolutionary Trace
GT	Ground Truth
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

List of Figures

List of Tables

Content

1 Introduction

Similarly to 'words', which need to be "assembled into sentences, paragraphs, chapters and books" (Sali et al. 2003) to tell a story, 'protein structures' need to be assembled into protein complexes to perform a specific task. To form complexes, proteins interact with other proteins, DNA, RNA and small molecules using their interface residues. All those types of interactions are under intense scrutiny by the research community, each of them defining a distinct field of research. In this study we are focused on protein-protein interactions (PPIs) and prediction of their interfaces. Modifications in PPIs affect the events that take place within cells which may lead to critical diseases such as cancer (Sali et al. 2003; Nebel 2012). Therefore, knowledge about PPIs and their resulting 3D complexes can provide key information for drug design.

A number of experimental techniques have been used to decipher PPIs. Not only are these methods expensive and time-consuming, they also produce datasets which are incomplete and inconsistent (Ezkurdia et al. 2009). Therefore many computational methods have emerged and are used for predicting PPI sites, also named as interfaces. An interface is a set of residues on one protein which interacts with residues from other proteins to form a complex. Knowledge of interfaces is important to fully understand proteins molecular mechanism and to identify potential drug targets. However, in order to elucidate biological process involved in pathways and signal transduction cascades, atomic structures of the 3D complexes are required (Nebel 2012). Docking is a popular computational method which predicts the possible structure of the complex produced by two proteins using the known 3D structure of the individual proteins. However, docking of two proteins can result in a large number of different conformational models the

majority of which is far from correct. This highlights one of the main limitations of docking. Therefore, scoring functions have been proposed which are used to re-score and re-rank docked conformations in order to detect near-native models. One way to distinguish native-like models from false docked poses is to use knowledge of protein interfaces. If one knows the possible location of interface residues on each individual protein, docked complexes which do not involve those interfaces can be rejected. Therefore, accurate prediction of protein interfaces can assist with detection of native-like conformations.

In this thesis, we will explore the realm of protein interface prediction and their application in distinguishing native-like complexes by re-ranking docked poses. We propose novel approaches to identify protein binding sites and demonstrate performance better than current state-of-the-art methods.

In this introductory chapter, we first present the context of our research in sections 1.1 and 1.2. This is followed by a summary on our aim and objectives in 1.3 and our scientific contribution in 1.4. Section 1.5 outlines the structure of the whole thesis.

1.1 Protein Interface

When two proteins interact residues from one protein create contact with residues of the interacting partner. These residues are called interface residues and are located at the binding site of proteins. These sites display specific chemical and physical properties which is important for bimolecular recognition and interaction with other proteins (Aytuna et al. 2005). Several experimental techniques have been used to identify interface residues. Mutagenesis has been particularly powerful in their detection. In mutagenesis, mutation of interface residues will influence the interaction, while mutation of non-interfaces does not have any effect (Van Dijk, Boelens, et al. 2005). Methods such as yeast two-hybrid, Plasmon resonance and phage displays can also be used to study if a residue mutation has affected the complex formation. To identify a residue for mutation, either it can be selected based on prior knowledge (such as a highly conserved residue) or it can be performed on all residues using alanine scanning (Van Dijk, Boelens, et al. 2005) . Alanine-scanning measures the contribution of each residue to formation of the protein-protein complex, by analysing the drop in the

binding free energy when the target residue is mutated to alanine. Those studies have also shown that not all interface residues contribute equally to binding free energy (Aytuna et al. 2005).

A significant energy contribution to the stabilisation of a protein-protein complex has been seen by interface residues which are named Hot spot (Burgoyne & Jackson 2006). Mutation of hot spots to alanine have caused a significant drop in binding affinity (Bogan & Thorn 1998; Clackson & Wells 1995). These evolutionary conserved residues are located in the centre of interface patches (Hu et al. 2000; DeLano 2002), and are surrounded by a hydrophobic O-ring of energetically less important residues (Fernández-Recio 2011). Since experimental methods are costly, , computational methods have been proposed to detect both interface residues and hot spots (Ofran & Rost 2007b). These computational methods are introduced in chapter 2 and for the sake of being exhaustive both interface and hotspot predictors are discussed.

In order for those predictors to be able to distinguish interfaces on a target protein, they need to rely on interface patterns extracted from experimentally determined complexes in the Protein Data Bank (PDB) (Berman et al. 2000), the single worldwide repository of large biological molecular 3D structures . There are three main ways of defining interface residues in a protein complex:

1) The first approach measures the distance between atoms of one protein chain to atoms of its interacting partner. If the distance between is smaller than a defined threshold (usually between 5-8 Å (Cazals 2010)) then the residues containing those atoms are defined as belonging to the interface. Most methods use the definition of the Critical Assessment of PRedicted Interactions (CAPRI) (Janin et al. 2003) for this threshold: an interface residue is defined as an amino acid whose heavy atoms are within 5Å from those of a residue in a separate chain. In this thesis we have adopted this definition for identifying interface residues.
2) Another method measures the difference in the solvent accessible surface area (ASA) when individual protein chains are separated from their complex form. ASA is the area around the atom defined as its van der Waals radius plus a water probe radius. If an atom ASA in complex is smaller than in monomer by a specific threshold, the residue containing that atom is defined

as belonging to the interface (Yan et al. 2004). Figure 1.1 illustrates this idea. Interface residues can be further divided to buried or exposed depending on how accessible they are after forming a complex.

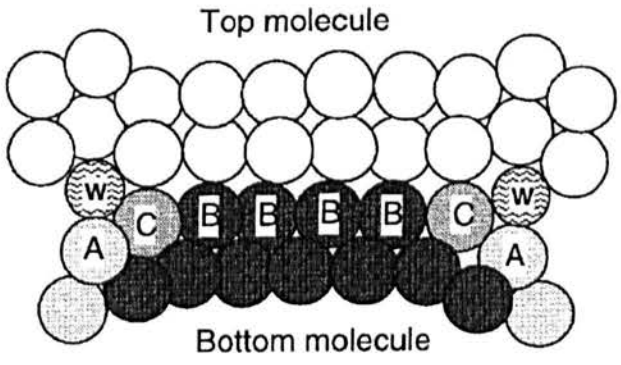


Figure 1.1: Example of interface definition using ASA. Atoms of the 'bottom molecule' which have lost accessibility to solvent due to interaction with the 'top molecule' are referred as interface residues (marked as A, B and C). B is totally buried while A and C are partial accessible. The water probe is marked as W. Taken from (Conte et al. 1999).

3) Finally, Voronoi diagrams have been used to select interface residues (Poupon 2004). The Voronoi diagrams (VD) divide the space into several convex polytopes here named as cells. Each cell contains one atom; two atoms are called neighbours if they share a VD edge. If the VD of two atoms from two different protein chains share a common edge then residues of those atoms are defined as interfaces. See Figure 1.2 for an example.

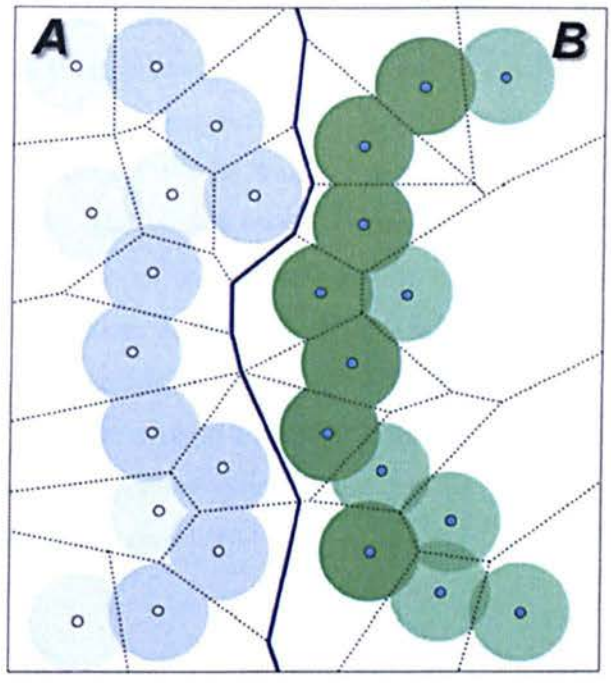


Figure 1.2: 2D representation of Voronoi diagrams between two protein chains A and B. Atoms of A and B are shown in blue and green circles, respectively. The Voronoi cells are shown in dotted lines between two neighbouring atoms and the solid blue line shows the Voronoi geometrical interface between the two chains. A cell is known as interface if it has an edge with the Voronoi geometrical interface line. In this image the interface atoms are shown darker than other atoms. Taken from (Gong et al. 2005).

Since the interface size and location defined by the methods above are identical or have a large overlap, experiments have shown that the method used to define the interface does not have an impact on predictors performance (Ezkurdia et al. 2009). Although the choice of method for defining interfaces is not important, the thresholds for defining distances and comparing ASA are critical for selecting specific features of interfaces (De Vries & Bonvin 2008).

To define interfaces, some methods first detect surface residues and then select interfaces with one of the methods above. Surface residues, which are exposed residues, are selected when their relative solvent accessibility (RSA) is above a threshold. This threshold has varied among methods from 16% to 5% (Ezkurdia et al. 2009), where a higher threshold lowers the number of selected surface residues. Although higher threshold may lead to information loss, it has been shown to improve some classifier-based approaches (Ezkurdia et al. 2009).

While interfaces defined as above may be composed of irregularly dispersed residues on a protein, some predictors have used protein patches which are continuous and generally form a circular area on a protein surface. The early work of Jones & Thornton (Jones & Thornton 1997a) has defined patches by first selecting surface residues with RSA>5%. Then, each surface accessible residue was used to define a surface patch. A patch consists of a central surface accessible residue with n nearest surface accessible neighbour, defined by the distance between their c-alpha atoms. While Jones & Thornton (Jones & Thornton 1997a) have used the number of residues (n) as a constraint, a recent method (Gamliel et al. 2011) has defined a sphere with radius r=7Å around the β-carbon atom of the surface accessible residue to select neighbouring residues (see Figure 1.3).

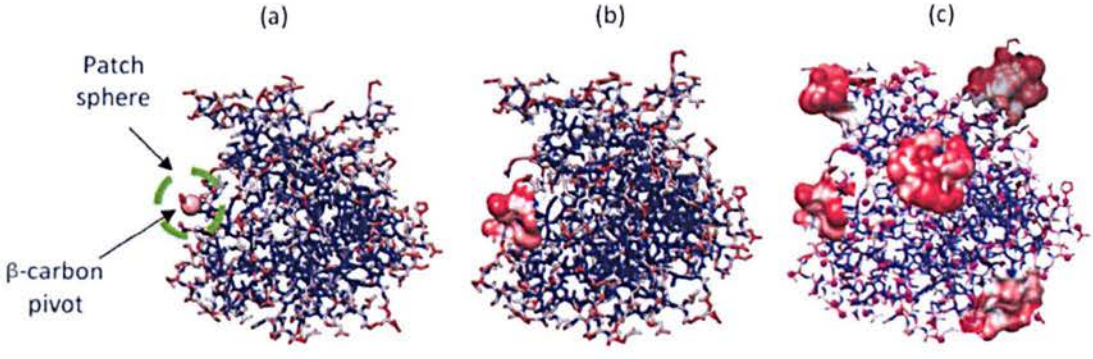


Figure 1.3: Example of surface patch definition. (a) Surface accessible residues are coloured in blue. A surface patch is defined with a central exposed β-carbon atom (large sphere) and the exposed residues within a radius r=7Å (in green). (b) Residues defining the surface patch are shown in red. (c) All patches of the protein. Taken from (Gamliel et al. 2011).

In practice, there are fewer interface predictors which use patch definition than those relying on the residue-based interface definition (see chapter 2). Regardless of the methods used, prediction of interface residues has been essential in assisting detection of native-like docked poses (discussed in chapter 2).

1.2 Protein Docking

3D complexes of proteins have provided valuable insight to structural mechanism of protein interactions and their individual role in a complex. Experimental techniques such as NMR, X-rays and yeast-2-hybrid are expensive and time consuming approaches to generate protein complexes. Therefore, in-silico approaches, or docking,

have been proposed to predict complexes from individual protein chains. Docking methods started with methods which search through the conformation space and generate a large number of docked poses for two protein chains (Figure 1.4 shows a schematic representation of docking). This search is guided by fast scoring functions based on energy potentials and shape complementarity. However, among the generated docked poses many contain models far from the native model and energy-based scoring functions fail to detect near-native configurations.

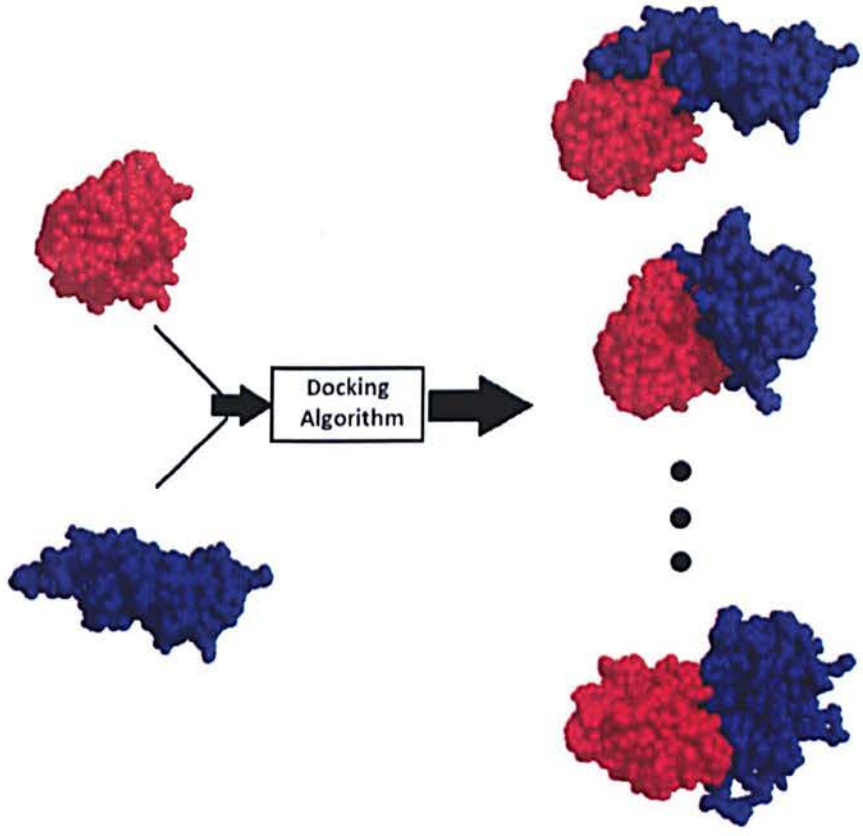


Figure 1.4: Docking algorithm. Two input chains are shown in red and blue. Several docked poses are generated by the docking algorithm.

Therefore, more elaborated scoring functions have been proposed to re-score and re-rank docked poses. Methods aiming at re-scoring docked poses can be broadly grouped into: (i) knowledge-based potentials scoring functions (ii) usage of learning strategies and (iii) usage of predicted interfaces. Among these methods, usage of predicted interface residues has shown great popularity (Xue et al. n.d.; Gottschalk et al. 2004; Huang & Schroeder 2008; Khashan et al. 2012) since: 1-knowledge-based

potentials have shown low correlation to binding affinities (Kastritis & Bonvin 2010) and 2- usage of learning strategies has not significantly improved the knowledge-based potentials scoring functions (Zhao et al. 2011).

With the increase in the available 3D complexes in PDB, template-based docking methods have emerged (Ghoorah et al. 2011; Tuncbag et al. 2011; Kundrotas & Alexov 2006; Zhang et al. 2013). They use available 3D complexes in PDB as templates to generate new complexes. Since the previous above mentioned docking approaches do not use any templates, they are called 'template-free docking'. While template-based docking has enabled large scale PPI prediction, template-free docking still remains highly important. First, many proteins in the cell do not correspond to the energetically stable crystallised templates (Vakser 2013). Second, still some PPI lack templates or the quality of the template is really low (Movshovitz-Attias et al. 2010). Therefore, in this thesis we have focused on 'template-free' docking and we simply refer to it as 'docking'. In this research we aim at detecting near-native docked poses by re-ranking them using knowledge of protein interfaces.

1.3 Aim and Objectives

The overall aim of this research is to improve the prediction of protein interfaces and to distinguishing near native docking models by means of re-ranking docked configurations. Our main objectives are summarised below:

- Detection of native-like docked models using 3D motif:
  Although docking algorithms are capable of producing native-like models, identifying them in a pool of decoys is still challenging. Therefore methods have used experimental mutagenesis studies on a short list of low energy docked conformations to identify the native-like model. Since energy-based ranking has proven not to be reliable (Kastritis & Bonvin 2010), we have introduced structural 3D motifs of protein binding sites to provide more reliable ranking before employing mutation studies.

- Improving interface prediction by considering nature of interacting partners:

  Current state of the art methods predict interfaces for a specific protein, by generating a consensus interaction patterns among available complexes taken as templates from PDB (for details see chapter 2). One of the chains in these template complexes is either homologue or structurally similar to the protein of interest. Despite their huge contributions to prediction of protein interfaces, these state-of-the-art methods do not consider the diversity of interacting partners of the templates. It is important since some proteins display different binding sites depending on their interacting partners, whereas others interact through a single binding site. If, among templates, the same binding site interacts with a diverse set of ligands, it has to be rewarded since it displays a general pattern. One of the main objectives of this thesis is to distinguish protein interfaces considering the interacting partners of the homologues/structurally similar complexes.

- Improving interface prediction by detection of the best homologue:

  While current methods generalise the interaction pattern looking at all possible complexes, detection of the best representative complex can improve the interface prediction. Therefore, we have investigated this idea using the binding site transitivity concept. Given a query protein pair the binding sites of the homologues of the first query protein (QP) are structurally compared against the interfaces of the homologues of the partners of the second QP.

- Re-ranking of docked models using predicted protein interfaces:

  We have investigated the use of our interfaces prediction frameworks in order to rescore docked models. The use of interface knowledge in re-ranking docked conformations depends on the quality of the interface predictor. Therefore, a more accurate interface predictor should be able to improve re-ranking of docked conformations.

- Development of a publicly accessible tool providing interface prediction and re-ranking of docked models:
  This tool will allow the research community not only to perform prediction and ranking, but also to compare their methods with ours. This tool is written in C# for Windows operating system.

1.4 Scientific Contribution

This thesis tackles the essential problems of protein interface prediction and detection of native-like models among docking solutions. The scientific contributions which are based on our novel ideas are summarised as below:

- In chapter 2, we first provide an extensive review of protein interface predictors and discuss the motivations behind these methods. We explore the strengths and weaknesses of the different prediction methods. Then, we provide an overview on docking methods and describe approaches for re-ranking docking conformations. In particular, we look at the usage of predicted interfaces into detection of native-like docked conformations. This chapter provides a comprehensive discussion of the fields of protein interface prediction and docking approaches.

- In chapter 3, we have introduced 3D motifs of structural binding sites for ranking docked conformations and selection of near-native model. A 3D motif is a structural descriptor of a protein binding site, which describes the 3D pattern a specific protein uses to bind to its partner. Creation of a 3D motif is achieved by, first, extracting all homologues complexes of the protein of interest from PDB and, secondly, detecting interacting residues to create the motif. In order to detect native-like docked poses, experimental mutation studies have been applied on a short list of low energy docked poses (Sivasubramanian et al. 2006; Van Dijk, Boelens, et al. 2005). However, since energy ranking is not sufficiently reliable to produce a short list containing a native-like pose, 3D motifs have been used to improve docked poses ranking. As a proof of concept, the 3D

motif was used to investigate the mode of interaction between an antimicrobial peptide and a lipoprotein receptor.

- Although prediction of protein interfaces is a very active field of research, the current state-of-the-art methods do not fully consider the diversity of the interacting partners in their predictions. Therefore, in chapter 4 we have proposed, T-PIP, a protein interface predictor which uses a scoring strategy to detect interface residues by looking at homologous proteins of the QP. The two main elements involved in scoring are: Homology between the QP and its homologues and the diversity between the interacting partners of these homologues. Comparison of T-PIP to the state of the art methods has shown better interface prediction performance.

- Since predicted interface residues have assisted the detection of near-native models, in chapter 4, we have proposed PioDock which uses T-PIP predicted interfaces to re-rank all the docked conformations of two protein chains. Performance of PioDock on standard benchmarks has shown to be superior to the state of the art. Chapter 4 concludes by showing that that detection of the best homologue and mapping its interface residues on the QP could potentially improve interface prediction.

- Finally, in chapter 5 we investigated the detection of best homologue using the binding site transitivity concept. This is achieved by structural interface comparison of the first QP homologues and the binding site of the homologues' partner of the second QP. For a reasonable number of proteins this approach could detect a good quality homologue and therefore improve interface prediction. It is demonstrated that the combination of this strategy with the interface predictor introduced in chapter 4 could significantly improve performance obtained by those individual predictors.

1.5 Thesis Outline

This thesis is divided into six chapters:

This chapter has provided an overview of our research and highlighted our scientific contribution.

Chapter 2 presents an extensive literature review of protein interface predictors followed by a review on docking methods.

Chapter 3 has proposed binding site structural 3D motifs for detection of native-liked docked conformations. This chapter is based on a journal paper published in Protein Peptides and Letters (Esmaielbeiki et al. 2012).

Chapter 4 has proposed a novel framework, T-PioDock, for prediction of a protein complex 3D structure from the structures of its components. T-PioDock first predicts protein interface and then identifies near-native conformations from 3D models that docking software produced by scoring those models using the predicted interfaces. Part of this chapter is based on a work published in Open Access Series in Informatics (OASICS), German Conference on Bioinformatics 2012 (Esmaielbeiki & Nebel 2012). This was further extended and has been accepted 'due to changes' by the BMC Bioinformatics journal. T-PioDock is freely available for public use from http://manorey.net/bioinformatics/wepip/. T-PioDock has also taken part in the latest CAPRI competition, round 28.

Chapter 5 has introduced a framework for protein interface prediction based on detection of the best homologues. A manuscript based on this chapter is in preparation for submission to a bioinformatics journal.

Finally, chapter 6 concludes the undertaken research in this thesis and suggests further research possibilities.

2 Literature Review

2.1 Introduction

This chapter first investigates protein interface predictors and discusses their strengths and limitations (section 2.3). These methods are broadly divided into two groups: (i) intrinsic-based predictors, which use the query protein (QP) sequence or structural features to perform predictions (discussed in section 2.3.1) and (ii) template-based predictors which take advantage of structurally similar proteins to the QP (see section 2.3.2). Then, in section 2.4 docking methods for prediction of protein complexes are investigated. They are also divided into two groups, i.e. template-free (section 2.4.1) and template-based approaches (section 2.4.2). In particular, we discuss the use of predicted interface residues in re-ranking template-free docking conformations (section 2.4.1.4.3). Finally, evaluation metrics used for comparing protein interface predictors and docking methods are presented in section 2.5.

2.2 Overview

Proteins function by interacting with other biological units and therefore, analysis of protein-protein interactions (PPIs) is essential for comprehending cellular processes. Since any alternation in PPI can result in critical diseases, identification of PPI and their 3D complex provides key information for drug design. Experimental techniques such as Y2H (Brückner et al. 2009), phage display (Pande et al. 2010) and affinity purification (Kim et al. 2010) have played an important role in deciphering protein interaction networks. Despite these efforts only 10% of the human interactome

has been experimentally determined (Venkatesan et al. 2008). Moreover, experimental methods produce a large number of false positive (Kuzu et al. 2012; Reš et al. 2005). Consequently, computational methods are required for predicting and verifying protein interactions.

While knowledge of the protein interactions is important to fully elucidate biological processes, 3D structure of protein complexes are essential to understand their atomic interaction pattern. Experimental techniques such as X-ray crystallography and nuclear magnetic resonance are available to produce atomic structure of protein interactions. However, their high costs in term of time and resources have limited their practical use. Since a large number - above 90,000 in August 2013 – of protein-protein complexes are available in the Protein Data Bank (PDB) (Berman et al. 2000) and PQS (Henrick & Thornton 1998), they can be used for modelling other PPIs (Zhang et al. 2012).

One of the main computational methods for predicting complex 3D structures is docking. Docking algorithms can be divided into two groups (Kundrotas et al. 2012), i.e. template-based and template-free docking. With the increase in the number of 3D structures template-based docking has become particularly popular using experimentally determined structures as templates to generate new complexes. Template-based docking is particularly attractive since, unlike template-free docking, its low computational cost makes it suitable for interactome scale predictions. Since not all proteins can be modelled using templates, template-free docking still remains highly important (section 2.4.3).

Template-free docking investigates the whole configuration space along with energy-based cost functions to produce a large set of possible conformations. Although they have been shown to produce native-like poses but identifying them in a large pool of poses is still a challenging task. Therefore, to distinguish these models, docking algorithms include scoring functions which integrate constraints to filter and rank docking models. Scoring functions for re-ranking docking conformations are broadly divided in three groups of knowledge-based potential, machine learning functions and knowledge of predicted interfaces. The first group has shown low correlation to binding affinities (Kastritis & Bonvin 2010) and machine learning methods have not significantly improved upon knowledge-based potential (Zhao et al. 2011). Therefore,

The use of protein interfaces as a constraint to filter docking poses have been the focus of studies in the past (Gottschalk et al. 2004; Huang & Schroeder 2008) and recently (De Vries & Bonvin 2011; Li & Kihara 2012).

In this chapter, first progresses of interface predictor methods are discussed (section 2.3). Second, we review the use of constraints such as predicted interfaces in ranking docking conformations (section 2.4).

2.3 Protein Interface Prediction

Many computational methods have been proposed for identifying interface residues of proteins. Interface predictors are broadly divided into two non-exclusive categories based on their use of protein information. The first approaches are based on the specific features of the protein's sequences and/or structures (intrinsic-based approach). The second ones explore proteins which are either sequentially or structurally related to the QP to study their pattern of interactions (template-based approach).

In sections 2.3.1 and 2.3.2 we discuss intrinsic-based and template-based approaches, respectively.

2.3.1Intrinsic-based Predictors

These predictors use the sequence and/or structure of the query protein to detect potential binding residues. Sequence features are properties of residues in the protein's amino acid sequence, such as composition and propensity of interface residues (Ofran & Rost 2003b), physico-chemical properties (Ofran & Rost 2007a; Chen & Li 2010), sequence evolutionary conservation (Xue et al. 2011; Pazos et al. 1997) or predicted structural characteristics (Ofran & Rost 2007a), whereas structural features are those associated to the atomic coordinate of proteins, such as secondary structure (Jones & Thornton 1997a; Talavera et al. 2011), solvent-accessible surface area (Šikić et al. 2009; Chung et al. 2005), geometrical shape of the protein surface (Šikić et al. 2009) and crystallographic B-factor (Jones & Thornton 1997a). Some methods use solely proteins sequence and their sequential features (section 2.3.1.1), while others require the 3D structure of the input protein in order to generate structural features (section 2.3.1.2) which may be combined with sequential features.

To benefit from a combination of sequence/structure features instead of focusing on only one, two broad approaches (De Vries & Bonvin 2008) have been introduced (a detailed review on these techniques can be found in (Zhou & Qin 2007)). The first approach, uses a physically relevant parametric function (De Vries & Bonvin 2008) (such as: empirical scoring function (Liang et al. 2006) or PLS regression (Kufareva et al. 2007)) for combining features either linearly or nonlinearly. The second one, uses machine learning techniques such as support vector machine (SVM) (Koike & Takagi 2004; Bradford & Westhead 2005; Chung et al. 2005) , neural networks (NN) (Ofran & Rost 2003b; Fariselli et al. 2002; H Chen & Zhou 2005; Wang et al. 2006), Random Forest (RF) (Šikić et al. 2009; Li et al. 2007; Li et al. 2012) and Bayesian networks (Neuvirth et al. 2004; Bradford et al. 2006).

The prediction output format of these methods can also be divided into two groups: patch-based or residue-based approaches. In the first one, continuous patches of residues are predicted as interface. In the second group, which is the most common, a list of residues which are not necessary continuous on the protein surface is predicted as interface. Although many patch-based interface predictors methods (Bradford & Westhead 2005; Bradford et al. 2006; Pettit et al. 2007) have been introduced, comparative study (De Vries & Bonvin 2008) between patch- and residue-based approaches shows a low sensitivity in patch-based predictions (around 20%). Therefore, more attention has been paid to residue-based prediction.

Section 2.3.1.1 first discusses interface prediction methods which require only the QP sequence to perform intrinsic-based prediction and in section 2.3.1.2, we investigate those which are based on the QP's 3D structure. For the sake of being exhaustive, we will also mention methods which aim at predicting hot-spots.

2.3.1.1 Sequence-based Predictors

A small number of methods predict interfaces using solely the query protein sequence information with no need to use its 3D structure. The main idea behind these methods is that their approach can be applied to a wider range of proteins; but the main difficulty is to achieve high specificity. Therefore, to improve prediction, they take advantage of properties such as evolutionary information or predicted structural features.

In section 2.3.1.1.1, we first discuss methods which use the properties of the residues in the QP's sequence to make predictions. Then, in section 2.3.1.1.2 we discuss those sequence-based methods which use evolutionary information derived from comparing a QP sequence to homologous sequences. Finally, in 2.3.1.1.3 we discuss methods which combine evolutionary information with residue intrinsic information to improve prediction performance.

2.3.1.1.1 Intrinsic Features-based Predictors

Methods in this category use the QP sequence and study its sequential features in order to detect potential interfaces. One of the early studies using only sequence information (Gallet et al. 2000) detects binding sites (called "receptor-binding domains" (RBDs)) by analysing hydrophobicity distribution of the protein sequence. To achieve this, a sliding window of N residues is moved along the stretch of the protein sequence. In each window, the mean hydrophobic moment and mean hydrophobicity is calculated to detect the RBD. The concerning points about this method are the following. First, the performance of this method depends highly on the dataset used for evaluation (Sensitivity ranging between 59 to 80%) (Fernández-Recio 2011). Second, false positive and false negative are not discussed (Ofran & Rost 2003b). Third, studies have shown that hydrophobicity alone is not a sufficient characteristic to distinguish protein interfaces (Koike & Takagi 2004; Jones & Thornton 1997a).

In contrast, Ofran and Rost's (Ofran & Rost 2003b) work increased specificity by using amino-acid composition. They have previously shown that residues at the interface have a totally different composition than others (Ofran & Rost 2003a). Therefore, this information was used to train a NN, where a window of size 9 traversed across the protein sequences of a training set comprising 222 complexes. If the central residue of the window belonged to the interface, it was labelled as interaction site. This framework was only used on transient interaction between two non-identical chains and was tested on 111 complexes. Although they outperformed the RBD method in term of accuracy (62%), sensitivity was extremely low, down to 0.5% (Ofran & Rost 2003b).

However, they observed that interface residues tend to form a cluster on the protein sequence which supports the claim that prediction can be done using sequence information alone. Based on this observation, Yan et al. (Yan et al. 2004) attempted to

improve specificity by introducing a two-stage classifier using SVM and Bayesian network. In the first stage, the SVM was trained using the same approach as Ofran and Rost (Ofran & Rost 2003b). This SVM accepts a vector of 9 residues as input and outputs a Boolean which indicates if the central residue and its neighbours belong to the interface. In the second stage, the Bayesian network classifier is trained to predict the central residue of the window as interface/non-interface based on the class label of its neighbours. They have proven that the use of the two-stage methods results in better predictions than the SVM alone (~27% in specificity). The main difference between this method and the previous sequence-only methods is that when defining interface on the training set, instead of classifying all residues as interface or non-interface, only *surface* residues are considered. On a dataset of 77 proteins, the two-stage classifier achieved 58% specificity, 39% sensitivity and 30% MCC while RBDs displayed 31%, 37% and -2%, specificity, sensitivity and MCC, respectively.

The main limitation of these methods is that their specificity is low. It should be noted that although Yan et al. (Yan et al. 2004) has reached 58% specificity, but this number is biased due to the nature of their training set: unequal numbers of positive and negative samples affect specificity score (Baldi et al. 2000). For example, if a classifier is trained on a training set with 80% non-interface residues, then if the classifier predicts all residues as non-interface it will achieve 80% accuracy. To further improve sequence-based predictors, Yan et al. and Ofran and Rost have suggested the integration of evolutionary information and predicted structural features into their frameworks.

2.3.1.1.2 Sequence Evolutionary-based Predictors

Methods have been developed which generate interfaces using only evolutionary conservation information (see (Lovell & Robertson 2010) for a review). These methods are based on the concept that interface residues are more conserved than the rest of the protein surface (Pazos et al. 1997). This was investigated in one of the early studies (Pazos et al. 1997) which detected potential PPIs by investigating correlated mutations in Multiple Sequence Alignments (MSAs): it is based on the intuitive assumption that residues which interact are under co-evolution pressure.

Later on, they introduced TreeDet (del Sol Mesa et al. 2003) which uses evolutionary information and sequence conservation to detect functionally important

residues on proteins. This is achieved by dividing protein families into subfamilies. These conserved residues named, "specificity determining positions" can be used to detect protein-protein binding sites (Rausell et al. 2010). A recent study, Xue et al. (Xue et al. 2011) performed a study on more than 300,000 MSAs to analyse the extent that interface conservation in homologous proteins can be exploit to provide a correct prediction. Based on this, they developed HomPPI a sequence homology-based method where proteins homologous to a QP are divided into three zones: Safe, Twilight and Dark Zone. They range from homologues with highly conserved interfaces in the Safe zone to moderate and poor conservation in the Twilight and Dark zones. HomPPI attempts first to predict interfaces by retrieving homologues from the Safe zone using BLASTP. If such homologous proteins are not available, the Dark and Twilight zones are then explored.

They present two variants of HomPPI: 1) NPS-HomPPI: which predicts the interfaces of a QP without the knowledge of its interacting partner and 2) PS-HomPPI: which predicts the interfaces of the QP knowing its interacting partner. In NPS-HomPPI, the homologous are retrieved using BLASTP. If at least one homologue is found in the Safe Zone that will be used for prediction otherwise Twilight and Dark zones are investigated. If no homologue is found in any zones NPS-HomPPI fails to make a prediction. For each homologue generated by BLASTP, an Interface Conservation (IC) score is calculated. IC is based on a combination of BLASTP statistics generated from the alignment of the homologous protein with the QP. At most, the top k homologous proteins (k=10 in their study) with the highest IC score are kept and a MSA containing the QP is created. Interface and non-interface residues of the homologous are marked on the MSA. Then, for each position of the QP on the MSA, a prediction score is calculated. This is achieved simply by dividing the number of interface residues by non-interface residues suggested by the homologous proteins in that position. On the QP, positions with prediction score ≥ 0.5 are predicted as interface (an example can be seen in Figure 2.1). In PS-HomPPI, given A as the QP and B as its interacting partner, homo-interologues of A-B are retrieved using BLASTP. This means only complexes which involve homologues of both A and B are used for predicting the interfaces. PS-HomPPI only investigates homologous in the Safe and Twilight Zones and the scoring and prediction strategy is similar to NPS-HomPPI. Comparison on a

transient dataset shows PS-HomPPI outperforms NPS-HomPPI, which indicates that knowledge of the QP's interacting partner improves the interface prediction.

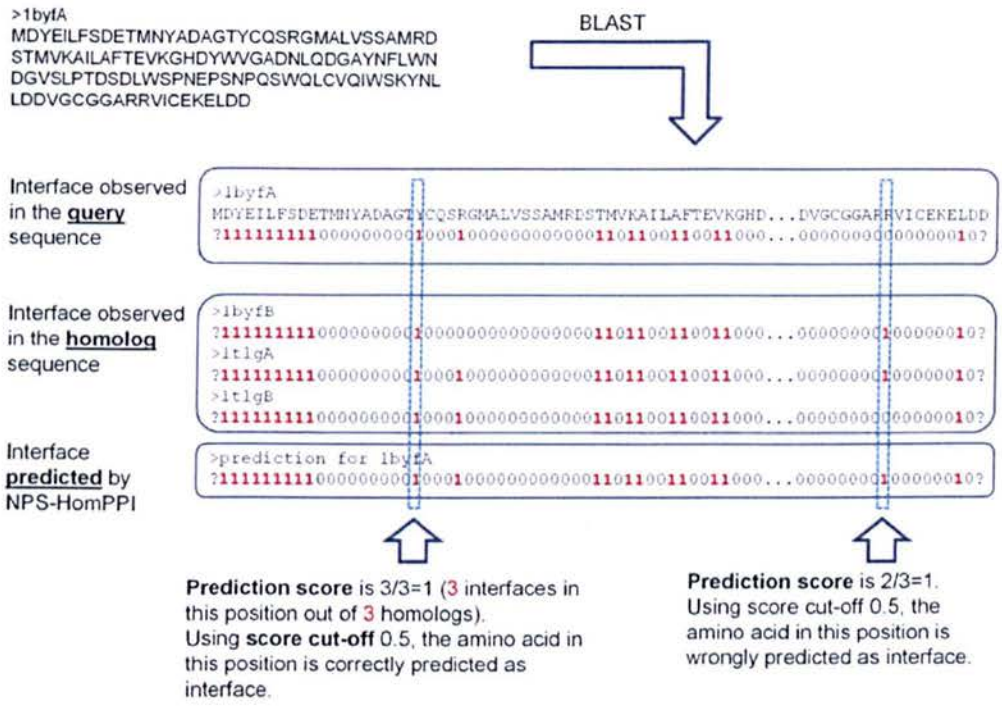


Figure 2.1: An example of NPS-HomPPI . Homologues of the QP, 1byf chain A, are generated using BLAST. Only 3 homologues have been detected in the Safe and Twilight Zone. Observed interfaces of homologues are mapped on their sequences. Red 1 and black 0 represents interface and non-interface residues. If majority of residues are marked as interface, that position will be predicted as interface for QP. Taken from (Xue et al. 2011).

2.3.1.1.3 Predictors Combining Evolutionary and Intrinsic Information

Methods in this category use evolutionary information combined with residues intrinsic information. Res et al. (Reš et al. 2005) improved the intrinsic-based work of Ofran and Rost (Ofran & Rost 2003b) by combing it with evolutionary information and using SVM instead of NN. They were the ***first*** people to use evolutionary information along with intrinsic features to perform interface prediction. Their work was inspired by the success of using evolutionary phylogenetic information for predicting functional sites (Lichtarge et al. 1996; Madabushi et al. 2004). They tested their method on 77 interacting protein chains with sequence identity <30% (Yan et al. 2004) and achieved 64% accuracy which is 6% higher than Ofran and Rost's (Ofran & Rost 2003b) on the

same dataset. Another method focused on hetero-complexes by Wang et al. (Wang et al. 2006) trained an SVM using a combination of amino acid evolution conservation and sequence profile. They were calculated using homologous MSA and phylogenetic tree, respectively. Interesting results shows that the prediction performance stays the same using both the combination of these two features or each one on its own. No comparison has been made with other methods.

In addition to evolutionary information, some sequence-based methods add predicted structural information. Ofran et al. introduced ISIS (Ofran & Rost 2007a), an improvement to their previous work (Ofran & Rost 2003b), which trains a NN integrating both evolutionary profiles and knowledge of predicted structural features, i.e. surface accessibility and secondary structure (Ofran & Rost 2007a). The performance of their ISIS reached 61% specificity and 20% sensitivity. This shows that the use of predicted structural information significantly improves the sensitivity which was 0.5% in their previous work. ISIS prediction of few residues (low sensitivity) with high accuracy, suggests the importance of these residues in binding which are referred as hot-spots residues (Fernández-Recio 2011; Ofran & Rost 2007b). Another method which uses predicted structural features is PSIVER (Murakami & Mizuguchi 2010) which for the first time trained a Naïve Bays classifier (NBC) and kernel density estimation (KDE) with two sequence features (position-specific scoring matrix (PSSM) and predicted accessibility). PSSM is a sequence profile which provides a measure of conservation based on the composition of a sequence and of its homologues identified by PSI-BLAST (De Vries & Bonvin 2008). PSIVER is motivated by the fact that, since in PPI the correlations between different features are not well understood, NBC, which ignores dependency between the input parameters, may perform better in classifications. PSIVER, outperforms ISIS with a F-measure of 32.5 %, and 26.3%, respectively.

To further advance sequence-based methods, work has been done to improve the training sets of the machine learning classifiers. The reason behind this is that the numbers of positive samples (interacting residues) are smaller than the negative samples which create an imbalance between training sets. Therefore, to deal with this problem, Chen and Jeong (Chen & Jeong 2009) introduced a RF based classifier which uses a much larger number of features (1050 features per residue) extracted from physicochemical property, evolutionary conservation score, amino acid distances, and

PSSM. They argue that since the properties which highlight the interfaces are not fully understood, taking a large number of features can improve the prediction. Comparison with Wang et al. (Wang et al. 2006) and Yan et al. (Yan et al. 2004) shows an improvement in prediction while dealing with unbalanced data in their training set (Based on a leave-one-out ROC Curve for specificity rate of 70%, the sensitivities of Yan's, Wang's and Chen and Jeong are 30%, 39% and 73%, respectively). Another approach by Chen and Li (Chen & Li 2010) uses ten-SVMs trained on the residues' hydrophobic and evolutionary information of different subsets of hetero-complexes proteins. If the prediction output of all the SVMs for a specific residue is above a threshold, that residue is considered as belonging to the interface. To deal with training imbalance, they create several subsets of positive and negative training data of similar size having no overlap. Recently, they improved their work by introducing PPI-OD (P. Chen et al. 2012) which targets removing interface and non-interface residues outliers from the training data. Comparisons with their previous work (Chen & Li 2010), PPI-OD improves the MCC and F1 by 4.4% and 3.6%, respectively. Wang et al. introduced another predictor (Wang et al. 2012) using an SVM trained with a feature vector of 17 amino acid generated from AAindex database (Kawashima & Kanehisa 2000). They use a ten-fold cross-validation strategy and to overcome training imbalance, negative samples were selected randomly based on the size of positive samples. On a dataset of hetero-complexes used in their previous study (Wang et al. 2006) they achieved a sensitivity and specificity of 57.9% and 65.0%, respectively. While in their previous work they achieved globally lower performance - 49.7% specificity - despite better sensitivity, i.e. 66.3% sensitivity (Wang et al. 2006). Unfortunately, no comparison with other methods is provided in their study.

Another trend in sequence only prediction takes into account knowledge of the interacting partner. Ahmad and Mizuguchi (Ahmad & Mizuguchi 2011) trained a two-stage NN in which, instead of a single residue, a residue pair PSSM and propensity are considered. Comparison to PSIVER shows a larger area under the RocCurve (63.8 to 54.1). For large scale applications, PIPE-Sites (Amos-Binks et al. 2011) was introduced. PIPE-Sites not only predicts binding site, but also predicts if two QPs interact. PIPE-Sites (Amos-Binks et al. 2011) uses a sliding sequence window which traverses both protein sequences separately. For each pair of windows, PIPE-Sites searches for their

co-occurrence in a sequence database of known protein-protein interactions. Using all the pairs of windows and their co-occurrence, a 3D landscape is generated where high peaks correspondents to the interacting sub-sequences. If a landscape is showing a flat distribution PIPE-Sites will assume that the QPs do not interact.

2.3.1.2 Structure-based Predictors

Methods in this group require the 3D structure of the QP in order to perform predictions. A small number of methods mainly use evolutionary information of the sequences to detect interfaces while the 3D structure is required in the final step to detect potential binding patches (see section 2.3.1.2.1). The majority of structure-based methods require the 3D structure to export structural features used to detect interfaces. These methods can be combined with sequential information (see section 2.3.1.2.2). Another trend in this group combines docking with structural features to generate models (see section 2.3.1.2.3).

2.3.1.2.1 3D Evolutionary-based Predictors

Methods in this group infer interfaces from homologous proteins by analysing evolutionary information and sequence conservations. The 3D structure of the QP is required at the last stage for clustering the conserved residues in a final patch of interface residues. Similar to methods in section 2.3.1.1.3, these methods are inspired by the success of using evolutionary phylogenetic in (Lichtarge et al. 1996). Lichtarge et al introduced Evolutionary Trace (ET) (Lichtarge et al. 1996) which uses a phylogenetic tree built from a Multiple Sequence Alignment of homologous proteins to calculate residue conservation scores. Then, residues with the best scores are mapped on the QP's 3D structure and clustered to infer binding patches. ConSurf (Ashkenazy et al. 2010) is a variant of the ET method, which takes advantage of empirical Bayesian and maximum likelihood to calculate the conservation score. Joint Evolutionary Trace (JET) (Engelen et al. 2009) is an extension of ET, which is designed for large scale applications and is able to make predictions even with weak signals. JET differs from the ET method by improving sequence alignment and consideration of residues' conservation and physio-chemical properties in the clustering stage. Comparison on Huang dataset (Caffrey et al.

2004), shows 34.2% sensitivity and 85.1% specificity for ET and 39.8% sensitivity and 86.9% specificity for JET.

2.3.1.2.2 Predictors Combining Evolutionary and 3D Intrinsic Information

Methods in this category use 3D structural features or their combination with sequence properties to predict interfaces. The story of analysing contact residue goes back to 1975 with the pivotal work of Chothia & Janin (Chothia & Janin 1975) on a small number of proteins. They discovered that hydrophobicity is a key element to stabilising protein-protein interactions. Later, Jones and Thornton (Jones & Thornton 1997b), studied a larger dataset, investigating the surface of protein complexes for six parameters; solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area. These properties where then used to detect patches responsible for protein interactions with 66% (39 out of 59) success rate (Jones & Thornton 1997a; Jones & Thornton 1997b). A later study, SHARP2 (Murakami & Jones 2006) used a parametric combination of the same features to provide a fast and robust server for Jones & Thornton's (Jones & Thornton 1997b) work.

To investigate the importance of 3D information for predicting interface residues, Koike and Takagi (Koike & Takagi 2004) trained a SVM using both sequence profile only and a combination of sequence and structure data such as accessible surface area (ASA) and patch flatness. Using sequence-only information they achieved precision and recall of ~40% and 39%, respectively, while a combination with structure information improved significantly performance displaying precision and recall of ~56% and 50%, respectively. Also, Sikić et al (Šikić et al. 2009) integrated both sequence information only and a combination of sequence and structure data (such as accessible surface area, protrusion and depth index) into Random Forest. Their sequence-based prediction scheme achieved a precision of 84% with a 26% recall and an F-measure of 40%, whereas using structure, the overall prediction performance increased to, respectively, 76%, 38% and 51%. Moreover, the overall precision-recall curve is higher when 3D information is used. Those results were confirmed by (Zhou & Qin 2007) who argued that, since in sequence-based methods non-interface residues

cannot be trivially eliminated, fewer interface residues information is available for prediction.

As presented above usage of structural properties improves sequence-based predictors. However, there has not been a single property which could completely discriminate interface residues form other residues. Therefore, predictors have based their predictions on combining multiple input properties from several residues to predict if a residue belongs to the interface. Therefore, structure-based methods differ from one another depending on what features they use and what method they use to combine them. These predictors are broadly divided in two groups (Zhou & Qin 2007) of (i) numerical value and (ii) probabilistic based predictors where they both require training set. Below each of these groups and their techniques to perform prediction is discussed.

2.3.1.2.2.1 Numerical Value-based Predictors

In this group for an input data x_i of residue i, which is the residue under study, and its spatially neighbouring residues j, with values x_j, a scoring function is defined as:

$$S_i = f(x_i, x_{j \epsilon n_i}, c)$$

Where x represents the structural (and sequential) features defined for each residue and c is a number of coefficients that is optimised through training. Using this function i is determined as interface or non-interface based on a numerical value; for example i will be considered as interface if S_i is above a threshold (Zhou & Qin 2007). Numerical value-based methods for interface prediction are discussed below.

To calculate S_i some methods (Li et al. 2006; Kufareva et al. 2007) use a linear combination of input data. PIER (Kufareva et al. 2007) uses linear regression (partial least squares (PLS)) of statistical properties of protein surface at atomic level. They also investigated the prediction performance by adding the evolutionary information to the properties set. They concluded that this has only 7-10% contribution while atomic features have 90-93% contribution. Although linear combination is a really simple method, general predictors using this approach have produced lower performance in comparison to other predictors (Zhou & Qin 2007).

Therefore to improve prediction, a set of numerical value-based methods use scoring functions which are inspired by empirical energy functions. These scoring functions model contribution terms of different input data and therefore allow better

discrimination between interface and non-interface properties in comparison to linear methods. InterProSurf (Negi et al. 2007) uses interface residue propensity derived from a dataset of structures along with ASA in a parametric scoring function to make predictions. Similarly, WHISCY (De Vries et al. 2006) predicts interfaces using a scoring function with parametric combination of interface propensity and surface conservation. Taking advantage of the fact that interface residues have higher side-chain energy (Cole & Warwicker 2002), in addition to interface propensity and residue conservation, Liang et al. (Liang et al. 2006) introduced an empirical energy function based on a linear combination of side chain energy score into their system, PINUP. One drawback of these scoring functions is that due to their complicated terms, which make them a better predictor than linear combination, they require deep physical insight (Zhou & Qin 2007).

Therefore, machine learning technique was introduced which allow a simpler strategy for combining input data. One of the widely used numerical value-based machine leaning methods is SVM. In SVM non-linear independent input data are mapped to a feature space and a hyperplane is fitted to optimally separate the interface and non-interface data. Figure 2.2 shows an example in which SVM is used to separate two classes of data which are shown with red circles and blue rectangles. The aim is to find a hyperplane (Solid line in Figure 2.2) which divides these points as far as possible. In other words SVM wants to increase the margins. The data points which are located on the margin are called support vectors (shown by numbers 1 to 5 in Figure 2.2.A). Therefore, in Figure 2.2.B among the proposed hyperplanes (i.e. H1, H2 and H3), H1 is the best because H3 does not separate accurately between the two classes and although H2 does, its associated margin is smaller than H1's. Using this concept, Bradford and Westhead (Bradford & Westhead 2005) proposed PPI-Pred which is a patch-based SVM based method trained on seven surface properties: shape index and curvedness, residues conservation score, electrostatic potentials, hydrophobicity, interface propensities and Solvent ASA. Comparing PPI-Pred with the patch-based method of Jones & Thornton (Jones & Thornton 1997b), on a dataset of 47 proteins from (Jones & Thornton 1997b), shows a sensitivity of 72% and 64% for both methods, respectively. Another advantage of PPI-Pred is that it can be used for both obligate and transient proteins, whereas Jones & Thornton's (Jones & Thornton 1997b) requires different

scoring functions to make prediction for proteins from different types. At the same time a *residue-based* predictor, Bordner and Abagyan (Bordner & Abagyan 2005) employed an SVM-based predictor trained with evolutionary conservation information and local surface properties (hydrophobicity, solvation Energy, and ASA). To generate the interface propensity the above mentioned methods focused on one type of complex since different complexes have different propensities. While Dong et al. (Dong et al. 2007) calculates propensity using several complex types (homo-permanent, homo-transient, hetero-permanent, and hetero-transient complexes) and along with sequence profile, evolutionary information and accessible surface area trains a SVM-based predictor. They observed that usage of propensity especially through binary profile generated from PSI-BLAST has a significant positive impact on prediction performance especially if they are combined with evolutionary information.

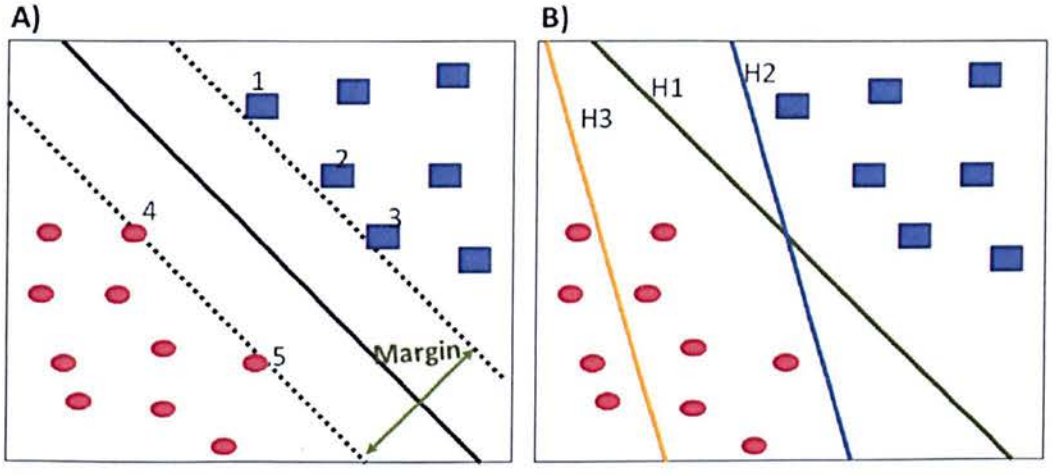


Figure 2.2: Concept behind Support Vector Machine. Classes of data are shown with red circles and blue rectangles. Hyperplanes are shown in solid lines. A: Support Vectors are shown by numbers from 1 to5. Margin is represented by the green line. B: three hyperplanes are shown in solid lines.

Comparisons of Bordner and Abagyan (Bordner & Abagyan 2005) SVM-based method with PIER (Kufareva et al. 2007) which is a linear value-based method has shown that although PIER has a better precision by ~53% but its recalls drops by ~23%. Therefore, predictors with a balance in their precision and recall are required. Since too many features can be the cause of this imbalance, Nguyen and Rajapakse (Nguyen & Rajapakse 2006) proposed a SVM-based classifier only using two features: ASA and PSI-BLAST sequence profile . Combination of the 2 features has better prediction

performance than using each one individually. In accuracy, specificity and sensitivity they achieve 2.8%, 18%, 29% higher than Yan et al. (Yan et al. 2004) method. To further improve the prediction, Li et al. (Li et al. 2008) introduced a novel neighbouring profile using both sequential and spatial neighbours to train an SVM-based predictor. For each residue eight features were used: physicochemical characteristics, hydrophobic index, relative accessible surface area, secondary structure, sequence conservation, side-chain environment, sequence distance, and spatial distance. Also, they classed each residue into four groups: interior, core interface, rim interface, and non-interface. An interaction area consists of a core interface surrounded by rim interfaces (Bahadur et al. 2003). A core interface is defined as a residue which is accessible to solvent in unbound form while in a bound form it is zero ASA. Interaction residues which are still solvent accessible in the bound form are rim interface. Composition of a core interface area resembles the protein interior (hydrophobic core) while the rim region composition is similar to the protein surface (Higa & Tozzi 2008). Li et al. SVM classifier was only trained using the core residues. As discussed before, one of the problems with SVM is that its results are impacted by the imbalance of positive and negative data in the training set. To solve this problem, Deng et al. (Deng et al. 2009) used an ensemble of bootstrap re-sampling, SVM-based fusion classifiers and weighted voting strategy. This allowed them to use a wider selection of sequential and structural features providing 8 different feature spaces for training. They achieve sensitivity and specificity of 76% and 78%, respectively. While on the same dataset, Wang et al. (Wang et al. 2006) shows sensitivity and specificity of 69% and 66%, respectively. In general SVM-based methods have outperformed the scoring function based predictors, but there is still room for improvement. Especially that SVM framework performs a non-parametric combination of input data which are assumed to be independent (Browne et al. 2010; Lo et al. 2005). Therefore, methods introduced NN in their predictions where input data are linearly combined and coefficients are assigned to inputs based on the training sets. In addition, NN has the ability to find complex pattern even among large number of input data (Fariselli et al. 2003).

Therefore, Zhou et al. (Zhou & Shan 2001) introduced PPISP which makes predictions from a 3D structure using a NN that is trained with protein's surface sequence profiles and solvent accessibility of neighbouring residues. The main

advantage of this method is its insensitivity to structural differences by achieving similar performance with bound and unbound proteins. Independently, Fariselli et al. (Fariselli et al. 2002) also proposed a NN based system to learn from exposed residues on protein surfaces which are involved in interaction. Although PPISP and Fariselli et al. methods used different training sets, homology cut-offs and surface definitions, they achieved similar accuracy around 70%. PPISP was further improved using a larger training data set which led to increased accuracy, 80%, at the cost of lower coverage, -17% (H Chen & Zhou 2005). Over and under prediction of protein interface was the main concern of PPISP therefore it was extended to Cons-PPISP (H Chen & Zhou 2005) which performed a consensus prediction from multiple NN models with various accuracy and coverage. Cons-PPISP showed an improvement of 3–8 percentage points in accuracy in comparison to the individual models. A recent method (Y. Chen et al. 2012) uses radial basis function (RBF) neural networks using sequence profiles, entropy, relative entropy, conservation weight, accessible surface area and sequence variability features. Six different sliding windows of size 1, 3, 5, 7, 9 and 11 capturing these features for each amino acid is created. Then, six RBF-NNs are trained with these windows providing six groups of predictions. These predictions are then integrated using decision fusion (DF) and Genetic Algorithm based Selective Ensemble (GASEN), to make the final prediction. Using RBF-NN provides better results than NN and SVM but no comparison results with other methods are presented.

Some studies have used NN to investigate the impact of specific features in protein binding site prediction. Porollo and Meller (Porollo & Meller 2006), studied *relative solvent accessibility (RSA)* as a fingerprint for detection of interface residues and comparison shows its significance to other features such as evolutionary conservation, physicochemical characteristics and features used by other methods. Therefore, they introduced SPPIDER which combines RSA with 19 other features (which the combination achieved best performance on the training data) with a SVM and a NN. Comparison with Bordner and Abagyan (Bordner & Abagyan 2005) and PPI-Pred (Bradford & Westhead 2005) shows that SPIDDER achieves better performance (13% higher Matthews correlation coefficients (MCC) than other methods). Similarly, Patch Finder Plus (Stawiski et al. 2003) focused on detection of large *electrostatic* patches of proteins using residue conservation, frequency,

composition, ASA, surface concavity and H-bond potentials. The main aim was to detect DNA-binding areas but the method can also be used for protein binding site prediction. An important issue regarding data processing is that some input data might be missing or redundant (Lin et al. 2004). Therefore methods have used RF which is a classification method and has shown not only to deal with missing data but also handle heterogeneous input data (Browne et al. 2010). VORFFIP (Segura et al. 2011) uses a 2-step RF ensemble using residue features such as structural, energy terms, sequence conservation, and crystallographic B-factors. That predictor relies on a novel definition of residue environment using Voronoi Diagrams (Figure 2.3). VORFFIP shows better performance in comparison to SPIDER (Porollo & Meller 2006), WHISCY (De Vries et al. 2006) and Sikić et al. (Šikić et al. 2009) in which residue environments are defined by sequence sliding window or Euclidian distance. Recently, they expanded their framework to Multi-VORFFIP (Segura et al. 2012), which aims at detecting protein-, peptide-, DNA- and RNA binding sites on a query protein.

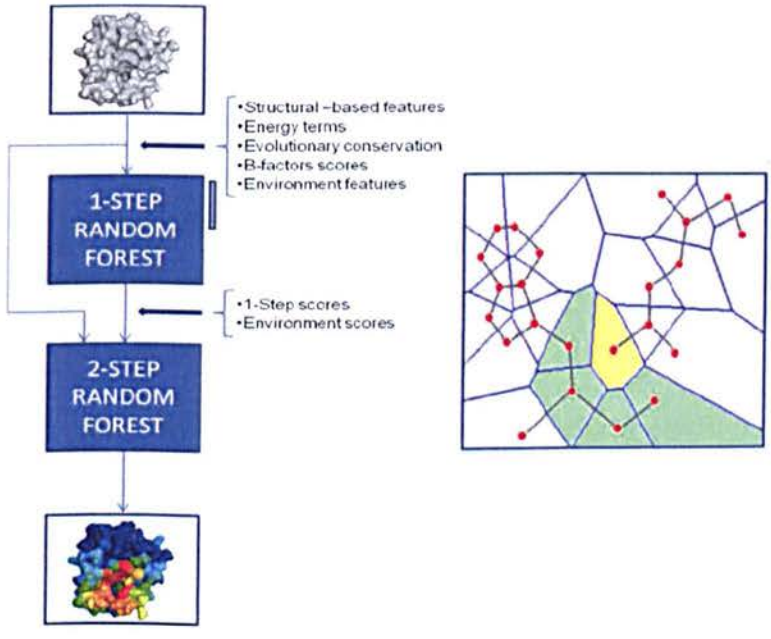


Figure 2.3: VORFFIP prediction pipeline and an example of Voronoi Diagram. The 2-step approach of VORFFIP is displayed on left, on right the Voronoi Diagram of two neighbouring residues are shown. Red dots display heavy atoms and coloured cells show atoms which are interacting with another atom of the neighbouring residue. Taken from (Blas et al. 2012).

A larger set of input data generated only from the 3D structure was used in a recent study (Qiu & Wang 2012) which uses RF along with 28 properties (such as:

solvation energy, hydrophobicity, depth and protrusion index and B-factor). They perform a residues-based prediction followed by a clustering to produce patch-based prediction. Comparison with other patch-based methods ((Bradford & Westhead 2005; Bradford et al. 2006; Higa & Tozzi 2008)) shows an improved success rate. Also, the residue-based method shows improvements in terms of sensitivity–specificity relation: for example, with a specificity rate of 70%, they achieve 78% sensitivity while Yan (Yan et al. 2004),Wang (Wang et al. 2006), Chen and Jeong (Chen & Jeong 2009) are 30%, 39% and 73% , respectively. Independently, Li et al. (Li et al. 2012) developed an RF algorithm with a Minimum Redundancy Maximal Relevance (mRMR) method followed by incremental feature selection (IFS). They use physicochemical/biochemical properties, sequence conservation, residual disorder, secondary structure and solvent accessibility along with five 3D features (Protrusion index, depth index (DPX), ASA, molecular surface area and surface curvature (SC)). They discovered that integrating the 3D information improves the prediction and DPX and SC contribute the most. Comparison to Sikić et al (Šikić et al. 2009) shows an improvement in precision, specificity, accuracy by approximately 1%, 3% and 2%, respectively.

All the above mentioned numerical value based methods classify an input in a binary manner as interface or non-interface. Although usage of SVM, NN and RF has improved the prediction performance over linear regression and scoring function-based methods, at the same time transparency is lost in the prediction (Zhou & Qin 2007).

2.3.1.2.2.2 Probabilistic-based Predictors

The aim of the probabilistic-based methods is to find the conditional probability $p(s|x_1, \ldots, x_k)$ of s being interface or non-interface, where i is the residue under study with x_1 to x_k properties. If $p(s = interface|x_1, \ldots, x_k)$ is above a certain threshold then s is considered as interface. Conditional probability can be generated from the training sets using several methods such as: naïve Bayesian, Bayesian Network, Hidden Markov Model and Conditional Random Forest.

Neuvirth et al. (Neuvirth et al. 2004) studied the properties of the transient protein–protein hetero-complexes interfaces both in bound and unbound forms. They trained a patch-based naive Bayesian, ProMate, with a combination of different properties such as evolutionary conservation, secondary information, hydrophobicity,

chemical component, geometric properties and information from relevant crystal structures. For 70% of the proteins ProMate was able to correctly predict the location of the binding site (Neuvirth et al. 2004). A prediction is considered as correct if more than half of the predicted interface of a continuous patch is an interface in ground truth. Comparison of ProMate with the SVM-based predictor of Li et al. (Li et al. 2008), on a dataset of 50 proteins, shows that Li et al. perform better than Promate, with 60.7% sensitivity and 41.9% specificity while ProMate resulted in 9.9% sensitivity and 28.1% specificity. While ProMate assumes that input data are independent, Bradford et al. (Bradford et al. 2006) were the first to introduce dependency between data using a Bayesian networks trained on surface patches properties. In a leave-one-out cross validation the Bayes classifier achieves a success rate of 82% in comparison to their SVM-based approach which was 76%. This patch-based framework can handle missing data and even without evolutionary information it still achieves a performance comparable to the original Bayes classifier. Another Bayesian classifier-based method using core and rim interface definition was introduced by Higa and Tozzi (Higa & Tozzi 2008). They use a two-stage Bayesian classifier with 28 chemical and structural properties where the first stage detects the core residues and the second stage expands the core residues and detect rim residues around the core. Their patch-based predictor made successful prediction in 82.1% of the cases which is similar to Bradford et al. (Bradford et al. 2006) Bayesian method.

Li et al. (Li et al. 2007) argues that the pervious methods which look at the interface prediction as a classification problem, do not consider the relationship between interface and non-interface of neighbouring residues. To overcome this problem they use sequence labelling using conditional random fields (CRFs) trained with sequence profile and residue accessible surface area. Comparison with NN and SVM methods shows CRF-based methods performs better. Also adding the residue conservation feature to the CRF-based method did not change the results. To improve the performance further, for the sequential labelling task they used hidden Markov support vector machine (HM-SVM) (Liu et al. 2009) and comparing to their CRFs-based method they achieved better performance. On a mixed data set of hetero-complexes and homo-complexes, the F1 score for the NN, SVM, CRF and HM-SVM base methods are 44.7%, 48.1%, 49.9% and 51.0%, respectively. Since HM-SVM has shown the best

performance, recently Savojardo et al. (Savojardo et al. 2012) has introduced ISPRED2. This method uses relative solvent accessibility (RSA) used by SPPIDER (Porollo & Meller 2006), with evolutionary information (PSSM) to make predictions. Comparison to Wang et al. (Wang et al. 2006), Nguyen and Rajapakse (Nguyen & Rajapakse 2006), Deng et al. (Deng et al. 2009), Liu et al. (Liu et al. 2009) shows an increase of 4%, 19%, 2% and 15% in F1 measure, respectively.

The different approaches for combining interface properties into one prediction framework, was discussed above. Each one has their pros and cons and generally they do not significantly improve each other's performance. Since combinations of classifiers tend to perform better than individual ones, as it was shown for protein structure predictions (Huiling Chen & Zhou 2005), a few interface predictor methods have integrated interface predictors into one meta framework.

PI2PE (Tjong et al. 2007) combines cons-PPISP (H Chen & Zhou 2005), DNA-Interaction Site Prediction from a List of Adjacent Residues (DISPLAR) (Tjong & Zhou 2007) and Weighted Ensemble Solvent Accessibility (WESA) (Huiling Chen & Zhou 2005). WESA itself is a meta-predictor of ASA using protein sequence. Unfortunately, there are no results available on the performance of PI2PE in comparison to the individual methods. While in PI2PE only one protein inetrafce predictor was used, to benefit from more interface predictors meta-PPISP (Qin & Zhou 2007b), combines the scores of PINUP, Cons-PPISP and ProMate using linear regression analysis. As expected it outperformed individual methods by increasing their accuracy by an additional 4.8 to 18.2% points. Similar experiment conducted by Zhou et al. (Zhou & Qin 2007) using Docking Benchmark 2.0 (Mintseris et al. 2005)confirmed the superiority of Meta-PPISP over PINUP (Liang et al. 2006), ProMate (Neuvirth et al. 2004) and Cons-PPISP (H Chen & Zhou 2005) with accuracy of 50, 48, 38 and 36%, respectively. On the same dataset the accuracy for PPI-Pred and SPIDDER were 36% and 38%, respectively (Zhou & Qin 2007). In addition to PINUP and ProMate, MetaPPI (Huang & Schroeder 2008) combined PPI-Pred, PPISP, and SPIDER into one framework. Their main aim was to study usage of meta interface prediction for ranking docking models. MetaPPI improved success rates by 15% in comparison to the best individual predictors. While meta predictors are an easy way to combine several

predictors, one should consider the contribution of each individual method when analysing the results (Fernández-Recio 2011).

The aim of all the above mentioned numerical value and probabilistic based methods is to combine various features to distinguish interface and non-interface residues. However, as mentioned by Neuvirth et al. (Neuvirth et al. 2007) the combination of features are informative if they are orthogonal. This means that merging features can result in a better predictor once they are less correlated. That is the reason why adding new properties to already available methods have little impact on the their performance; because combined features are not orthogonal enough (De Vries & Bonvin 2008). Therefore, interface prediction has reached saturation where combination of the same features cannot improve the predictions. Apart from that, classifiers performance and robustness increases when a smaller set of complementary properties are used instead of large amount of dependent noisy features (Zellner et al. 2012; Hermes & Buhmann 2000). Therefore, a recent study introduces Pres-Cont (Zellner et al. 2012) a robust SVM-based predictor based on ASA, hydrophobicity, conservation, and the local environment of each surface residues properties. Using only four features Pres-Cont performs as good as the SPPIDER, ProMate, and meta-PPISP, which are more complicated. These methods perform differently depending on the protein type while Pres-Cont prediction quality remains the same for both permanent and transient complexes.

2.3.1.2.3 3D Docking Predictors

A different approach uses protein-protein docking to generate potential interfaces. Since docking methods have shown to generate native-like models among their solution space, they can be used to study the interface used for creating complexes. Therefore, Fernández et al. (Fernández-Recio et al. 2004) uses the 100 lowest-energy docking models to compute a normalized interface propensity score for identifying interface residues. Performance of this method on 21 unbound PPI has a positive predicted value of 81%, while the sensitivity is quite low. But since the accuracy is high, these results suggest that the predicted interfaces are hot-spots (Fernández-Recio 2011). Since desolvation has been shown to be important in protein binding, ODA (docking-based prediction is Optimal Docking Area)(Fernandez-Recio et al. 2005)

based its prediction on finding surface patches of optimal docking desolvation energy (Figure 2.4). ODA picks a set of starting points which are the central coordinates of each residues side-chain. For each staring point different patch sizes are calculated iteratively. Desolvation energy is calculated for each patch until the most energetically favourable patch is detected. On a dataset of non-obligate protein hetero-complexes in 80% of the cases ODA prediction corresponds to real hotspot residues. One of the main limitation of ODA is that it fails in cases where electrostatic plays a more important role than desolvation (Fernández-Recio 2011).

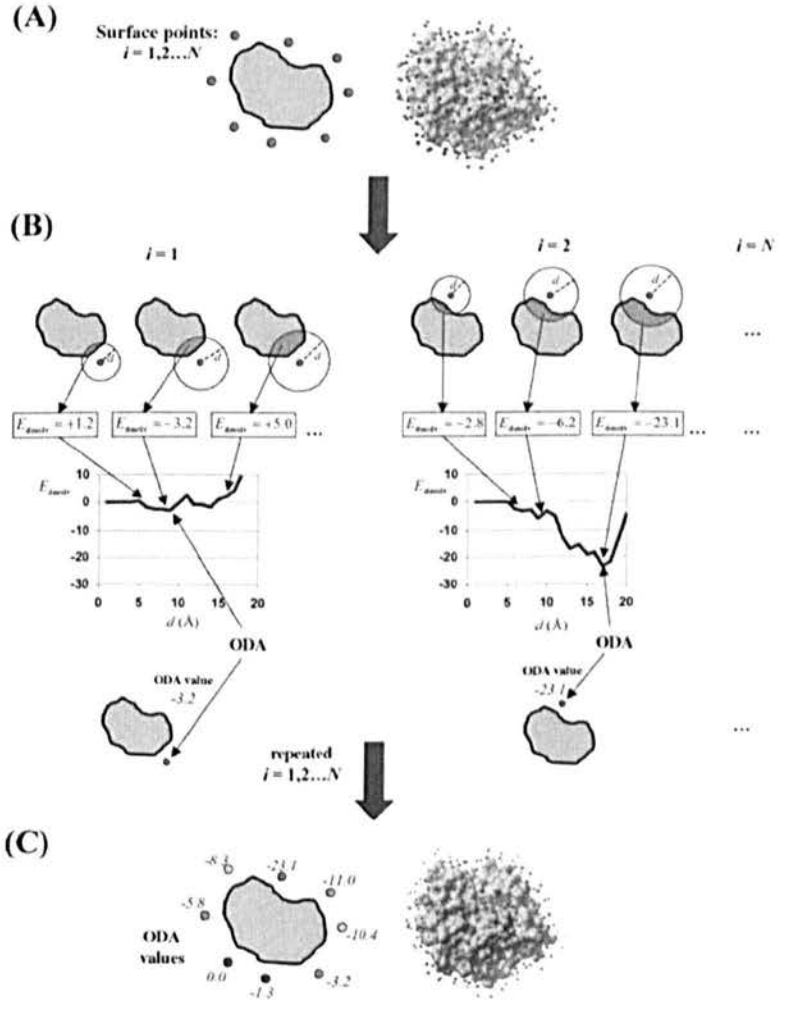


Figure 2.4: Schematic representation of Optimal Docking Area (ODA) predictor. A) Starting points are defined on the protein side-chains and shown around the protein. B) For each starting point different interface patch sizes are calculated. Desolvation energy is calculated for each patch until the most energetically favourable patch is detected. C) Each starting point will be labelled with its best ODA score. Taken from (Fernandez-Recio et al. 2005).

Similar to Fernández et al.'s (Fernández-Recio et al. 2004) approach, DoBi (Guo et al. 2012) first generates docked configurations and then studies their interfaces. DoBi enumerates all the possible configuration of two proteins and scores them based on an Atomic Contact Energy (ACE) function. The best scoring configurations are selected and their binding residues are predicted as interfaces. Comparison between DoBi and Fernández et al.'s (Fernández-Recio et al. 2004) shows DoBi has a better performance by achieving 44.3%, 70.5%, 39.6%, 0.54 for accuracy, coverage, success rate and F-measure, respectively, while Fernández et al. achieves 39.3%, 72.7%, 37.2% and 0.51 for the same measures, respectively. On a benchmark of 41 complexes from DBMK 2.0 and 27 targets of CAPRI, DoBi achieves an F-score of 0.55, which outperforms MetaPPI, meta-PPISP and PPI-Pred (F-scores of 0.35, 0.43 and 0.32, respectively). On another dataset comprising 57 non-homologous proteins, DoBi outperforms with an F-score of 0.56, while PINUP and ProMate have the F-scores 0.43 and 0.21, respectively. Therefore, DoBi shows a performance better than many strong 'Predictors Combining Evolutionary and 3D Intrinsic Information', but at the same time its application is limited since DoBi requires two protein chains in order to perform predictions and also generation of docking models is computationally expensive.

2.3.2Template-based Predictors

Predictors in this group use available experimental 3D structures as templates to predict interfaces. Some of them take advantage of the structures homologous to the QP. (see section 2.3.2.1) while others use structural neighbours which have the same fold as the QP structure (see section 2.3.2.2). These methods are discussed in more details below.

2.3.2.1 Homologous Template-based Predictors

Early studies (Ma et al. 2003; Hu et al. 2000) showed that residues which are structurally conserved at the binding sites (in particular polar residues) of a protein family belong to a hot-spot. They discovered that using these structurally conserved residues a binding site can be distinguished from the surface patches of the protein. Also, comparing protein complexes in the same classification of SCOP (Andreeva et al. 2008) shows homologous proteins interact with their partners keeping the same

orientation (Aloy et al. 2003). Moreover, binding site localisation within each family has proven to be similar regardless of the similarity of their binding partner (Korkin et al. 2005). A more detailed study of the physico-chemical properties of the interface residues shows higher similarity in homologous proteins than non-homologous ones (Martin 2010). Similar studies have been investigated binding site similarities at protein domains (Shoemaker et al. 2006; Han et al. 2006; Kim & Ison 2005; Littler & Hubbard 2005). Therefore, integration of homologous structural information into interface predictors can improve performance.

Exploiting results of these studies, Chung et al. (Chung et al. 2005) were first to investigate binding site detection based on structurally conserved residues. For each protein, the homologous structures were generated from the same SCOP family level. Conserved residues were selected using a conservation score based on structure alignment and weighted by a normalised B-factor. To detect interfaces, they trained a SVM based system combining sequence profiles of neighbouring residues, ASA and conservation scores. Then they clustered residues returned by the SVM based system as potential interface residues to reject those which are isolated. On a dataset of 274 non-redundant chains of hetero-complexes, respectively 52%, 77% and 21% of the binding sites were predicted precisely, correctly and partially. Precisely and correctly predictions means more than 70% and 50%, respectively, of the ground truth interfaces were predicted in the binding site. A partial prediction refers to identification of only a few ground truth interface residues at the binding site. In addition, comparison of the above trained SVM with an SVM which does not involve the structural conservation score shows that integration of conservation score improves the correct prediction by 17%.

In addition to structural conservation, physico-chemical conservation was explored using a graph-based approach by Konc and Janezic (Konc & Janezic 2007) for detection of interaction sites. In this approach only one or two homologues structures with high sequence identity to QP were explored. Later on, the algorithm was extended to take advantage of a larger number of homologues (Carl et al. 2008). Although these methods confirm that the structural binding site conservation is an important factor, further work is required to increase the sensitivity of their methods.

A combination of sequence and structure conservation scores, was introduced in IBIS (Shoemaker et al. 2010; Tyagi et al. 2012) to detect potential binding sites. IBIS starts with generating homologous complexes with at least 30% sequence similarity to the QP. Structures are superposed on the QP, retaining the ones whose binding sites have an overlap of at least 75% with the QP structure (Figure 2.5). Using the alignment, a Structure-based-MSA (S-MSA) is created which allows highlighting the interface residues of homologues. Then, using the S-MSA a Binding site similarity matrix is generated by comparing the structure and sequence of each homologue against all other homologues. This matrix is used to cluster similar binding sites from different homologues into the same group. Clusters are then ranked based on a weighted combination of sequence similarity score, conservation score, average of contact and PSSM score. The inferred binding site of the best rank group is then mapped onto the QP (more details are provided in chapter 4). Comparison between IBIS and HomPPI demonstrates the value of using structural data to generate an MSA, since on a dataset of 188 proteins introduced in (Q. C. Zhang et al. 2010) IBIS with 69.7 precision and 72.0 recall performs significantly better than HomPPI with 62.8 precision and 50.4 recall.

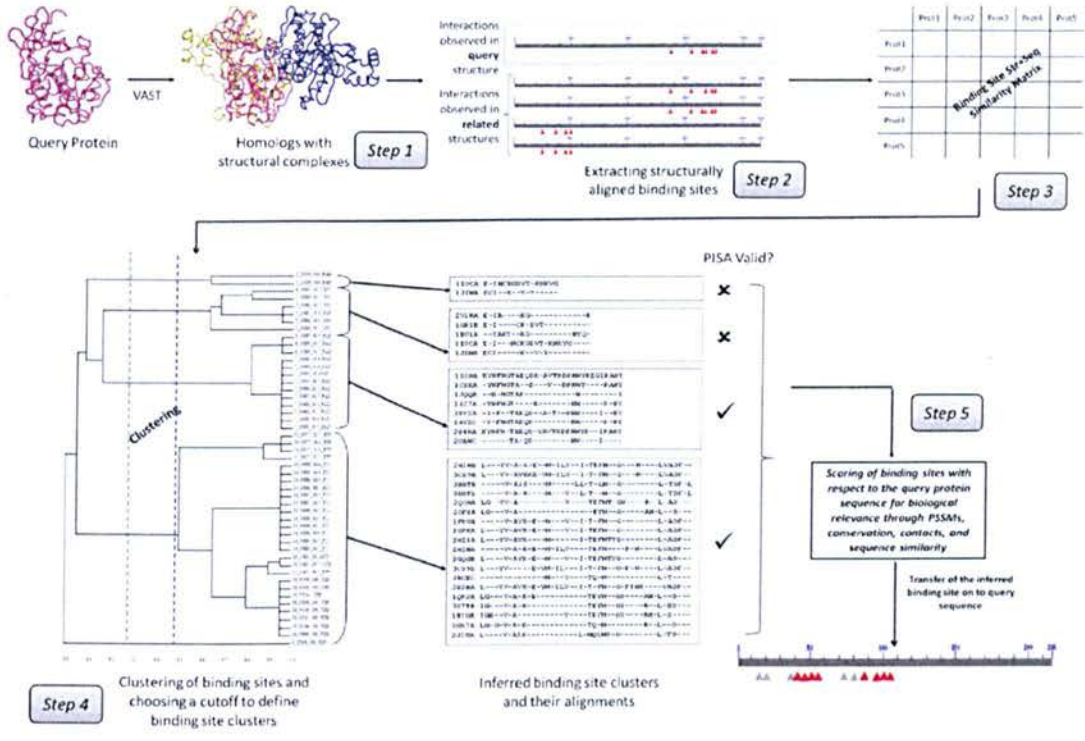


Figure 2.5: IBIS predictor Pipeline. For a Query Protein a set of homologues are generated. Step1: QP and homologues are structurally aligned. Step2: S-MSA is built and interface residues are marked on the S-MSA. Step 3: a similarity matrix of homologues is created. Step4: Homologue's binding sites are clustered. Step5: Scores are given to each cluster and the best scored one is mapped onto the QP. Taken from (Tyagi et al. 2012).

2.3.2.2 Structural Neighbour-based Predictors

As we saw in section 2.3.2.1, homologous structures have been used to predict protein interfaces, function and complexes. These studies have confirmed the significant structural conservation of binding sites among proteins within one family. However, although closer homologous proteins increase the reliability of predictions, at the same time it limits the number of proteins which a prediction can be made for: homologous template-based predictors are limited to the availability of homologues structures of the QP.

Petrey et al. (Petrey et al. 2009) demonstrated that for those proteins which homologous structural information fails to provide a functional classification, functional relationship can be detected by remote structural neighbours. Structural neighbours refer to proteins which are structurally similar to the QP even if they do not have any evolutionary relationship. Russell et al. (Russell et al. 1998) discovered that proteins

with similar folds but low sequence identity interact with their ligands using the same location. Based on this discovery, remote structural neighbours have been used for protein-ligand interaction prediction (Brylinski & Skolnick 2008). This was further extended to protein-protein interaction prediction, in which for a QP, Konc et al. (Konc & Janežič 2010a; Konc & Janežič 2010b) and Carl et al. (Carl et al. 2010) search through the whole PDB looking for protein which are structurally similar to the QP. Similarly to their previous work (Carl et al. 2008), a graph-based approach has been used to detect interactions sites which are both structurally and physico-chemical conserved among structural neighbours. Since these methods provide ranking score for conserved residues they have been used to detect protein binding sites. Prediction on a dataset of 39 protein resulted in a sensitivity of 43.9% while ConSurf showed 38. 1% (Konc & Janežič 2010a; Konc & Janežič 2010b). However, in their exploration for structural neighbours, they only kept highly structurally similar proteins which still prevent prediction for proteins without close structural neighbours. Zhang et al. further studied (Q. C. Zhang et al. 2010) the extent that remote structural neighbours can be exploited for detection of PPIs. They discovered structurally similar proteins (structural neighbours) display similar interaction sites which are evolutionary conserved (Q. C. Zhang et al. 2010) and even among remote structural neighbours a significant level of interface conservation has been recognised. They also confirmed that the use of structural neighbours increases the accuracy and coverage significantly in comparison to methods which do not use 3D information. Recently, in a large scale study Kundrotas et al. (Kundrotas et al. 2012) investigated the extend of structural templates availability in PDB to model PPIs. Based on their study for 33,537 complexes out of 33,840 (99%) structural templates are available to model the PPI. Comparing the templates to the experimental complexes, show that one-third of the templates are good quality. Kundrotas et al. (Kundrotas et al. 2012) investigations confirm Zhang et al.'s (Q. C. Zhang et al. 2010) study on the completeness of PDB (Berman et al. 2000) to model PPIs.

Based on these observation, Zhang et al. (Q. C. Zhang et al. 2010) introduced PredUs which extracts proteins which are structurally similar to the QP and maps interacting residues from structural neighbours onto QP even if they do not display any homology. PredUs structurally aligns the QP on the structurally similar protein, counts

the interfaces at each position and builds a contact map. Then a contact frequency map is built using the weighted sum of the individual contact maps. That frequency map is used to predict QP interfaces. Figure 2.6 shows a schematic view of this process. Comparison of PredUs with Promate, cons-PPISP and PINUP on DBMK 3.0 (Hwang et al. 2008) and CAPRI dataset (Lensink & Wodak 2010). The precision of PredUs is similar to others but the recall has improved significantly. This shows that PredUs, by taking advantage of the 3D structures of structural neighbours, can detect interface residues which are not distinguishable by methods which only explore the 3D structural features of the QP. Recently, PredUs (Zhang et al. 2011) has used contact frequencies along with SVM to predict if a surface residue belongs to an interface or not. On DBMK3, SVM-based PredUs reaches precision and recall of 50% and 58% while their previous approach (Q. C. Zhang et al. 2010) achieved 44% and 46%, respectively.

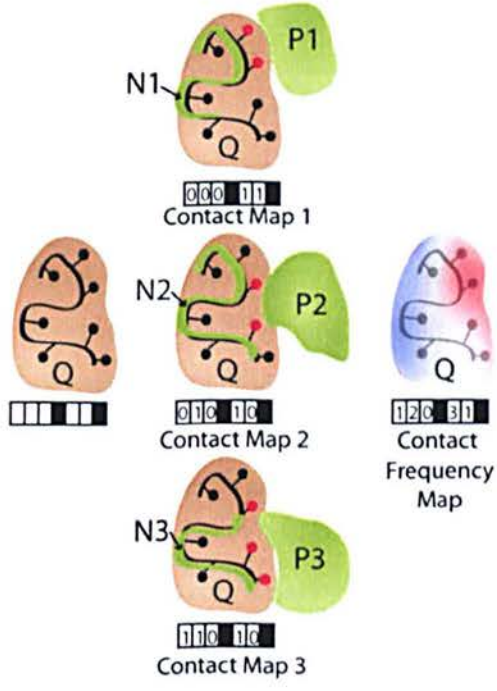


Figure 2.6: Schematic representation of PredUs contact and contact frequency maps. The QP is coloured in brown and has seven residues with five on the surface. The green line represents the Ni structural neighbour and Pi represents its interacting partner. Interfaces of Ni are mapped on the QP and the contact map is updated. Black squares on contact map are non-surface residues and 1 and 0 represent interaction or non-interaction respectively. Contact frequency map is built using the individual contact maps. Taken from (Q. C. Zhang et al. 2010).

Whereas PredUs depends on the existence of global structural neighbours, PrISE proposes to deal with this limitation by predicting interface residues from local

structural similarity only (Jordan et al. 2012). To achieve this, they created a repository of structural elements (SE) generated from PDB (Berman et al. 2000). For each SE of the query protein (consisting of a central residue and its neighbours) a set of similar SEs are extracted from the repository and a weight is assigned to them based on their similarity to QP. Finally, that central residue of QP is predicted as interface if a weighted majority of its similar SEs is classified as interface residue. Although PrISE is more general than PredUs, it displays comparable performance. In comparison with Carl et al. (Carl et al. 2010), PrISE has demonstrated an increase of ~25% in recall, precision and F1 score. Also, comparisons with Promate, PINUP, Cons-PPISP, and Meta-PPISP confirmed the superiority of template-based methods.

2.3.3Conclusion

Various methods have been proposed for predicting protein interfaces as mentioned above. These methods and their main features are summarised in Table 2.1. A high number of methods investigate protein sequential or structural features in order to characterise protein interfaces. Usage of 3D structural properties has improved the sequence-based predictions. Moreover, evolutionary conservation was shown to be an important property. For example, HomPPI using sequence-based evolutionary information outperforms the best performing structure-based methods such as PIER and Promate based on review studies in (Zhou & Qin 2007; De Vries & Bonvin 2008). Therefore, methods have integrated various structural features along with evolutionary information to increase performance.

The combination of different features using various techniques has been investigated by intrinsic-based predictors. However, it seems that these methods have reached their saturation, and combination of more properties does not improve their prediction performance (Neuvirth et al. 2007; De Vries & Bonvin 2008; Zellner et al. 2012). Combining properties is only useful if they provide complementary information rather than adding redundant information. Also, too many properties will cause over fitting of classifier methods (Zellner et al. 2012).

On the other hand, many studies have investigated the 3D structure of binding sites among protein families. They discovered that the binding site localisation and

structure are conserved among homologous. These properties have improved the detection of functional residues and protein-ligand binding sites. Therefore, predictors took advantage of structurally conserved residues among homologous proteins to improve binding site predictions as exemplified in the comparison between IBIS and HomPPI, which, in addition from sequence conservation, takes into account structural conservation (Tyagi et al. 2012).

Although homologous template-based predictors improve the predictions, they are limited to those proteins whose homologous structure exists. Therefore, methods have extended their search for templates to structural neighbours, since interface conservation exists even among remote structural neighbours. In addition, with the increase in experimentally determined 3D complexes good quality templates can be found for many proteins (Kundrotas et al. 2012). Therefore, usage of structural neighbours is the current focus of template-based protein interfaces predictors.

Although, template-based methods are currently the predictors under the main focus, one of their main limitations is their dependency to availability of the QP 3D structure. Also, these predictors have not investigated the contribution of interacting partners of structural neighbours in the prediction. In addition, since these methods perform structural comparisons their computational time is high which limits their application to high-throughput predictions.

Table 2.1: Protein interface predictors. Method column refers to the technique used to combine features. Cells with * refer to predicted structural features. Cells annotated with ψ highlight indirect use of intrinsic features which is based on sequence co-occurrence.

				Input			Main Knowledge Source (properties)				Intrinsic-based			Template-based		Output	
Section	Predictor	year	Method	Sequence	Structure	Both	Sequence	Structure	Both	additional	Evolution Info.	Intrinsic features	Both	Homologous Str	Str Neighbour	Residue-based	Patch-based
2.3.1.1.1	Gallet et al. (RBD)	2000	/	■			■					■				■	
	Ofran&Rost	2003	NN	■			■					■				■	
	Yan et al.	2004	SVM	■			■					■				■	
2.3.1.1.2	TreeDet	2006	/	■			■				■					■	
	HomPPI	2011	/	■			■				■					■	
2.3.1.1.3	Res et al.	2005	SVM	■			■						■			■	
	Wang et al.	2006	SVM	■			■						■			■	
	ISIS	2007	NN	■			■			*			■			■	
	PSIVER	2010	NBC+KDE	■			■			*			■			■	
	Chen&Jeong	2009	RF	■			■						■			■	
	Chen & Li	2010	SVM	■			■						■			■	
	PPI-OD	2012	SVM	■			■						■			■	
	Wang et al.	2012	SVM	■			■						■			■	
	Ahmad&Mizuguchi	2011	NN	■			■						■			■	
	PIPE-Sites	2011	/	■			■						ψ				■
2.3.1.2.1	Lichtarge et al.	1996	/		■		■				■						Clustering
	JET	2009	/		■				■								Clustering
	Consurf	2010	Bayesian			■	■				■					■	
2.3.1.2.2.1	Jones&Thornton	1997	Linear		■				■				■				■
	PIER	2007	Linear		■				■				■			■	
	SHARP2	2006	Scoring Function		■				■				■				■
	InterProSurf	2010	Scoring Function		■				■				■			■	
	WHISCY	2006	Scoring Function		■				■				■			■	
	PINUP	2006	Scoring Function		■				■				■			■	
	Koike & Takagi	2004	SVM		■				■				■			■	

Section	Predictor	year	Method	Input			Main Knowledge Source (properties)				Intrinsic-based			Template-based		Output	
				Sequence	Structure	Both	Sequence	Structure	Both	additional	Evolution Info.	Intrinsic features	Both	Homologous Str	Str Neighbour	Residue-based	Patch-based
	PPI-Pred	2005	SVM		■				■				■				■
	Bordner&Abagyan	2005	SVM		■				■				■			■	
	Dong et al.	2007	SVM		■				■				■			■	
	Nguyen&Rajapakse	2006	SVM		■				■				■			■	
	Li et al.	2008	SVM		■				■				■			■	
	Deng et al.	2009	SVM		■				■				■			■	
	Pres-Cont	2012	SVM		■				■				■			■	
	PPISP	2001	NN		■				■				■			■	
	Fariselli et al.	2002	NN+SVM		■				■				■			■	
	Cons-PPISP	2005	NN		■				■				■			■	
	Y. Chen et al.	2012	RBF-NN		■				■				■			■	
	SPPIDER	2006	NN		■				■				■			■	
	Sikić et al. (seq+str)	2009	RF		■				■				■			■	
	VORFFIP	2011	RF		■				■				■			■	
	Multi-VORFFIP	2012	RF		■				■				■			■	
	Qiu&Wang	2012	RF		■				■				■			■	
	Li et al.	2012	RF+IFS		■				■				■			■	
2.3.1.2.2.2	ProMate	2004	Naive Bayesian		■				■				■				■
	Bradford et al.	2006	Bayesian Network		■				■				■				■
	Higa&Tozzi	2008	Bayesian Network		■				■				■				■
	Li et al.	2007	CRF		■				■				■			■	
	Liu et al.	2009	HM-SVM		■				■				■			■	
	ISPRED2	2012	HM-SVM		■				■				■			■	■
	Meta-PPISP	2007	meta														
	PI2PE	2007	meta														
	MetaPPI	2008	meta														
2.3.1.2.3	Fernández-Recio et al.	2004	Docking		■			■				Propensity				■	
	ODA	2011	Docking		■			■				Desolvation Energy				■	
	DoBi	2012	Docking		■			■				Residues Count				■	
·	Chung et al.	2005	SVM		■			■					■	■		■	

				Input			Main Knowledge Source (properties)				Intrinsic-based			Template-based		Output	
Section	Predictor	year	Method	Sequence	Structure	Both	Sequence	Structure	Both	additional	Evolution Info.	Intrinsic features	Both	Homologous Str	Str Neighbour	Residue-based	Patch-based
	Konc&Janezic	2007	graph based		■			■				■		■			■
	Carl et al.	2008	graph based		■			■				■		■			■
	IBIS	2010, 2012	Parametric		■				■				■	■		■	
2.3.2.2	Konc&Janezic	2010	graph based		■			■				■			■	■	■
2.3.2.2	Carl et al.	2010	graph based		■			■				■			■	■	■
2.3.2.2	PredUs	2011	SVM		■			■							■	■	
2.3.2.2	PrISE	2012	Parametric		■			■							■	■	

2.4 Protein Docking

Protein-Protein interactions play a critical role in the molecular processes that take place in the living cell. Therefore, structural knowledge on PPI are really important for drug design (Kundrotas et al. 2012). Experimental techniques such as Crystallographic (X-ray) and nuclear magnetic resonance (NMR) have been used to determine complexes' structures (Ritchie 2008). However, since these methods are slow and require expensive resources, they cannot be used for high-throughput analysis of protein-protein complexes (Russell et al. 2004). Therefore, docking techniques are required to uncover genome-wide PPI.

Protein-protein docking aims to computationally predict the 3D structure of a protein-protein complex using the individual structures of its components. Performances of docking algorithms are compared biannually in the CAPRI (Critical Assessment of Predicted Interaction) competition (Janin & Wodak 2007; Fleishman et al. 2011) and are evaluated against larger protein docking benchmarks (Mintseris et al. 2005; Hwang, Vreven, Janin, et al. 2010; Hwang et al. 2008), designed in such a way that predictions can be assessed according to their level of difficulty.

Most of the designed docking methods perform binary docking, i.e. it involves only two protein chains. A number of methods have introduced multimeric protein complex prediction, but due to their complexity not many studies have investigated them. Some of these docking methods are restricted based on the complex type (such as homomeric and/or symmetric complexes) (André et al. 2007; Berchanski & Eisenstein 2003; Comeau & Camacho 2005; Schneidman-Duhovny et al. 2005b) while others can predict any type of complexes (Karaca et al. 2010; Esquivel-Rodríguez et al. 2012; Inbar et al. 2005). Since these methods are built upon pairwise protein docking, in this chapter we will focus only on binary docking methods and by 'docking' we will be simply referring to binary docking.

Docking algorithms can be divided into two groups (Kundrotas et al. 2012): (i) template-free docking: in which the whole configuration space is investigated. (ii) template-based docking: in which already known similar complexes to the proteins of interest are used as a template to generate new complexes. Here we first discuss the methodology behind template-free docking and then discuss the importance of template-based docking to produce large-scale PPIs.

Although numerous docking methods have been proposed, their comparison in terms of quantitative measures is a difficult task. First, many different factors such as backbone RMSD, interface RMSD and fraction of contacts contribute toward quality of one model. Also, docking methods may target different difficulties in docking, for example one might only deal with side-chain flexibility while others solve backbone flexibilities. Therefore, in this section instead of using quantitative measures to compare methods their performances in CAPRI predictions are mainly considered.

2.4.1Template-free docking

These methods referred as *ab initio* produce protein complexes (Kastritis & Bonvin 2010) following two main consecutive procedures. The first step is **sampling** in which a large number of docked orientations (decoys) are generated (Ritchie 2008). Two main approaches have been used to generate the rigid-body docked solutions in this search space (Halperin et al. 2002). One uses grid or Connolly representation of the protein and investigates the possible conformational space using techniques such as geometric hashing and fast Fourier transform (FFT) (Ritchie 2008). The other one

screens part of the solution space using either Monte Carlo (MC) simulated annealing, molecular dynamics (MD), or genetic algorithms (GA). Although current docking methods are successful at generating near-native conformations in the sampling procedure, the main issue is to detect those conformations among a pool of decoys

Therefore, docking methods have introduced a second stage, **refinement,** in which docked models are scored and re-ranked using elaborated scoring functions. Moreover, possible backbone and side-chain flexibility are considered at this stage. The multi-stage docking procedure is illustrated in Figure 2.7. Note that experimental data can also be used to assist docking models (section 2.4.1.3).

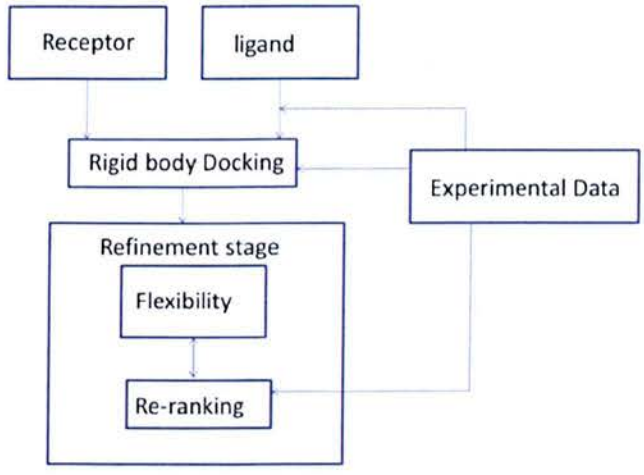


Figure 2.7: Template-free docking methods general pipeline. In the first stage, docking methods generate rigid body docking of the ligand-receptor. Then, at the refinement stage models are scored and re-ranked using elaborated scoring functions. Backbone and side-chain flexibility are also considered at the refinement stage. Experimental data can be used at different stages to assist the docking procedure.

An important factor in docking procedures is the requirement of computational efficiency. As a consequence, rigid-body docking uses a low-resolution (course-grained) protein representation along with fast energy scoring functions (Vakser 2013). Course-grained is a simplified representation which range from residue representation to selection of side chain centroid. Low resolution models implicitly account for local side-chain conformational flexibility (Gray et al. 2003) by allowing some overlapping between them (see section 2.4.1.2.1). In the refinements stage, more expensive scoring functions can be used to score docked models. Ideally, a scoring function based on

physico-chemical principles, along with atomic representation of atoms can discriminate between the native-like and non-native models. However, such scores are both computationally expensive and error prone since they are sensitive to any conformational changes upon binding (Andreani et al. 2013). Therefore, coarse-grained knowledge-based potentials have been used both in fast scoring functions and in refinement stage. They are less sensitive to structural inaccuracy such as x-ray unbound models and can also model proteins for high-throughput interaction network. They have shown to preserve most of the useful information which can be detailed in an atomic-base scoring function (Fitzgerald et al. 2007; Zhang et al. 2004).

In the next few sections we first discuss how template-free docking approaches search through the conformational space (section 2.4.1.1) and the scoring functions they use for detecting plausible conformations (section 2.4.1.1.1). Then, we look at different approaches of introducing flexibility in docking (section 2.4.1.2). The use of experimental data to drive docking is further discussed (section 2.4.1.3), and finally we look at how more detailed scoring functions will access the detection of near-native models (section 2.4.1.4).

2.4.1.1 Rigid Body Docking

Docking of two proteins started with methods which are based on geometry and shape complementarity criteria. Many docking methods have adopted those criteria to perform geometric-based rigid body. They can be divided into two groups based on their approach to protein representation and conformational space exploration.

In the first approach, methods project the protein shape on a 3D Cartesian grid and classify each cell as surface, interior and exterior. Then, various orientations are generated and scored by the grid overlap between the receptor and ligand (here ligand refers to the interacting partner). Since the search consists of a blind six degree translation and rotation a vast number of models would be produced (in order of billions) (Ritchie 2008). Therefore, to cover this large space in an efficient time, Fast Fourier Transform (FFT) techniques, which were first introduced in the work of Katchalski-Katzir et al. (Katchalski-Katzir et al. 1992), has been proposed. In this technique (shown in Figure 2.8), Fourier transform is calculated for each protein represented in a grid. Then Fourier correlations are calculated where high correlation

corresponds to good surface complementarity. Best models are saved and then the ligand orientation is changed and FFT calculation starts again.

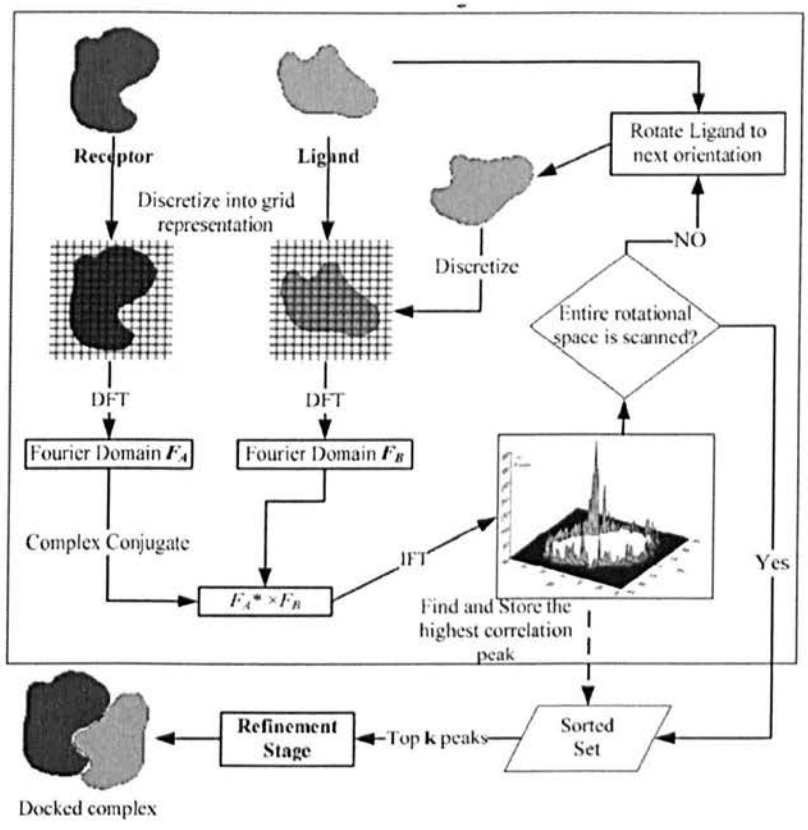


Figure 2.8: fast Fourier transform technique introduced by of Katchalski-Katzir et al. (Katchalski-Katzir et al. 1992). First, protein shapes are projected on a 3D Cartesian grid and Fast Fourier transforms are calculated for each protein. Second, Fourier correlations are computed and high correlations which correspond to good surface complementarity are stored. Finally, ligand orientation is changed and FFT calculation starts again

In the second approach, proteins are represented by their surface geometric features. The bases of this method were established by Connolly (Connolly 1983) which introduced a new surface representation (see an example in Figure 2.9). In this representation, protein surface is processed to detect a few hundred sparse critical points. These points are named 'caps', 'pits' and 'belts' (shown in Figure 2.9), which respectively belong to one, two and three atoms (Janin 2013; Duhovny et al. 2002). Then, a graph representation is induced using these points which display the local geometric shape of the protein surface (concavities, convexities and flats). Geometric hashing (GH) which was first introduced in computer vision for object recognition

(Lamdan & Wolfson 1988) is used along with this representation. A hash table storing the ligand positions based on each triple point is search by the receptor graph points. A clique-algorithm is used to match these geometric shapes on the receptor and ligand surface and generate docking poses which are then evaluated using a force field function. Examples of these docking methods are: PatchDock (Duhovny et al. 2002), 3D-Garden (Lesk & Sternberg 2008) and SKE-Dock (Terashi et al. 2007).

In comparison to FFT, geometric hashing is faster since a smaller set of docking poses are investigated while FFT searches the whole space. On the other hand, geometric hashing may miss correct solutions and also requires pre-processing of protein surfaces (Ritchie 2008). Therefore, more methods have focused on using FFT which, unlike geometric hashing, can be easily integrated with other features such as electrostatic and hydrophobicity (Janin 2013).

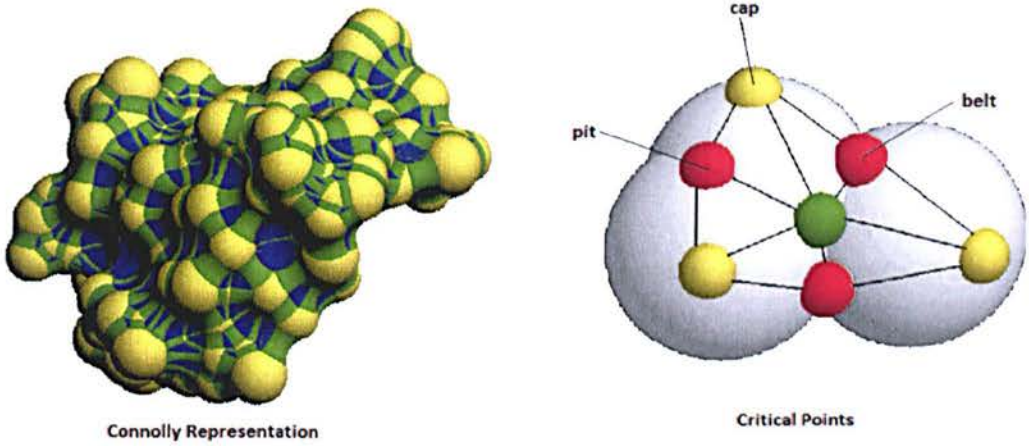


Figure 2.9: Connolly representation (on left) and its critical points (on right). 'caps', 'pits' and 'belts' belonging to one, two and three adjacent atoms, respectively are represented by yellow, red and green spheres. While FFT and GH explore the whole conformational space, some methods have taken advantage of GA (Gardiner et al. 2001; Morris et al. 1998; Jones et al. 1997) and MC (Gray et al. 2003; Hart & Read 1992) simulated annealing algorithms to sample part of the conformational space. In GA a random start position is selected which further produces new generations of models using crossovers and mutations. In MC rigid-body docking, a random position is selected and one partner is translated and rotated around the other one. New models will be kept if they meet the Metropolis condition (accepting configurations which have a lower energy than the original conformation, and also accepting those with a higher

energy with a probability which decreases with increasing energy), allowing selection of models with low energy located at local minima. Then the procedure is repeated using a new starting conformation. This method allows flexibilities by side-chain and backbone dihedral angles movements.

These shape complementarity based methods are reliable for scoring bound–bound conformations (Lawrence & Colman 1993). But bound structures which are taken from already known complexes are not of predictive value (Janin 2013). For the more difficult and common case of unbound-unbound docking, shape complementarity on its own is not reliable (Katchalski-Katzir et al. 1992). In addition, shape complementarity scores fail to detect near-native models among the large number of docked solutions. Therefore, rigid docking methods should be "soft" enough to consider structural flexibility but at the same time to distinguish the correct poses. Therefore, to deal with detection of native like models, new terms have been introduced in the shape complementarity scoring function of the rigid-body methods (see section 2.4.1.1.1 for more details).

2.4.1.1.1 Fast Scoring Functions

Shape complementarity on its own is not reliable (Katchalski-Katzir et al. 1992) to detect near-native models. Moreover, electrostatic has shown to be important in protein interactions (Sheinerman et al. 2000; Tobias 2001). Therefore, docking methods have integrated electrostatic forces into scoring function. FTDOCK (Gabb et al. 1997), uses Coulombic electrostatic potential to reject models which are geometric fit but whose polar interactions are unfavourable charged. This simple electrostatic model improved the ranking of near-native model. Mandell et al. (Mandell et al. 2001) claims that FTDOCK is not very discriminative since all models which are energetically favourable are similarly treated (while they may have very different electrostatic magnitude). Therefore, they introduced DOT (Mandell et al. 2001), which integrates electrostatic using Poisson-Boltzmann equation within their FFT-based docking method. They found including electrostatic will only improve models whose binding is largely mediated by their electrostatic potentials (Mandell et al. 2001). Instead of calculating the interaction energy of each docked pose, MolFit (Heifetz et al. 2002) estimates the tendency of protein chains to form a energetically favourable complexes.

Therefore, the electrostatic areas of each protein are defined as positive, neutral, or negative patches. These patches are then integrated in the protein 3D representation to perform geometric-electrostatic FFT-based docking. They noticed that integration of electrostatic in scoring functions shows an increase in the number of correct solution, in comparison to using shape fit alone. But this combination does not improve the ranking of the best solution.

Sheinerman et al. have argued that since interactions take place in water to effectively incorporate electrostatic potentials into docking, desolvation must be taken into account (Sheinerman et al. 2000). Desolvation calculates the energy of breaking protein-water bonds and creating protein-protein and water-water bonds (also known as hydrophobic effect) (Nussinov & Schreiber 2010). Knowledge-based Potentials (also known as statistical potentials) have been successfully used in protein structure predictions. Therefore, various docking methods have introduced those desolvation models in their scoring function. These potentials measure the importance of an observed atom or residues contact in comparison to a reference (Nussinov & Schreiber 2010). In addition to electrostatic and desolvation, ICM-DISCO (Fernández-Recio et al. 2003; Fernández-Recio et al. 2002) and RosettaDock (Gray et al. 2003) combined hydrogen bonding in their MC-based approaches. A comparison to FTDOCK, which does not include desolvation, shows improvement in producing more native-like orientations (Fernández-Recio et al. 2003; Gray et al. 2003). However, Cheng et al. (Cheng et al. 2007) argued that electrostatic and desolvation terms are the main contributors to the scoring models and hydrogen bonding has no impact.

Therefore, focusing on desolvation terms, ZDOCK (Mintseris et al. 2007), one of the top performing methods as estimated by CAPRI (Hwang, Vreven, Pierce, et al. 2010), introduced a new atomic statistical potential (atomic contact potentials (ACP)) based on contact propensities seen in transient protein complexes. A potential is shown by a KxK interaction matrix (for K atom types) and measured as sum of K correlation functions. The main drawback of the ACP is that despite increasing accuracy it is computationally expensive (usually k=20). To improve the speed ZDOCK 3.0.2 took advantage of a 3D convolution library which allows for non-cubic rectangular grids representation which speeds up the FFT correlation computation (Pierce et al. 2011). Another approach toward optimising these potentials is to take advantage of Principle

Component Analysis (PCA). Sumikoshi et al. (Sumikoshi et al. 2005) simply applied PCA on cross terms of ACP and took the 2 most important eigenvectors as main contributors of energy. PIER FFT-based docking (Kozakov et al. 2006) also uses PCA along with a new knowledge-based potential scoring function, where they calculate the pairwise frequency of atom contacts in complexes in correspondence to the frequency of the contact in a large dataset of docked decoys (DARS). PCA of DARS cross terms allows the detection of the main energy contributors which are then used on FFT correlation. Comparison to ZDOCK, which at the time was one of the top performing docking methods, shows improvement by generally generating 50% more near-native solutions.

Although these fast scoring functions have improved the generation of orientations close to near-native models, detection of near-native models among a large set of decoys remains a challenge. Therefore, many docking methods include a more sophisticated scoring function to re-rank models in a refinement stage (Tovchigrechko & Vakser 2005).

2.4.1.2 Introduction of Flexibility in Rigid Docking

Flexibility can be introduced in docking methods in three different manners (Bonvin 2006). 1) Smoothing the geometric criteria and allowing penetration in rigid docking (soft docking). 2) Performing multiple run of docking on ensembles of conformations (cross/ensemble docking). 3) Allowing explicit backbone or/and side-chain flexibility during docking or refinement. Below we will discuss these in more details.

2.4.1.2.1 Flexibility by Soft Docking

The simplest way to introduce flexibility has been by softening the geometric criteria, which means to allow overlap between interacting surfaces. Jiang and Kim (Jiang & Kim 1991) were the first people to exploit softness to accept minor conformational changes during the docking procedure. They used a grid cubic representation of the protein in which a docked model was produced by matching the surface cubes only rejecting the overlap of interior parts of protein.

Several geometric representations have been used to assist this overlap process. Vlaker (Vakser 1996) used a coarse grid representation in which only c-alpha atoms are used. In order to allow more overlap between protein surfaces, HEX (Ritchie & Kemp 2000) uses a grid-free version of FFT (grid-free spherical polar Fourier (SPF)) in which rotational correlation is calculated instead of translation (Ritchie 2008). In these approach spherical hydrophobic volumes of protein surfaces are compared by spherical harmonic functions (Halperin et al. 2002), which, apart from softening the docking SPF, increases speed since the search space is largely reduced. Therefore, HEX is much faster than FTDOCK (Halperin et al. 2002). HEX was further improved to FRODOCK (Garzon et al. 2009) which uses Fast Rotational Method instead of SPF. Allowing grid overlap using AND Boolean operator has been investigated in BIGGER (Palma et al. 2000) which performs a real-space grid search with bit mapping that is optimised heuristically for speed.

Van der Waals volumes have also been used to measure geometric fitness allowing overlaps in methods such as FTDOCK (Gabb et al. 1997) and RosettaDock (Gray et al. 2003). In FTDOCK instead of calculating charge-charge interactions, dispersed point charges of grids are used which also simulates side-chain movements. Instead of using volumes, Heifez and Eisenstein (Heifetz & Eisenstein 2003) considered overlap by trimming the protein's side chain. This is achieved by a weighted protein surface representation in which side-chains contributes less to the complementarity score. Another approach to detect flexible side chains in a grid representation is to use MD simulation of the protein (Ma et al. 2003). Cells which are occupied by all MD conformations are kept for docking and other cells are assumed to be mobile.

The main drawback of these soft docking approaches is that when full-atom representations of low-resolution models are created severe steric clashes are inevitable. In addition, they address only slight side-chain flexibility, which does not account for backbone movements.

2.4.1.2.2 Flexibility by Ensemble Docking

Flexibility can be introduced through rigid-body docking of ensemble of conformations (Bonvin 2006). These ensembles are taken from X-ray or NMR structures or generated using computational methods such as MD simulations, normal

modes and loop modelling. One way of docking ensembles is to dock them one by one (cross-docking) (Figure 2.10) but since it is computationally expensive methods such as mean-field approach have been used (Andrusier et al. 2008).

Two studies of Smith et al.(Smith et al. 2005) and Grünberg et al. (Grünberg et al. 2004) investigated the use of ensembles docking by using MD simulations along with 3D-DOCK and HEX docking methods. They discovered an increase in the number of native like solutions in the pool of docked conformations. However, at the same time, scoring became more difficult since wrong solutions were given higher ranks.

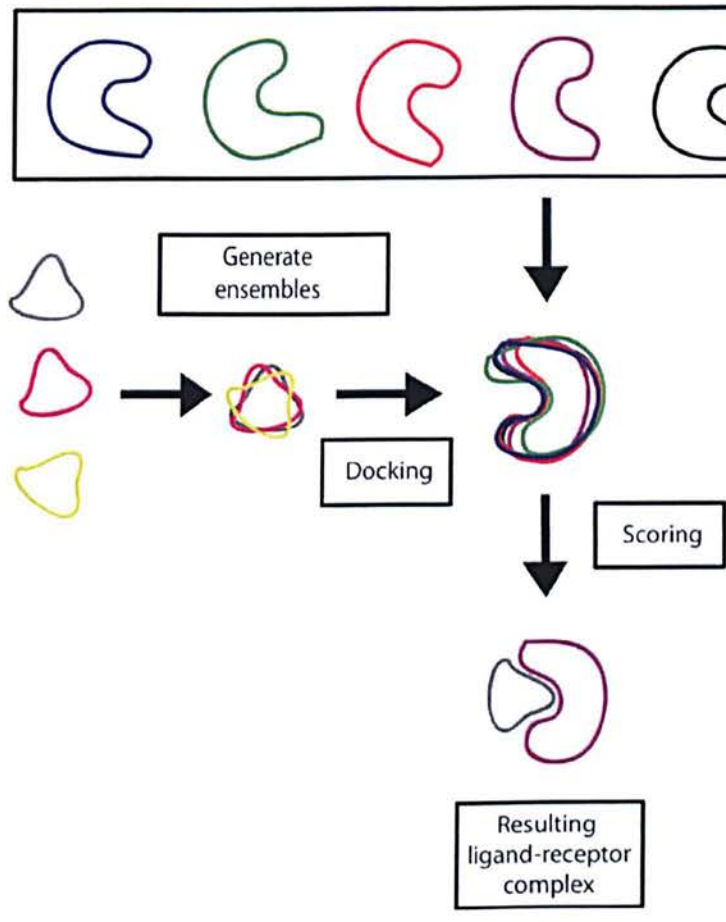


Figure 2.10: Ensemble Cross-docking strategy. Different ensembles of receptors and ligands are docked separately. Taken from (Bronowska 2013).

2.4.1.2.3 Side-Chain and Back-bone Flexibility

To introduce flexibility during docking or at refinement stage most methods use energy minimization (EM) techniques such as MD simulated annealing or MC

simulated annealing. Some of these methods not only deal with side-chain flexibility but also introduce back-bone flexibility.

One of the early methods, ICM, introduced by Totrov and Abagyan (Totrov & Abagyan 1994) used Monte-Carlo simulated annealing to allow side-chain movements at the same time as docking. Their method was computationally expensive due to using atomic details of protein structures along with a force field scoring function. To improve the computational time, methods based on EM usually adopt two step approaches in which first rigid-body models are produced using a simplified protein model and a coarse force field, second, models at the local minimum are further refined using the detailed atomic representation of proteins (Janin 2013). Table 2.2 provides a general overview of different methods introducing flexibility on side chain and backbones.

Table 2.2: side-chain and backbone flexibility in docking methods.

Method	Side-Chain flexibility	Backbone Flexibility	Stage
RDOCK	ABNR EM	None	Refinement
ICM-DISCO	biased probability MC	None	Refinement
HADDOCK	MD simulated annealing	MD simulated annealing	Refinement
3D-Dock	Rotamers Library + mean field	None	Refinement
FiberDock	Rotamers Library + ILP	MC simulated annealing	Refinement
RosettaDock	Rotamers Library + MC simulated annealing	MC simulated annealing	Refinement
ATTRACT	Multi-copy Loops + mean field	Normal Modes EM	During Docking
FlexDock	Hinge Bending	Hinge Bending	During Docking

To introduce flexibility in ZDOCK docked models, RDOCK (Li et al. 2003) simply uses Adopted Basis Newton-Raphson (ABNR) energy minimisation to remove clashes and optimise polar and charge interactions. ICM-DISCO (Fernández-Recio et al. 2003; Fernández-Recio et al. 2002) uses a MC rigid docking with fast-soft energy interaction function to rapidly produce docked decoys. Then, models are refined by changing the ligand's side-chain torsion angles and optimising using biased probability MC to select the model with best side-chain energy. Comparison of ICM-DISCO to

FDOCK and BIGGER shows a better selection of native-like models. While these methods introduce flexibility only in ligand side-chains, HADDOCK (Dominguez et al. 2003) takes into account both receptor and ligand side-chains flexibility. This happens at refinement stage using MD simulated annealing along with backbone flexibility. Recently a new method (SwarmDock) (Torchala et al. 2013) has proposed to use hybrid particle swarm optimisation to introduce flexibility into docking.

Some methods use library of rotamers to introduce side chain movements. At refinement stage, 3D-Dock (Jackson et al. 1998; Carter et al. 2005) creates a probability-based conformational matrix of side-chain movement using the library which is refined by mean field approach (MultiDock). More advanced approaches combine rotamer library with energy minimisation. In FireDock (Andrusier et al. 2007) the optimal solution of rotamers are found by integer linear programming (ILP) technique. In FiberDock (Mashiach et al. 2010) this technique was combined with backbone flexibility using MC minimization. Alternatively, RosettaDock (Gray et al. 2003) starts with a low-resolution rigid-body MC search using residues-based potentials. Then, side chains are added to docked poses using a library of rotamers. Both side-chain and backbones are then optimised using simulated annealing MC.

Loops are another section of the proteins which goes under conformational changes upon binding. ATTRACT (Bastard et al. 2006; Zacharias 2003) tackles this issue by using multi-copies of loop conformations along with low-resolution protein representation to introduce side-chain and backbone flexibility during docking. Best models are then selected for further EM. A similar approach was introduced before (MC2) (Bastard et al. 2003) but since full atom representation was used, it proved computationally too expensive for high-throughput docking.

Proteins parts such as domains can rotate around regions called hinge (Emekli et al. 2008). Hinges are usually formed by several residues and can go under large conformational changes while keeping the conformation of rotating parts unchanged. FlexDock (Schneidman-Duhovny et al. 2005a) uses hinge identification on one of the docked proteins to allow for flexibility. Therefore, these methods require a prior knowledge of the location of hinges (Wolfson et al. 2005) which can either be predicted (Emekli et al. 2008) or generated by 3D structural comparison among homologues structures (Bonvin 2006). Using hinges a protein can be cut-downed into several rigid

sections (subdomains). These subdomains of one interacting partner are separately docked (using PatchDock) onto the other protein. Then, using a graph assembly and docking score, subdomains are assembled to create a final complex. Hinge bending has shown to deal with large conformational changes.

2.4.1.3 Data-Driven Docking

One of the drawbacks of the rigid-docking methods is the search through the whole conformational space (Dominguez et al. 2003) which can produce models far from the possible solution since fast scoring functions not always being efficient: they may fail to distinguish near native from non-native poses. Therefore, to constrain the search space, the wealth of PPI experimental data such as NMR, Mass spectrometry and mutagenesis (Van Dijk, Boelens, et al. 2005) can be used to focus on conformations encountered in nature. In data-driven docking those data are used to guide the docking itself, either by accounting for additional energy terms in the fast scoring function (Van Dijk, De Vries, et al. 2005), introducing anchor points (Fahmy & Wagner 2002) or up weighting specific residues in the FFT docking (Ben-Zeev et al. 2003; Schneider & Zacharias 2012).

For example, in TreeDock (Fahmy & Wagner 2002) orientations are limited using anchor points selected by NMR chemical shift perturbation and mutagenesis data. Anchor points are pair of atoms on the protein chains which are always in contact (Schuck 2007). ZDOCK have used biological and structural information in literature along with homology search to detect important residues (Zhang et al. 2005). They introduce "blocking" residues to avoid their appearance in the interface of conformational decoys by giving zero desolvation energy. HADDOCK have introduced NMR information to define ambiguous interaction restraints (AIRs) term along with electrostatic and van der Waals to score docking conformations. AIR boosts docking once atoms of interest come close together at the interface. This strategy proved very successful since HADDOCK showed impressive results in CAPRI rounds 3-5 (Méndez et al. 2005) and 13-19 (De Vries et al. 2010) for protein-protein and protein-RNA docking, respectively. However, reliance on experimental data means that when sufficient experimental data are not available, HADDOCK performed poorly in comparison to other state-of-the-art methods (De Vries & Bonvin 2011). However, this

limitation could be overcome by integrating interface predictors. HADDOCK introduced CPORT (HADDOCK-CPORT) (De Vries & Bonvin 2011) a consensus meta predictor based on WHISCY, PIER, ProMate, cons-PPISP, SPPIDER, and PINUP. CPORT prediction was added to AIR to drive the docking, which showed 72% increase of high quality models in the top 400 solutions in comparison to HADDOCK without CPORT. While HADDOCK has incorporated interface information into its energy scoring function, Kanamori et al (Kanamori et al. 2007) have used Evolutionary Trace (ET) of the residues calculated from homologues MSA to weight the degree of their shape complementarity. Some approaches have used data information as a pre-scan filter to limit the search space before performing docking (Schneidman-Duhovny et al. 2003; Kowalsman & Eisenstein 2009). Li et al. (Li & Kihara 2012) used various predicted interfaces constraints and noticed that pre-processing highly depends on the accuracy of predicted residues. Therefore, they have integrated predicted interface into their docking procedure. Although one might argue that, interface predictions themselves contain false positives, it has been shown that (Zhou & Qin 2007) this does not affect docking performance. However, what can negatively affects the results is when at least part of the true interfaces are not covered by interface predictors (Zhou & Qin 2007).

One of the main drawbacks of using experimental data or interface predictions as a constraint before docking or for driving docking methods, is that docked conformations must meet at least one of the constraints (Schneider & Zacharias 2012). Therefore, these methods are highly sensitive to any incorrect data (De Vries et al. 2010), and may achieve poor performance. At the same time, studies confirm that the use of experimental data and interface information into docking methods results in more reliable structures (Van Dijk, Boelens, et al. 2005). A recent critical assessment (Shih & Hwang 2013) on data-driven docking have shown that success rate depending on the content of interface information used to drive docking. Residues contact information gives better docked performance than simply using the interface or non-interface state of residues. To reduce the sensitivity of docked poses to experimental data, methods have used energy-based scoring functions to first generate docked poses and then took advantage of experimental data to re-score docking conformations during the refinement stage.

2.4.1.4 Re-ranking Docking Conformations

As discussed above, the most recent rigid-docking strategies are able to produce several native like conformations (Ritchie 2008) and the main difficulty lies in detection of these models (Lensink & Wodak 2010). Therefore, identifying the native pose by re-ranking docking poses is an active field of research (Li et al. 2003; Pierce & Weng 2007; Vreven et al. 2011). Since comparative studies (Lensink et al. 2007; Janin et al. 2003) have shown that energy-based scoring functions are error-prone, knowledge-based (statistical) methods have been proposed. Here we are focusing on scoring function used at refinement stage to re-rank docking poses.

In general the idea of knowledge-based functions is to gain general statistic and knowledge of the protein binding sites properties in order to distinguish correct binding sites in a set of docked decoys. Knowledge-based functions can be broadly divide into two groups: 1) knowledge-based Potential Functions: which use a databases (Keskin et al. 2004) such as PDB (Berman et al. 2000) or Dockground (Douguet et al. 2006) to learn about the potentials involved between interacting interfaces. These potentials can be atom-based or residues-based/coarse-grained. As discussed above, since atom-based potentials are vulnerable to conformational changes and to account for a faster processing speed, coarse-grained potentials are mainly used. Methods have also taken advantage of Machine Learning techniques to learn about potentials using larger number of energy terms. 2) Interface-based Functions: experimental data or predicted interfaces are the two main data used in these methods.

Moreover, methods have used the combination of these functions to perform predictions. For example, Interface -based Functions may include potential terms in their scoring functions.

2.4.1.4.1 Knowledge-based Potential Functions

Knowledge-based potentials have been successfully used in predicting protein structures (Sippl 1995). These methods measure the free energy of the protein by sampling the surface potential energy (Bernauer et al. 2007).

Different atomic knowledge-based potentials have been used to re-rank ZDOCK decoys, all of them producing better results than ZDOCK alone. ZRANK (Pierce & Weng 2007) proposed the use of a combination of three atom-based terms, i.e. van der

Waals, electrostatics and desolvation (atomic contact potentials (ACP)). Using docking benchmark 2.0, Pierce et al. (De Vries & Bonvin 2011) applied ZRANK to rank complexes generated by ZDOCK and managed to find a hit among the top 100 solutions for 83% of the cases containing a near-native pose. Another atom-based function (CFPScore) uses (Liu et al. 2006) a combination of packing density, contact size, atom-based potential of mean force (PMFScore) and geometric complementarily to re-rank models produced by ZDOCK. This score with an error rate of 5% can detect the true biological interface from crystal artefacts. Better result was achieved by RDOCK (Li et al. 2003) which performs energy minimisation followed by electrostatics and desolvation scoring functions to re-rank ZDOCK models. Usage of minimisation and generation of local minima models before re-ranking were shown to improve results (Viswanath et al. 2012). Similar approach was also used in PyDock (Cheng et al. 2007) which used electrostatic and atom-based desolvation to re-rank FTDOCK and ZDOCK conformations. Their investigation showed that these two terms are the main contribution in detecting good quality models. While PyDock takes into account both electrostatic and desolvation, GB-rerank uses desolvation energy alone to re-rank F2DOCK top 2000 models. F2DOCK (Chowdhury et al. 2013) is a FFT based docking method using electrostatic, interface propensity and hydrophobicity fast scoring function. Comparisons to ZDOCK 3.0.2 (Pierce et al. 2011) on Docking benchmark 4.0 shows that their performance are complementary to each other which suggests better results can be achieved by combining the results of these two docking tools (Chowdhury et al. 2013).

Atom-base scoring functions are vulnerable to any error or conformation change through the docking process. In addition, most docking models use coarse-grain representation of complexes to account for speed and flexibility. For example, PyDock speed was increased in pyDockCG (Solernou & Fernandez-Recio 2011) where coarse-grained potential was used instead of full-atom. On the other hand, ranking ZDOCK models using DFIRE - an atom-based statistical energy function (Zhang et al. 2005) - only produced good ranks for near-native models without atomic clashes. As a consequence, residue-based/coarse-grain scoring functions were designed. 3D-Dock (Jackson et al. 1998; Carter et al. 2005) uses RPScore (Residue level Pair potential Score) which relies on empirically residue pair potential derived from statistical analysis

of protein-protein interfaces. Murphy et al. (Murphy et al. 2003) combined atom-based ACP and residue-based RPScore to re-rank models. They concluded that those scores are complementary since RPScore creates more hits, while ACP gives higher rank to near-native models. Therefore, integrating residue- and atom- based potentials, an extension of ZRANK, IRAD (Integration of Residue- and Atom-based potentials for Docking) (Vreven et al. 2011), was introduced. Since these potentials can handle conformational changes, IRAD outperforms ZRANK when dealing with complexes of medium docking difficulty (Vreven et al. 2011). While IRAD combined three different residue potentials (H. Lu et al. 2003; Glaser et al. 2001; Tobi & Bahar 2006) DOCK/PIERR (Viswanath et al. 2012) used PIE (Ravikant & Elber 2010) residue potential along with atomic potential to re-rank rigid docking poses with minimised side-chain. Re-ranking performed by a combined atomic+residue potential scoring function performs better than ranking produced by each term separately. DOCK/PIERR compares favourably to leading docking methods such as ZDOCK+ZRANK, Cluspro, and PATCHDOCK+FIBERDOCK.

Practically, coarse-grained representations have been used to account for speed in docking methods. However, better accuracy requires to consider high-order interactions (Johansson & Hamelryck 2013). The above mentioned potentials are mainly based on pair-wise interface contacts (two-body) where nearest neighbour or contact residues are defined by arbitrary distance criteria (Krishnamoorthy & Tropsha 2003). These functions leave out multi-body and long-range interactions which are the contributors to the stability of protein complexes (Khashan et al. 2012). Therefore, scoring functions taking into account three or four-body statistical potentials have been formulated as probability tables (Khashan et al. 2012; Krishnamoorthy & Tropsha 2003; Feng et al. 2007; Li & Liang 2005; Ngan et al. 2006) similarly to previous pair-wise potentials (Yan et al. 2013; Ravikant & Elber 2010; Mintseris et al. 2007; Zhang et al. 1997).

DECK (Liu & Vakser 2011) and ITScore-PP (Huang & Zou 2008) two coarse grained, multi-body knowledge-based distance-dependent scoring functions have been developed for ranking docked protein complexes. One of the main difficulties in knowledge-based methods is to introduce a reference state. To address this problem, DECK uses five different reference states from a set of non-native matches (decoys),

and ITScore-PP performs iterations on 851 PDB protein complexes. ITScore-PP has relied on Delaunay tessellation representations (Krishnamoorthy & Tropsha 2003; Huang & Zou 2008) while DECK uses atomic environment where all interactions are taken into account (Summa et al. 2005) (an schematic example in Figure 2.11). Tessellation which is based on amino acid representation has been shown to provide valuable structural information such as molecular recognition (Bernauer et al. 2007). Both scores improve the near-native hits in comparison to ZDOCK and in the CAPRI ranking competition for target 32 (Lensink & Wodak 2010) DECK has shown to be the best scorer by, ranking 2 acceptable models in 3rd and 5th position out of 10.

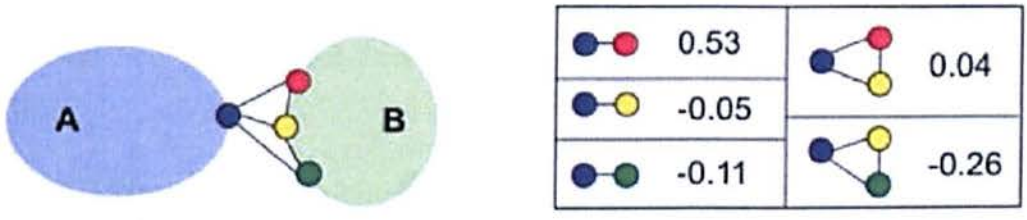


Figure 2.11: A schematic representation of Delaunay tessellation representations of protein surface on the left. On the right is an example of 2-body scoring (left in the table) and 3-body scoring (right in the table) of residue contacts. Taken from (Andreani et al. 2013).

Scoring functions have also been used alongside clustering of the docked poses at refinement stage. Although clustering itself is not a scoring function, it was shown that clustering of low energy conformations docking provides a simple mechanism for detection of near-native orientation (Ritchie 2008). For example, Cluspro (Comeau et al. 2004) works by initially calculating 70,000 docking models using DOT docking program (Comeau et al. 2004). As refinement stage, 1000 best models are selected using a statistical desolvation and electrostatic energy function. Cluspro performs additional clustering to select models with the most neighbours within a 9 Å C-alpha RMSD cut-off and scores them based on their cluster size. Recent improvements to Cluspro (Kozakov et al. 2010) have been achieved by first generating 1000 best energy models using PIER. Then these models are clustered and the 30 largest clusters are kept for stability analysis and structural refinement using Monte Carlo simulation and medium range optimization method, respectively. Using this method in rounds 13–19 of CAPRI, Cluspro outperformed 53 of 63 participating groups. Unlike Cluspro, in F2DOCK (Chowdhury et al. 2013) clustering takes place before scoring to penalise a

pose if a similar pose exists with a better score. Then knowledge-based potentials based on Lennard-Jones potential, the number of steric clashes, interface area, interface propensity and residue-residue contact preferences are used to re-score the top 2000 poses.

The aim of these scores is to detect the near-native docking models which are distinguishable by low binding affinity. These scores calculate the free energy of the complexes which relates to the binding affinity (Shen & Sali 2006). However, a comparison study (Kastritis & Bonvin 2010) of knowledge-based potentials in re-scoring docking conformations has shown poor correlation to binding affinity. Therefore, improvements of current scoring functions and possibility creating consensus and learning functions are required for describing binding affinities (Kastritis & Bonvin 2010).

2.4.1.4.2 Statistical and Machine Learning Functions

Since experimentally determined complexes provide useful information, statistical and machine learning techniques have been used to learn from available 3D structures. These techniques can handle large number of independent sets of input data to characterise protein binding sites.

Some approaches have used statistical learning to characterise binding sites based on their properties in a complex. Bernauer et al. (Bernauer et al. 2007) have used low-resolution Voronoi representation of protein surfaces and generated 84 parameters describing surface geometry and physico-chemical properties. These data was used by a genetic algorithm to create an efficient scoring function that optimises the area under the curve. For 4 out of 5 CAPRI targets Bernauer et al. achieved better ranking than HADDOCK. To further improve their work, Voronoi representation was used with an optimised scoring function using probabilistic multi-classifiers adaptation (Bourquard et al. 2011). In the CAPRI ranking competition, at least one acceptable or better solution was found within the top 10 models for 4 out of 8 targets.

Recently, Othersen et al. (Othersen et al. 2012) have used Mutual Information (MI) to select structural features which discriminate between near and far-native conformations. Structural features in their study can be grouped into two main categories: (i) pairwise interface contacts of amino acids, and (ii) size of the solvent-

accessible surface area. They have proven that MI is dimensionless and can be applied to a wide variety of structural features and different size datasets. Experiments on Dockground (Douguet et al. 2006) dataset using a fivefold cross validation on 11 structural features showed they could detect correct or near-native structure in 41 cases out of 60.

Other approaches have used supervised machine learning, i.e. SVM, trained by various properties. But instead of just classifying models as native or non-native they have used probabilistic SVM which allows them to generate a ranking of docked poses. This can be achieved by the position of a complex relative to a hyperplane of the SVM which estimates the probability of that complex to be native or not. Mertin et al. (Martin & Schomburg 2008) trains a probabilistic SVM from proteins in DBMK 2.0 (Mintseris et al. 2005)using properties such as evolutionary relationship, interface propensities, gap volume, buried surface area, residue- and atom-based pair potentials on residue and atom level. Re-ranking using this SVM-based score has shown better performance than knowledge-based potentials of ACP and RPScores. Being different in the training properties, PROCOS (Fink et al. 2011) uses a probabilistic SVM which combines van der Waals and electrostatic energies and knowledge based pair potentials. Comparison to ZRANK shows more near-native hits are detected within the top 10 models.

Unlike PROCOS and Mertin et al.'s approach which are all supervised, Zhoe et al. (Zhao et al. 2011) proposed a semi-supervised (TSVM) along with a supervised (SVM) method trained using interface features generated from interaction interfaces of protein complexes. The features can be grouped into (i) residue type statistics, (ii) surface patch parameters, (iii) secondary structure statistics, and (iv) number of contacts. They generated docking conformations using PatchDock (Duhovny et al. 2002) and RosettaDock (Gray et al. 2003) for 124 proteins for docking benchmark 3.0 (Hwang et al. 2008) and performed two test. First, they classified near-native and non-native structures by leave-one-out cross-validation experiments which showed that SVM and TSVM performed similarly with accuracies around 80%respectively. Second, they tested the ability of SVM/TSVM to re-rank docking conformations and compared it with the original ranking provided by the docking methods. They managed to improve the rank of the hit-pose for PatchDock in 22 targets with an average increase in rank of 3 and 5 positions for SVM and TSVM, respectively. These results suggests that SVM

and TSVM are good classifiers of native and non-native structures but do not provide significant improvement in re-ranking models.

2.4.1.4.3 Knowledge of Predicted Interfaces

As discussed in data-driven docking procedures, usage of experimental data in pre-filtering and during docking causes limitations. Therefore, methods were developed to exploit those data after docking, i.e. for ranking docking models (Pons et al. 2010; Schneidman-Duhovny et al. 2012). Although these techniques improve the detection of near-native, they are limited to availability of experimental data. Therefore, in their absence, prediction of interface residues (De Vries & Bonvin 2011; Li & Kihara 2012) may be a solution. As discussed in the data-driven docking, interface residues information can be exploited in two complementary ways (Qin & Zhou 2007a) for providing better rankings of docking. They can be applied as pre-filtering (Korkin et al. 2006), to narrow the initial search space of the docking models, or alternatively as post-filtering, for re-ranking docking conformations by calculating the similarity between the interfaces of the docked models and the predicted ones. On one hand, pre-filtering limits the search space from the start but is less practical to use since the constraint should be imposed on the algorithms of the docking methods (Qin & Zhou 2007a). On the other hand, post-filtering is more practical since it can be combined with other scores (such as energy or knowledge based) to re-rank all generated models from several docking algorithms.

Computationally predicted protein interfaces have significant impact on exploration of interactomes (Zhang et al. 2012; Mosca et al. 2012; Vakser 2013). Therefore, many methods have investigated the use of binding site predictions (alone or in combination with knowledge-based potentials) to score docking conformations. Further insight into the value of using interface knowledge was gain when Zhoe et al. (Zhou & Qin 2007) performed a comparison between ZDOCK rankings and those methods taking advantage of simulated interface predictions where a fraction of real interfaces were substituted randomly by non-interface residues. Experiments show that knowledge of a few correct interface residues – up to 60% of the interface substitution – is sufficient to outperform ZDOCK rankings. In spite of those encouraging results, very

few of the interface prediction methods introduced in the previous section have been used to rank docking models.

As discussed in Protein Interface Prediction section, residues are more conserved and can be detected by analysing the evolutionary rates among protein families. Therefore, conservation score have been integrated into re-scoring function. Trees et al. (Tress et al. 2005) used evolutionary trace (ET) to re-rank docking conformations of GRAMMX and HEX. 3 acceptable models were detected for 7 CAPRI targets of CAPRI round 3-5 (Méndez et al. 2005), which was a significant result for a predictor based on sequence only information. Later on, GRAMM-X (Tovchigrechko & Vakser 2005) included the degree of interface residues evolutionary conservation along with knowledge-based potentials to re-rank models at refinement stage which resulted in 2 medium accuracy predictions of CAPRI round 5. Heuser et al. (Heuser et al. 2005) investigated the conservation of a smaller set of residues (Phe, Met, and Trp and their polar neighbour residues) to detect binding site and re-rank models. While these methods use a set of conserved residues, DockRank (Xue et al. n.d.; Xue et al. 2010) investigated the use of PS-HomPPI (Xue et al. 2011) in ranking docking conformations since PS-HomPPI was shown to be one of the best performing evolutionary-based predictors using MSA of QPs. DockRank was tested on docking models generated by Cluspro for docking benchmark 3.0 and comparing its rankings with those provided by Cluspro shows that in 61 cases out of 64, DockRank identifies better models. While these methods have been looking at the evolutionary rate of a site in a MSA, other methods use evolutionary information generated from a residue pair in complex. SCOTCH (Madaoui & Guerois 2008) have investigated mutations of the protein **complex** interfaces which disrupt the physicochemical complementarity of protein interfaces in order to detect native-like contacts. Compared with mutation scores which calculate conservation score of each protein individually (evolutionary rate of a position in a MSA) (Pazos et al. 1997; Afonnikov et al. 2001), SCOTCH shows improvement in re-ranking FTDOCK decoys. A more sophisticated score, InterEvScore (Andreani et al. 2013), takes advantage of evolutionary conservation based on interface co-evolution which is combined with two- and three coarse-grained statistical potentials to rank docked models. Evolutionary conservation was not only based on MSA but also considered interface co-evolution among contacting residues. InterEvScore showed

significant improvement in scoring docking models in comparison to ZDOCK and ZRANK. This can be explained by complementary information gained from both conservation and two- and three statistical potentials. A totally different approach of generating conserved residues has been used by CONSRANK (Oliva et al. 2013). In this approach the two protein chains are first docked and then conservation of inter-residue contacts is investigated among the decoys ensemble. The conservation is simply based on the frequency of contacts at each site among the ensemble decoys. Then, these decoys are re-scored where higher ranks are given to decoys which represent the most frequently observed contacts at their binding sites. The reason behind CONSRANK is that docking algorithms are capable of generating a large numbers of near-native docking poses which can be used as a guide toward the correct pose. Comparisons to scorer groups involved in CAPRI (Lensink & Wodak 2010), CONSRANK improves the fraction of near-native hits in the top 10 models by 9.9.

With the increase in the number of 3D structures new paradigm of interface predictors have been used for ranking docking models. Gottschalk et al. (Gottschalk et al. 2004) produced a pivotal study extending their interface prediction method, Promate (Neuvirth et al. 2004) (see Protein-Protein Interface Prediction section), to re-ranking docking models. By calculating the tightness of fit of the two docked proteins against the predicted binding sites, they obtained a near-native solution in 77% of the cases. Similarly, Qin et al. (Qin & Zhou 2007a) showed that usage of cons-PPISP Interface prediction improves the ranking for 8 out of 20 CAPRI targets. These results show that true positives in interface predictions with a low rate of false positives are capable of providing reasonable rankings. Therefore PINUP with better performance than Promate and cons-PPISP (Zhou & Qin 2007), (see Protein-Protein Interface Prediction section), was used to re-ranking Z-DOCK and Rosetta decoys producing a significant increase in detection of near-native conformations (Liang et al. 2009). Since combinations of interface predictors improves individual ones in MetaPPI, Huang et al. (Huang & Schroeder 2008) exploited the use of MetaPPI predictor for ranking BDOCK docking models. As discussed in the Protein Interface Prediction section MetaPPI combines predictions of Promate, PPI-Pred, PPISP, PINUP, and SPPIDER. The re-scoring results on a benchmark of 63 complexes showed that for 16 out of 20 cases of enzyme-

inhibitors Huang et al. achieved hits with low RMSD compared to the native structure while the results were worse for non enzyme-inhibitors.

The above methods model interacting residues' patterns using two-body (or three or four-body) terms while multi-residue interactions are the ones which contribute to the stability of protein complexes. Therefore, to overcome the limitation of the number of terms for representing the pattern of interacting residues Khashan et al. (Khashan et al. 2012) proposed SPIDER, which relies on the frequency of interaction patterns taking place between interfacial residues of protein complexes. This relied on the creation of a library of contacts where each protein complex (here Dockground is used) was converted into a graph of residue contacts and then using sub graph mining methods, common interfacial patterns were generated. The score of a docking model depends on its interface geometric similarity to the patterns in the library and the frequency of that pattern. Although, SPIDER was shown to outperform ZRANK and was ranked 6 out of 28 in CAPRI round 21, InterEvScore performed better. The main reason is that statistics on interface contact are limited in the multi-body docking potentials, whereas InterEvScore is based on a reliable two- and three-body interaction statistics database (InterEvol (Faure et al. 2012)).

2.4.2Template-based Docking

Although many improvements have been achieved using template free docking approaches, scoring functions are still not good enough to detect near-native models from a pool of false positives. Moreover, increase in the size of proteins negatively affects the computational time of docking methods (Pons et al. 2010). Therefore, these docking methods are not suitable for large-scale PPI. With the increases in the number of experimentally determined protein-protein structures, a new trend of docking methods emerged. These methods take advantage of the 3D structures available in PDB as a template to generalise the model of interactions between two proteins. In addition, these methods can predict whether two protein chains interact or not (Ogmen et al. 2005). For detecting templates the search strategy of those template-based methods can be based on 1) sequence similarity (Aloy et al. 2003; Kundrotas et al. 2008; Launay & Simonson 2008) 2) Threading sequence on structure (Lu et al. 2002; Chen & Skolnick

2008; Mukherjee & Zhang 2011) or 3) structural alignments either locally or globally (Zhang et al. 2012; Tuncbag et al. 2012; Günther et al. 2007).

Template-based docking started with the pivotal work of Aloy et al. (Aloy & Russell 2002) using the homology concept. Using a known 3D complex and the homologous sequences of each interacting protein, putative protein pairs are scored using empirical potentials derived from the 3D complex. This method provides a mechanism for deciphering whole interaction networks but can only be used when template and query sequences are from the same Pfam family. This method combined with experimental techniques, was used to decipher the domain–domain complexes in Yeast (Aloy et al. 2004). In the same line, Davis et al. (Davis et al. 2006) predicted the higher-order complexes of protein in *Saccharomyces cerevisiae* using the sequence similarity to structurally known complexes. This method for the first time introduced higher-order complexes prediction based on structural templates.

Intrinsic structural information has been combined with sequence alignment similarities, in order to detect templates. HOMBACOP uses profile-to-profile alignment combined with structural information of the template interfacial residues and query protein predicted residues (Kundrotas et al. 2008). Even for query proteins of which highly homologous templates does not exist, reasonable models can be produced using this profile-to-profile alignment (Kundrotas et al. 2008). Similarly, Launay and Simonson (Launay & Simonson 2008) have improved Needleman–Wunsch alignment of target proteins on templates by including solvent accessibility of interfaces residues. Then they added energy calculation to select the best model. While these methods look at sequence homology to detect templates, KBDOCK (Ghoorah et al. 2011) creates a 3D domain-domain interaction database in which for each Pfam domain family homologous complexes are structurally aligned and binding sites are analysed. Using KBDOCK, complex templates were detected for 45 out of 73 complexes and, for 24 complexes, templates were found for only one of the chains. It is suggested that this information can be used to guide template-free docking.

The main problem with the above mentioned methods is that they are limited to available homologous sequence and structure. These methods can only account for ~20% of the know PPI (Kundrotas et al. 2008). Therefore, methods have gone beyond homology, using two different techniques for finding templates: usage of threading of

protein sequences on template structures, and usage of the 3D structures of the query proteins to search for templates with similar structure either globally or at interface level.

Single threading has been widely used to predict protein 3D structure, in which the sequence of query protein is aligned on a library of folds and the best fit hit is selected to model the query protein structure. Since the physical principals of protein folding and binding are similar, PPI methods have adopted threading (Vakser 2013). MULTIPROSPECTOR (Lu et al. 2002) (Figure 2.12) has extended single threading to multimer threading, in which the sequence of the target protein chains are separately threaded on a library of protein-protein complexes. A template is selected by not only how well both chains fit on the template but also based on statistical interfacial pair potentials and stability of the target complex. This method allows exploiting more distantly related protein templates and has been used to predict *Saccharomyces cerevisiae*S protein interaction network in (L. Lu et al. 2003). In these methods, since the query chains are aligned separately on the template, the cooperative relationship between chains such as burial residues, conformational changes and binding specificity are not taken to account during the threading and alignment procedure. To deal with this issues, following a threading step, M-TASSER (Chen & Skolnick 2008) performs backbone and orientation refinements plus energy calculations. Instead of using a refinement stage, COTH (Mukherjee & Zhang 2011) uses a co-threading technique in which both protein chains are simultaneously aligned on protein-protein templates. Methods based on threading generate models using the sequence of the target protein while docking methods are based on the knowledge of the protein structure. Therefore, methods have used protein complexes with similar structure to the QP chains as templates to model the QP complex. This similarity can be either global or local at interface level, but in both cases methods using structural similarities cover more predictions than threading methods (Vreven et al. n.d.).

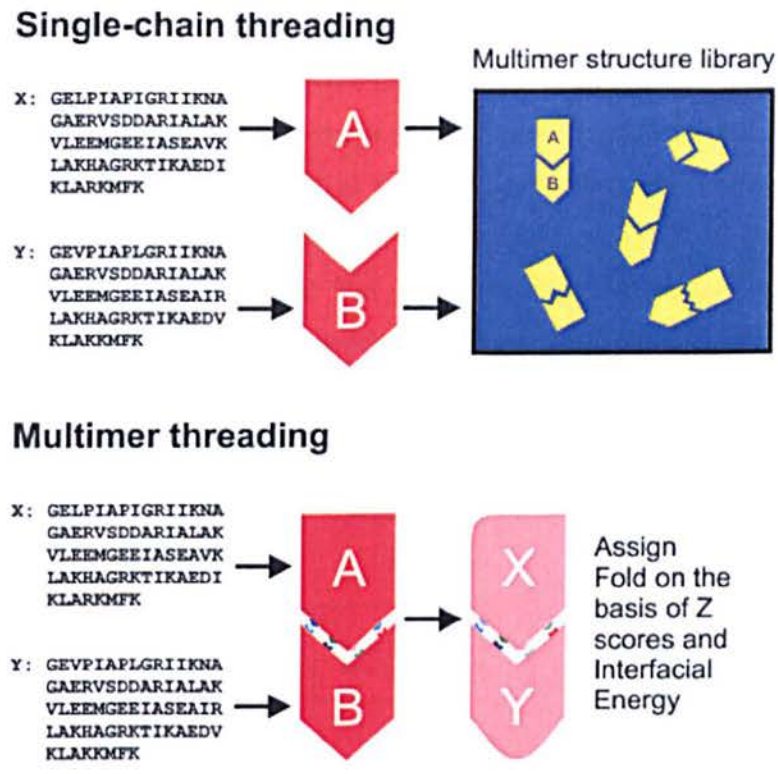


Figure 2.12: MULTIPROSPECTOR principle. First each chain is separately threaded on dimers available in a library of templates. Once both chains are threaded on a dimer, statistical interfacial pair potentials are used to evaluate the energy of that dimer complex. Taken from (Lu et al. 2002).

Since similar interface architectures have been detected among proteins displaying different functions and structures (Keskin & Nussinov 2007), it has been proposed Template-based docking using interface similarities but ignoring global sequence and structure similarities. Moreover, since accuracy of high-throughput modelling relies on correctly modelling the binding sites (Kundrotas & Vakser 2010), it could be shown that complexes created by the alignment of interfaces on templates are more accurate than those relying on the whole structure (Sinha et al. 2010). Based on these results, ISEARCH (Günther et al. 2007) uses surface patches of target proteins to search in a patch-based library of domain-domain interactions. Once a hit is found target interfaces are aligned onto the template. To impose more accurate detection of the correct template interface, PRISM (Tuncbag et al. 2012) (Figure 2.13) combined geometric similarities with evolutionary conserved residues (hotspots) (Ogmen et al. 2005). The target surface is structurally aligned on the library of protein-protein interfaces to detect the best aligned templates. In this alignment at least one hot-spot

residue of template should be aligned with a target residue. Once templates are detected, monomer chains are transformed onto the template and FireDock (Mashiach et al. 2010) is used to resolve clashes and rank putative models based on energy. This is followed by a refinement stage to add flexibility to side-chain and backbone. This stage makes PRISM different from other template-based methods which only use rigid-body alignment. PRISM is computational less expensive in comparison to ZDOCK and PatchDock since it can remove false-positives by using structural templates (Tuncbag et al. 2011). As a consequence, PRISM has been used for constructing signalling pathways (Kuzu et al. 2012).

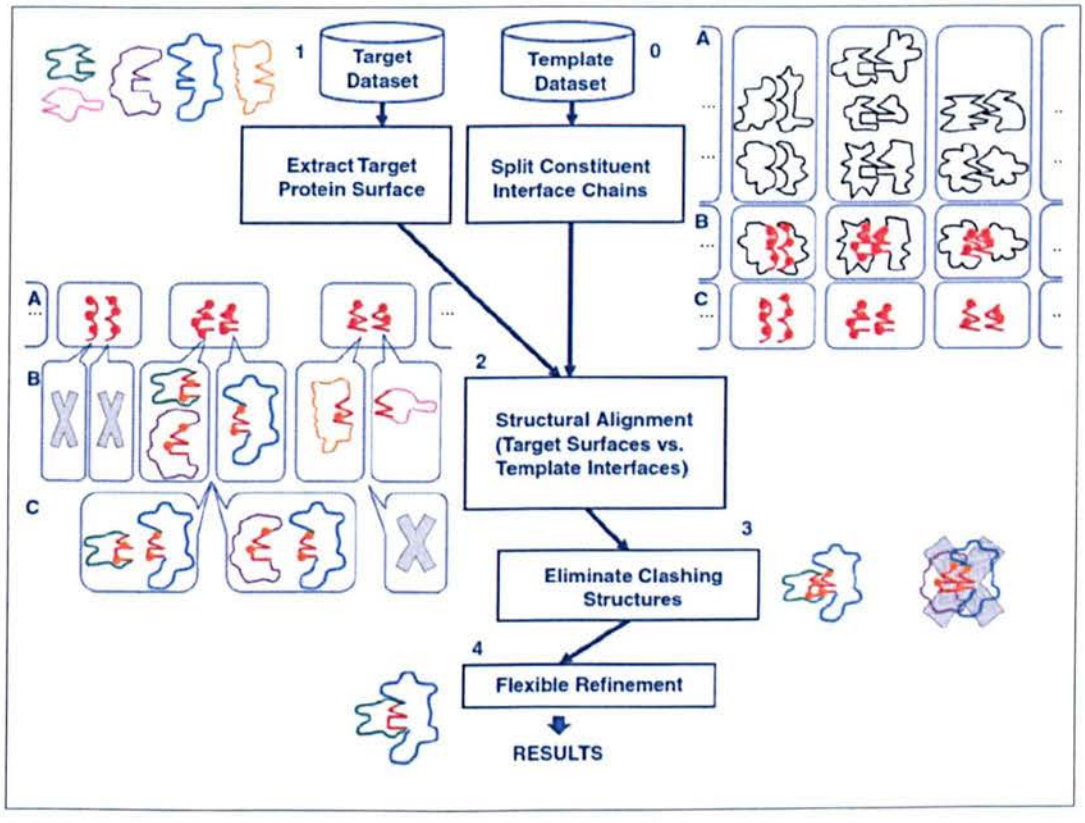


Figure 2.13: PRISM pipeline for template-based docking using interface similarities. Step 0: a Template Dataset is created. A) Available complexes are retrieved and similar structures are clustered together, B) one representative of the clusters is selected and the interfaces are detected, and C) only the interfaces are stored in the dataset. Step 1: The surfaces of the target proteins are identified. Step2: Target surfaces are structurally aligned on the template's interfaces and possible new complexes are evaluated. Step 3: FireDock is used to remove clashes. Step 4: backbone and side-chain flexibilities are added. Taken from (Kuzu et al. 2012).

A recent method, PrePPI (Zhang et al. 2012), has deciphered ~2 million PPIs including ~60 000 yeast PPIs and ~370 000 human PPIs(deposited in PrePPI database

(Zhang et al. 2013)). The two main highlights of this method are: (i) usage of structural neighbours: they can be found for most proteins in PDB and even remote ones have shown similar binding properties (Q. C. Zhang et al. 2010). (ii) usage of homology modelling for the input targets whose X-ray model is not available to cover more predictions. The pipeline of PrePPI (Figure 2.14), which has shown to be as accurate as experimental methods, is described below. First, PrePPI searches for close or remote structural neighbours using experimental or homology models of the input targets. Once a complex is found which contains chains from the pool of structural neighbours on both sides, that complex is taken as template. After aligning the target chains on the template, five scores are generated and combined in a Bayesian framework to predict the likelihood of two protein interacting. These five scores are based on structural features generated from global and interfacial geometric fitness of targets on templates plus interface properties (such as evolutionary conservation, residue type and statistical probability of being in interface) generated from known PPIs.

There are three main reasons for the accuracy of this method (Zhang et al. 2012). 1) Considering both homology-modelling and structural neighbours allows the exploitation of larger number of models. 2) Scoring is efficient and at the same time has a high discriminative power on closely related family members. 3) Bayesian framework can perform a reliable prediction combining even with weak and independent interaction signals. To account for speed, PrePPI performs a structure-based sequence alignment of targets on the template. Therefore, this method does not provide any 3D structure of the new complex.

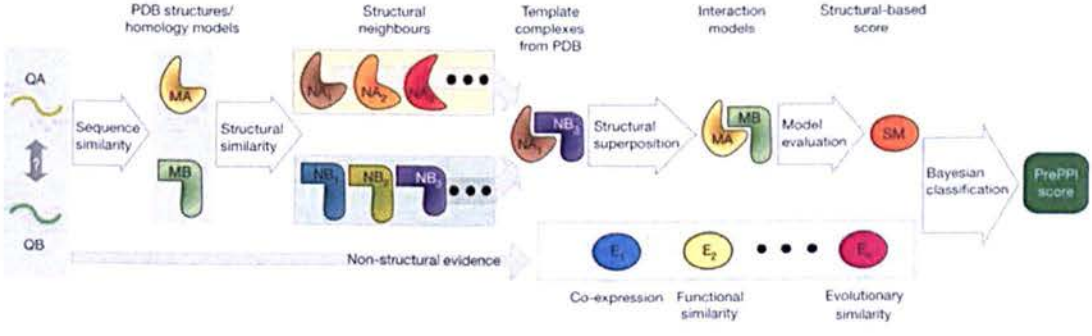


Figure 2.14: PrePPI Pipeline for large-scale template-based docking. PrePPI searches for close or remote structural neighbours of the input targets. Once a complex containing both sides is found, that complex is taken as template. Target chains are aligned on template and five scores are generated and combined in a Bayesian framework to predict the likelihood of two protein interacting.

As demonstrated above, the accuracy of Template-based docking depends on the availability of templates. A recent large-scale study (Kundrotas & Vakser 2010) confirmed that for almost all PPI a template is available in PDB, where one-third are of good quality (interface RMSD<5 Å). Also, since a library of interfaces is close to complete (Gao & Skolnick 2010), a representative binding site can be found for any given interface (Zhang et al. 2012; Q. C. Zhang et al. 2010). Therefore, with the availability of templates, low-resolution template-based methods are effectively used for large-scale PPI. It should be highlighted that template-based docking for prediction of new PPI is possible only when the structure of the individual components are available or if they can be generated by homology modelling (Kundrotas et al. 2012; Zhang et al. 2012). To summarise, Figure 2.15 shows the results for five genomes which have the largest known number of PPIs (Vakser 2013). The red sections represent protein complexes whose X-ray is available. Green section shows complexes in which protein sequences are used as templates to generate new complexes from the X-ray ones. Blue represents complexes which are generated by structural templates (also including ones which the individual chain can be modelled by homology). For 99% of the PPIs where the structures of the individual components (or their models built by homology modelling) were available, a structural complex template was available in PDB. As it is shown, using structural templates a larger set of PPI can be modelled in comparison to using sequences as templates. 'no template' indicates that structures were not available for both of the individual components (Vakser 2013).

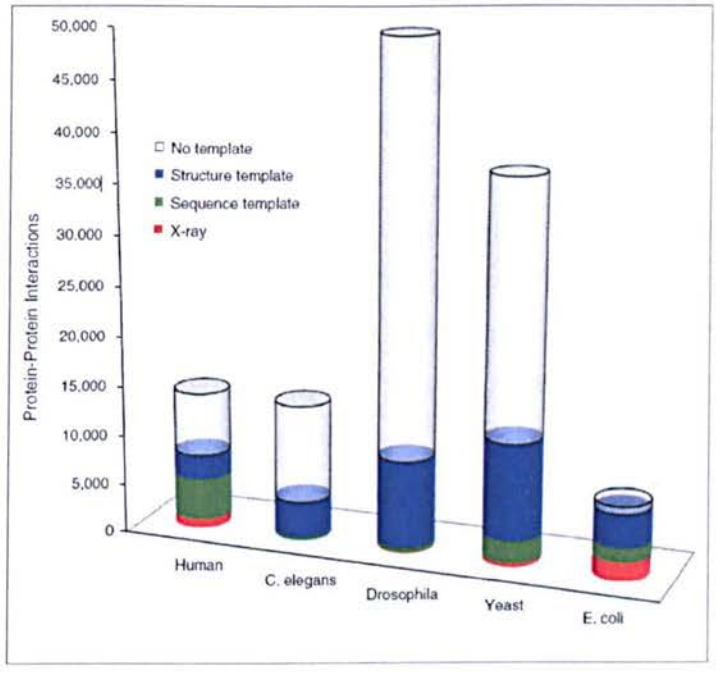


Figure 2.15: PPI for five genomes. Red represents complexes with a x-ray structure. Green refers to complexes modelled by sequence templates. Blue are complexes which are modelled by a template structure. In this group either the x-ray of individual proteins are use if available or homology modelling is used. Taken from (Vakser 2013).

2.4.3Conclusion

With the increase in experimentally determined structures and the emergence of template-based docking , genome-wide PPI deciphering can happen in the foreseeable future (Vakser 2013). Since these high-throughput methods are low-resolution, "The first credible model of a cell will be low resolution" (Vakser 2013). Eventually, high resolution modelling should be used to get a clearer picture of what is happening inside the cell (Vakser 2013).

Although template-based docking has enabled large scale PPI prediction, template-free docking still remains highly important. First, many proteins in the cell do not correspond to the energetically stable crystallised templates (Vakser 2013). Second, although there has been an increase in the number of templates (Kundrotas et al. 2012), still some PPI lack templates or the quality of the template is really low (Movshovitz-Attias et al. 2010). Third, free-docking approaches with their refinement stage have made it possible to generate high resolution structures (Vakser 2013) which are important for understanding the molecular mechanism of protein contacts. Moreover,

results obtained from template-based and template-free docking are complementary and combining their docked models has shown to improve the detection of near-native models (Vreven et al. n.d.). This shows that both methods have their own strength and limitations.

As discussed above, template-free docking methods predicting protein complexes compete against each other in CAPRI competitions (Janin et al. 2003). A comparison of these methods in the latest CAPRI rounds 22-27 is displayed in Table 2.3. These rounds consist of 10 Targets, T46 to T58 which T55 and T56 were only for scorer and T52 was cancelled. 'Stars' are given based on the quality of the predicted complex in comparison to the native complex: * (acceptable), ** (medium), and *** (high) whereas no star means incorrect prediction. Ranks are given to each group based on their overall performance. In total more than 40 groups participated, here we show the results for the top 12 ranked groups. Based on this comparison Cluspro (Comeau et al. 2004; Kozakov et al. 2006), which was ranked 9th, is the best performing docking *server*. Therefore, in this thesis we have used Cluspro which is also a really fast docking method to generate docked poses. It should be noted that our work in this thesis is not limited to Cluspro and can be used to re-rank any docked model regardless of the tool used to generate them.

HADDOCK has also ranked as the second best *server*. The first ranked group (Bonvin) is composed of the developers of HADDOCK, which shows that the use of HADDOCK along with human interpretation of docked models can significantly improve docked model prediction.

Table 2.3: Template-free docking performance in CAPRI rounds 22-27.

Rank	Group	T46	T47	T48	T49	T50	T51	T53	T54	T57	T58	Summary: #Targets / *** + ** + *
1	Bonvin	*	***	*	*	**	*	**		**	*	9 / 1 *** + 3 ** + 5 *
2	Bates		**	*	*	*	*	*		*	**	8 / 2 ** + 6 *
3	Vakser		***	*	*	*			*	*	*	7 / 1 *** + 6 *
4	Vajda		***	**	*	**		***		**		6 / 2 *** + 3 ** + 1 *
5	Fernandez-Recio		***	*	*	**		**			**	6 / 1 *** + 3 ** + 2 *
5	Shen		***	**	**	*		**	*			6 / 1 *** + 3 ** + 2 *
7	Zou		***	**	*	*				**	*	6 / 1 *** + 2 ** + 3 *
8	Zacharias		***	*	*	*		*		*		6 / 1 *** + 5 *
9	ClusPro		**	**	*	**		**		*		6 / 4 ** + 2 *
10	Eisenstein		***	**	*	**		*				5 / 1 *** + 2 ** + 2 *
10	Grudinin		***			**		*		*	**	5 / 1 *** + 2 ** + 2 *
12	Gray		***					*		*	**	4 / 1 *** + 1 ** + 2 *
12	HADDOCK	*	***		*					**		4 / 1 *** + 1 ** + 2 *
12	Seok		***					**		*	*	4 / 1 *** + 1 ** + 2 *
12	Weng		***		*	*		**				4 / 1 *** + 1 ** + 2 *

Template-free docking approaches have been successful in generating near-native docking poses. However, detection of native-like conformation among a pool of decoys remains a challenge. Therefore, several scoring functions have been proposed to re-rank docking conformations at the refinement stage. Among them, usage of predicted interface residues has shown great popularity since other methods have limitations: (i) knowledge-based potentials show low correlation to binding affinities (ii) experimental data are not available for all proteins and (iii) usage of learning strategies has not significantly improved the knowledge-based potentials scoring functions. With improvement in protein interface prediction methods, re-ranking based on that knowledge can improve detection of near-native models. With the increase in experimentally determined proteins and the recent developments in template-based interface predictors, further improvements in re-ranking methods using predicted interfaces should be further investigated.

2.5 Datasets and Metrics

2.5.1 Protein Interface Prediction Evaluation

In order to compare the performance of interface predictors it is important to calculate their True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) rates (Yan et al. 2004). TP refers to interface residues correctly predicted as interface, FP to non-interface residue wrongly predicted as interface, TN to non-interface residues correctly predicted as non-interface and FN to interface residues wrongly predicted as non-interface. The correctness and wrongness of predictions are calculated in respect to the ground truth (GT), which is defined as the X-ray structure of the target protein in its complex form. To summarise these four figures into a single performance measure a few metrics have been proposed. Below these metrics are introduced where each one evaluates the predictor from a different aspect. Note that these metrics are expressed as percentage by multiplying by 100 but for simplicity *100 is not demonstrated in the formulas below.

To study the quality of predicted interface residues in respect to GT interfaces, *recall* is used:

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

In other word recall (also called sensitivity) evaluates the percentage of correctly predicted interfaces. A complement measure to recall is introduced by *precision* which evaluates how many of the predicted interfaces are actually a GT interface:

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

To summarise, high recall means that the predictor has correctly predicted most of the GT interfaces while high precision means that the list of predicted interfaces contains more correct predictions than wrong prediction. The two examples below show how *precision* and *recall* complement each other:

Example 1: Assume a protein with 10,000 residues which has 100 interfaces. If a predictor, predicts all residues as non-interface except one which is correctly predicted

as interface then it will achieve a *precision*~100% while this predictor is not of any use. The *recall* of this predictor will be ~1% which will highlight this problem.

Example 2: if for the above example the predictor, predicts the entire residues as interface then *recall* will be ~100% while *precision*~1%.

Since, precision and recall do not capture the information of TNs, *specificity* was introduced which explains the percentage of non-interfaces which are correctly predicted as non-interfaces:

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

None of the above mentioned metrics consider the four figures of (TP, TN, FP and FN) at the same time which can bias the performance comparison. Therefore, metrics which integrate all the four figures were introduced (Baldi et al. 2000).

Accuracy has been one of the widely used metrics which express the ration of correctly predicted interface and non-interface residues to the total number of cases:

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

But one drawback of using accuracy is that in cases such as above Example 1, the accuracy will be ~99%. Therefore, another measure, F1 score was introduced which, calculates the harmonic mean of *precision* and *recall*:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Since F1 score does not consider true negative rate, Matthews correlation coefficient (MCC) was introduced:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}}$$

MCC has shown to be effective especially for predictors which are biased because of the imbalances in their training set. For instance in situations similar to Example 1 mentioned above, MCC has shown to discriminate between predictor performances.

A widely used method to express specificity against recall in a binary classification predictor is receiving operator characteristic (ROC) plots. Roc curves, express a predictors false positive rate (1-specificity) relative to true positive rate (recall) at various threshold values. This curve can be expressed by calculating the Area under the Roc Curve (AUC) which evaluates if positive samples are ranked higher than negative samples. Therefore, for a perfect predictor AUC will be 100% and for a predictor no better than random will be 50%. It is also possible to get a curve worse than random which means that the predictor has a negative correlation with the actual answer.

The above mentioned metrics capture different aspects of the predictors and therefore, all of them are required to give an insight to the predictor's performance. In this thesis we have exploited all these metrics to compare interface predictors.

2.5.2 Docking Algorithm Evaluation

The performances of docking algorithms are evaluated in CAPRI (Critical Assessment of Predicted Interactions) (Janin et al. 2003), which is a biannual community-wide experiment for docking. In CAPRI competition the aim is to dock protein-protein complexes from the individual components –extracted from the PDB (Berman et al. 2000)- of the final target. Up to now (August 2013) CAPRI has had 28 rounds, modelling 59 targets. It should be noted that some of the targets have been protein-RNA complex, such as T33 (T stands for Target), and some of the individual components might require homology modelling, such as components of T51 and T59, since their structures are not available in PDB.

Predictors can submit up to 10 docked models to CAPRI, which will be evaluated against the unpublished experimental structures (Janin 2013). In addition to prediction assessment, in recent years CAPRI has introduced experiments for scorer groups to evaluate their re-ranking performance. In this part, for each target, the

predictors will upload hundreds of their docked poses creating a big repository of docked models. Then, the scorers will use their scoring methods to re-rank these models and submit at most the top 10 models.

To assess the submitted models of predictors three main criteria is used by CAPRI (Janin 2005). Assume that receptor (R) is the larger component and ligand (L) is the smaller component:

1) $L-RMSD$: measures the root mean square displacement (rmsd) of the backbone atoms (C-alpha atoms) of the target and model's L, after optimally aligning the R of target and model.
2) $I-RMSD$: measures the root mean square displacement (rmsd) of the backbone atoms (C-alpha atoms) of all interface residues of the target and model's L, after optimally aligning the R of target and model. Interface on L is defined as all residues that have atoms less than 5 Å apart from the R.
3) f_{nat} : measures fraction of native contacts using $f_{nat} = n_c/N_c$ where N_c is the total number of residue pair contacts in the target complex and n_c is the number of native contacts of the target which is also present in the docked model.

Based on these three criteria CAPRI evaluates the listed inequalities available in Table 2.4 (going from top to bottom) to give starts to the quality of each docked model: no star (incorrect), * (acceptable), ** (medium), and *** (high).

Table 2.4: CAPRI Assessment Criteria as shown in (Lensink & Wodak 2010).

Incorrect		$f_{nat} < 0.1$	**OR**	$L-RMSD > 10.0$	**AND**	$I-RMSD > 4.0$
Acceptable		$f_{nat} \geq 0.3$	**AND**	$L-RMSD > 5.0$	**AND**	$I-RMSD > 2.0$
	OR	$(0.1 \leq f_{nat} < 0.3)$	**AND**	$(L-RMSD \leq 10.0$	**OR**	$I-RMSD \leq 4.0)$
Medium		$f_{nat} \geq 0.5$	**AND**	$L-RMSD > 1.0$	**AND**	$I-RMSD > 1.0$
	OR	$(0.3 \leq f_{nat} < 0.5)$	**AND**	$(L-RMSD \leq 5.0$	**OR**	$I-RMSD \leq 2.0)$
High		$f_{nat} \geq 0.5$	**AND**	$L-RMSD \leq 1.0$	**AND**	$I-RMSD \leq 1.0$

In this thesis, we require to compare the performance of different scorer methods in re-ranking docked conformations. These docked models are evaluated by CAPRI

criteria but the four categories of CAPRI (incorrect, acceptable, medium and high) do not provide a continuous ranking of all models. Therefore, $L-RMSD$ and $I-RMSD$ are used separately to provide two gold standard rankings of docked models. f_{nat} has not been used to generate another ranking list since f_{nat} can only discriminate between relatively good configurations- all models failing to predict a single interface residue receive a score of 0.

As it was proposed by (Xue et al. n.d.), different scorer methods can be evaluated by calculating the Pearson's chi-squared statistic between a gold standard ranking of models and rankings generated by scorer methods. The chi-squared statistic (x^2) determines the goodness of relationship between a set of observed and a set of expected values:

$$x^2 = \sum_{k=1}^{n} \frac{(observed_k - expected_k)^2}{expected_k}$$

Here, $expected_k$ is the rank of the model k in the gold standard and $observed_k$ is the rank assigned to model k by a ranking method. Since the number of docked models may differ between protein pairs, the chi-squared statistic is normalized using the total number of docked models produced for that protein pair, m:

$$normalizedx^2 = \frac{x^2}{m}$$

$normalizedx^2$ represents the similarity between two ranking lists by giving higher weights to the models that are ranked higher based on the gold standard: correct ranking is more important for top-ranking models than lower-ranking models. Perfect ranking would return a value of 0.

2.5.3Datasets

Three standard benchmark datasets which are widely used by protein interface and docking predictors, are used in this thesis: Ds56unbound (Janin & Wodak 2007), Docking Benchmark 3.0 (DBMK3.0) (Hwang et al. 2008) and Docking Benchmark 4.0 (DBMK4.0) (Hwang, Vreven, Janin, et al. 2010) . In this thesis, Ds56unbound has been mainly used as training set while, DBMK3.0/4.0 has been used for evaluating the interface predictors and docking model ranking approaches. These datasets contain high-resolution protein structures both in their unbound and bound forms.

Ds56unbound is comprised of 56 unbound chains generated from 27 CAPRI targets, T01~T27 (Janin & Wodak 2007). In total, it contains 12173 residues including 2112 interacting ones (Jordan et al. 2012). Since interface residues are not explicitly provided in DS56unbound, they were generated from the interface residues in their bound form (DS56bound). The DS56bound contains 12123 residues including 2154 interacting ones. Interfaces are defined using the same definition as CAPRI's, i.e. all residues of a protein chain that have atoms less than 5 Å apart from the interacting partner.

DBMK was originally introduced as DBMK 0.0 for the evaluation of ZDOCK (Chen & Weng 2002) docking method. At that point it consisted of 54 complex targets (22 enzyme-inhibitor complexes, 16 antibody-antigen complexes, 10 complexes with other function, and 6 difficult cases) where only 29 of them had both the receptor and ligand in their unbound form. Difficult cases are complexes which have large conformational changes between their bound and unbound forms. Later on, DBMK 1.0 (Chen, Mintseris, et al. 2003) was introduced which added 5 new targets to the previous benchmark resulting in a total of 59 targets (22 enzyme-inhibitor complexes, 19 antibody-antigen complexes, 11 other complexes, and 7 difficult cases) with 31 targets being unbound-unbound. By 2005, DBMK 2.0 (Mintseris et al. 2005) was introduced, which only involved unbound-unbound targets and redundancy was removed among chains using SCOP classification (Murzin et al. 1995). In DBMK 2.0, targets are classified into three categories - rigid body, medium difficulty and difficult - based on their degree of conformational change between the bound and unbound forms. This resulted in 72 unbound–unbound cases in DBMK2.0, with 52 rigid-body, 13 medium difficulty, and 7 high difficulty cases. In 2008, DBMK3.0 (Hwang et al. 2008) extend to

124 unbound-unbound targets from 309 protein chains with 88 Rigid-body, 19 medium, and 17 difficult cases. The most recent benchmark, DBMK4.0 (Hwang, Vreven, Janin, et al. 2010), is an extension of DBMK3.0 with 53 new targets (total 176 targets: 123 Rigid-body, 29 medium, and 24 difficult.). In total, DBMK 4.0 contains 52 enzyme-inhibitor, 25 antibody-antigens, and 99 other complexes. These datasets contain essentially dimers, but there are also a few trimers and tetramers. In this thesis we have focused on DBMK3.0 and DBMK4.0 to compare between interface predictors and docking methods.

3 Binding Site 3D Motif for Docking Model Evaluation

3.1 Introduction

Docking algorithms have been used to model the complex formed by protein chains and are capable of producing native-like models. However, their energy-based scoring functions are not reliable enough to detect native-like poses among a large pool of decoys (Gray 2006; Kastritis & Bonvin 2010). Therefore, scoring functions utilising experimental mutation studies have been proposed to assist this detection (Van Dijk, Boelens, et al. 2005). This is usually achieved by conducting mutation analysis on the top ranked models in terms of low energy (Sivasubramanian et al. 2006). However, native-like models may fail to be part of that short energy based list which leads to poor predictions. To address this, in this chapter we propose to create a more reliable list by introducing 3D motifs of structural binding sites for re-ranking predicted docking models (Esmaielbeiki et al. 2012). As a proof of concept we investigate the mode of interaction between an antimicrobial peptide and a lipoprotein receptor.

The rest of the chapter is organised as follows, Section 3.2 investigates the use of experimental mutation studies in detection of native-like docked poses. Section 3.3 provides an introduction to antimicrobial peptide and lipoprotein receptor. The proposed methodology is described in details in section 3.4 and evaluated and discussed in section 3.5 and 3.6. Finally, section 3.7 concludes the chapter.

3.2 Related Work

Protein-protein docking aims to computationally predict the 3D structure of protein complexes using the unbound structures of its components (Lensink et al. 2007; Wodak & Méndez 2004; Janin 2010). By the year 2000, several docking algorithms had been introduced and, with the constant increase of the number of protein structures determined by high-throughput X-ray and NMR experiments, docking had become a popular method for generating new complex models (Janin 2013). From then, a new scientific question emerged: should the quality and accuracy of docked models be trusted? Therefore, to regularly assess different docking methods, in 2001 a biannually blind prediction competition for comparing the performances of docking algorithms, the Critical Assessment of Predicted Interaction (CAPRI) (Janin et al. 2003), was introduced. In this competition, predictors – humans and servers – are required to dock protein structures taken from the PDB (Berman et al. 2000) and the produced models are compared against the unpublished experimentally determined structures of the relevant complexes (Janin et al. 2003). A similar blind prediction challenge in a smaller scale was conducted in 1996 (Strynadka et al. 1996), for predicting the complex of β-lactamase inhibitory protein (BLIP) and TEM-1 β-lactamase. Only 6 different computational docking algorithms compete against each other and their produced models were compared against the unpublished ground truth complex. The results showed that despite significant local and global conformational changes in the ground truth docking methods correctly predicted the general mode of binding of BLIP and TEM-1. By 2011, CAPRI (Janin 2010; Janin 2013) had 22 rounds targeting 43 complexes with an average of 45 predictors in which 70% of the targets received good quality models from different docking predictors (Janin 2013). CAPRI results show that the quality of docked models has steadily increased and docking algorithms have produced models which correspond to native complexes (Janin 2010; Wodak & Méndez 2004). For example, target T37 which was a prediction between the G-protein Arf6 and the LZ2 leucine zipper of JIP4 resulted in good quality models by five different groups (Janin 2010; Wodak & Méndez 2004). Figure 3.1 displays the superposition of the best predicted model on the native structure. The predicted complexes of other groups are

shown as the centre of mass of LZ2 in dots where yellow and cyan represent incorrect and acceptable-quality models, respectively.

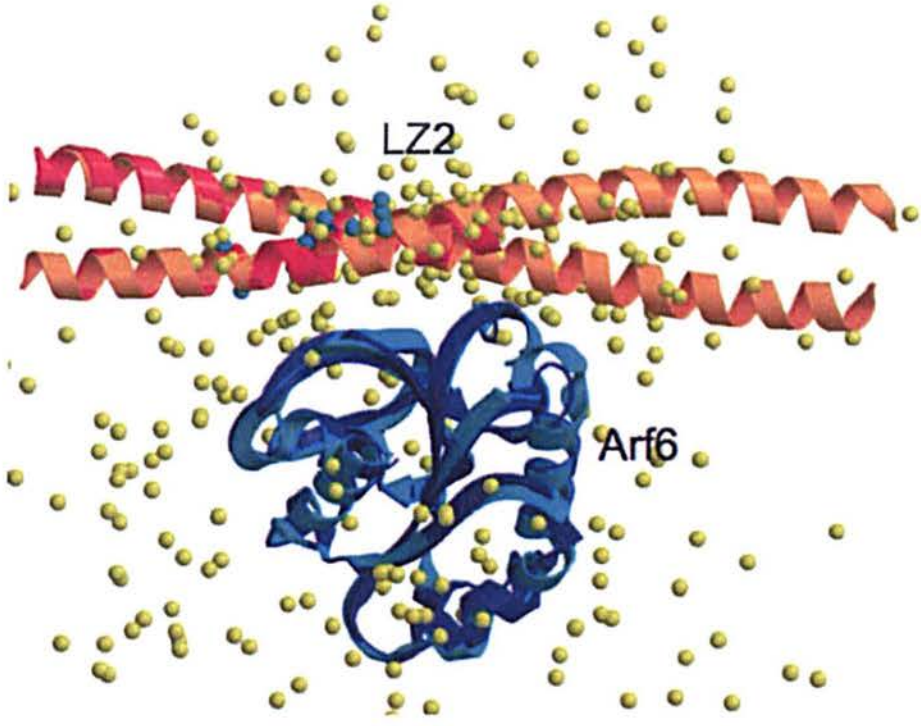


Figure 3.1: A successful docking prediction of target T37 in CAPRI competition, i.e. a complex between the G-protein Arf6 and the LZ2 leucine zipper of JIP4. The successful prediction is superposed on the x-ray native structure. Other models submitted by other groups are shown by a dot at the centre of mass of the LZ2. The dots are coloured cyan and yellow for acceptable-quality and incorrect models, respectively. Taken from (Janin 2010).

Although docking algorithms are able to produce correct complex conformations, they have to be identified among a large number – often hundreds - of putative models. Therefore scoring functions are required to select the most plausible solutions (see Chapter 2). Hence, alongside docking competitions, CAPRI also evaluates scoring methods. After each prediction round, predictor groups provide hundreds of their docked models which are merged together to create a repository of thousands of docked models. Then, scorer groups aim to identify the best top 10 models within that repository. Interesting results shows that scorers provide more accurate models than the predictors (Janin 2013). Since scorers select models from the repository created by predictors, this shows that docking methods are capable of creating good

quality models but may fail distinguishing them due to incorrect rankings. Therefore, scoring functions are required to create reliable rankings of docked models.

Combining experimental biochemical and biophysical data such as Mutagenesis, Mass spectrometry and NMR studies into docking methods not only result in better quality prediction, but also provide useful biological insight (Van Dijk, Boelens, et al. 2005). Experimental data can either be used to drive the docking procedure or to score docking conformations (see Chapter 2). However, as discussed in Chapter 2 section, post-filtering docking results is more robust.

Usage of mutagenesis studies as a post-filtering approach has been used in real applications. One early study aiming at understanding nucleocytoplasmic communications (Azuma et al. 1999) required the prediction of the complex of the small GTP-binding protein Ran and its regulator RCC1 (which is a chromatin associated guanine nucleotide exchange factor). Using GRAMM docking and mutation studies, the predicted Ran-RCC1 complex clarified the binding sites and the mechanism of nucleocytoplasmic transport (Scheffzek, Klaus and Wittinghofer 2001; Azuma et al. 1999). In another study, prediction of C1q in complex with C-reactive protein and IgG (Gaboriaud et al. 2003), also using docking and mutation studies, highlighted the two possible mode of interaction of C1q. A recent study exploits experimental information to create docking models of monoclonal antibody (mAb) 806 with epidermal growth factor receptor (EGFR) for real application (Gray 2006; Sivasubramanian et al. 2006). Docked models produced by RosettaDock (Gray et al. 2003) were filtered using residues important for mAb 806 binding as identified by mutagenesis studies. Eventually, among the three plausible models (Sivasubramanian et al. 2006), further computational mutagenesis identified a single candidate. This model emphasises the hypothesis that mAb 806 only binds to cancerous cells.

To conclude, these studies confirm that docking algorithms are capable to produce models which correspond to the real conformations. However, since docking algorithms provide an energy-based ranking of their large set of predicted models, which have proven not to be reliable (Gray 2006; Kastritis & Bonvin 2010) and fail to distinguish the near-native models, scoring techniques are required to rank and detect these models. Therefore, experimental mutation studies has been used to assist this detection (Van Dijk, Boelens, et al. 2005). To efficiently achieve this detection,

mutation analysis are performed on a short list of the top ranked models with lower energy (Sivasubramanian et al. 2006). Therefore, to create a reliable shortlist of docked models an extra step is required prior to mutation analysis.

In this chapter we introduce 3D motifs of structural binding sites for ranking predicted docking models and selection of near-native configurations. As a proof of concept we investigate the mode of interaction between an antimicrobial peptide and a lipoprotein receptor, i.e. a α-defensin and a Low Density Lipoprotein Receptor (LDLR), by the mean of predicted structural models (Nassar et al. 2002; Chang et al. 2005; Fuentealba et al. 2010; Nakashima et al. 1993; Higazi et al. 2000). First, we produce a novel 3D motif which describes the binding characteristics of LDLR-ligand interactions. Then, the motif is used as constraint to re-rank LDLR-α-defensin complex models generated by state of the art docking software. Finally, using mutagenesis studies and energy calculation, the most plausible models are selected.

3.3 Biological Background

Human antimicrobial peptides (AMPs) have come under intense scrutiny owing to their key multiple roles as antimicrobial agents against a range of bacteria, fungi and viruses. These roles have been reported to involve immunostimulation via chemotaxis, direct action on viral particles, and binding to, followed by internalisation, into mammalian cells where antimicrobial activity is manifested through inhibition of viral replication, via inhibition of protein kinase C signalling (Nassar et al. 2002; Chang et al. 2005; Fuentealba et al. 2010).

These molecules provide enormous scope for the investigation of mechanisms involved in infection, along with immune response events, and represent a reservoir of potential novel anti-infective agents. In this vein, the use of synthetic AMPs to treat HIV was reported as early as 1993 (Nakashima et al. 1993). Given the increase of drug resistant infections (Frieden et al. 1993; Cohen 1992) and the relative paucity of new clinically effective antimicrobial agents, further studies are warranted to optimise the activities of natural and synthetic AMPs.

One key step, which requires further study, is to optimise the binding of (synthetic) AMPs to mammalian cells to afford internalisation for intracellular defences to operate. Following the reported interaction of human α-defensins with a low density

lipoprotein receptor (LDLR) (Nassar et al. 2002; Chang et al. 2005; Fuentealba et al. 2010), a plausible approach is to study potential interactions between AMPs and the LDLR.

3.3.1Antimicrobial Peptides

Antimicrobial peptides (AMPs) (Diamond et al. 2009) are part of the innate immune system; they provide defences against microbial pathogens such as bacteria, fungi and viruses. They also act as immunomodulators and can activate specific immune cells by binding to host receptors (Wei et al. 2010). AMPs are short peptides with fewer than 60 amino acids; they are positively charged and are amphiphilic which means possessing the two attributes of being hydrophilic (water-loving) and hydrophobic (fat-loving).

AMPs are categorised into four distinct groups based on their structure, number of disulphide bonds and amino acid composition (Diamond et al. 2009; Schneider et al. 2005). One of the main groups is mammalian defensins which have a triple-stranded antiparallel β-sheet, stabilised by three intermolecular disulphide bonds. Based on their disulphide bond patterns, those peptides are classed into three sub-groups: α-defensins, β-defensins and θ-defensins.

Currently (up to June 2011), six different types of human α-defensins are known (Schneider et al. 2005; Van Wetering et al. 2005): four of them, i.e. HNP 1-4 (Human Neutrophil Peptides), are found in neutrophil[1] and the two others, i.e. HD-5 and HD-6, are present in intestinal Paneth cells[2] where HD-5 is also found in epithelial[3] cells of the female genital tract. Figure 3.2 displays the six human α-defensins sequences and their disulphide linkage.

[1] A type of white blood cell that kills and digests microorganisms.
[2] Paneth cells provide defense against microbes in the small intestine and are functionally similar to neutrophils. Paneth cells secrete antimicrobial molecules when they are exposed to bacteria.
[3] Epithelium is a layer of cells which covers and protects the surfaces of structures throughout the body.

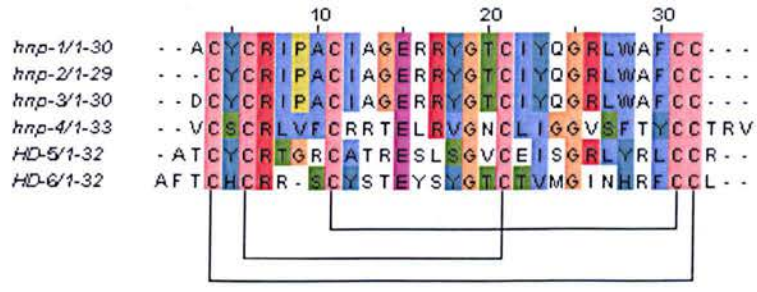


Figure 3.2: Sequence alignment of the six human α-defensins and their disulphide linkage. Ten conserved amino-acids can be seen which include six cysteines that are involved in the disulphide bonds. The conserved Cysteines are important for stabilising the structure.

3.3.1.1 Human α-Defensins Structure

The structural folds of these arginine-rich peptides are unique among the known AMPs. In 1991, Hill et al. (Hill et al. 1991) determined the first crystal structure of the human α-defensins, HNP3 (Figure 3.3), which is now considered as the archetype structure of this family (Diamond et al. 2009). Since residues essential for correct protein folding are conserved among α-defensins family (Figure 3.4), they all display a similar structure to HNP3 (Hill et al. 1991). The main features are: 1) six Cysteines which are important for stabilising the structure, and 2) the Gly18 residue which is part of the conserved β-bulge (Xie et al. 2005) and is essential for correct folding. These conserved residues are highlighted using arrows in Figure 3.4.

The HNP3 structure contains three- stranded antiparallel β-sheets which contain 60% of the residues and are held in place using three disulphide bonds. The most conserved feature among the human defensins structures is the β-bulge found in the middle of the second beta-sheet. This β-bulge initiates a β-hairpin between the second and third hydrogen-bonded anti-parallel β-sheets which is closed by the Cys10-Cys30 disulphide bond. Finally, the β-strand located at the N-terminus is hydrogen bonded to the β hairpin to produce a three-stranded β structure. In addition, the sequence of residues which are not involved in the β-sheet is stabilised by a Arg6-Glu14 salt bridge (Figure 3.4) which creates a rigid conformation.

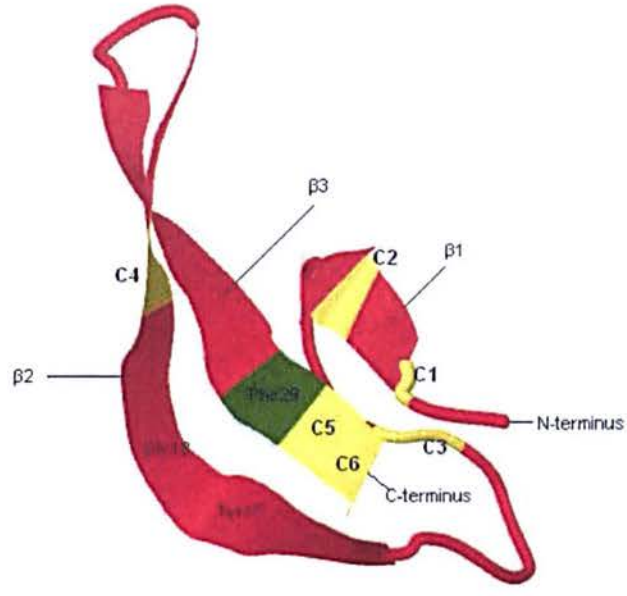


Figure 3.3: HNP3 monomer (PDB code: 1DFN). The three beta strand and the termini are shown on the figure. The Cys residues are coloured in yellow and C1, C2, C3, C4, C5 and C6 represent Cys3, Cys5, Cys10, Cys20, Cys30 and Cys31, respectively.

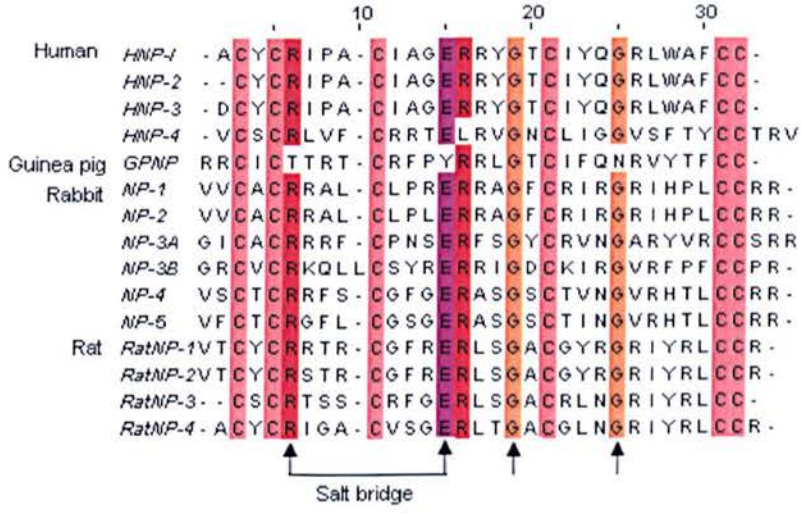


Figure 3.4: Sequences of defensins to show the conserved sections in comparisons to HNP3. Since the Cysteines are essential to stabilise the protein, they are conserved. Gly 18 which is important for the correct folding and is part of the conserved β-bulge is shown by arrow. Arg6-Glu14 salt bridge stabilises the sequence of residues which are not involved in the β-sheet. The colouring scheme is obtained by ClutalW. Each colour is associated with a {threshold, residue group} where 'threshold' means the minimum percentage of the presence of the 'residue group' in that location. In this image the colours and their associated {threshold, residue group} are: PINK {100%, C}, RED {+60%,KR},{+80%, K,R,Q}, MAGENTA {+60%,KR},{+50%,QE},{+85%,E,Q,D} and ORANGE {+0%, G}.

Crystal of HNP3 shows a symmetric dimer (Hill et al. 1991) which displays a basket shape structure (Figure 3.5). The core and top of the basket are respectively hydrophobic and hydrophilic, with six arginine residues forming an equatorial ring around the dimer. The twists and coils in the structure cause hydrogen bonds between the N-terminus of the two monomers and, therefore, form a mini-channel inside the dimer. The mini-channel and the hydrophobic patch allow the defensins to perform their antibacterial activity through interacting and entering the bacteria membrane (Hill et al. 1991) (See section 3.3.1.2).

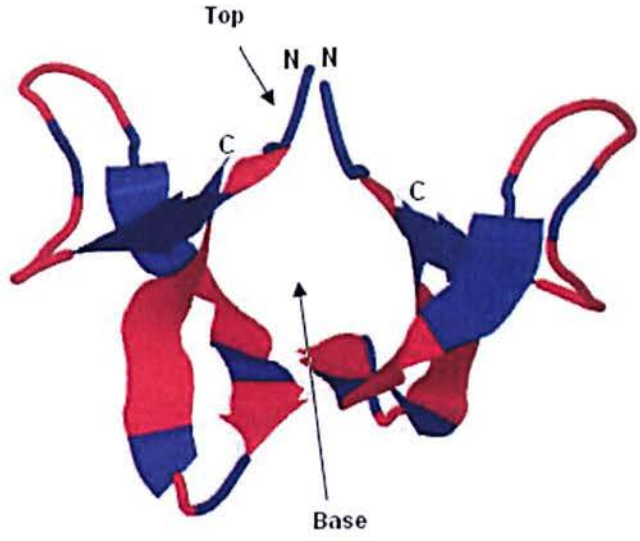


Figure 3.5: HNP3's basket shape with polar top and apolar base taken from PDB code: 1DFN. Polar and apolar residues are coloured in blue and red, respectively. The core and top of the basket are apolar and polar, respectively. The mini-channel inside the dimer allows the defensins to perform antibacterial activity.

3.3.1.2 Defensins Mechanism of Action

Defensins adopt various techniques for providing the immunity against bacteria and antivirus microorganism. The widely accepted defensins' bacteria killing technique is by disrupting the bacteria membrane and creating pore which results in the death of the cell (Kagan et al. 1990). This is because of the electrostatic attraction of cationic defensins and negatively charge membrane. Studies (Ericksen et al. 2005) have also shown that not only cationicity but also hydrophobicity of defensins governs their ability in bacterial killing. For example, HNP1 (with +3 charge) is more effective in

response to S.aureus (Gram-positive) than HBD3 (with + 11 charge). Further studies on α-defensins antibacterial activity (Leeuw et al. 2010) shows that they not only kill bacteria by targeting their membrane but also by interacting with their intracellular components. For example, HNP1 interacts with Lipid II which is a component of the bacteria cell wall.

Defensins respond to viral infection in two distinct ways (Klotman & Chang 2006; Chang et al. 2005): first, similar to their antibacterial activity technique by directly attacking the viron, or second, indirectly by interacting with the target cells. Figure 3.6, taken from (Klotman & Chang 2006), displays HNP1 dual role in blocking HIV-1. Whereas, in the absence of serum the interaction between HNP1 and the viral glycoprotein inhibits HIV-1 replication, in the presence of serum, HNP1 blocks HIV-1 at nuclear import and transcription stage by interfering with its affected cell protein kinase C (PKC) signalling pathway. This can be achieved by HNP1 interaction with G-protein coupled receptors (GPCR) or other cell surface receptors such as low-density-lipoprotein receptor (LDLR). HNP2 and HNP3 have shown similar activity against HIV-1 (Klotman & Chang 2006).

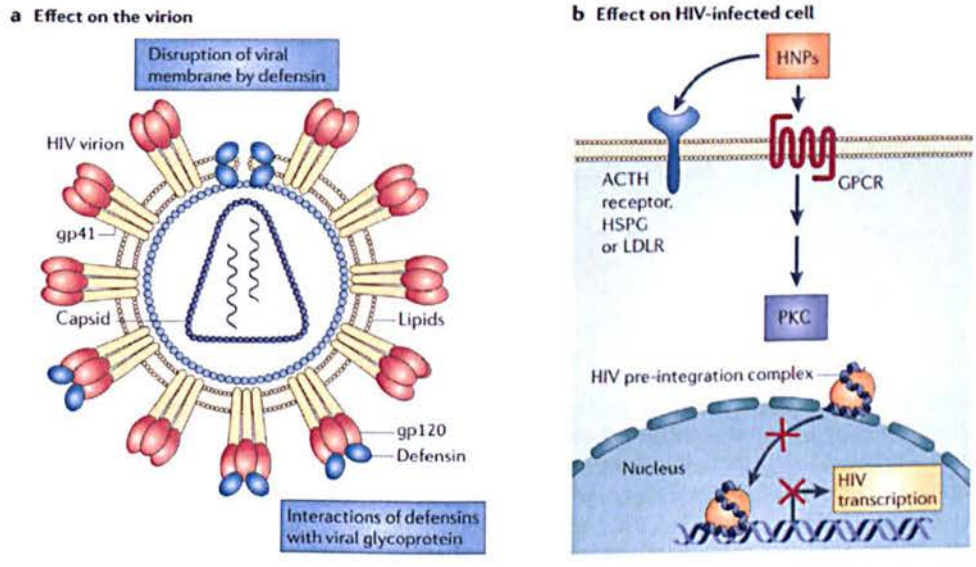


Figure 3.6: Dual antiviral mechanism of defensins. a) In the absence of serum HNP1 directly interacts with the viral glycoprotein which inhibits HIV-1 replication. b) In the presence of serum, HNP1 blocks HIV-1 at nuclear import and transcription stage by interfering with its affected cell protein kinase C (PKC) signalling pathway. Taken from (Klotman & Chang 2006).

3.3.1 Low Density Lipoprotein Receptor

The LDLR family contains seven homologous members and is responsible for mediating different types of ligands especially cholesterol into the cell (Blacklow 2007). Their structure is composed of several domains which include a ligand binding (LB) domain (Guttman, Prieto, Croy, et al. 2010) composed of ligand binding modules (LAs), also named complement-type repeats (CRs), a beta-propeller domain and transmembrane and cytoplasmic sections (Figure 3.7, top row).

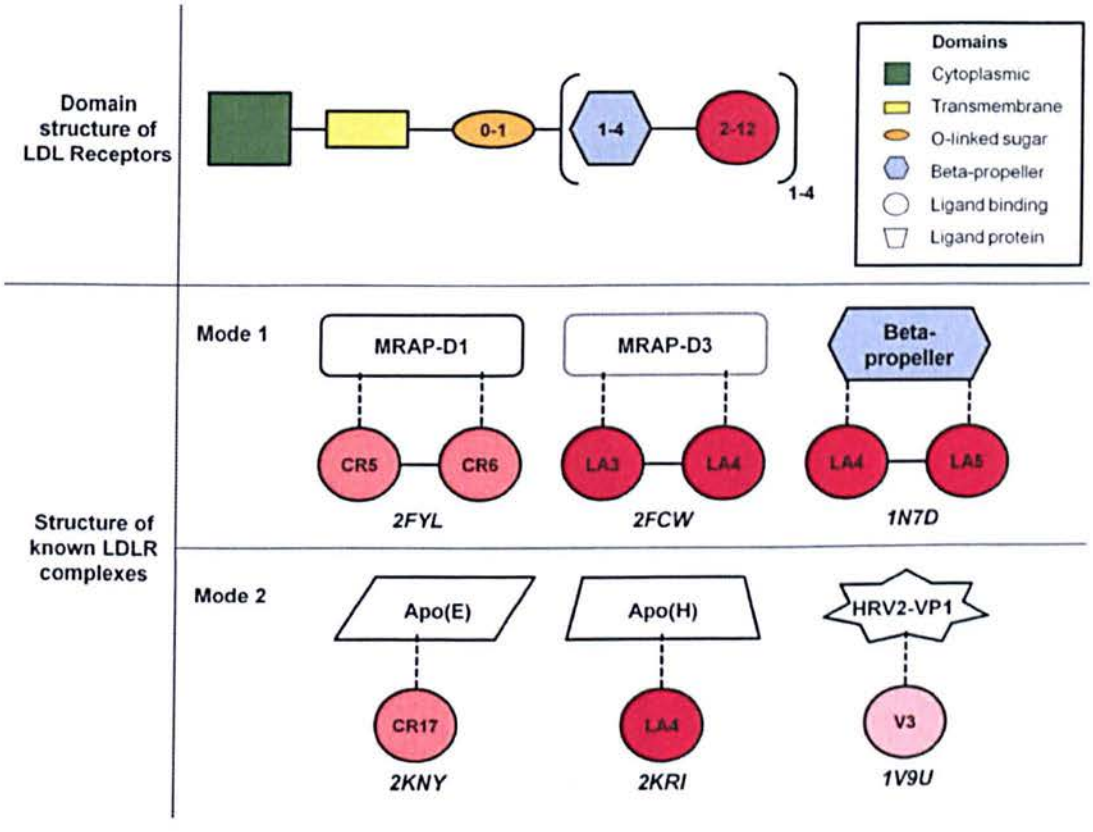


Figure 3.7: Modular structure of LDLR receptor family: general domain pattern (top) and schematic representation of the LDLR-ligand binding modes of known complexes (bottom). (top) The structure of LDLR family is composed of several domains including a ligand binding (LB) domain, composed of ligand binding modules, beta-propeller, transmembrane and cytoplasmic domains. (bottom) The two modes of interaction among LDLR-ligand complexes are shown. In mode 1, two ligand binding modules of LDLR are necessary to interact with the ligand, while in mode 2, only one ligand binding module is used.

The Low-density lipoprotein receptor family interacts with a wide variety of human and virion proteins (Fisher et al. 2006) through their homologous LA modules which are between 40-50 residues long (Beglov et al. 2009; Herz & Bock 2002) (Figure

3.8). LA's structure is stabilised by three disulphide bonds and a calcium ion (Guttman, Prieto, Handel, et al. 2010; Rudenko & Deisenhofer 2003). This ion is an essential element of the ligand binding domain conformation since it is required to establish interaction between LDLR and the ligand (Fisher et al. 2006; Dirlam-Schatz & Attie 1998).

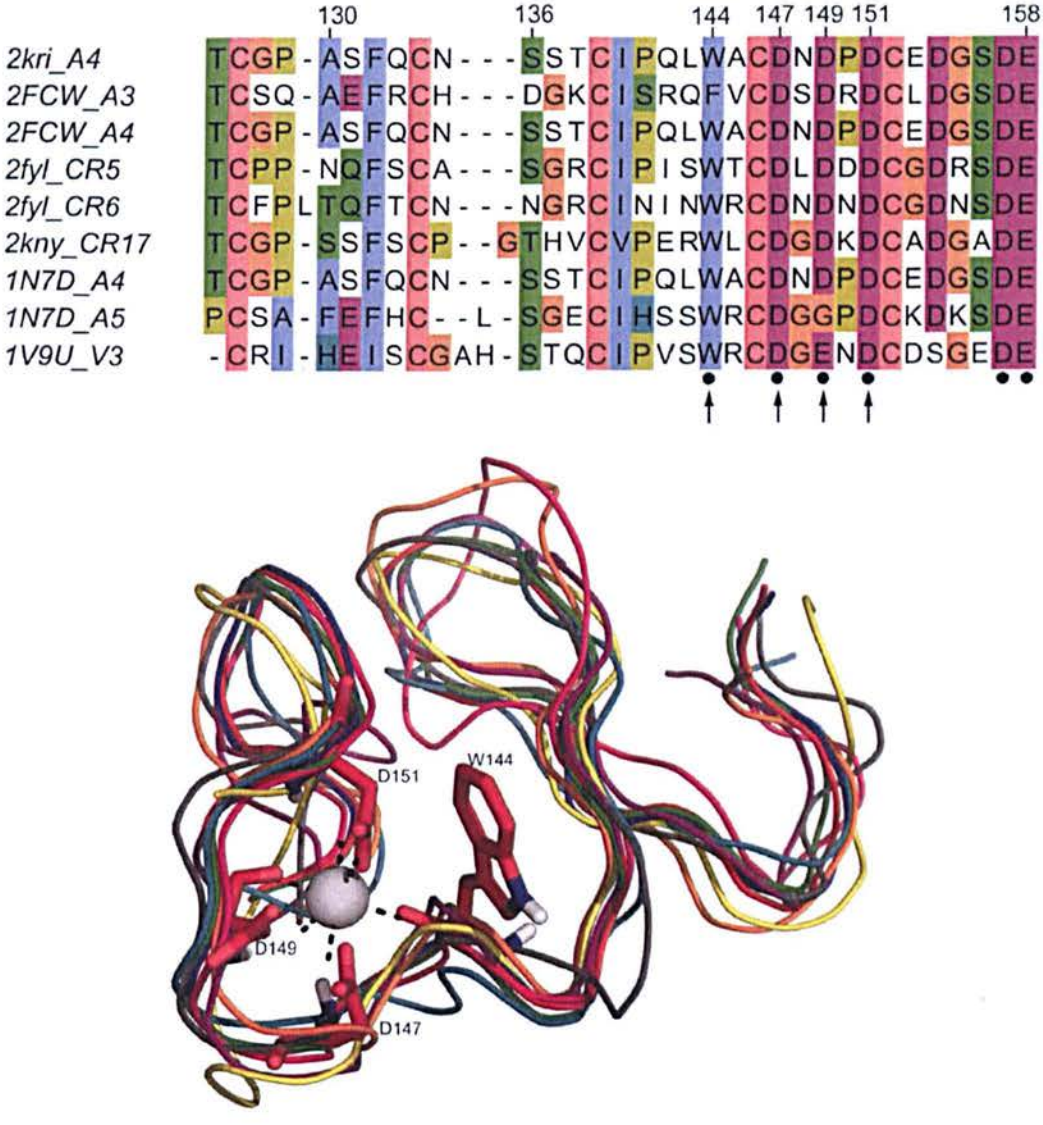


Figure 3.8: Multiple sequence and structure alignment of ligand binding domains of LDLR family complexes. In the sequence alignment, residues involved in calcium interaction are denoted by dots. The three conserved acidic residues and conserved tryptophan/ phenylalanine are highlighted with black arrows. Sequence numbering is based on 2KRI:B. In the structure alignment the three conserved acidic residues and conserved tryptophan are shown on 2KRI structure. Residues numbering is based on 2KRI:B. The ligand binding domains associated with each colour are: 2KRI-A4: red, 2FCW-A3:green, 2FCW-A4:dark blue, 2FYL-CR5:purple, 2FYL-CR6:yellow, 2KNY-CR17:orange, 1N7D-A4:deep teal, 1N7D-A5: grey, 1V9U-V3: pink.

High-resolution crystal structures of the available LDLR complexes have revealed that electrostatic forces play an important role in interactions (Fisher et al. 2006). This key function is captured by the minimal interaction motif described by Jensen et al. (Jensen et al. 2006) (Figure 3.9), which also highlights a hydrophobic element in the interaction. The receptor's conserved acidic residues (ASP/GLU) interact with a ligand's lysine through a salt bridge creating a hydrophobic environment for the side chain of a receptor's tryptophan (TRP). In addition, a hydrophobic side chain, ψ, (usually Leucine or Isoleucine) from the ligand sits next to TRP.

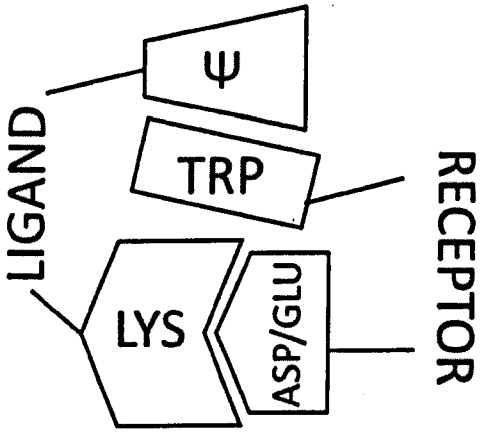


Figure 3.9: Minimal binding motif defined by Jensen et al. (Jensen et al. 2006). The motif highlights the importance of electrostatic and hydrophobic forces in LDLR complexes. The LDLR's (RECEPTOR) conserved acidic residues (ASP/GLU) interact with a LIGAND's lysine (LYS). This salt bridge creates a hydrophobic environment for the side chain of the receptor's tryptophan (TRP). Moreover, Ψ, a hydrophobic side chain from the LIGAND sits next to TRP.

3.4 Proposed Methodology

3.3.2Overview

Although docking methods usually produce correct configurations, their ranking is not reliable enough. Consequently, re-ranking models is required to enable the selection of a near-native model. Therefore, to achieve this we propose a method in which using binding site 3D motifs, docking models are scored and ranked. This method consists of two main modules: 1) creation of 3D motif (Figure 3.10), 2) Ranking docking conformations using this motif (Figure 3.11).

A 3D motif is a structural descriptor of a protein binding site, which describes the 3D pattern a specific protein (here, called receptor) uses to bind to its ligands (here, ligand refers to any interacting partner of a protein). 3D motifs are created by extracting the homologous complexes of a specific receptor target from PDB (Berman et al. 2000) (See Figure 3.10). Then the binding sites of the homologous complexes are analysed to identify the interface residues which are structurally conserved among the homologues. Afterward, a selected set of atoms are aligned and the 3D motif is created. This motif is evaluated using a leave-one-out strategy (see section 3.3.8) which means leaving one homologous protein out of the training set and generating the 3D motif based on the other homologous proteins. Docked models are created for the left-out homologue and its ligand which are already ranked by the docking tool scoring function. The 3D motif is then used to re-rank the docked poses. Another ranking list is generated based on comparing the RMSD of docked models with the ground truth. The 3D motif is judged unsatisfactory, i.e. fails, if its ranking does not improve the docking tool ranking in comparison to the ground truth. In that case, another subset of the current atoms are selected to create a new 3D motif.

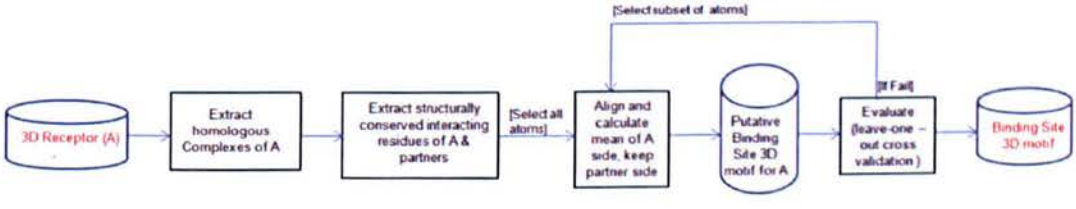


Figure 3.10: The pipeline for creating 3D motif. First, homologous complexes of the 3D Receptor (A) are extracted from PDB. Second, structurally conserved binding sites of the homologues and their partners are identified. Third, the positions of residues on homologues are averaged while their interacting partner residues are kept. This process results in generating a putative binding site 3D motif which is then evaluated with a leave-one-out cross validation.

Using the 3D motif created in the previous stage, docked receptor-ligand conformations are evaluated and re-ranked (Figure 3.11). Finally, the best models are selected using mutagenesis studies and energy calculation.

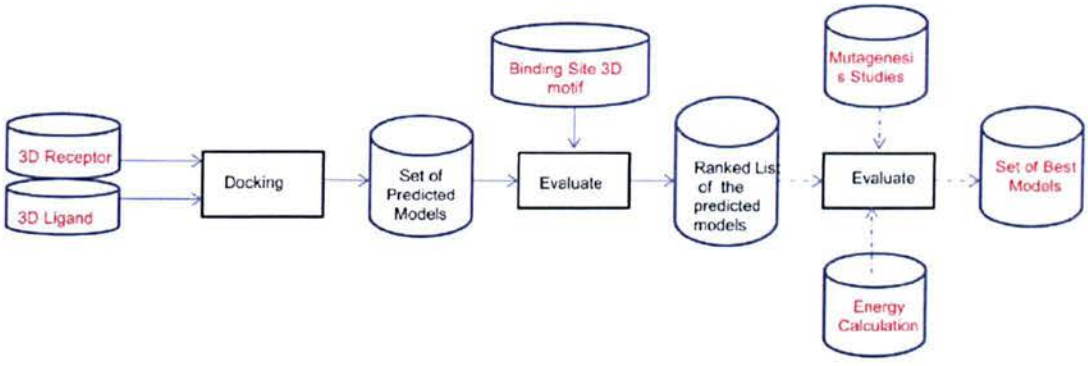


Figure 3.11: The pipeline for re-ranking docking models. Docked models Receptor and Ligand are generated which results in a set of docked predicted models. Then, the binding site 3D motif of the Receptor is used to evaluate and create a ranked list of docked models. This list is further assessed using mutagenesis studies and energy calculation.

In this study, we focus on predicting the configuration of the LDLR-HNP1 complex. First, a 3D motif capturing atomic interactions of LDLR binding interface is created. A 3D motif could not be created for HNP1, since no homologous complex it is known. Second, state-of-the-art docking software produces a set of LDLR-HNP1 complex models which are re-ranked based on the LDLR 3D motif. Models incompatible with mutagenesis studies are rejected and finally, binding energy estimations are used to identify the most stable of the remaining models.

3.3.3Dataset

Our study relies on the investigation of all 3D complexes involving a ligand binding domain of the LDLR family. Query of the RCSB Protein Data Bank (Berman et al. 2000) using BLAST (Altschul et al. 1997) on March 2011 revealed that the structures of six complexes have been resolved (Table 3.1). They all involve human proteins belonging to three members of the LDLR family, i.e. Low-density Lipoprotein receptor (LDLR), lipoprotein receptor-related protein 1 (LRP1) and Very low-density lipoprotein receptor (VLDLR). The sequences of their ligand binding domain were extracted from Uniprot (Suzek et al. 2007), where their accession numbers are P01130: LDLR, P98155: VLDLR and Q07954: LRP1, respectively. Although their ligand binding modules are named LA, CR and V for LDLR, LRP and VLDR, respectively, in this paper we use LA when referring to any of them.

Table 3.1: Known 3D structures of complexes involving members of the LDLR family. The PDB Code of the structures is provided along with the ligand binding domain name. The ligand complete name and its domain name are also given in the table.

PDB Code	Receptor and Domains	Ligand	Ligand Complete Name
2FCW (Fisher et al. 2006)	LDLR LA3,LA4	MRAP D3	Alpha-2-macroglobulin receptor-associated protein Domain 3
2FYL (Jensen et al. 2006)	LRP CR5,CR6	MRAP D1	Alpha-2-macroglobulin receptor-associated protein Domain 1
2KRI (Beglov et al. 2009)	LDLR LA4	Apo(H)	Beta-2-glycoprotein 1
2KNY (Guttman, Prieto, Handel, et al. 2010)	LRP CR17	Apo(E)	Apolipoprotein E
1N7D (Rudenko et al. 2002)	LDLR LA4,LA5	LDLR beta propeller	-
1V9U (Verdaguer et al. 2004)	VLDLR V3	HRV2 VP1	Human rhinovirus 2 Viral Protein 1

In agreement with the existing 2D motif (Jensen et al. 2006), sequence alignment of the LA modules (Figure 3.8) using ClustalW (Chenna et al. 2003) shows highly conserved acidic residues and a tryptophan/phenylalanine (TRP/PHE) - pairwise E-values were calculated using BLAST (Altschul et al. 1997) (Table 3.2). Structural conservation of the ligand binding domains of the receptors, i.e. LA3, LA4, LA5, CR5, CR6, CR17 and V3, are illustrated (Figure 3.8) and quantified (Table 3.2) using the 3D alignment tool Pymol (Schrödinger, LLC 2010).

Since LA4 is the domain which is the most common in this set - in three cases out of six – it is used as representative for the purpose of α-defensin docking. Similarly, among these AMPs, defensin Human Neutrophil Peptide-1 (HNP1), which has been specifically shown to interact with LDLR (Nassar et al. 2002; Higazi et al. 2000), is selected as representative. Sequences and structures of HNP1 and LA4 were extracted from the PDB (Berman et al. 2000): 3GNY (Wei et al. 2009) and 2KRI (Beglov et al. 2009) codes respectively.

Table 3.2: E-value between sequences of the Ligand binding domains and RMSD between their 3D structures. Sequence similarities are calculated using BLAST and structural RMSD is estimated using Pymol (Schrödinger, LLC 2010).

Sequence Similarity									
	2FCWA 3	2FCWA 4	2FYLCR 5	2FYLCR 6	2KRIA 4	2KNYCR1 7	1N7DA 4	1N7DA 5	1V9UVD 3
2FCWA3	1e-19	2e-09	3e-08	2e-08	2e-09	4e-08	1e-07	4e-07	2e-04
2FCWA4		1e-23	2e-08	6e-08	2e-21	1e-10	1e-18	9e-08	2e-06
2FYLCR5			3e-23	2e-11	2e-08	3e-09	3e-07	4e-08	1e-06
2FYLCR6				1e-22	6e-08	2e-06	1e-07	1e-08	2e-06
2KRIA4					2e-21	1e-10	1e-18	8e-08	4e-06
2KNYCR1 7						9e-30	6e-08	5e-05	4e-06
1N7DA4							2e-23	2e-07	3e-06
1N7DA5								1e-23	2e-07
1V9UVD3									8e-22
RMSD Between Structures (Å)									
	2FCWA 3	2FCWA 4	2FYLCR 5	2FYLCR 6	2KRIA 4	2KNYCR1 7	1N7DA 4	1N7DA 5	1V9UVD 3
2FCWA3	0	0.29	0.84	1.16	0.43	0.77	0.96	0.68	0.38
2FCWA4		0	0.6	3.18	0.43	0.77	1.07	0.67	0.64
2FYLCR5			0	3.19	0.73	1.04	1.24	0.85	1.00
2FYLCR6				0	3.89	3.01	4.62	4.27	3.47
2KRIA4					0	0.93	1.67	0.92	0.53
2KNYCR1 7						0	1.43	1.06	1.28
1N7DA4							0	1.41	1.20
1N7DA5								0	0.84
1V9UVD3									0

3.3.4Modes of Interactions of LDLR-Ligand Complexes

Within the known LDLR-ligand complexes, two modes of interaction between the LB module and the ligand have been identified (Figure 3.7, bottom rows). In the first mode, two ligand binding modules of LDLR are required to establish an interaction with the ligand. In 2FCW (Fisher et al. 2006) the third and fourth modules of the ligand binding domain (LA3,4) bind to MRAP domain 3 (MRAPD3). In 2FYL (Jensen et al. 2006), two modules of complement-type ligand binding repeats (CR5,6) interact with two different sections of MRAP domain 1 (MRAPD1). Similarly, LA4,5 of 1N7D (Rudenko et al. 2002) bind to two different sites of LDLR beta propeller.

In the second mode, only one ligand binding module of LDLR binds to the ligand. Apo(H) and Apo(E) bind to A4 in 2KRI and CR17 in 2KNY, respectively. In

1V9U (Verdaguer et al. 2004), the third LB module of VLDLR (V3) interacts with Human rhinovirus 2 (HRV2) viral proteins VP1.

As a whole, the available six structures describe 9 different binding sites, since three complexes operate in the first mode of interaction.

3.3.5Creation of 3D Motif

Nebel et al. (Nebel et al. 2007; Nebel 2006) proposed a method in which 3D motif of ligand binding sites were created using their consensus atom positions. Note that in Nebel's study, 'ligand' referred to a small molecule such as Adenosine triphosphate (ATP). First, ligand binding sites of proteins are aligned and compared against each other. A similarity matrix of these comparisons is used to cluster the binding sites. In each cluster a consensus 3D pattern is created by pair-wise superposition of the binding sites and only keeping atoms which are paired with the same chemical properties. Extending the existing LDLR 2D motif (Jensen et al. 2006) using that approach, we produce a 3D motif which describes the conserved 3D positions of the key atoms involved in LDLR-ligand interaction (Figure 3.12). Whereas in Nebel's approach, 3D motifs are created by superposing small ligands which share similar structures and focusing on clusters with similar chemical properties on the receptor side, here we focus on superposing homologous structures and detecting local conservation and then selecting representative residues on the ligand side. Therefore in LDLR-ligand 3D motif, from the receptor, the conserved acidic residues and TRP/PHE are represented by their alpha carbon atoms. In addition, in order to add a constraint regarding interaction with the calcium ion, we include the oxygen atom of the TRP/PHE carboxyl group with whom it interacts. On the ligand side, the basic residue interaction is expressed by the side-chain nitrogen atom(s) which form(s) hydrogen bond(s) with acidic residues of the receptor.

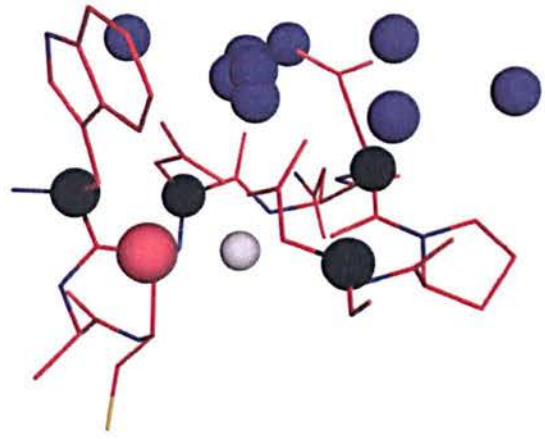


Figure 3.12: The 3D motif is represented by spheres. The blue ones show positions of N atoms from the ligand. The black ones are the C-alpha atoms of the ASP and TRP and the red sphere is the O atom of TRP. Location of the calcium is marked by a grey sphere. Image produced using Pymol.

The actual coordinates of the consensus atoms forming the 3D motif are calculated by the multiple structure alignments of these atoms using all available receptor-protein complexes. Here, only atoms from the receptor side are used as superimposition constraints. Since their 3D structures are very well conserved - their average RMSD is 0.28 Å - positions of all receptor atoms in the 3D motif are approximated by the average coordinates of the aligned atoms. On the other hand, given that every ligand displays a very different receptor binding sites, there is no consensus 3D position regarding the location of the nitrogen atom(s) of the basic residue(s). However, since there must be specific constraints in terms of their distance and orientation from the receptor, in the 3D motif, we express implicitly these constraints by storing all the actual nitrogen positions available in our training set.

Note that among the 9 binding sites of the available structures, we excluded that of 2FYLCR6 in the construction of our motif since its LA module is structurally different from the others as measured by an average RMSD of 2.97 Å (Table 3.2).

3.3.6Docking

Docking predictions are performed using the ClusPro 2.0 docking program (Comeau et al. 2004), which, in addition to be freely available for academic research, has demonstrated best performance at CAPRI 2009 (Critical Assessment of Predicted Interactions) (Shen et al. 2007; Comeau et al. 2007; Kozakov et al. 2010). ClusPro

works by initially calculating 70,000 docking models. Then, the 1000 models with the best energy conformation are selected and clustered using PIPER (Kozakov et al. 2006). Models with the most neighbours within a 9 Å C-alpha RMSD cut-off are chosen as cluster representatives and are qualified by the size of their associated cluster.

The ClusPro docking results are generated according to different constraints. For each category, software produces a set of predicted models ranked according to their cluster size. Since previous studies have highlighted the important role of electrostatic and hydrophobic interactions in LDLR complexes (Fisher et al. 2006; Beglov et al. 2009; Jensen et al. 2006), we only consider predictions generated under 'electrostatic favoured' and 'van der Waals + Electrostatic forces (VDW/elec)' modes. In this work default software parameters are used.

3.3.7Ranking of Putative Complexes

Docking models are scored and ranked using the fitting of the 3D motif to each predicted complex. 3D motif fitting is performed by rigid procrustes superimposition (Gower 1975; Goodall 1991) on the binding site of the predicted LDLR-ligand complex using receptor atoms as constraints. We define the quality of a prediction as the shortest distance between the nitrogen of the basic residue of the ligand and those present in the 3D motif.

3.3.83D Motif Evaluation Methodology

Our 3D motif was evaluated in docking prediction task using a leave-one-complex-out cross validation. First, a resolved 3D complex involving LDLR is selected. Secondly, a 3D motif is produced using all the other available LDLR complexes. Thirdly, the two chains involved in the complex are submitted to ClusPro which generates a set of putative complex models. Then, the fitting of the 3D motif to each model is used to score predicted complexes. Finally, the produced ranked list is compared with the list of models ranked according to their quality as expressed by their RMSD with respect to the actual resolved structure.

3.3.9LDLR-HNP1 Model Selection

Using the procedure previously described, LDLR-HNP1 complex estimates are generated by ClusPro and ranked using our 3D motif fitting measure. Then, the best

models according to that score are further analysed in order to establish which ones are in agreement with the literature.

Finally, the stability of the remaining modelled complexes is quantified by both calculating the number of intermolecular contacts and estimating pair wise interaction energies between the different chains involved in those complexes. Detailed information on residue-residue and atom-atom contacts is provided by the Contact Map Analysis server which is part of the software suite SPACE (Sobolev et al. 2005). In addition, since previous studies (Sánchez et al. 2008; Kiel et al. 2005) have shown good correlation between experimental measurements and energy calculations produced by the FoldX software (Guerois et al. 2002; Schymkowitz et al. 2005), its latest version, v3.0 beta5.1 (http://foldx.embl.de/) has been selected to evaluate binding energy between the two HNP1 monomers and between LDLR and each of the HNP1 chains.

3.5 Evaluation

3.3.10 3D Motif Validation

Our 3D motif, displayed in Figure 3.13, is evaluated against predictions of 9 binding sites. Results are reported in Figure 3.13, where the number of ranks required to achieve 100% recall, r100%recall, is expressed as a function of the number of top quality predictions, t. A perfect prediction evaluation scheme would place the t best predictions on the t top-most positions of the ranked list, whereas the worst evaluation scheme would require the whole list to recall the t best predictions.

Although Cluspro developers do not recommend judging the produced models according to their associated cluster size, software output shows models ranked according to that score which obviously influence user's usage of these models. Therefore, we also show on Figure 3.13 how cluster size would perform if used to rank models.

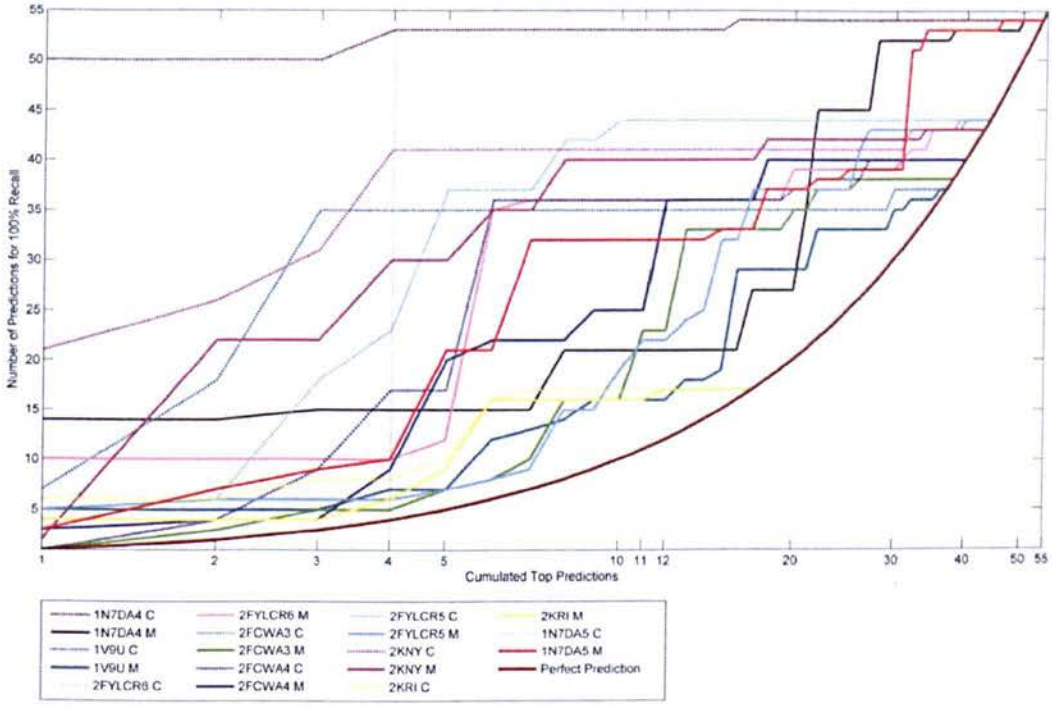


Figure 3.13: Number of ranks to achieve 100% recall of the top predictions (or recall of top predictions with top positions). In the legend the complexes names are followed by C and M for curves based on cluster size and 3D motif method, respectively. In all cases, the curves of ranking produced by the 3D motif are closer to the perfect prediction in comparison to Cluspro ranking. Therefore, 3D motif raking improves the ranking produced by Cluspro and as a result fewer models are needed to recall the top quality predictions.

In every case, ranking based on 3D motif fitting produces curves closer to the perfect prediction than those generated from cluster size ranking. As a consequence fewer models are needed to recall the top quality predictions when the 3D motif is used to access LDLR interaction predictions. If the LDLR-Apo(E) (2KNY) complex is excluded, our 3D motif allows the discovery of the 4 best quality models within a shortlist of 15. Usage of the cluster score would require listing 53 models to achieve the same outcome. The different behaviour displayed by 2KNY could be explained by the fact that this model is not a true complex since the fragment of Apo(E) has been fused with a linker to CR17 to ensure interactions between both domains (Guttman, Prieto, Handel, et al. 2010).

This experiment validates the usage of the LDLR 3D motif as a good indicator of model quality.

3.3.11 Literature Study of LDLR-HNP1 Complex

Since HNP1 has a hydrophobic and cationic face (Figure 3.14) that resembles the binding patch of ligands which interact with LDLR (Higazi et al. 2000; Quinn et al. 2008; Soman et al. 2010), its mode of interaction may be similar to those previously studied. In addition, this area belongs to a pocket detected by both Fpocket (Le Guilloux et al. 2009) and CastP software (Dundas et al. 2006) (Figure 3.15).

Regarding the hydrophobic aspect, Ala-scanning mutational study of HNP1 revealed tryptophan26 (W26) is a key residue in direct interaction with target proteins and enables the peptide to form dimmers (Wei et al. 2010). In addition, either W26 or phenylalanine28 (F28) mutation decreases HNP1 antibacterial activity. The importance of W26 is further highlighted by the fact it is either conserved or replaced by an amino acid displaying an aromatic side chain in other human α-defensins.

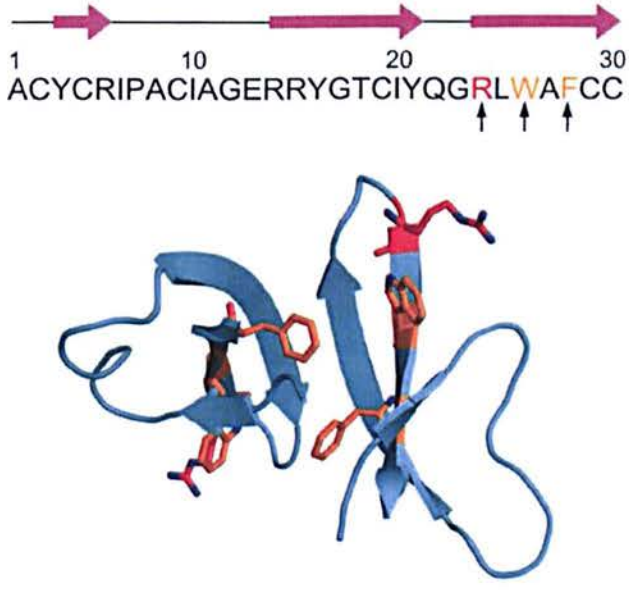


Figure 3.14: HNP1 sequence and 3D structure of the HNP1 dimer. The secondary structure of HNP1 is shown above the sequence. W26 and F28 are highlighted using arrows in the sequence and orange sticks in the 3D structure. R24 is also marked in red. Image produced using Pymol (Schrödinger, LLC 2010).

As for the cationic face, HNP1 sequence comprises four basic residues, i.e. arginines, which could play a role similar to the lysines present in the studied LDLR-ligand complexes. Among these basic residues, arginine24 (R24) has been reported as an important residue for interacting with bacterial lipids (Y. Zhang et al. 2010).

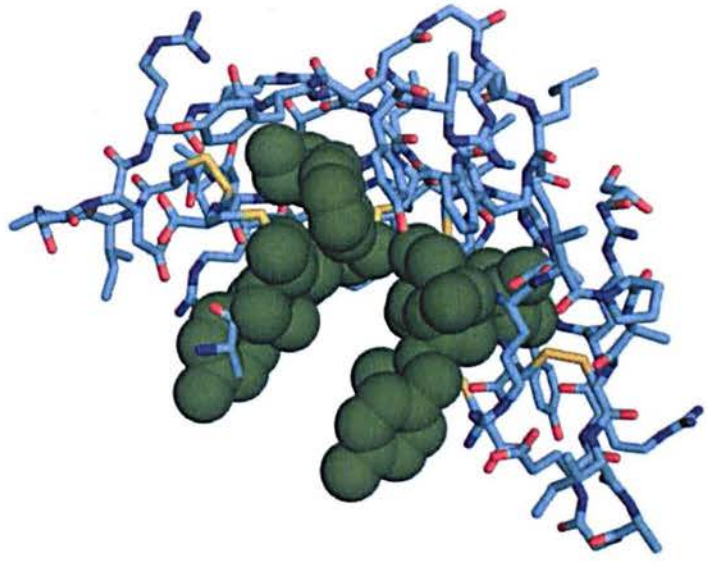


Figure 3.15: The pocket detected for HNP1 dimer using CastP software (Dundas et al. 2006). The green spheres represent the atoms located in the pocket of HNP1. The rest of the protein is shown using cyan sticks. Images produced using Pymol (Schrödinger, LLC 2010).

Although beta sheets are dominant in HNP1 and LDLR structures, the study of known LDLR-ligand complexes does not support the involvement of beta sheets in their interactions. Actually, this study suggests formation of a salt bridge between HNP1's R24 and LDLR acidic residues and that either W26 or F28 plays the role of ψ in the minimal motif (Figure 3.9).

3.3.12 Docking Prediction of LDLR-HNP1 Complex

Cluspro produced a total of 43 predicted models using both the electrostatic and VWD/elec categories. Those models were ranked using our 3D motif and, as suggested by our previous experiment, only the top 15 are considered for further analysis (Table 3.3). Since R24 and either W26 or F28 are expected to be involved in the interaction, only Model.002.01, Model.006.02 and Model.006.18 are in agreement with literature findings.

Pairwise structural alignment reveals high similarity between Model.006.18 and Model.002.01 (1.61 Å RMSD). This shows that Cluspro converged towards a specific docking configuration from two different sets of constraints. Model.002.01 is chosen as

representative of this configuration. In addition, as required by the minimal motif (Figure 3.9), Model.002.01 and Model.006.02 have candidates for the role of ψ since the TRP 144 of LA4 interacts with both W26:B and F28:A of HNP1 (Figure 3.16). Both models position their R24 N atoms at similar locations (RMSD < 0.2Å). However, there is approximately a 90-degree angle between the positions of the ligands which leads to a 13.13 Å RMSD between those two putative complex configurations.

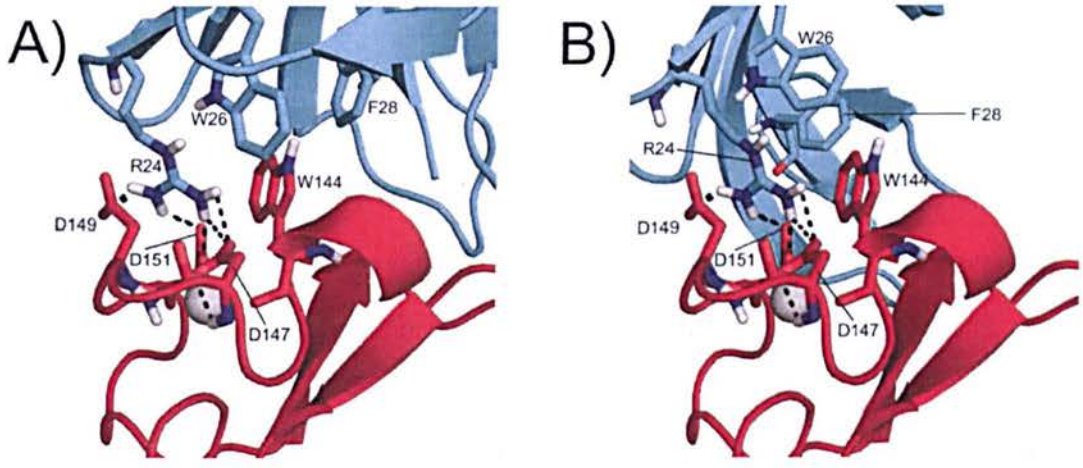


Figure 3.16: Proposed LDLR-HNP1 interaction models for (A) model.002.01 and (B) model.006.02. Structures of HNP1 and A4 are shown in cyan and red, respectively. Calcium ion is represented as a grey sphere. R24 creates salt bridge with the aspartic residues which are shown as black dashed lines. W144 of A4 and F28 and W26 of HNP1 provide the hydrophobic interactions. Images produced using Pymol.

Table 3.3: Residues involved in interaction between LDLR and HNP1 according to docking results. Model IDs starting by 002 and 006 are produced according to electrostatic and VDW+elec constraints, respectively. Contacts between residues are identified by SPACE (Sobolev et al. 2005).

Model ID			
	ASP Residue(s) (LDLR)	ARG residue: Chain (HNP1)	Hydrophobic residue: Chain (HNP1)
mutagenesis studies	**D147,149,151**	**R24**	**W26/F28**
model.002.01	D147,149,151	R24:B	W26:B, F28:A
model.006.19	D149,151	R15:A	W26:A,I6:A,L25:A
model.006.02	D147,149,151	R24:B	W26:B,F28:A
model.006.05	D147,151	R14:B	I10:B
model.006.28	D147,151	R14:B	-
model.006.03	D147,149,151	R24:B	I6:B
model.006.12	D147,151	R15:A	-
model.006.23	D149	R15:B	I20:A
model.002.15	D147,149	R14:A	W26 :B,F28:A
model.006.17	D149,151	R14:B	-
model.006.06	D149,151	R24:B	I6:B
model.006.18	D149,151	R24:B	W26:B,F28:A
model.006.00	D147,149	R14:B	-
model.006.13	D149,151	R14:A	-
model.006.22	D147,149,151	R15:A	-

Complex stability analysis based on FoldX binding energy calculations (Figure 3.17) reveals that Model.002.01 is a much more stable LDLR-HNP1 complex than Model.006.02. Although Cluspro energy values (-712.5 and -143.3 Kcal/mol for Model.002.01 and Model.006.02 respectively) are not particularly accurate (Ponomarev & Audie 2011) , they are in agreement with FoldX conclusions. In addition, the fact that Cluspro simulations based on two different sets of constraints led to the configuration exemplified by Model.002.01 supports the presumption of its higher stability. It is interesting to notice that, for this model, the strength of the LDLR-HNP1 bonds weakens the bond between the two HNP1 monomers (Figure 3.17).

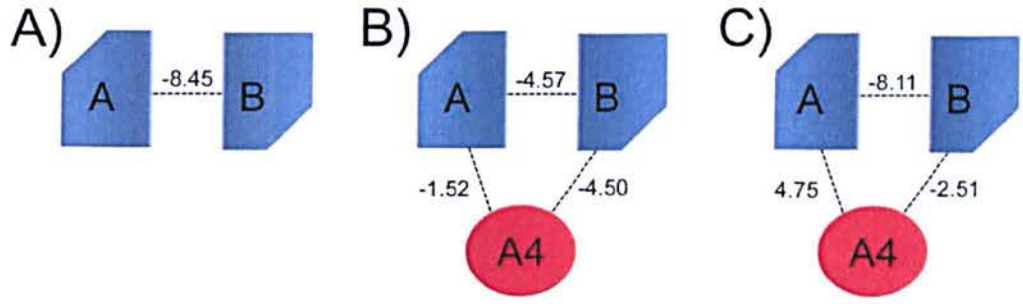


Figure 3.17: Complex stability expressed by interaction energy estimated by FoldX for the structures. HNP1 dimer and A4 are shown by rectangle and circle, respectively. (A) HNP1 dimer (PDB Code: 3GNY), (B) Model.002.01, (C) Model.006.02. Energies are in Kcal/mol.

3.3.13 Comparison with Model Selected by Cluspro Energy Function

Ranking provided by Cluspro energy function selects Model.004.05 as the putative model. Figure 3.18 displays the alignment between Model.002.01 and Model.004.05. In Model.004.05, both W26:B and F28:A interact with W144 but R24 only interacts with D151 out of the three main ASP residues of LDLR. This shows that the ranking provided by the energy function does not best describe a model which corresponds to the experimental results. Based on the energy ranking Model.002.01 is ranked 25 out of 84.

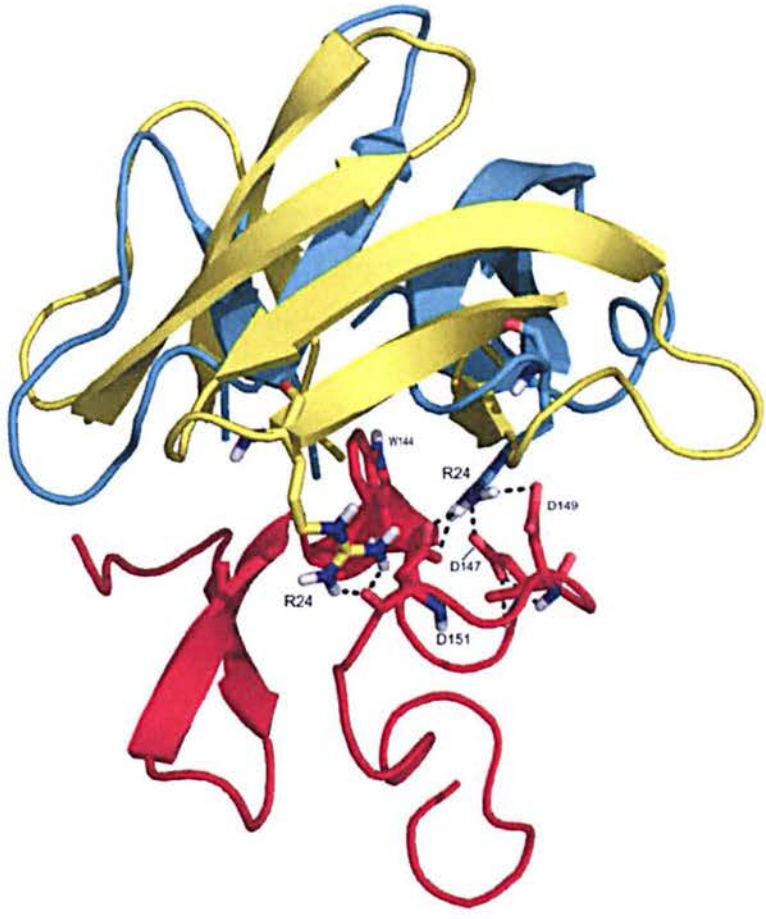


Figure 3.18: Proposed LDLR-HNP1 interaction model according to Cluspro energy function. Structures of A4 are shown in red. HNP structure of model.002.01 and model.004.05, are shown in cyan and yellow, respectively.

3.6 Discussion

The major objective of this investigation was to establish whether a LDLR-HNP1 interaction can occur based upon computational models. Previous reports of i) the versatility of ligand recognition exhibited by the LDLR family (Blacklow 2007), ii) an interaction between human α-defensins with LDLR (Nassar et al. 2002; Higazi et al. 2000), and iii) its role in internalising ligands (such as cholesterol and amyloid-beta (Nassar et al. 2002; Fuentealba et al. 2010; Lazaridis 2005) led to examination of the putative interaction.

The study relating to HNP1 dimer formation, conducted using SPACE (Sobolev et al. 2005), revealed that the structure contains 33 intermolecular contacts including 3

hydrogen bonds. The dimer binding energy calculation of -8.45 Kcal/mol (Figure 3.17.A) is commensurate with several models where stable interactions occur (Kiel et al. 2005; Ponomarev & Audie 2011).

The major observation from the modelling is that interactions between the different chains of Model002.01 are very strong, -10.59 Kcal/mol as a whole (Figure 3.17.B). For model.006.02, a very different scenario is depicted where the energy saving for interaction with the dimer is greatly diminished where the interaction with one monomer requires 4.75 Kcal/mol (Figure 3.17.A). This thermodynamically unfavourable scenario points to Model.002.01 as preferential.

The strength of binding seen in Model.002.01 is reflected in the levels of intra-molecular interactions. In addition to the contacts present in the dimer, binding to the receptor generates a further 48 contacts including 8 as hydrogen bonds and 3 as electrostatic interactions.

Within the complex, an intriguing feature is the modulation of dimer interaction energies depending on which model is studied. A considerably weaker level of dimer binding strength is observed for Model.002.01 which may have ramifications for internalisation should this step proceed through the monomer form. In contrast, for Model.006.02, the binding interaction for the dimer remains strong.

One aim of this study involved identification of the receptor binding mechanism for the purposes of informing the future design of synthetic HNPs to afford maximum internalisation. This chapter highlights the key putative contacts between HNP1 and the LDLR, and moreover, emphasises the potential importance of maintaining the HNP1 dimer form for binding and potentially for internalisation. Further computational studies are required to clarify the mechanism of internalisation and interaction with membrane (Fleming et al. 2008; Lazaridis 2005).

These insights, from computation study based drug design, provide a number of avenues towards novel synthetic antimicrobial peptides which can be synthesised and tested through conventional assays. Strengthening or weakening LDLR-HNP interactions may have synergistic or dysergistic effects on the two key aspects, namely docking and internalisation. In this vein, strengthening the links that make the HNP1 dimer, even to the extent of forming fixed permanent bond to anchor the dimer link, may be an avenue to greater efficacy in some forms of antimicrobial activity.

3.7 Conclusion

Since docking ranking using energy functions are unreliable and fail to distinguish near-native models, in this chapter we proposed a method to rank docking conformations based on a binding site 3D motif that describes the protein's binding site environment. We showed that 3D motif is able to provide ranking which corresponds to experimental results. This is achieved while Cluspro docking energy function fails to identify a plausible model.

The main value of 3D motifs is that they provide biological insight of protein interaction and can be used in real applications as shown in LDLR-HNP1 complex prediction. Another possible application of 3D motifs is to identify putative partners of proteins whose 3D structure is known. A ligand of interest can be evaluated by a 3D motif to see if it fits the expected binding environment.

Despite the advantages of 3D motifs, they have some limitations: First, dependence on literature studies prevents their application for high throughput analysis. Second, the creation of a 3D motif for a set of homologous complexes assumes that the site of interest has high binding specificity. However, homologous proteins may bind to very different ligands using the same binding site (Martin 2010). Third, in this study selection of 3D motifs' residues requires both sequence and structure conservation in terms of chemical properties. Such a constraint does not allow taking advantage of structural neighbours of a protein, which may have low or no sequence similarity, to produce useful information about their binding sites (Zhang et al. 2011; Jordan et al. 2012).

In Chapter 4 we will address the first two issues by proposing a framework suitable for high throughput protein complex prediction. This is achieved by developing a protein interface predictor method which considers ligand diversity at binding site. This predictor is then used to re-rank docking conformations. The third limitation is also partially addressed in Chapter 4 since it takes advantage of remote homologues. Moreover, this will be further investigated in Chapter 5 where structural neighbours are considered.

4 Protein Interface Prediction and its Application to Re-ranking Docking Conformations

4.1 Introduction

Since proteins function by interacting with other molecules, analysis of protein-protein interactions is essential for comprehending biological processes (Shoemaker & Panchenko 2007). Whereas understanding of atomic interactions within a complex is especially useful for drug design, limitations of experimental techniques have restricted their availability (Reš et al. 2005). Despite progress in protein interface and docking predictions, there is still room for improvement. In this chapter, we propose T-PioDock, a framework for prediction of a protein complex 3D structure from the structures of its components. T-PioDock supports the identification of near-native conformations from 3D models that docking software produced by scoring those models using binding interfaces predicted by T-PIP. T-PioDock is freely available for download at: http://manorey.net/bioinformatics/wepip/. The rest of the chapter is organised as follows, Section 4.2 presents the most relevant state-of-the-art methods in protein interface prediction (section 4.2.1) and its application to re-ranking docking conformations (section 4.2.2). The proposed methodology is described in details in section 4.3. First we investigate the principles behind T-PIP interface predictor (section 4.3.2) and then we discuss how this prediction can be used for re-ranking docked conformations (section 4.3.3). In section 4.4 T-PioDock framework is evaluated which

is followed by further discussion and a summary of chapter in section 4.5 and 4.6, respectively.

4.2 Related Work

4.2.1 Protein Interface Prediction

Protein-protein interaction (PPIs) is essential for the functionality of living cells. Alterations of these interactions affect biochemical processes which may lead to critical diseases such as cancer (Shoemaker & Panchenko 2007). Therefore, knowledge about protein interactions and their resulting 3D complexes can provide key information for drug design. A number of experimental techniques are available to identify residues involved in PPIs (Shoemaker & Panchenko 2007). Although they provide valuable contribution to PPI knowledge, their cost in terms of time and expense limits their practical use (Ezkurdia et al. 2009). Consequently, computational methods have been proposed to identify protein interfaces.

As discussed in chapter 2, when the 3D structure of the query protein (QP) is available, integration of structural information, e.g. residues secondary structure or solvent-accessible surface area, allows better predictions (Ofran & Rost 2007a; Šikić et al. 2009). Three approaches taking advantage of these properties are seen as state-of-the-art (Zhou & Qin 2007). ProMate combines 13 different properties, such as chemical component, geometric properties and information from relevant crystal structures, to generate a quantitative measure (Neuvirth et al. 2004). Protein interface residues are then predicted using a clustering process relying on mutual information. Alternatively, Cons-PPISP discriminates between residues using neural networks trained with protein's surface sequence profiles and solvent accessibility of neighbouring residues (H Chen & Zhou 2005). Finally, PINUP addresses the problem using an empirical energy function which is based on a linear combination of side chain energy score, interface propensity and residue conservation (Liang et al. 2006). Despite fundamental differences, these three approaches display very similar performance (Zhou & Qin 2007). However, each of them seems to capture different important aspects of residue interactions. As a result, a meta-predictor, Meta-PPISP, combining their scores using a

linear regression analysis, is able to outperform each of these individual methods in terms of accuracy (Qin & Zhou 2007b).

With the increasing number of experimentally determined protein 3D structures, template-based approaches (Chapter 2, section 2.3.2) have become the recent focus of interface prediction. They are guided by two main facts. First, homologous proteins tend to display structurally similar interaction sites (Tsai et al. 1996; Aloy et al. 2003). Secondly, protein's binding sites are evolutionarily conserved among structurally similar proteins (or structural neighbours) (Q. C. Zhang et al. 2010; Zhang et al. 2011; Konc & Janezic 2007; Konc & Janežič 2010a; Carl et al. 2008; Carl et al. 2010). Even remote structural neighbours have been shown to display a significant level of interface conservation (Q. C. Zhang et al. 2010). Consequently, template-based interface predictors rely on analysing proteins which are structurally similar to the query protein. Predictors in this category are either based on homologous template (Chapter 2, section 2.3.2.1) or structural neighbours (Chapter 2, section 2.3.2.2). IBIS (Tyagi et al. 2012) is the state-of-the-art from the first category and similarly PredUs (Zhang et al. 2011; Q. C. Zhang et al. 2010) and PrISE (Jordan et al. 2012) are the best predictors from the second category.

For each QP, IBIS (Tyagi et al. 2012) extracts homologous complexes which have at least 30% sequence similarity to the QP. First, using VAST algorithm (Gibrat et al. 1996), QP and the homologous complexes are structurally aligned. Homologues for which at least 75% of their binding sites are structurally aligned with the QP are kept. IBIS does not remove redundant structures and will keep all homologues which meet the above condition. Second, a Structure-based-MSA (S-MSA) of the homologues in reference to the QP is build and all interface residues of homologous complexes are mapped on it. Third, a similarity matrix of homologues is created by comparing each homologue against all other homologues. For scoring two pairs of homologue, A and B, only positions which are marked as interfaces on the S-MSA for either A or B is used. The score of each position, which shows the aligned residue of A and B in that position, is its BLOSUM score (Henikoff & Henikoff 1992). This score takes into account gap penalties and is normalised to allow comparison among different pair of homologues. Fourth, using the generated similarity matrix the homologues binding sites are clustered using a pseudo-free energy function which defines the cut-off value. Therefore, each

cluster will have a set of binding sites which are defined to be similar. Fifth, for each cluster with more than one non-redundant protein (threshold of 90%), 4 scores are calculated: their weighted sum provides a ranking score for that cluster. These scores for each cluster are: (i) positional conservation in the binding site, (ii) average of total number of contacts (iii) position specific score matrix (PSSM) and (iiii) average sequence identity of the cluster members to the QP . Finally, all clusters are ranked based on their ranking score and the interfaces of the best ranked cluster is mapped on the QP.

While IBIS prediction depends on the availability of homologues complexes, PredUs and PrISE take advantage of structural neighbours. PredUs, maps interacting residues from structural neighbours onto query proteins (QP) even if they are not homologous (Zhang et al. 2011). Whereas PredUs still dependents on the existence of structural neighbours of QP, PrISE proposes to deal with this limitation by predicting interface residues from local structural similarity only (Jordan et al. 2012). This is achieved using a repository of structural elements (SE) generated from the Protein Data Bank (PDB) (Berman et al. 2000). For each SE of the query protein (consisting of a central residues and it neighbours) a set of similar SEs are extracted from the repository. A weight is then assigned to them based on the local (SE similarities) and global (protein surface similarity) structural similarities. The central residues of the query protein's SE are predicted as interface if a weighted majority of its similar SEs are interface residues. Although more general, PrISE displays comparable performance to PredUs (Zhang et al. 2011). Both methods will be further discussed in chapter 5.

Those two approaches have significantly improved the ability of predicting interface residues; see section 4.4.2.2. However, they do not deal satisfactorily with the very heterogeneous nature of the PDB. First, the presence of duplicate complexes and/or homologues may bias predictions towards specific configurations, and can affect performance negatively. Second, confidence in the information provided by the interface of a structural neighbour should depend on its degree of homology with QP. Although PrISE acknowledges both issues, it does not address the first one (Zhang et al. 2011). On the other hand, PredUs deals with these matters in a binary fashion. Complexes involving structural neighbours are clustered and a 40% similarity cut-off is used to choose the representatives which will inform interface prediction. Third, both

PrISE and PredUs are able to make prediction only if a structure is available for the QP. This significantly reduces their applicability. Similarly, IBIS does not address the three above mentioned point. In addition, since it relies on close homologues (sequence identity above 30%), it does not fully take advantage of templates and fails to make predictions for some proteins.

In this chapter we introduce T-PIP (Template based Protein Interface Prediction) framework, a novel PIP approach based on homologous structural neighbours' information. T-PIP addresses the above mentioned limitations by quantifying, first, homology between QP and its structural neighbours and, second, the diversity between the ligands of the structural neighbours (here, ligands refers to the interacting partners of proteins). Finally, predictions can be performed for sequences of unknown structure if that of a homologous protein is available. T-PIP's main contribution is the weighted score assigned to each residue of QP, which takes into account not only the degree of similarity between structural neighbours, but also the nature of their interacting partners.

4.2.2Scoring Protein-Protein Docking Conformations

Protein-protein docking aims to computationally predict the 3D structure of a protein complex using the unbound structures of its components (Halperin et al. 2002; Smith & Sternberg 2002; Ritchie 2008; Bonvin 2006). Since docking software produces 100's to 1000's of putative models, their exploitation requires the ability to score them accurately (Li et al. 2003; Pierce & Weng 2007; Vreven et al. 2011).

As discussed in chapter 2, knowledge of predicted interface residues has proved particularly successful (Qin & Zhou 2007a) in comparison to knowledge-based potentials and machine learning functions. It has been applied to either constrain the initial search space of docking software (pre-filtering) (Van Dijk, De Vries, et al. 2005; De Vries & Bonvin 2011; Li & Kihara 2012) or score docking conformations (post-filtering) by calculating the similarity between the interfaces of the docked models and the predicted ones (Qin & Zhou 2007a; Xue et al. 2010). On one hand, pre-filtering limits the search space from the start but is less practical to use since the constraint should be imposed on the algorithm of the docking method (Qin & Zhou 2007a). On the other hand, post-filtering can be applied to models generated by any docking software

and can be combined with other scoring function. As a consequence, post-filtering has proved more popular and practical (Qin & Zhou 2007a).

Experiments aimed at gaining insight into the value of using interface information showed that knowledge of at least 40% of interface residues is sufficient to significantly improve rankings (Zhou & Qin 2007) of models generated by ZDOCK (Chen, Li, et al. 2003). As a consequence standard interface prediction approaches, such as cons-PPISP (H Chen & Zhou 2005), Promate (Neuvirth et al. 2004) and HomPPI (Xue et al. 2011), were extended to evaluate the fit of docked proteins against their predicted binding sites (Qin & Zhou 2007a; Xue et al. 2010). By combining five interface predictors, i.e. Promate (Neuvirth et al. 2004), PPI–Pred (Bradford & Westhead 2005), PPISP (Zhou & Shan 2001), PINUP (Liang et al. 2006), and SPPIDER(Porollo & Meller 2006) into one framework called MetaPPI (Huang & Schroeder 2008), success rates were improved by 15% in comparison to the best individual predictors. DockRank (Xue et al. 2010) using HomPPI (Xue et al. 2011), which has a prediction performance higher than Promate (see chapter 2), has shown to provided better ranking than Cluspro on Docking Benchmark 3.0. However, since DockRank is constrained by the availability of a homologous complex containing both chains of the query, it fails to provide ranking for new complexes. Finally, instead of representing interacting interfaces as a two-patch system, SPIDER (Khashan et al. 2012) evaluates multi-body interactions using a library of contacts containing graph representations of common interfacial patterns. Although SPIDER has claimed to outperform ZRANK (Pierce & Weng 2007), its usage is limited by the requirement of accurate and high resolution interfaces.

This chapter contributes to this topic by proposing T-PioDock (Template based Protein Interface prediction and protein interface Overlap for Docking model scoring), a complete framework for prediction of a complex 3D structure. T-PioDock supports the identification of near-native conformations from 3D models that docking software produced by scoring those models using binding interfaces predicted by T-PIP.

4.3 Methodology

4.3.1 Overview

As highlighted in the latest edition of CAPRI (Fleishman et al. 2011), despite progress in docking predictions, there is still room for improvement. In this study, we contribute to this topic by proposing T-PioDock (Template based Protein Interface prediction and protein interface Overlap for Docking model scoring), a complete framework for prediction of a complex 3D structure. T-PioDock aims at supporting the identification of near-native conformations from 3D models produced by any docking software by scoring those models. T-PioDock exploits template based predictions of complexes' binding interfaces to evaluate docking configurations.

T-PioDock's pipeline is described in Figure 4.1. The input to the system is the 3D structures of the query proteins. First, the T-PIP module (Template based Protein Interface Prediction) evaluates the complexity of the protein targets –i.e. 'trivial', 'homologous' or 'unknown'- in terms of homologue availability in the PDB (Berman et al. 2000) and predicts their interfaces using the most appropriate template-based method. These interfaces are then passed to the PioDock module (Protein Interface Overlap for Docking model scoring) which exploits them to score conformation models produced by any docking software. Finally, those scores can be used to help at the identification of near-native conformations by ranking available conformation models.

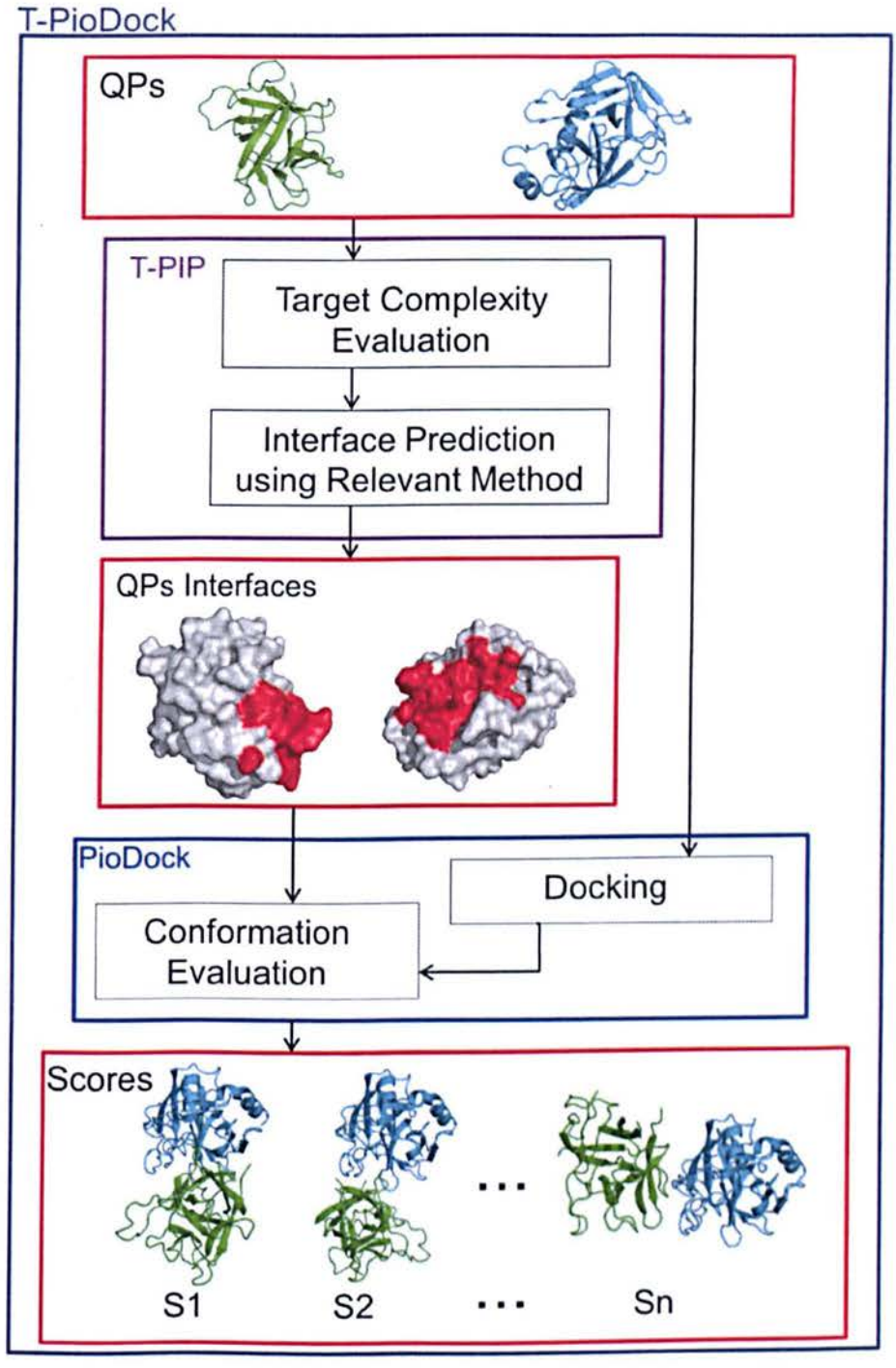


Figure 4.1: T-PioDock pipeline.The two query proteins, shown in green and cyan, are the inputs to T-PIP. T-PIP evaluates target complexity and predicts their interfaces using the most relevant method. In the figure, the predicted interfaces are coloured in red on the query protein surface which is in grey. PioDock exploits these interfaces to score models produced by standard docking. The result is a list of docked complexes with a score (S) associated to them.

4.3.2Template Based Protein Interface Prediction Principle

When two protein chains form a dimer, they bind through their interaction interfaces. We propose T-PIP which predicts the amino acids which are involved in binding interactions based on the 3D structures of the dimer partners. In this study dimer refers to any two protein chains involved in interaction. Not only does our approach predict the locations of interfaces when both binding partners are known, but it also infers the most likely binding interface of a single protein. T-PIP module, first, evaluates the complexity of a protein target in terms of availability of 3D structures of homologous proteins and, second, applies the most relevant template based interface predictor. Figure 4.2.A and Figure 4.2.B describe those interface prediction pipelines for single and pair protein queries respectively.

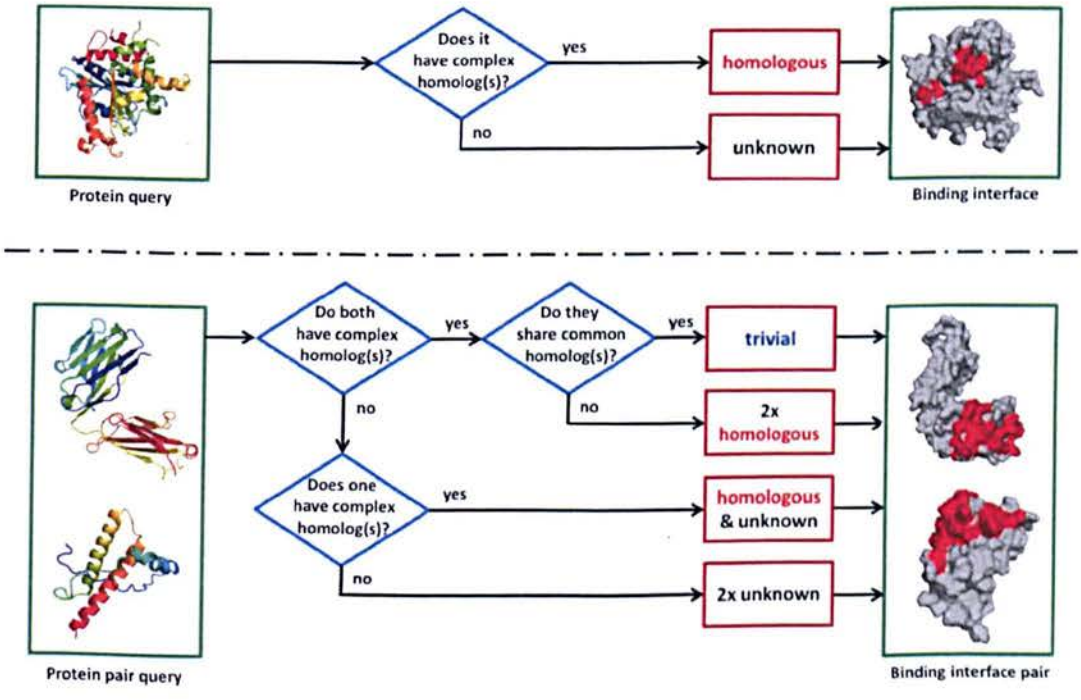


Figure 4.2: Interface prediction framework for single (A) and pair protein queries (B). A) For a single query, if at least one homologous complex exists, interfaces are predicted using 'homologous' otherwise 'unknown' is used. B) For a pair of proteins, depending on the existence of homologous complexes, interfaces are predicted by one or a combination of the 'trivial', 'homologous' and 'unknown' methods.

This methodology relies on discovering interaction patterns from the analysis of the 3D structures of complexes involving homologs of the dimer partners, called 'homologous complexes'. In this work, proteins are defined as homologous if their sequence similarity is expressed by an $E-value \leq 10^{-2}$ as suggested in (Xue et al. 2011). Initially, protein targets are categorised into three categories: 'trivial', 'homologous' and 'unknown'. This is achieved by, searching homologues complexes of the query proteins in PDB (Berman et al. 2000) using BLAST (Altschul et al. 1997). If among their homologous complexes both QPs share at least one common complex, the target is considered to be 'trivial', in that case T-PIP exploits them as templates to predict both interfaces jointly (see section 4.3.2.1). If each QP possesses a set of homologous complexes, but none of them belongs to both sets, the target is classified as 'homologous' and interfaces are predicted independently from the interaction partner (see section 4.3.2.2). Finally, if no homologous complex is found for at least one of the QP, the target is judged to be 'unknown'. In this case, a third party PIP software, such as PredUs (Liang et al. 2006), is required. In this work, we use PredUs when homologous complexes are not available, because, not only it is one of the best performing methods, but also it has been implemented as a Web server which can be used free of charge.

4.3.2.1 Trivial Targets

With rapid increase of experimentally determined structures, homology modelling of the whole 3D structure of a dimer is becoming more and more possible. It has been reported (Ghoorah et al. 2011) that high quality homologous models could be found for 62% of the protein complexes present in the standard Protein Docking Benchmark 4.0 (Hwang, Vreven, Janin, et al. 2010). Consequently, template-based docking methods have been proposed based on common dimer complexes, i.e. dimers where each chain is homologous to a sequence of the query dimer (Ghoorah et al. 2011; Kundrotas & Alexov 2006; Kundrotas et al. 2008). For example, 66 % of complexes generated by HOMBACOP were categorised as either acceptable or medium-quality models according to CAPRI assessment criteria (Kundrotas et al. 2008). Since these approaches have proved particularly accurate, for 'trivial' targets interfaces are inferred based on common complex homologs.

Homologous complexes of each sequence of the protein query pairs are extracted from the PDB using BLAST. Common homologous complexes are then selected and ranked by multiplying the E-values associated with the sequential alignments of each query chain with the homologous chain of the common complex. The common complex with the lower score is selected as the template from which the interfaces of the query chains are inferred. This is achieved by mapping the interface residues of the templates on the query chains according to their sequence alignments.

4.3.2.2 Homologous Targets

The idea behind our approach relies on the observation that interface residues are usually structurally conserved between evolutionary related proteins (Zhang et al. 2011). Following extraction of homologous complexes from the PDB using BLAST, the 3D structure of QP is structurally aligned with its homologues. In this study processing time is reduced by considering at most the 30 homologous complexes involving a chain whose E-value shows closest similarity to the QP. Alignment of multiple protein structures is performed by Multiprot (Shatsky et al. 2004), since it is a popular tool (Keskin et al. 2005; Halperin et al. 2004; Winter et al. 2006; Ogmen et al. 2005) that has already be used successfully in interface residue prediction (Keskin et al. 2005). Using this information, a structure-based multiple sequence alignment (S-MSA) is produced. Then, known interface residues of the homologous complexes are highlighted on the S-MSA. Figure 4.3 shows a schematic representation of the process. In agreement with the CAPRI definition (Janin & Wodak 2007), an interface residue is defined as an amino acid whose heavy atoms are within 5Å from those of a residue in a separate chain. Using this multiple alignment, an interaction score is calculated for each residue of the query protein (see section 4.3.2.2.1). Figure 4.4 shows a detailed example of S-MSA and the calculated interaction scores. Finally, the expected number of interface residues, n_{IA}, is predicted from known interfaces (see section 4.3.2.2.2). The n_{IA} residues with the highest scores are then returned as defining the interaction interface.

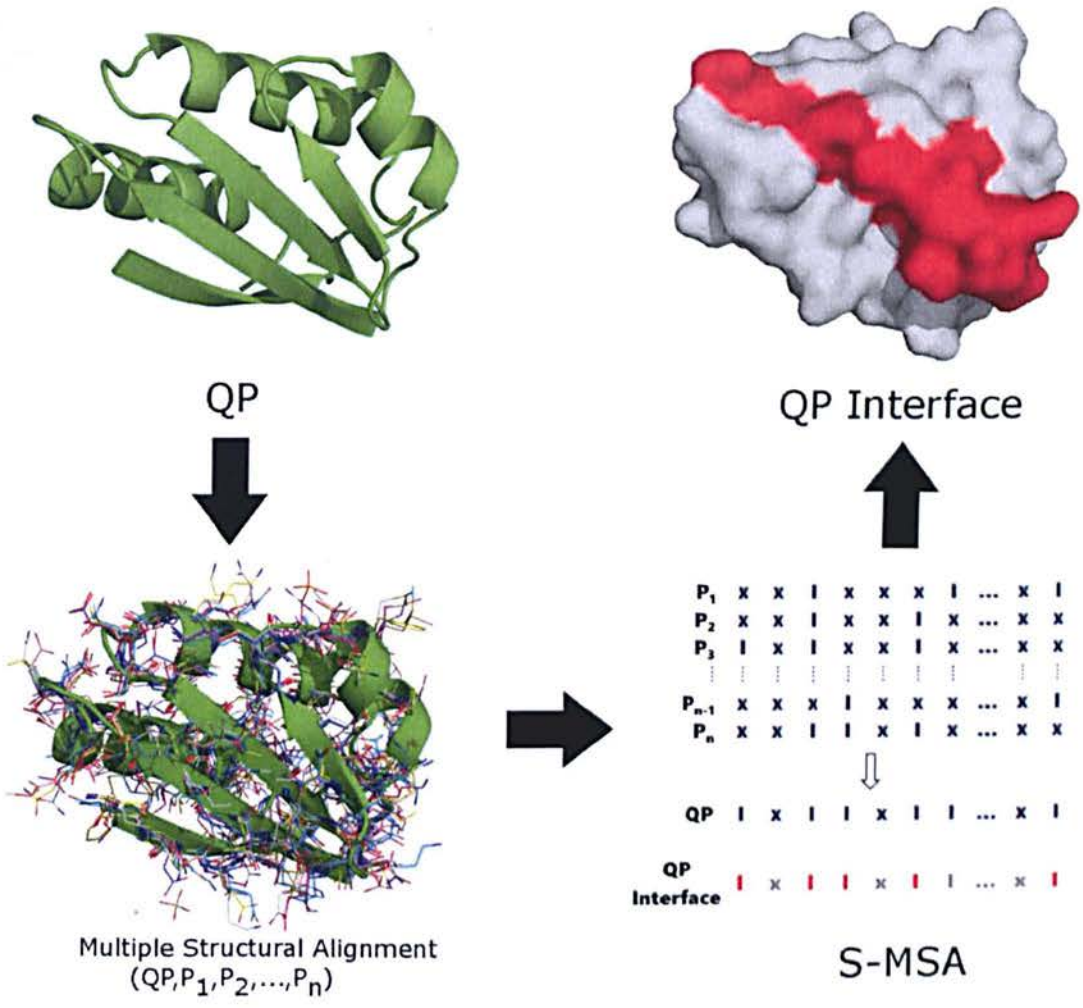


Figure 4.3: Application of the T-PIP on a homologous query protein (green). First, it is structurally aligned with its homologous complexes. Then, an S-MSA is produced where X and I represent non-interface and interface residues, respectively. Finally, interaction residues (red) of QP are predicted according to interaction scores and the estimated size of the interface. Note that residue weights are not shown here.

Although this method was initially designed for predicting interface residues for query proteins whose 3D structure is known, it can also be applied when only the sequence of the query is available. In this case, an initial S-MSA is created using only homologous complexes of the QP. Then, the QP sequence is integrated to that S-MSA using ClustalW Profile Alignment command (Thompson et al. 1994) to create a complete MSA.

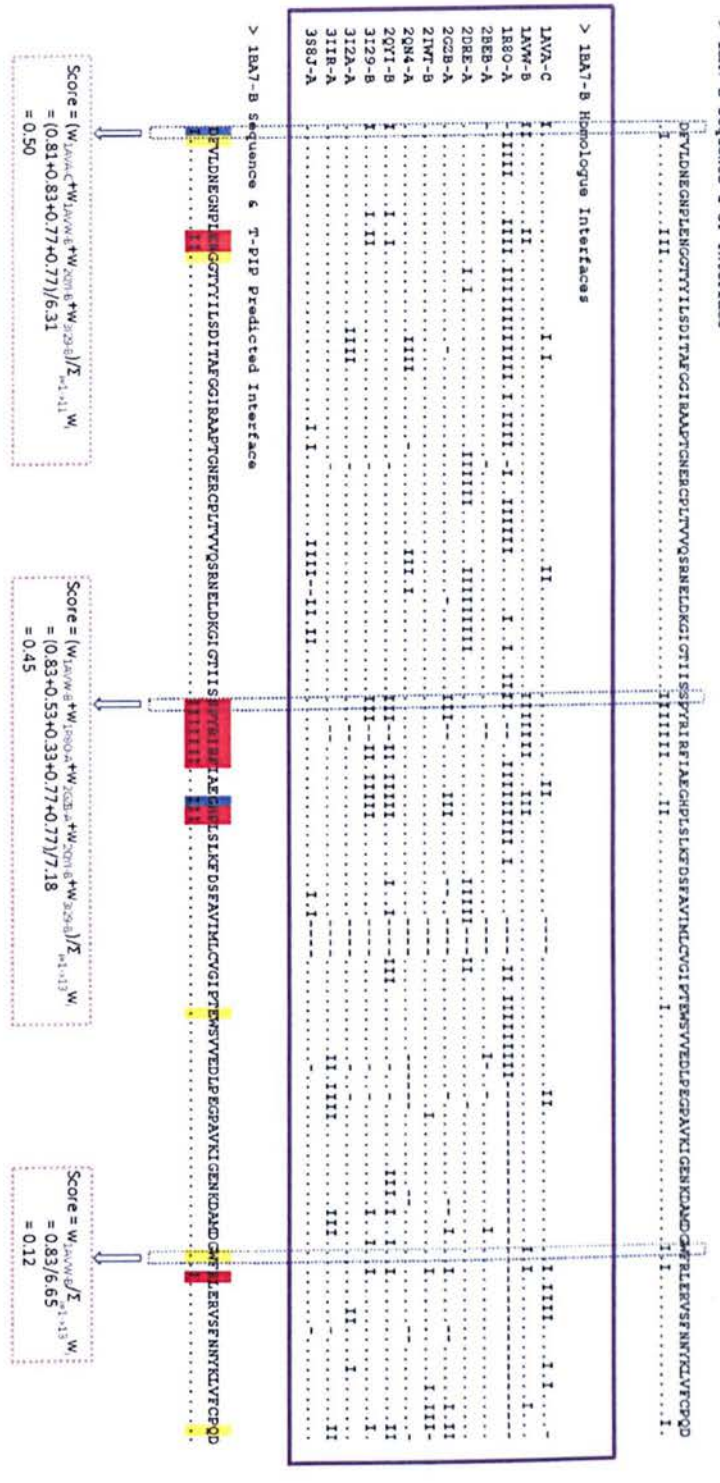


Figure 4.4: Example of interface prediction for 'homologous' target 1BA7-B. The amino acid sequence and the ground truth interfaces are shown at the top ('I' and '.' represent interface and non-interface residues, respectively). Actual interfaces of homologues of 1BA7-B are show in the purple box. Prediction made by T-PIP is displayed below the box, where red, yellow and blue highlights show residues which are correctly, missed and wrongly predicted as interface residues by T-PIP. 3D representation of this interface is provided in Figure 4.11. At the bottom of the figure, three examples of T-PIP weighting calculations are displayed.

4.3.2.2.1 Interaction Score

In order to identify residues likely to be involved in the dimer interaction, we propose a residue scoring function which relies on the S-MSA of QP and its complexed homologues. In principle, any QP amino acid aligned with a residue involved in a dimer interaction is a potential candidate. However, we express confidence in the association of interaction activity to a residue according to three factors: the degree of homology between the QP sequence and that of the protein from which the interaction is inferred, the nature of the ligand involved in the interaction with the homologous protein and the number of homologous proteins suggesting interaction.

First, the more homologous a complexed protein, k, is to QP, the more informative is that protein regarding which residues are likely to be involved in the dimer interaction. We express this information by the query weight, x_k, Equation 4.1:

$$x_k = \begin{cases} 1 - 10^{-200}, \text{if } E_k < 10^{-200} \\ 1 - E_k, \text{if } 10^{-200} \leq E_k \leq 10^{-2} \\ 0, \quad \text{if } E_k > 10^{-2} \end{cases} \qquad \textbf{Equation 4.1}$$

Where E_k is the E-value of protein E_k against QP as estimated by BLAST. The threshold are based on the constraints for retrieving homologues complexes ($E-value \leq 10^{-2}$) suggested in (Xue et al. 2011).

Secondly, since none of the homologous complexed proteins interacts with the query ligand, diversity of ligands has to be rewarded given that they increase generalisation of interaction patterns. A second weight conveys this requirement by penalising homologous proteins, whose ligands are similar to each other. This is estimated by the average sequence identity between the sequence of a ligand and all the other as expressed by the arithmetic mean of the pair wise E-values. Given a homologous complex protein, *k*, interacting with a ligand, L_k, and the other $N-1$ homologous complexed proteins interacting with their respective ligand, L_j, the ligand weight, y_k, is formulated as Equation 4.2:

$$y_k = \begin{cases} \frac{\sum_{j=1, j\neq k}^{N} E_{(L_k, L_j)}}{N-1}, \text{if } N > 1 \\ 1, \quad \text{if } N = 1 \end{cases} \qquad \textbf{Equation 4.2}$$

Where $E_{(L_k,L_j)}$ is set to 1, if $E_{(L_k,L_j)} > 1$, and $E_{(L_k,L_j)}$ is set to 10^{-200}, if $E_{(L_k,L_j)} < 10^{-200}$.

The y_k score is designed so that the presence of complex duplicates does not bias predictions towards their configuration. For example, if a QP has 3 complex homologs, A, B and C where L_A is unrelated to L_B and L_C, but L_B and L_C are identical, $E_{(L_A,L_B)} = 1$, $E_{(L_A,L_C)} = 1$ and $E_{(L_B,L_C)} = 0$. Therefore, $y_A = y_B + y_C$, i.e. interface configurations of A and B/C will have the same weight.

The weighted score of the residue i of protein, k, is expressed by the product of these two weights (Equation 4.3):

$$w_{ik} = \begin{cases} x_k y_k, & \text{if i interacts with } L_k \\ 0, & \text{otherwise} \end{cases} \qquad \textbf{Equation 4.3}$$

Finally, since it was shown that usage of non-interface information improves prediction performance (Zhang et al. 2011; Jordan et al. 2012), the score for residues i of QP is calculated in (Equation 4.4) as the sum of the weights of the interface residues in the homologs over all the interface and non-interface residues which are 3D aligned with i:

$$S_i = \frac{\sum_{j=1}^{N} w_{ij}}{\sum_{j=1}^{N} x_j y_j} \qquad \textbf{Equation 4.4}$$

Note that for non-interface residues the ligand which is geometrically the closest is used to calculate their weight y_k.

4.3.2.2.2 Estimation of the Number of Interface Residues

After calculating S_i for all the amino acids of QP, it is necessary to estimate the expected number of interface residues of its interface, n_{IA}. Studies have shown that despite variability in ligand structures, the binding location between homologous structures and their ligands is conserved (Keskin & Nussinov 2007). This suggests that the number of interface residues between homologues should remain quite stable even when the binding partners vary. Therefore, T-PIP uses the weighted average number of interacting amino acids of all QP's homologues (n_{IA}) to estimate the number of interface residues of QP (Equation 4.5):

$$n_{IA} = \sum_{i=1}^{R} S_i \qquad \text{Equation 4.5}$$

Where R is the number of residues in the QP sequence plus the number of gaps added to allow alignment with its homologues. Finally, once n_{IA} has been calculated, the predicted interface is defined as the n_{IA} residues with the highest scores.

4.3.2.2.3 Comparison of Prediction to Ground Truth

After predicting the interfaces using T-PIP and other interface predictors, their performances are compared against the GT of the benchmark datasets using metrics provided in 2.5.1.

For each target, a GT sequence is generated using their X-ray structure: interacting residues are replaced by the symbol 'I', whereas non-interface ones are substituted by the symbol 'N'. Note that interfaces are defined using CAPRI's definition, i.e. all residues of a protein chain that have atoms less than 5 Å apart from the interacting partner. For example in Figure 4.4, on the top, the amino acid sequence of 1BA7-B is shown. Underneath it, the ground truth interfaces and non-interfaces are displayed as a sequence of Is and dots ('I' and '.' represent interface and non-interface residues, respectively). Then, each interface prediction is expressed similarly as a sequence of 'N' and 'I' symbols. Again in Figure 4.4, below the purple box, the predicted interface and non-interface are shown as a sequence of symbols. Finally, comparison of a prediction with the GT sequence allows calculating the TP, FP, TN and FN: matches between I (or N) at the same position are marked as TP (or TN), whereas mismatches, I in GT and N in prediction (respectively, N and I), are noted as FN (respectively, FP). In Figure 4.4, TP, FN and FP are shown in red, yellow and blue highlights, respectively, on the predicted sequence.

4.3.2.2.4 Calculating the Significance of the Predictions

After calculating the performance of T-PIP and other interface predictors, differences among them are evaluated to establish their statistical significance. This is achieved by calculating for each metric their associated standard deviations and P-values. Given a mean value (μ) of metric for all the targets in a dataset, the standard deviation (σ) is calculated as below:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Where n is number of targets in the dataset and x_i is the calculated performance value for target i. A smaller standard value shows that the distributions of the data are closer to the mean value. Standard deviation can be used to measure the uncertainty and large standard deviations can show that differences are not significant among different datasets.

P-values are calculated between T-PIP and each predictor (on different metrics) by applying a two-tailed t-test with a null hypothesis of "there is no relationship between two performances of the predictors". A higher P-value indicates that the probability that the observed difference among two phenomena has occurred by chance is higher. P-values smaller than 0.05 are considered as a statistically significant threshold (Darnell et al. 2007). Therefore, any P-value smaller than 0.05 shows that the two dataset do not happen by chance and the null hypothesis can be rejected. P-values are calculated using the T-TEST function of Excel.

4.3.3 Protein Interface Overlap for Docking Model Scoring

PioDock scores docking conformations according to their consistency with interfaces predicted by T-PIP. Given the putative docking conformation of a complex A-B, the model is assigned a score on the basis of the overlap between its interface residues and those predicted for each of its components, i.e. A and B. We define the complex overlap score of A-B, $\text{complexOverlap}_{A-B}$, as the average between two overlap scores (overlap) calculated for A and B separately:

$$\text{complexOverlap}_{A-B} = \frac{\text{overlap}_A + \text{overlap}_B}{2} \qquad \textbf{Equation 4.6}$$

Where the overlap score for A in regard to B, overlap_A, is calculated using the following formula proposed by Kuo et al. (Kuo et al. 2011):

$$\text{overlap}_A = \frac{\text{interface } A_{Docked} \cap \text{interface } A_{T-PIP}}{\sqrt{(\text{intrfaces } A_{Docked}).(\text{interfaces } A_{T-PIP})}} \qquad \textbf{Equation 4.7}$$

Where interface A_{Docked} and interface A_{T-PIP} in the numerator of the formula represent, respectively, the sets of the residue in the interfaces of docked model and the ones predicted by T-PIP. While interface A_{Docked} and interface A_{T-PIP} in the denominator represent the number of residues in the interface of docked model and the ones predicted by T-PIP, respectively.

complexOverlap scores of native complexes should equal to 1, whereas completely incorrect conformations should be assigned a value of zero. In this study, complexOverlap score was used to rank all conformational models generated by docking software for a given complex. When experiment was conducted to evaluate the PioDock module on its own, actual target interfaces were used instead of their predictions.

Note that when no interface prediction could be performed for one of the two docking partners, the overlap score for that protein is equal to zero and complexOverlap score is calculated using only the overlap score of the other protein.

4.3.3.1 Evaluation of Docking Model Scoring

In order to allow any evaluation it is necessary to have some gold standard or ground truth. However, comparison of two docked models is far from being a straightforward task since CAPRI has to use three differences measures to assess the docking model quality (Lensink & Wodak 2010): l-rmsd measures the RMSD between the backbones of the two complexes ligands, i-rmsd restricts its evaluation to interface residues, whereas f_{nat} is the fraction of native contacts within the interface. Since f_{nat} can only discriminate between relatively good configurations – all models failing to predict a single interface residue receive a score of 0, only i-rmsd and l-rmsd are used in this study. The normalised Pearson's chi-squared statistic ($normalizedx^2$) (see chapter 2, section 2.5.2 for details) was used to compare between a gold standard ranking of models and rankings generated by those methods. A lower $normalizedx^2$ represents a ranking which is more similar to the ranking of the gold standard. Perfect ranking would return a value of 0.

4.4 Evaluation

4.4.1 Datasets and Tools

Three standard benchmark datasets were used in this study: Ds56unbound (Janin & Wodak 2007), Docking Benchmark 3.0 (DBMK3.0) (Hwang et al. 2008) and Docking Benchmark 4.0 (DBMK4.0) (Hwang, Vreven, Janin, et al. 2010) . While Ds56unbound has been our training set, DBMK3.0/4.0 has been used for evaluating the interface predictors and docking model ranking approaches. These datasets contain high-resolution protein structures both in their unbound and bound forms.

Ds56unbound is comprised of 56 unbound chains generated from 27 CAPRI targets, T01~T27 (Janin & Wodak 2007). In total, it contains 12173 residues including 2112 interacting ones. This dataset is used to perform evaluation of all interface prediction methods of interest. Since interface residues are not explicitly provided in DS56unbound, they were generated from the interface residues in their bound form. The process is illustrated in Figure 4.5. First, interface residues are detected on the bound complexes. Then, the unbound sequences are aligned with the bound sequences. Finally, the interfaces are mapped from the bound sequences onto the unbound ones.

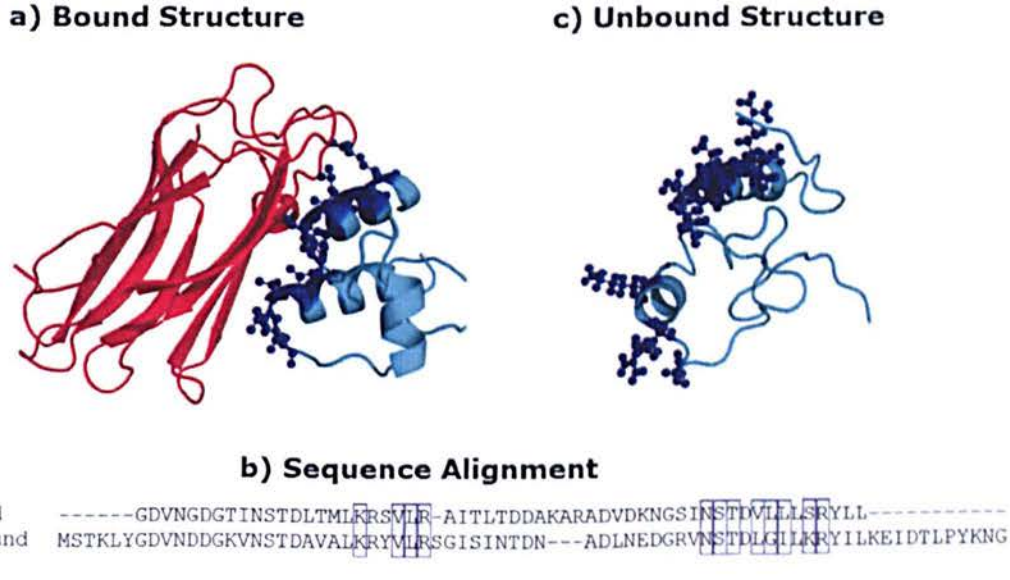


Figure 4.5: Generation of ground truth interface residues. a) Interfaces residues (blue spheres) are identified on the bound structure (cyan). The interaction partner is in red. b) The unbound and bound sequences are aligned to infer interfaces of the unbound structure. Mapping is shown by blue rectangles. c) Inferred interfaces are shown as blue spheres on the unbound structure (cyan).

DBMK3.0 and DBMK4.0 were originally introduced for the evaluation of protein docking methods. DBMK3.0 contains 124 unbound-unbound targets and 309 protein chains, whereas DBMK4.0 is an extension with 53 new targets. Targets are classified into three categories - rigid body, medium difficulty and difficult - based on their degree of conformational change between the bound and unbound forms. These datasets contain essentially dimers, but there are also a few trimers and tetramers. Since there is no agreed methodology for the evaluation of predicted interfaces when dealing with complexes involving more than two chains, those oligomers were excluded from our experiments to ensure fair and consistent comparisons. Moreover, in order to allow comparison with PredUs, we only considered a subset of DBMK3.0, where the chains share at most 40% sequence similarity and their lengths are above 50 amino acids. As a consequence, we produced two subsets of DBMK3.0 and DBMK4.0, i.e. DS120 and DS236 (See Additional files 1 and 2 for details), with 120 and 236 dimer targets respectively. The most promising interface prediction methods were further evaluated on those datasets and docking experiments were conducted on DS236.

In this study, initial docking predictions were produced using the ClusPro 2.0 docking server (Comeau et al. 2004), which performed best at CAPRI 2009 (Kozakov et al. 2010). For a pair of proteins, Cluspro generates hundreds of conformational models usually containing at least one near native model. Although these models are evaluated by an energy-based scoring function and clustering, rankings performed according to those scores have proved unreliable (Esmaielbeiki et al. 2012; Xue et al. 2010).

4.4.2Evaluation of Interface Prediction Method

4.4.2.1 Evaluation of the Interaction Score

All chains from Ds56unbound were processed by our interface prediction framework. Following initial homolog search where the Ds56bound complexes were excluded from the BLAST results, homologous complexes were returned for 51 chains. According to T-PIP, 27 chains were classified as 'trivial' (Ds27unbound), 24 as 'homologous' (Ds24unbound) and 5 as 'unknown'(Ds5unbound). Table 4.1 provides detailed performance of our system using standard measures. As expected, the more the method is able to exploit homology, the better is the interface prediction. Moreover, the

table reveals that T-PIP prediction for 'homologous' targets is quite conservative: it displays relatively low recall value compared to precision. Figure 4.6 illustrates qualitatively those results by displaying representative predicted interfaces compared to ground truth.

Table 4.1: Detailed performance of the WePIP framework. DSxunbound: this means x chains out of the 56 unbound chains are solved by this specific predictor.

Predictor & Categories	Precision	Recall	F1	Accuracy	MCC	AUC
T-PIP DS56unbound	**53.9**	**48.5**	**49.6**	**84.0**	**41.1**	**72.9**
Trivial (DS27unbound)	68.8	60.5	63.2	87.0	56.0	77.1
Homologous (DS24unbound)	43.3	35.6	38.2	82.3	28.8	70.1
Unknown (DS5unbound)	23.6	45.5	30.1	75.8	19.4	63.2
Homologous + Trivial (DS51unbound)	56.8	48.8	51.5	84.8	43.2	73.6

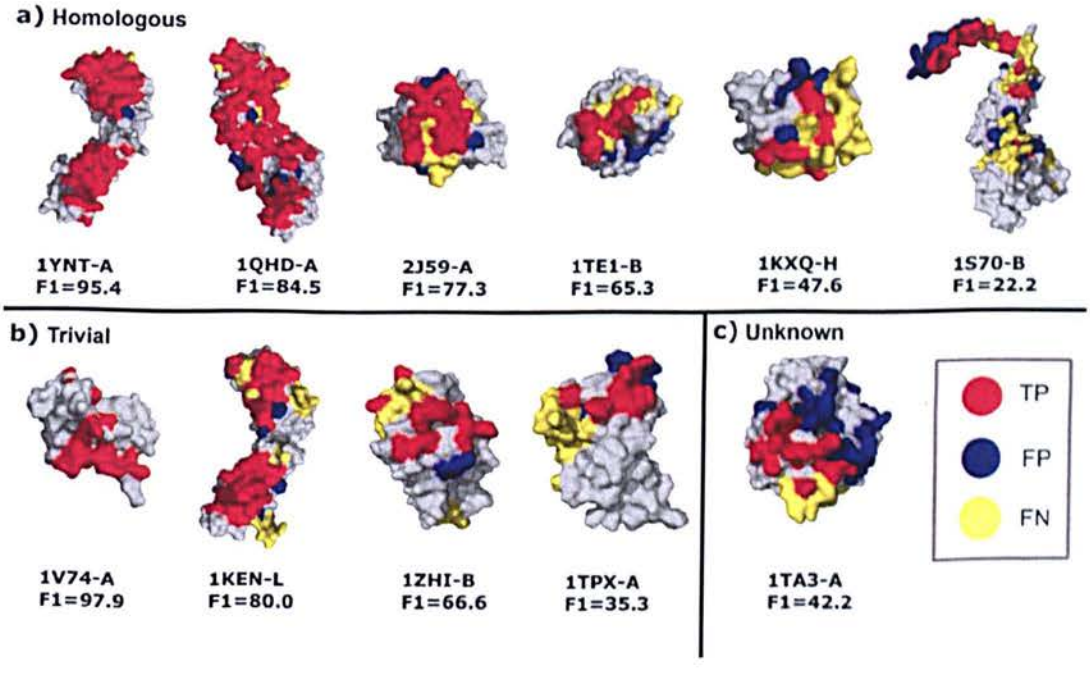


Figure 4.6: Interface predictions generated by the T-PIP framework using either a) Homologous, b) Trivial or c) Unknown. On each PDB target, true interface residues are coloured in red, whereas false positives and false negatives are shown in blue and yellow respectively. Corresponding F1 scores are also provided.

In Figure 4.7, we provide the receiver operating characteristic (ROC) curves of T-PIP interface predictions for the six homologous targets represented in Figure 4.6.A. To draw these curves all predicted interface residues are ranked based on their T-PIP

score where higher values represent higher possibility of a residue to be part of the interface. Then, only the top scored T-PIP residue is selected as prediction and the FP and TP values according to GT are calculated. This will be the first point on the ROC curve. Then, we select the top two scored T-PIP residues, calculate TP and FP and plot the point on the curve. This process continues until the last step which contains all the protein's residues. Based on results in Figure 4.7, first, curves are in agreements with model ranking based on F1 score. Second, accuracy regarding the number of estimated residues in the interface is highly correlated with the AUC. Finally, actual numbers of residues tend to be close to the curve's optimal cut-off point (Schisterman et al. 2005). This point is located on the ROC curve where the distance is the largest to the random diagonal. This suggests there is scope for improving the estimation of expected interacting residues.

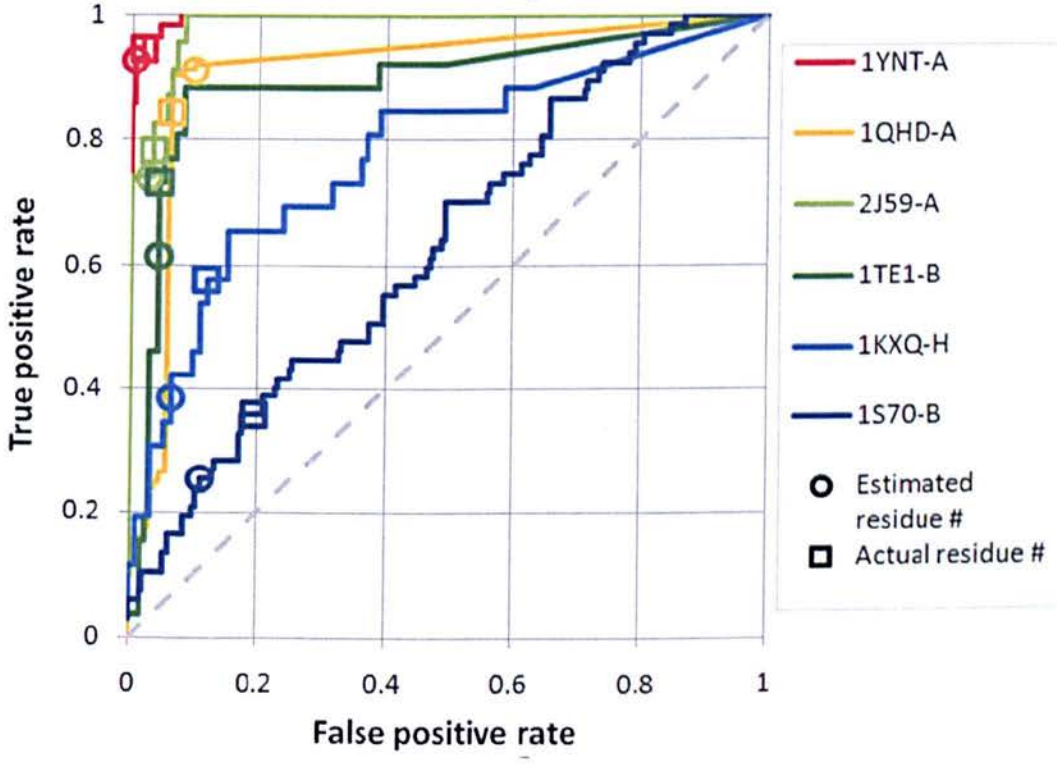


Figure 4.7: Receiver operating characteristic of T-PIP interface predictions for the six homologous targets of Figure 4.6.A. The dotted line shows the curve which would have been produced by a random predictor. Each point on the curves represents the number of FP and TP in comparison to the GT based on the top n scores provided by T-PIP. For each protein, the actual number and the number of interface residues predicted by T-PIP are shown by squares and circles, respectively.

In order to validate the formulation of the weights used by T-PIP, interface predictions for Ds24unbound were also estimated by setting those weights to 1. As shown in Table 4.2, results confirm the added value provided by our weighted schema. While usage of the query weight, x_k, only provides modest improvements when compared to a non weighted approach, the proposed ligand weight, y_k, offers a more significant increase of performance. Finally, when both weights are combined, most performance indicators improve further. These results confirm that taking into account both homology of QP and ligands leads to better interface predictions.

Table 4.2: Validation of weights used by Homologous module of T-PIP on DS24unbound.

Homologous (*DS24unbound*)							
Query weight	**Ligand weight**	**Precision**	**Recall**	**F1**	**Accuracy**	**MCC**	**AUC**
1	1	40.6	32.5	37.0	81.6	25.5	70.3
x_k	1	40.8	32.7	38.7	82.4	26.2	70.4
1	y_k	42.4	34.6	37.3	**82.6**	28.1	**70.9**
x_k	y_k	**43.3**	**35.6**	**41.7**	82.3	**28.8**	70.1

4.4.2.2 Evaluation of Prediction Performance

In the first set of experiments, performance of state of the art methods was performed using the Ds56unbound dataset. In addition, to evaluate interface prediction without knowledge of the QP structure, we also produced results where the QP sequence, instead of its structure, was aligned with an S-MSA of its homologues. Those results are labelled as T-PIPQPseq+S-MSA. Since the IBIS server may provide several interfaces for a given protein, performance is calculated here by selecting the interface with the highest score. Note that two targets could not have their interface predicted using IBIS.

Results presented in Table 4.3 confirm that template based approaches, i.e. IBIS, PrISE, PredUs and T-PIP, outperform feature based ones. Moreover, T-PIP displays either best or second best results competing with PrISE (Jordan et al. 2012) and PredUs (Q. C. Zhang et al. 2010; Zhang et al. 2011), depending on the metric considered. Comparison between standard T-PIP and T-PIPQPseq+S-MSA suggests that availability of QP structure only marginally increases performance and is, therefore, not

required for interface prediction. Nevertheless standard T-PIP is used in all remaining experiments.

Table 4.3: Evaluation of interface prediction methods using the Ds56unbound dataset.

Predictor (*DS56unbound*)	**Precision**	**Recall**	**F1**	**Accuracy**	**MCC**	**AUC**
Promate	28.7	27.3	28.0	76.6	14.0	62.7
PINUP	30.4	30.1	30.2	76.9	16.4	60.0
Cons-PPISP	37.4	34.5	35.9	79.5	23.8	71.2
Meta-PPISP	38.9	24.0	29.7	81.1	20.2	71.5
IBIS	48.2	29.3	34.4	82.5	27.9	-
PrISE	43.7	44.0	43.8	81.2	32.6	75.5
PredUs	43.3	53.6	47.9	73.2	30.4	72.9
T-PIP	**53.8**	**48.5**	**49.6**	**84.0**	**41.1**	**72.9**
T-PIP$_{QPseq+S-MSA}$	**53.4**	**48.1**	**49.2**	**83.9**	**40.7**	**72.4**

Further tests were conducted on the most promising approaches, i.e. IBIS, PrISE, PredUs and T-PIP, using DS120 and DS236 datasets. Note that, for DS120, PredUs and IBIS failed to process 1 (1ZK0-B) and 9 proteins, respectively. For DS236, IBIS did not make any prediction for 32 proteins and T-PIP for 2 proteins (1H20-A and 1QFD-A do not have any structural neighbour). Since PredUs used DS120 chains for training, its performance on an independent dataset is likely to be lower (results on DS236 were not available). When using the PRISE server, query chains were removed.

As shown in Table 4.4, T-PIP also displays the best performance on DS120 and DS236. Interestingly, T-PIP displays similar performance on Ds56unbound, DS120 and DS236 even though DS236 contains more structures from the difficult and medium-difficulty categories. Table 4.5 displays T-PIP results in each of those categories. As expected, better performance is achieved when targets have fewer conformation changes upon binding.

Table 4.4: Comparison of interface predictors' performance on DS120 and DS236.

Predictor & dataset	Precision	Recall	F1	Accuracy	MCC
T-PIP DS120	**52.6**	**56.1**	**52.5**	**85.4**	**45.1**
PredUs DS120	47.3	58.2	48.5	69.4	24.4
PrISE DS120	38.5	48.9	40.9	80.7	31.2
IBIS DS120	42.6	37.4	37.4	83.8	29.9
T-PIP DS236	**53.2**	**55.3**	**52.1**	**85.3**	**44.8**
PrISE DS236	41.2	47.5	41.5	81.0	32.0
IBIS DS236	40.9	36.9	36.2	83.6	28.8

Table 4.5: T-PIP performance on DS120 and DS236 according to DBMK categories.

Predictor & Categories	Precision	Recall	F1	Accuracy	MCC
T-PIP DS120	**52.6**	**56.1**	**52.5**	**85.4**	**45.1**
Rigid Body (86chains)	57.1	61.3	57.3	86.7	50.7
Medium-Difficulty (18 chains)	42.0	50.8	44.5	84.5	35.9
Difficult (16 chains)	42.9	34.0	35.8	79.2	26.2
T-PIP DS236	**53.2**	**55.3**	**52.1**	**85.3**	**44.8**
Rigid Body (156 chains)	56.8	59.4	56.2	86.7	49.5
Medium-Difficulty (44 chains)	45.1	52.2	47.0	85.6	39.3
Difficult (34 chains)	46.9	37.6	38.5	78.4	28.6

Statistical significance of T-PIP's results is first evaluated by providing standard deviations of predictors presented in Table 4.4. Table 2.1. reveals that in general PrISE displays the lowest standard deviations while T-PIP has the highest. Associated P-values are presented in Table 4.7, where figures highlighted in red displays a P-values>0.05). Superiority of T-PIP's performance is shown to be generally statistically significant in comparison to PrISE and IBIS's.

However, comparison with PredUs is less straight forward since, although accuracy and MCC metrics show T-PIP's performance is significantly better, precision, recall and F1 measures do not display any significant difference. Based on the results in table Table 4.4 T-PIP has the best performance except in recall where PredUs displays a higher value. However, comparison of the P-values of T-PIP and PredUs for recall metric in Table 4.7 shows that the difference is not statistically significant. There is a

high chance of 60% that better performance of PredUs compared to T-PIP in terms of recall metric has happened randomly.

As a conclusion, T-PIP performs significantly better than PrISE and IBIS. Regarding PredUs, out of the 5 metrics, no statistically significant difference can be found for 3 of them (Precision, Recall and F1). However, statistically significant better performance in terms of Accuracy and MCC suggests that T-PIP is a slightly better predictor than PredUs.

Table 4.6: Standard deviations of interface predictors' performance on DS120 and DS236. Numbers in the table display the standard deviation for every metric across the specified dataset.

Predictor & dataset	Precision	Recall	F1	Accuracy	MCC
T-PIP DS120	0.32	0.32	0.31	0.13	0.35
PredUs DS120	0.31	0.29	0.26	0.17	0.29
PrISE DS120	0.22	0.21	0.19	0.10	0.21
IBIS DS120	0.36	0.35	0.33	0.13	0.37
T-PIP DS236	0.32	0.32	0.30	0.12	0.34
PrISE DS236	0.23	0.22	0.20	0.11	0.22
IBIS DS236	0.36	0.34	0.32	0.12	0.36

Table 4.7: Significance of T-PIP predictions: comparison with other predictors on DS120 and DS236. Numbers represent P-values. Those highlighted in red are not judged significant (P-values>0.05).

T-PIP, Predictor & dataset	Precision	Recall	F1	Accuracy	MCC
T-PIP, *PredUs* DS120	0.17	0.60	0.27	2.11E-14	1.43E-06
T-PIP, *PrISE* DS120	6.6E-05	0.04	0.0006	0.003	0.0002
T-PIP, *IBIS* DS120	0.02	2.98E-05	0.0004	0.38	0.001
T-PIP, *PrISE* DS236	3.72E-06	0.0023	9.84E-06	6.01E-05	2.46E-06
T-PIP, *IBIS* DS236	0.0001	6.58E-09	8.87E-08	0.13	1.69E-06

Finally, experiments also confirm that homology information benefits interface prediction. As seen in Table 4.8, predictions for the 'trivial' category are a whole class above the others. Moreover, interfaces for the 'homologous' category display higher quality than those for the 'unknown' category: although recall performance remains

stable (the method used for processing the 'unknown' category, PredUs, has a particularly good recall, see Table 4.4), F1 and accuracy measures are better by around 10%, precision by 15% and MCC by a third.

Table 4.8: T-PIP performance on DS120 and DS236 according to target complexity. In the DSx notation, x in the number of chains in the category.

Predictor & Categories	Precision	Recall	F1	Accuracy	MCC
T-PIP DS120	**52.6**	**56.1**	**52.5**	**85.4**	**45.1**
Trivial DS63	64.9	67.5	66.0	89.1	60.5
Homologous DS48	36.5	43.7	38.2	82.9	29.6
Unknown DS9	31.6	41.8	34.5	73.3	19.2
T-PIP DS236	**53.2**	**55.3**	**52.1**	**85.3**	**44.8**
Trivial DS128	65.2	63.8	62.3	88.6	57.0
Homologous DS93	39.7	44.9	40.3	82.5	31.1
Unknown DS13	32.3	46.3	36.6	74.1	22.1

Processing of 'homologous' targets relies on extracting the relevant interacting residues from the interfaces of homologous proteins. In order to evaluate this process, for each protein from the 93 'homologous' targets defined in Table 4.4, the F1 score that would have been obtained using simply the interface of a homologue is computed. This shows how much the interface of a given homologue complex is representative of the solution binding site. In addition, for a given target, the average of its homologues F1 and its T-PIP F1 score is calculated. Figure 4.8 shows the quality of T-PIP predictions in respect to target homologues. Note that query proteins are identified using their association to their target employing the following notation: ABCD:WXYZ-E, where ABCD is the PDB code of the complex target and WXYZ-E is the query protein PDB code-chain, e.g. 1ZM4:1XK9-A.

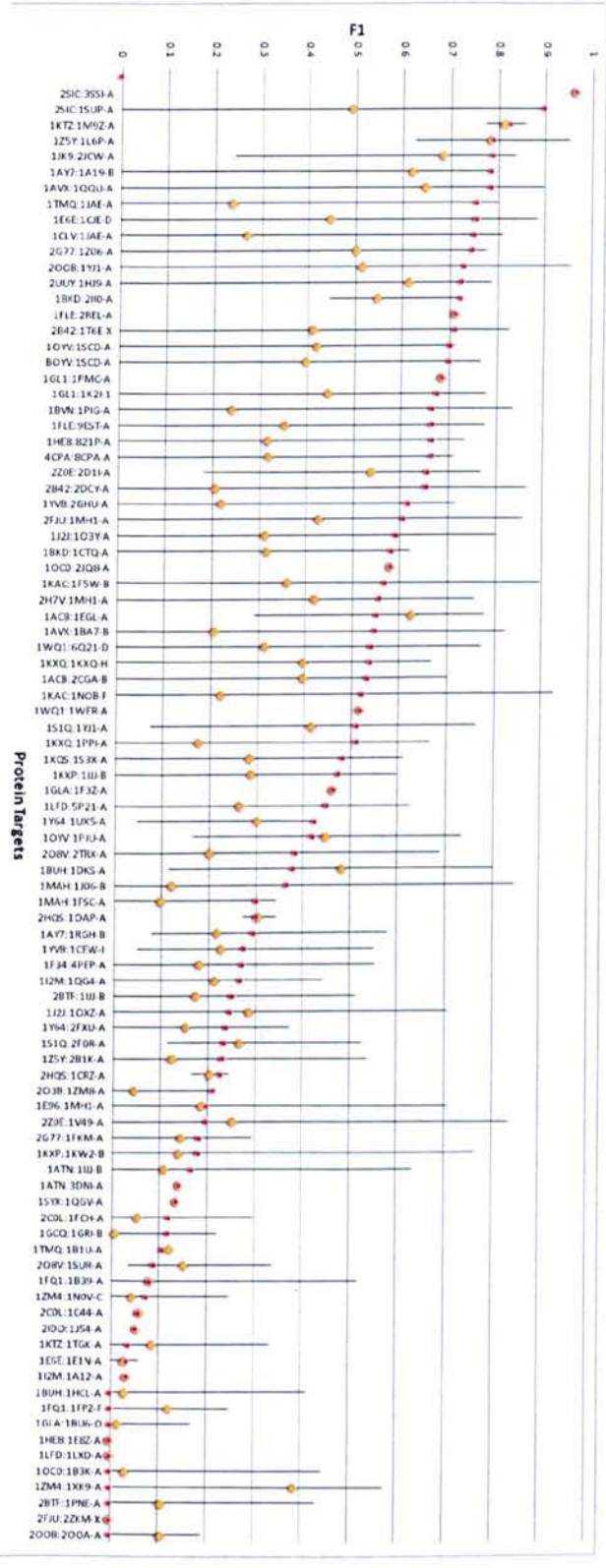


Figure 4.8: Interface F1 of 'homologous' targets in respect to available homologues. Horizontal line connects the maximum and minimum F1 calculated for homologues of a given target. Average homologue F1 and T-PIP F1 are shown by yellow diamonds and red squares, respectively.

In most cases the quality of T-PIP predictions is above average which confirms its ability to extract relevant information from a homologue set. Moreover, the figure also reveals that T-PIP improves on the best homologue interface in four cases: 2SIC:1SUP-A, 1BKD:2II0-A, 1BKD:2II0-A and 2O3B:1ZM8-A.

We analyse some of the targets in more detail. First, we focus on a couple of cases where T-PIP performed extremely badly. Figure 4.9 and Figure 4.10 explain prediction failures for 1ZM4:1XK9-A and 2FJU:2ZKM-X, respectively. Query chains are displayed in grey using solid surface representation. Representatives of interacting partners of proteins homologous to QPs are displayed as cartoons. Red, yellow and dark blue patches on solid surfaces represent correctly, missed and wrongly predicted surface residues, respectively. In the 1ZM4:1XK9-A case, as illustrated in Figure 4.9 with two representatives of the interacting partners of proteins homologous to the QP, the target has two distinct interfaces one of which corresponds to the interface involved in the complex of interest. Unfortunately, T-PIP selected the other one in its prediction.

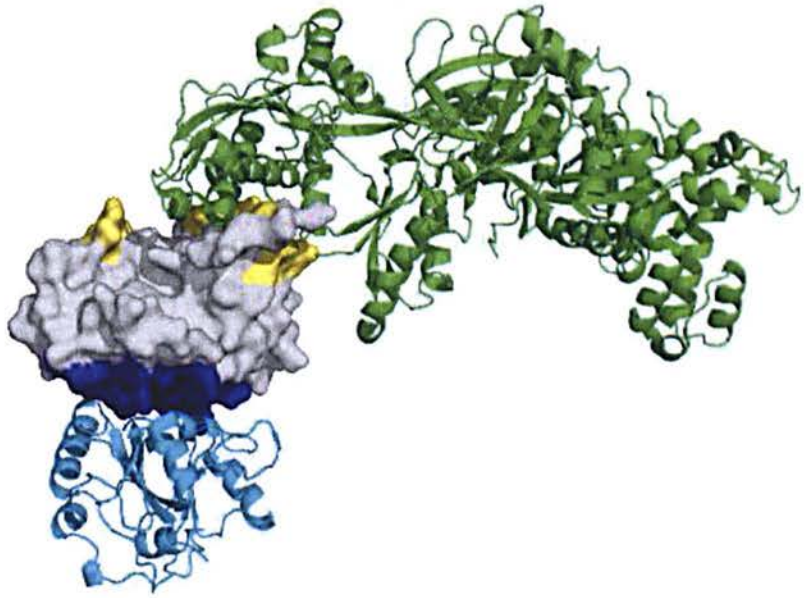


Figure 4.9: Example of failed interfaces predicted by T-PIP for 1ZM4:1XK9-A. Query chain is displayed in grey solid surface representation. Two representatives of interacting partners of the QPs homologues are displayed as cartoons (in cyan and green colours). Yellow and dark blue patches on solid surfaces represent missed (FN) and wrongly (FP) predicted surface residues, respectively.

In the 2FJU:2ZKM-X case, Figure 4.10, the actual interface of interest does not have a single representative among homologous complexes. As a consequence, T-PIP is not able to make any relevant prediction and suggests the most consensual binding site.

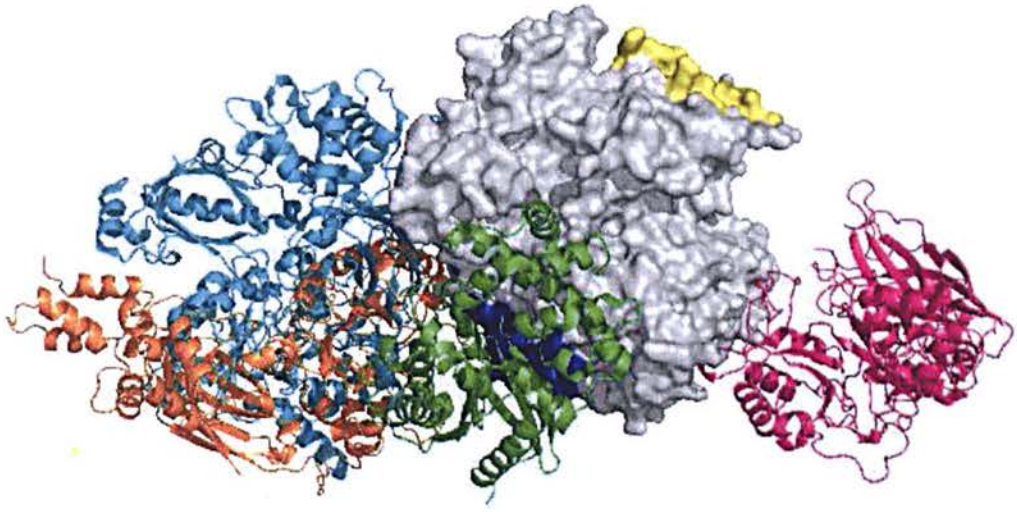


Figure 4.10: Example of failed interfaces predicted by T-PIP for 2FJU:2ZKM-X. Query chain is displayed in grey solid surface representation. Four representatives of interacting partners of the QPs homologues are displayed as cartoons (in cyan, orange, green and pink colours). Yellow and dark blue patches on solid surfaces represent missed (FN) and wrongly (FP) predicted surface residues, respectively.

Figure 4.11, shows a successful case, 1AVX:1BA7-B, where T-PIP extracts binding information from a set of 14 homologous complexes – only 3 representatives are illustrated – and, using appropriate weightings, manage to predict quite accurately the interface of the target.

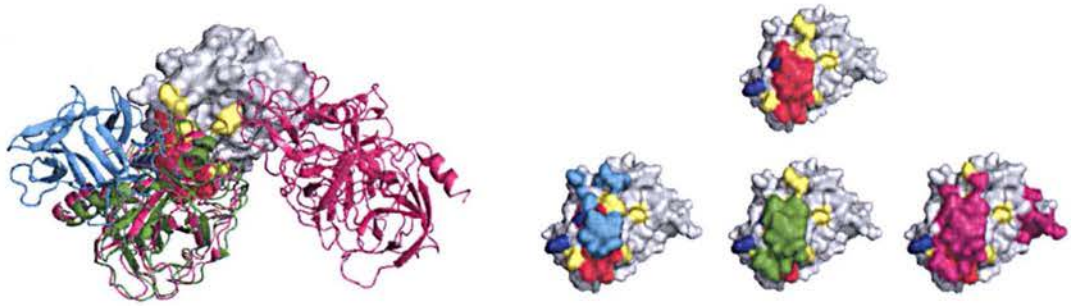


Figure 4.11: Example of successful interfaces predicted by T-PIP for 1AVX:1BA7-B. Query chain is displayed in grey solid surface representation. Three representatives of interacting partners of proteins homologous to QPs are displayed as cartoons (in cyan, green and pink). Red, yellow and dark blue patches on solid surfaces represent correctly (TP), missed (FN) and wrongly (FP) predicted surface residues, respectively. In A) cyan, green and pink patches correspond to the actual binding sites of the interacting partners of 1BA7-B's homologues.

4.4.3 Evaluation of Ranking Docking Conformations

After generation of possible docking conformations, model scoring allows identifying the most plausible conformations. This is evaluated by ranking those models and evaluating how close those rankings fit an 'ideal' ranking or 'ground truth' (GT) according to the configuration of the native complex. Two Capri criteria (Lensink & Wodak 2010), i.e. interface (i-rmsd) and ligand (l-rmsd) rmsds, were used to produce alternative ground truths for the 'homologous', DS93, and 'trivial', DS128, subsets of DBMK4.0 as defined in Table 4.8. Since the interfaces of 'unknown' category are predicted by a third party server, we have not considered them in this experiment. As seen in the 'x-rmsd' column of Table 4.9 and Table 4.10, where ranking generated by one GT criterion is evaluated against the other criterion's ranking, although their $normalized x^2$ are not 0, they are quite low which means they agree quite well with each other. Those values are used as reference scores in further evaluations. Note that comparison to DockRank (Xue et al. n.d.) was not performed since their web server was not available at the time.

Table 4.9: Performance of docking model rankings according to ground truth criterion (DS93 dataset) based on average normalizedx^2. 'x-rmsd' is calculated by evaluating ranking generated by i-rmsd (/l-rmsd) criterion against ranking produced by l-rmsd (/i-rmsd) criterion.

Ground truth criterion	Ranking method applied to DS93								
	x-rmsd	Interfaces +PioDock	T-PioDock	IRAD	ZRANK	SPIDER	SVM	TSVM	MI
i-rmsd	5.2	11.6	30.0	39.5	43.3	49.1	60.7	61.4	67.8
l-rmsd	6.0	12.5	29.7	39.5	44.2	50.6	63.9	64.5	70.9

In a first experiment, T-PioDock was compared to other state-of-the-art methods using the 52 targets of DS93. In addition, we evaluate the PioDock module by applying it on the ground truth interfaces of the target complexes (Interfaces+PioDock) instead of their T-PIP predictions. Table 4.9 displays the average $normalizedx^2$ between the GT and rankings produced by each method. First, although Interfaces+PioDock is not based on interface prediction, but actual interfaces, its $normalizedx^2$ is worse than the reference scores (here, it is the double). This can be explained by the fact that since docking interfaces are treated as two sets of interface residues without any pairwise knowledge (patches), which is the output of current interface predictors, they could perfectly overlap even if the position of a binding partner was rotated around the centre of the patches. Second, the table demonstrates that T-PioDock is superior to all other methods whatever the criterion used to generate the GT rankings. Moreover, relative performances between other methods are in agreement with previously reported results (Pierce & Weng 2007; Kuo et al. 2011; Zhao et al. 2011; Khashan et al. 2012).

Table 4.10: Performance of docking model rankings according to ground truth criterion (DS128 dataset) based on average normalizedx^2. 'x-rmsd' is calculated by evaluating ranking generated by i-rmsd (/l-rmsd) criterion against ranking produced by l-rmsd (/i-rmsd) criterion.

Ground truth criterion	Ranking method applied to DS128		
	x-rmsd	Interfaces +PioDock	T-PioDock
i-rmsd	5.9	13.6	23.3
l-rmsd	6.4	15.0	23.4

In a second experiment, T-PioDock is evaluated on DS128, see Table 4.10. As expected, better interface predictions for this 'trivial' dataset leads to better quality of rankings.

In order to have some insight regarding ranking as a mean of identifying near native configurations, Figure 4.12 displays the i-rmsd of the best produced docking model versus the i-rmsd of the model ranked number one by T-PioDock and Interfaces+PioDock on DS93 and DS128. First, the figure reveals the heterogeneous quality of the best docking model generated for a given target. On this set, the i-rmsd varies between an excellent 0.6 Å to a very poor 17.0 Å with a 4.9 Å average and a large standard deviation of 3.7 Å. Moreover, 13 targets did not have a single model below a 10 Å i-rmsd. Second, this figure shows good correlation between the quality of the best docking model and the ability of both T-PioDock and Interfaces+PioDock to detect that model, correlations of 0.65 and 0.81 respectively. A similar pattern is obtained using l-rmsd as GT.

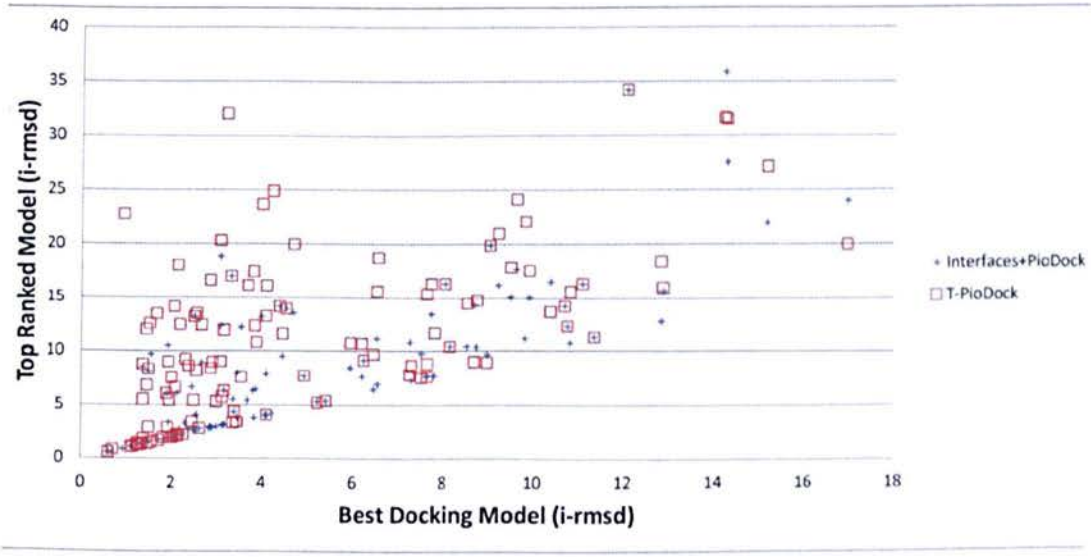


Figure 4.12: Correlation between the best model produced by docking and the best ranked model according to Interfaces+PioDock and T-PioDock. Each square point represents the i-rmsd of the best produced docking model versus the i-rmsd of the model ranked number one by T-PioDock. Each + point displays the i-rmsd of the best produced docking model versus the i-rmsd of the model ranked number one by interface+T-PioDock.

Since the quality of the best docking model is very unequal, it is interesting to quantify how it affects model ranking by T-PioDock. In order to study this, best models from the 'homologous' target set were clustered using K-means clustering into three

groups, i.e. good, average and bad, after normalisation. In Table 4.11, the average $normalizedx^2$ per group shows that T-PioDock produces significant better ranking when better quality model is available. This suggests that progress in docking would lead to better performance by T-PioDock.

Table 4.11: T-PioDock ranking performance (average normalizedx^2) based on the quality of the best model.

Ground truth criterion	Quality of the best model			
	All	Good	Average	Bad
i-rmsd	30.0	23.7	35.7	39.4
l-rmsd	29.7	21.5	37.1	47.6

Finally, since one of the goals of T-PioDock is to recognise near native models among all predictions, we conducted an experiment where the actual target structure was included in the list of possible models. After ranking, for each target, the relative rank of the native pose among all produced models was extracted. The histogram in Figure 4.13 shows that the native pose tends to be present in the top of the ranking lists. For example, 16% of the native models are within the 5 first positions. Actually, the frequency of finding the native model can be interpolated by a decreasing monotonic polynomial.

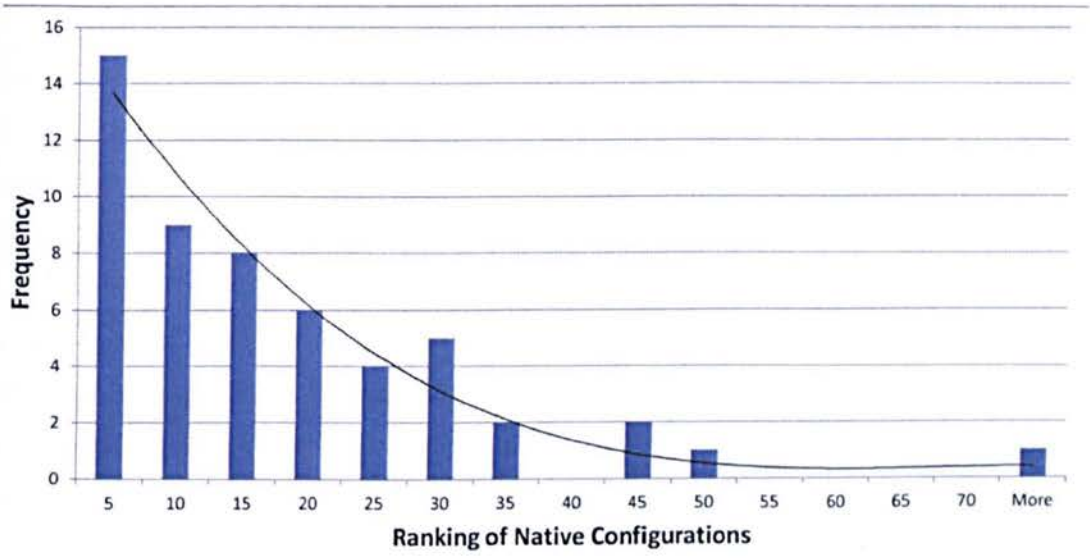


Figure 4.13: Histogram of the relative T-PioDock rank of the native configuration among all docked models. Each blue bar shows how many times T-PioDock ranks the native configuration within the specified interval (two successive values on the x axis). The first bar shows that for 15 targets T-PioDock placed the native configuration in the top 5 of its ranking list, whereas the second bar shows that 9 times it is placed between the 6^{th} and 10^{th} position. Those frequencies can be models by a decreasing monotonic polynomial.

4.4.4T-PioDock Tool

T-PIP and PioDock software are freely available from http://manorey.net/bioinformatics/wepip/. T-PIP program accepts the pdbcode or sequence of the protein of interest and predicts its interface. PioDock, accepts a set of protein- protein docking models and their T-PIP predicted interfaces as input and outputs a ranking list of docked models with their corresponding scores.

4.5 Discussion

This study has confirmed that despite sustained activity in the field, the prediction of a complex 3D structure remains a challenge. First, docking software may not be able to produce any near native conformation among the generated set of putative models. Second, identification of the best conformations remains a difficult task. In this work, we have contributed to this topic by offering a complete pipeline, T-PioDock, for scoring docking models according to the overlap of their components' predicted interfaces.

Experiments evaluating the proposed scoring process, PioDock, on actual interfaces (Interfaces+PioDock system) showed that the treatment of docking interfaces as patches instead of sets of residue interactions affects the quality of the model selection process: two patches can perfectly overlap even if all binary residue interactions are incorrect. Unfortunately, there is currently no promising alternative since current state of the art in interface prediction is not able to work at such a level of details even if this has started to be explored (Khashan et al. 2012; Esmaielbeiki et al. 2012). Although this is an important issue, the study has revealed that the main source of scoring inaccuracy resides with the quality of predicted interfaces, see Table 4.10. Exhaustive evaluation of interface prediction methods demonstrated that T-PIP generates best performance; moreover comprehensive comparison of state of the art methods for ranking docking models supported its integration within the T-PioDock framework since it outperformed all its competitors. However, as Table 4.3 and Table 4.4 showed, performance of interface predictions remains unsatisfactory: most metrics returns values within the 40-60% range, with the notable exception of 'accuracy', ~85%, which benefits from the low ratio between interface and non-interface residues. Although there is no doubt that the sustained growth of the PDB (Berman et al. 2000) will benefit template based methods and T-PIP in particular, Figure 4.8 also highlighted that T-PIP prediction generally does not outperform the best available homologue interfaces. This may be explained by the fact that residues are selected independently from each other, whereas homologues present interfaces where residues belong to a consistent interaction network. While experiments reported in Table 4.3 have demonstrated the superiority of template based methods over feature based ones, one would expect that analysis of local features could complement initial template based prediction by bringing local contextual information. Another approach, which potentially could improve significantly interface prediction, is based on detection of the best homologue. To support this hypothesis, Table 4.12 shows performance where either the best homologue or the average score for all homologues of a given target has been selected. The best homologous protein for each target is selected according to the highest F1 score calculated compared to the ground truth (GT). 'Average homologues' is calculated by averaging all targets average homologue which is simply the average of all the scores among homologues of that target. These results not only confirms that T-

PIP performs significantly better than the 'Average homologues' method, but also suggest that if it was possible to select automatically the best homologous protein, interface prediction would improve significantly.

Table 4.12: Results based on best homologues and average of homologues compared to T-PIP performance.

Homologous DS93	Precision	Recall	F1	Accuracy	MCC
Best Homologous	56.6	62.2	56.8	86.6	49.9
T-PIP	39.7	44.9	40.3	82.5	31.1
Average homologues	29.8	34	29.4	79.6	19.7

4.6 Conclusion

In this chapter, we have presented a novel framework, T-PioDock, for prediction of a complex 3D configuration from the structures of its components. It aims to support the identification of near-native conformations by scoring models produced by any docking software. This is achieved by exploiting predictions of complexes' binding interfaces.

Exhaustive evaluation of interface predictors on standard benchmark datasets has confirmed the superiority of template base approaches and has showed that the T-PIP methodology performs best. Moreover, comparison between T-PioDock and other state-of-the-art scoring methods has revealed that the proposed approach outperforms all its competitors.

Despite our contribution to the field, accurate identification of near-native conformations remains a challenging task. Although availability of 3D complexes will benefit to template based methods such as T-PioDock, we have identified specific limitations which need to be addressed. First, docking software are still not able to produce native like models for every target. Second, current interface predictors do not explicitly refer to binary residue interactions which leaves ambiguity when assessing quality of complex conformations. To address the second limitation, in the next chapter the concept of binding site transitivity is introduced for selection of best homologue

through interface alignment. This will result in an interface prediction which also describes the contact environment.

5 Structural-transitivity of Protein Binding Sites for Protein Interface Prediction

5.1 Introduction

Protein interface prediction is important to gain insight into biological processes and deciphering signalling pathways. In chapter 4, T-PIP interface predictor was introduced. Its evaluation revealed better performance than other state-of-the-art methods. We demonstrated that T-PIP performance could be improved by mapping the best homologue interface onto the query protein (QP). Therefore, in this chapter we introduce ICPIP (Iterative Closest Point Interface Predictor) which uses the binding site transitivity concept to detect the best homologue. This is achieved by structural interface comparison of the first QP homologues and the binding site of the homologues partner of the second QP.

The remainder of this chapter is organised as follows. Section 5.2 introduces interface predictors based on structural similarities. Then in section 5.3, the binding site transitivity concept and the ICPIP method are introduced in details. Then, the results of evaluation are presented in 5.4 and further discussed in 5.5. Finally the chapter is concluded in 5.6.

5.2 Related Work

Proteins interact with each other using their binding sites to perform a specific function. Therefore, knowledge of their interface residues is important for drug design.

As discussed in chapter 2, many methods have been developed for predicting protein interfaces. But with the increase of experimentally determined structures, template-based predictors (see chapter 2) have become the main focus of current prediction methods. Two main trends have emerged in using structure as template : the first one uses structures of protein homologous since they tend to display structurally and physico-chemically similar binding sites (Aloy et al. 2003; Tsai et al. 1996). IBIS one of best performing in this category, structurally aligns homologous complexes with at least 30% sequence similarity to the query protein (QP). A sequence and structure similarity matrix is then used to cluster binding sites and ranks are given to each cluster based on criteria such as average of contact or conservation.

These methods are limited to the availability of homologues; therefore the second trend in template-based predictions uses structural neighbours: proteins which are structurally similar to the QP. Protein structures have shown significant level of interface conservation not only among evolutionary related proteins but also between remote structural neighbours (Q. C. Zhang et al. 2010). Therefore, methods in this category cover more QPs since they are limited to the availability of homologuous structures. PredUs (Q. C. Zhang et al. 2010; Zhang et al. 2011) and PrISE (Jordan et al. 2012) are the top performing methods in this category. PredUs detects structural neighbours by global structural alignments using Ska (Petrey & Honig 2003) program. Comparing to QP structure, only structural neighbours with structural distance (Yang & Honig 2000) (a measure of structural similarity) smaller than a specific threshold are kept which allows capture of both evolutionary related and remote neighbours. Among these proteins the ones involved in PDB or PQS complexes are ranked based on their structural similarity to the QP. For each structural neighbour, they are first aligned on top of the QP and second their interfaces are mapped onto the QP residue. A vector contact map is associated for each alignment, where each cell corresponds to a residue on the QP. Every time an interface residue is mapped on the QP, its dedicated cell in the contact map is increased by the sequence identity between the QP and the structural neighbour. To avoid over counting highly similar structures, protein chains are clustered using a 40% sequence identity. If two structural neighbours belong to the same cluster and their interacting partners also cluster together, then only the structure with is structurally more similar to the QP is kept. Once all structural neighbours are mapped

on the QP, the sum of the individual contact maps are used to build the contact frequencies map which shows the predicted QP interfaces (Q. C. Zhang et al. 2010). PredUs server (Zhang et al. 2011) uses contact frequencies along with SVM to predict if a surface residue belongs to an interface or not. For each surface residue and its 14 spatially nearest surface residue, their contact frequency and solvent accessible surface areas (ASA) are inputs to the SVM. SVM creates a hyperplane to separate interface and non-interface residues where each residue will receive a score for its probability to be an interface. Any QP residue with score above zero is predicted as interface by default. While PredUs, uses global structural similarities, PrISE (Jordan et al. 2012) focuses on local interface similarities. Therefore, while PredUs is limited to the availability of structural neighbours, PrISE can perform predictions for a larger number of QPs. For a given QP, PrISE detects structurally similar interfaces by searching through a repository of structural elements (SE) generated from PDB complexes. A SE is defined by a central surface residue along with its neighbouring surface residues (any residue with an atom distance of $\leq$1.5 Å from central residue atoms). Each SE is represented by four features: (1) name of the central residue of the SE, (2) ASA of the central residue of the SE, (3) ASA of all the residues in the SE and (4) a histogram of atoms in the SE. Therefore, the numbers of SEs of a QP corresponds to the number of its surface residues. For each SE of the QP, similar SEs are detected from the repository. Two SEs are similar if: (1) they are from different proteins, (2) their central residues are the same and (3) the difference between their ASAs and also the ASA of their central residues are smaller than a threshold. The central residue of a QP's SE is predicted as interface if a weighted majority of the central residues of similar SEs are interfaces; otherwise they are labelled as non-interface. The weights are calculated using local (SE similarities) and global (protein surface similarity) structural similarities. The combined global and local similarities have shown better results than the use of each one individually. Finally, in PrISE any QP residue with score above 0.34 is predicted as interface by default. Both PrISE and PredUs have shown comparable performance.

In Chapter 4, we introduced T-PIP which uses homologues structures but unlike IBIS which only focuses on close homologues with sequence identity of 30%, T-PIP also investigates remote homologues. Similarly to PredUs, T-PIP mapping of homologues' interfaces onto the QP is scored by sequence similarity of the QP to its

homologue. However, the T-PIP score also takes into account the nature of the interacting partner. Finally, unlike IBIS, PredUs and PrISE which rely on QP structure, T-PIP can predict interfaces of a QP whose structure is not available. Although prediction methods taking QP sequence as input have been previously developed (see Chapter 2), they do not take advantage of structural template in their prediction and consequently do not perform as well.

T-PIP has shown to perform better than IBIS, PrISE and PredUs (except in recall). These methods along with T-PIP predict a residue as interface or non-interface by calculating a consensus score generated from all available templates. In Chapter 4, we demonstrated that detection of the best homologue and mapping its interface residues on the QP could potentially improve T-PIP performance by ~35% in F1 score. 'Best homologue' is a homologue or structural neighbour whose binding site is similar to the ground truth measured as the highest F1 score.

Protein interfaces have shown to have similar architecture and properties even among structurally and functionally diverse proteins (Tsai et al. 1996; Gao & Skolnick 2010). These similarities have been used by methods to explore interface predictions (e.g. PrISE and PredUs) and protein-protein interaction (PRISM(Ogmen et al. 2005)) prediction. They have shown that structural geometric alignment of interfaces can capture these similarities. Inspired by these methods, in this chapter we propose ICPIP (iterative closest point interface predictor) an interface prediction method based on detection of the best homologue. This is achieved by local interface comparison between a QP homologue and the interacting partner of its QP pair homologues. ICPIP relies on an algorithm aiming at minimising the distance between two clouds of points, i.e. ICP (iterative closest point) (Besl & McKay 1992), to detect potential interface similarities which enables the detection of the 'best homologue'. This computer graphics algorithm has already been applied to protein structure alignment (Ellingson & Zhang 2012; Xu et al. 2007; Bertolazzi et al. 2010) and protein function predictions (Ellingson & Zhang 2011). Xu et al. (Xu et al. 2007) have used ICP to perform global structure alignment among proteins. Comparison to well-known structural alignment methods such as VAST (Gibrat et al. 1996), DALI (Holm & Sander 1993) and CE (Holm & Sander 1993) has shown comparable results in the number of aligned atoms and RMSD obtained. ICP has also been used for protein interface alignment in order to

predict protein functions. Indeed, protein function can be inferred from binding site similarity to a protein with a known function (Ellingson & Zhang 2011). Moreover, a recent ICP-based interface alignment method, TIPSA (Ellingson & Zhang 2012), has shown to perform favourably as other state-of-the-art interface alignment methods such as SiteEngine (Shulman-Peleg et al. 2004), SiteBase (Gold & Jackson 2006) and MultiBind (Shulman-Peleg et al. 2008) which are based on geometric hashing.

In the next section, we will first discuss the ICPIP framework and then we will investigate the use of ICP for binding site structural alignment. We will also compare ICPIP results to T-PIP and discuss how the combination of these methods can significantly improve the prediction.

5.3 Methodology

5.3.1Overview

In nature a protein (called A) can form a complex with another protein (called B), by mimicking the binding site pattern of the known binding partners of protein B (Martin 2010). We call this concept 'binding site transitivity' and it is the main idea behind ICIPIP methodology. Examples of binding site transitivity in nature are shown in Figure 5.1 and Figure 5.2. Figure 5.1 (top figure) displays the final target complex 1FLE:EI which is an enzyme-inhibitor and can be modelled using binding site transitivity concept. Assuming that the complex structure 1FLE:EI is unknown, 1EAI:AC and 2Z7F:EI can be used to model it. Complex 1EAI:AC consists of 1EAI:A (in green), which is homologous to 1FLE:E, and its binding partner IEAI:C (in red). Similarly, complex 2Z7F:EI consists of 2Z7F:I (in cyan), which is homologous to IFLE:I, and its binding partner 2Z7F:E (in pink). Since the binding site of 1EAI:A (green) has a similar structural pattern to the binding site of 2Z7F:E (in pink), structural alignment of those two sites allows to display the configuration of a putative complex formed by interaction between 1EAI:A and 2Z7F:I. As these chains are homologues to the target chains (1FLE:E and 1FLE:I), the structure of 1FLE:EI can be inferred using the configuration of 1EAI:A-2Z7F:I by superimposition of the unbound structures of 1FLE:E and 1FLE:I, i.e. 9EST:A and 2REL:A, respectively (not displayed on the image), on their homologues.

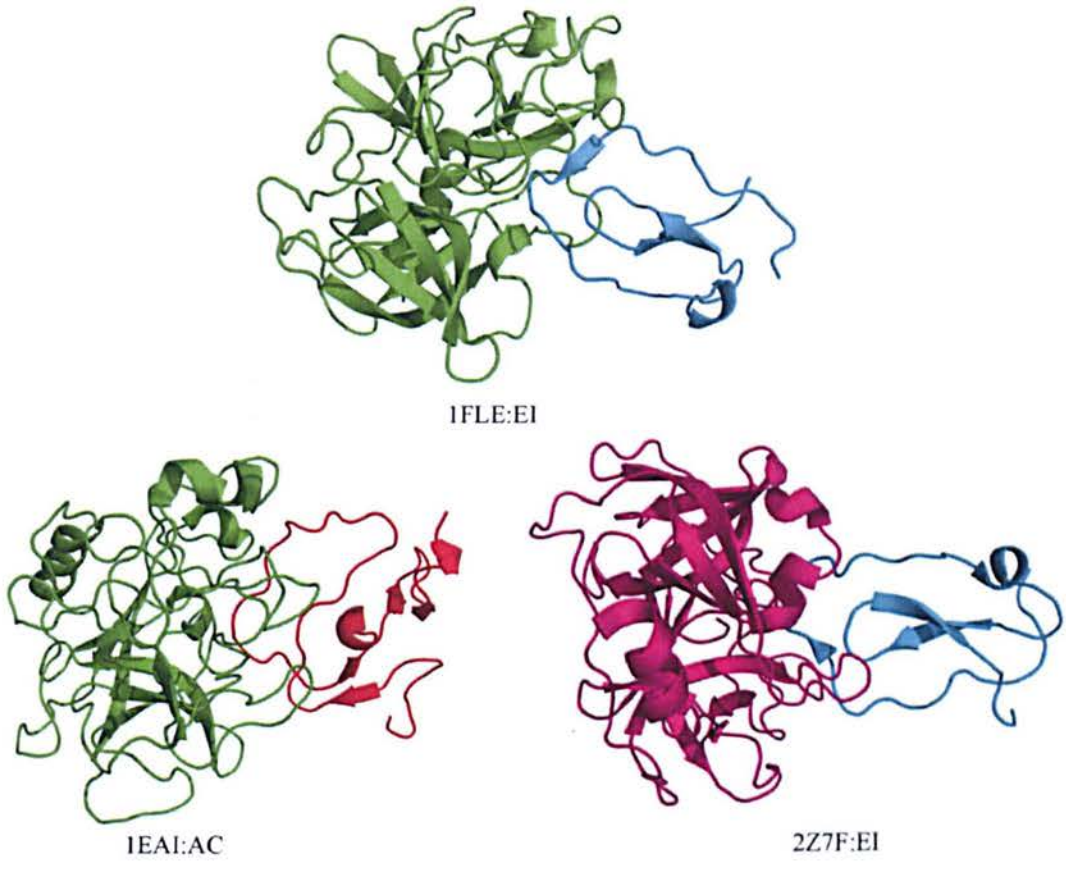


Figure 5.1: Binding site transitivity in nature: enzyme-inhibitor. 1FLE:EI is the final target modelled using binding site structural alignment of 1EAI:AC and 2Z7F:EI. Homologous chains are shown with similar colours. Note that 1EAI:C (in red) and 2Z7F:I (in cyan) have only 20% sequence similarity. Similarly, 1EAI:A (in green) and 2Z7F:E (in pink) have 40% sequence similarity.

Another example is shown in Figure 5.2 with a ras-kinase complex, where the target is 1HE8:AB (top figure). 1HE8:AB has been categorised in the difficult category of DBMK4.0. Binding site of 3IHY:A (in green) in interaction with 3IHY:E (in red) can be structural alignment on 1LFD:C (in pink) binding site to create the target model. The PDB code of the unbound structure of, 1HE8:A and 1HE8:B are 1E82:A and 821P:A,

respectively.

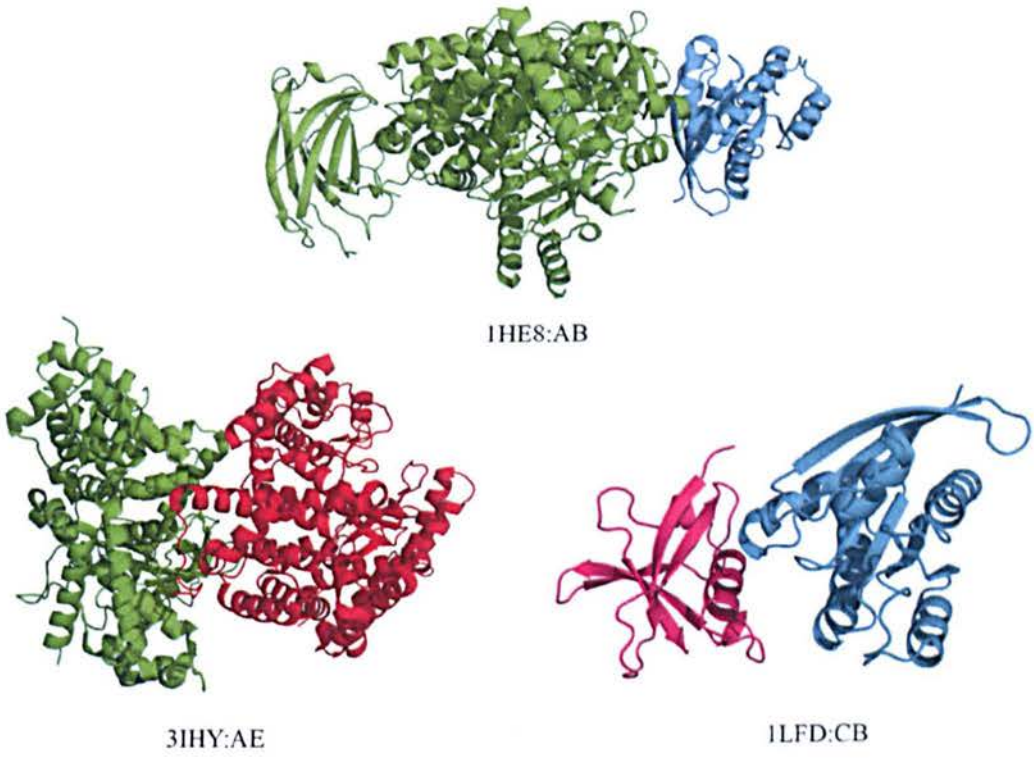


Figure 5.2: Binding site transitivity in nature: Ras-kinase. 1HE8:AB is the final target modelled using binding site structural alignment of 3IHY:AE and 1LFD:CB. Homologous chains are shown with similar colours.

ICIPIP is inspired by this concept of binding site transitivity which is schematically illustrated in Figure 5.3. The assumption is that protein A1 and B1 have been shown to interact but the complex that they form is unknown. The aim is to use the binding site transitivity concept to model the structure of the complex of A1-B1. In order to achieve this, known complexes of A1-ligand and B1-ligand are investigated looking for binding site similarities. Note that here ligand refers to interacting partner rather than a small molecule. Let's assume that A1-F and B1-G are complexes the binding sites of which display structure similarities. In this example, interface of A1 is structurally similar to interface of G (interfaces circled in red: 'a' and 'g' in Figure 5.3). Therefore, if 'a' is superposed on top of 'g' a configuration of the structure of the complex A1-B1 can be proposed. In this example, complex A1-B1 can also be

generated by structurally aligning F interface on B1 interface since they are structurally similar.

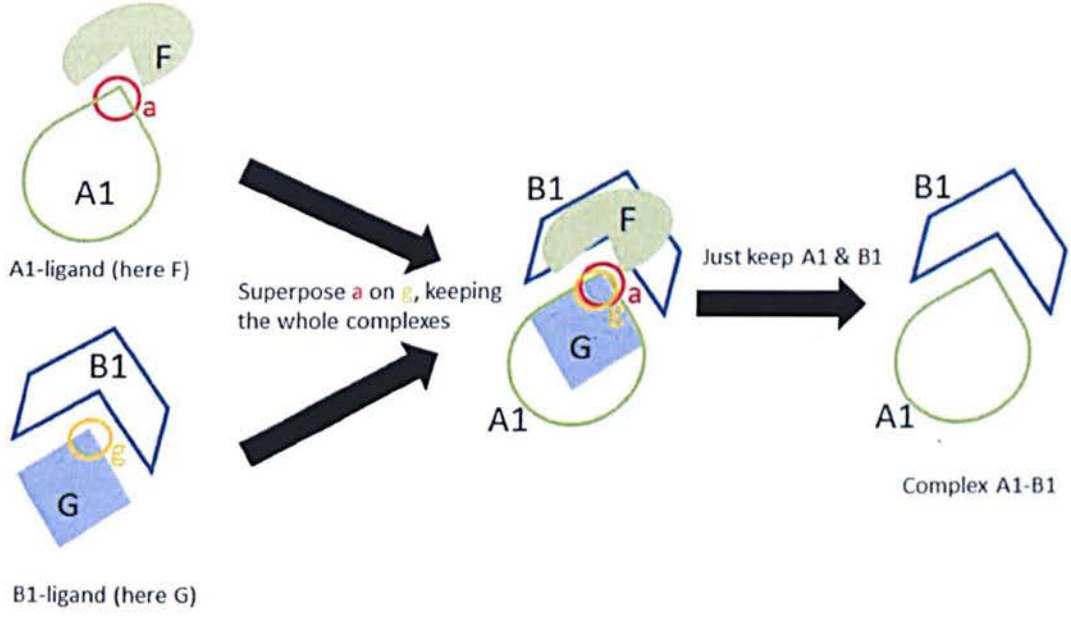


Figure 5.3: Binding site transitivity concept. The aim is to model A1-B1 complex using the binding transitivity concept. Available complexes of A1-ligand (here A1-F) and B1-ligand (here B1-G) are used. . The binding site of A1 and G are shown by red and orange circles and are named 'a' and 'g', respectively. Binding sites 'a' and 'g' are structurally similar so they can easily be superposed on top of each other. By keeping the whole complexes on both sides during the superposition process, the final complex A1-B1 can be deduced.

The example illustrated in Figure 5.3 is the simplest case. In practice A1 and B1 may have several interfaces used to interact with many different chains and also several complexes of A1-ligand and B1-ligand can exist. As a consequence, more elaborated interface comparison is required. In addition, it is not always possible to find complexes involving A1-ligand and B1-ligand. Therefore, homologues complexes of A1 (called Ai) and B1 (called Bj) should also be considered. In such cases where several Ai-ligand and Bj-ligand complexes are available the detection of the most similar binding sites requires more detailed comparison. Therefore, if binding sites can be compared and the most similar pairs can be distinguished, the best homologues of A1 and B1 can be identified and used to model A1-B1. This is the principle followed by ICIPIP.

Therefore, ICIPIP uses the binding site transitivity concept to detect the best homologue which can be used to model the complexes. The steps of finding homologues complexes of the QP are exactly the same as T-PIP. But once homologues

complexes are found, instead of using all complexes, ICIPIP attempts to identify the best homologue which should lead to better interface prediction accuracy as shown in Table 4.12. Binding site transitivity assists the detection of the best homologue.

To explain ICPIP pipeline in details, we refer to a simple schematic example (Figure 5.4). The aim of ICIPIP is to predict the interfaces for query protein A and B. In this example we only discuss the interface prediction for A since the same method can be repeated for B. Also, it should be noted that ICIPIP requires as input two QP chains - binding site transitivity would not make sense if there was a single chain input. In addition homologues complexes should be available for both QP chains.

To predict interfaces of A in Figure 5.4, first, for each pair of query proteins, A and B, homologous complexes are extracted as described in Chapter 4. For simplicity in this example only two homologues are considered for each query protein. Homologues of A and B are denoted respectively as Ai and Bj. The interacting partners of the two homologous proteins of A (i.e. A1, A2) and B (i.e. B1, B2) are represented in solid coloured shapes. Second, based on binding site transitivity concept if the interfaces of A1 and A2 (here a1, a2 and a3) resemble to one of the interfaces of the binding partners of either B1 or B2 (here pb1, pb2 and pb3), then that interface can be used to model Ai-Bi complex which will eventually result in a prediction of the A-B complex. In addition, the resulted Ai-Bi allows identifying the homologue the binding site of which has the most similarity to the final target.

To detect if two binding sites are similar each individual interface of Ai (a1, a2 and a3) are structurally aligned on each interface of the interacting partners of Bj (pb1, pb2 and pb3) (not shown in the figure). In this example this will result in 9 different pair-wise interface alignments. Note that Ai interfaces are first mapped on A and then the structural alignment takes place. Details about the structure alignment are provided in section 5.3.2.

Third, to detect the best alignment of interfaces, each aligned pair of interfaces is scored and a ranking list of all pairwise interfaces is generated (not shown in the figure). Details of the scoring and ranking are provided in section 5.3.3. The best interface alignment will allow identifying the best homologue which will then be used to predict interface residues of A. In this example, the superposing of a1 and pb1 has the best score since they have the most structure similarity (as shown qualitatively on Figure

5.4). Therefore, the homologue for A (here A1) belonging to the first ranked pairwise alignment (here a1 and pb1) is predicted as the best homologue. Finally, the interface of the best predicted homologue is mapped onto the QP, i.e. A and is used as predicted interfaces of ICIPIP. To detect the best homologue for B this process is repeated by aligning interfaces of Bj with those of the interacting partners of Ai.

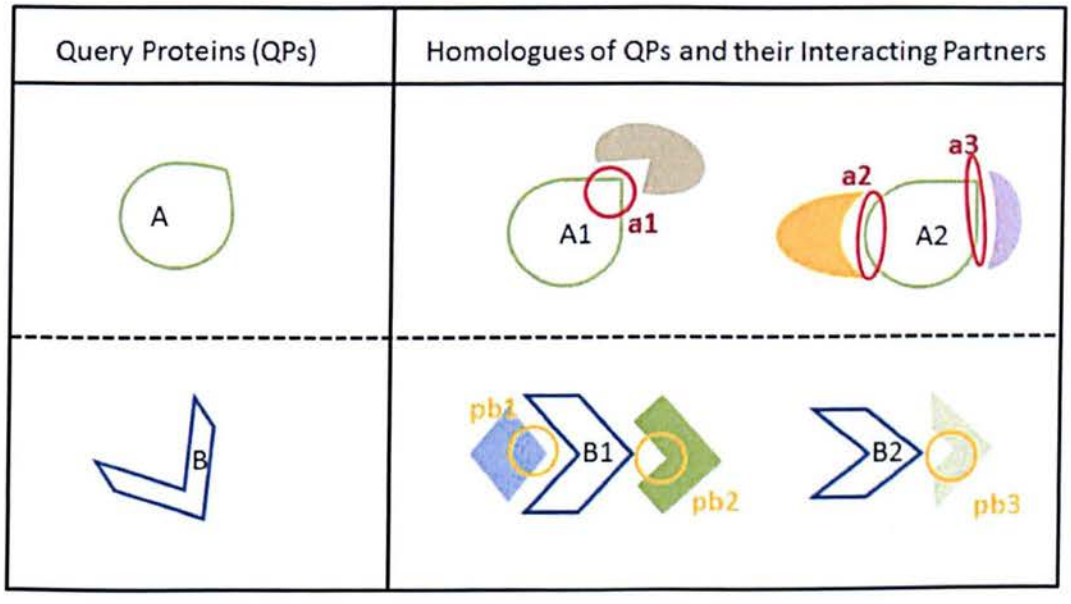


Figure 5.4: QP pair homologous complexes and their binding sites. A and B represent the QP pair. Ai and Bj (i,j=1,2) are homologues of A and B, respectively. Their interacting partners are displayed in solid coloured shapes. Interfaces on Ai in interaction with its binding partner, denoted as ai (i=1,2), are shown by a red circle. Interfaces on the binding partners of Bi, denoted as pbj (j=1,2,3)are shown by an orange circle. Here, the aim is to predict the interfaces of A. To achieve this ICIPIP structurally compares all interfaces of ai to all interfaces of pbj. The best alignment will point at the best homologue of A. The interfaces of the best homologue are mapped on A and are considered as the predicted interface for A.

To summarise, Figure 5.5 displays the ICPIP pipeline for predicting the interface of QP1. In order, to predict QP2 interfaces the same pipeline is repeated but this time the order of the input QPs is changed. A consequence of such a scheme is that ICPIP predictions are limited to QP pairs where homologous complexes are found for both interaction partners.

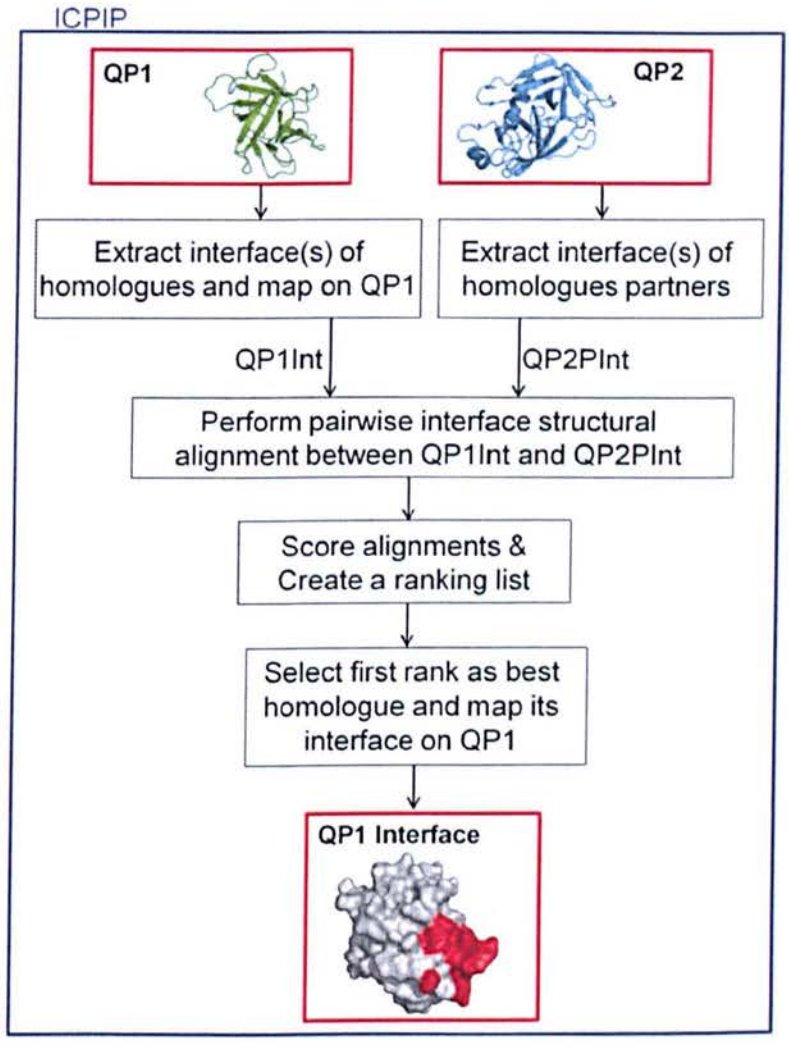


Figure 5.5: ICPIP Pipeline for Prediction of QP1 Interface. First, homologues complexes of QP1 (in green) is detected and their interfaces are extracted (QP1Int is the collection of these interfaces). Concurrently, homologues complexes of QP2 (in cyan) are detected and the interfaces of the binding partners of these homologues are extracted (QP2PInt is the collection of these interfaces). Then, each interface in QP1Int is structurally aligned on all interfaces in QP2Pint resulting in QP1Int*QP2Pint pairwise alignments. These pairwise alignments are then scored and a ranking list is created (called ICPIP ranking list). The first pairwise alignment in the list is selected as the best alignment (here QP1IntBest- QP2PIntBest). The homologue to which QP1IntBest is associated is considered as the best homologue (called ICPIP hit) and its interface (which is QP1IntBest) is mapped on QP (shown as red patches).

In section 5.3.2 we first discuss how the interface alignment is performed. Then in section 5.3.3, we detail the scoring method used for evaluating the alignments and selecting the best possible homologue. In this chapter 'ICPIP hit' refers to the best predicted homologue which has the first rank in ICPIP ranking list of structurally aligned interfaces. Also, 'best homologue' refers to the homologue whose F1 score

according to the ground truth (GT) is the highest among all the homologues of a specific target.

5.3.2Interface Structural Alignment

5.3.2.1 ICP Concept

To perform interface structural alignment we have used ICP (Besl & McKay 1992). ICP is a technique to minimise the distance between two sets of unlabelled point clouds through two steps of association and registration. For each point in the first point cloud, ICP looks for the closest point on the second set of point cloud (association). Once the points on both clouds are associated to one another, a rotation matrix and a translation vector are calculated based on a mean square error function to optimally align the two sets of points (registration). These two steps are repeated till changes in the root mean square error is below a certain threshold or pre-defined number of iteration is reached. Since the root mean square error corresponds to RMSD, in this chapter we simply call it RMSD. Figure 5.6, displays the iterative concept behind ICP for two clouds of points, M and S, coloured in blue and red. The green lines represent the calculated association from the previous step (except in step0 which is the initial association), while the pink lines show the new associations. 'Rot' and 'Tran' are the optimal rotation and translation matrices for each step which allow the subject (Si) to move closer - as defined by the associations - to the model (Mi).

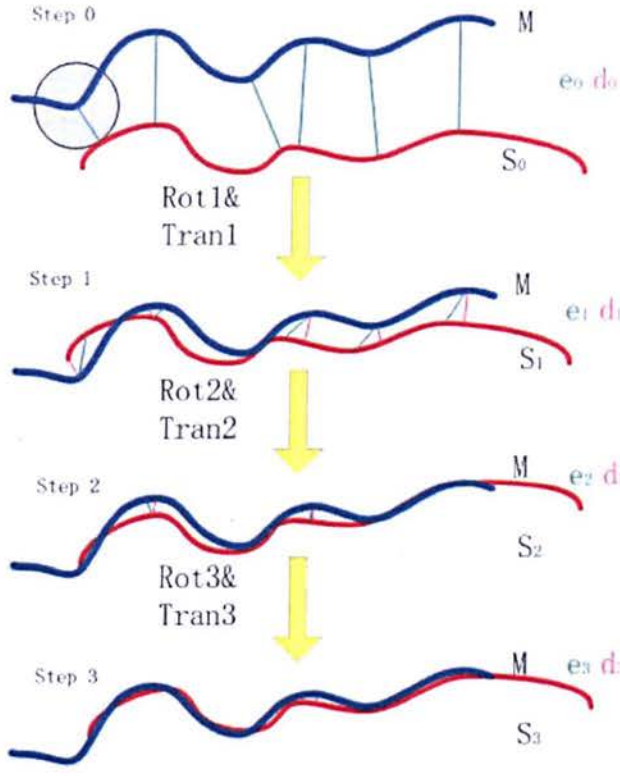


Figure 5.6: Iterative concept of ICP algorithm. Here, a 4-step iteration is shown. Two clouds of points, M (the model) and S (the subject), are coloured in blue and red. Rot and Tran are the optimal rotation and translation transformations in each step. For each point in Si, ICP associates it to the closest point on M. In step0 the initial association between point clouds are shown by green lines. This allows the calculation of Rot1&Trans1 which results in S moving closer to M (shown as S1 in step1). In Step1,2 and 3 pink lines show the new associations while the green line is the association from the previous step. In step3, RMSD reaches its minimum and ICP stops. Taken from(Wang 2012)

5.3.2.2 Improvements to ICP

To target the limitations of ICP (discussed below) improvements have been made to the standard ICP algorithm. First, an initialisation stage is considered before performing the standard ICP. Second, the association step of the ICP was edited to only perform unique and partial correspondences between the point clouds and also to detect outliers. These two improvements are discussed in details in section 5.3.2.2.1 and 5.3.2.2.2, respectively.

5.3.2.2.1 Initialisation Stage

ICP is a deterministic method whose final result depends on the initial alignment and can get stuck in a local minimum (Ellingson & Zhang 2012). To address this issue, it was proposed to run the algorithm many times each time changing its initialisation: that is the centre of mass of the two objects of interest are superposed and a large number of initial rotation states are generated (Besl & McKay 1992). Since this strategy has proved efficient, we decided to adopt it by introducing an "alignment initialisation" stage before performing ICP. Figure 5.7, displays the complete pipeline used in ICIPIP for protein interface alignment. The input to this pipeline is a pair of protein interfaces, IA and IB, which are going to be structurally aligned. The "alignment initialisation" stage is shown as a red box in Figure 5.7 and is as below:

For each pair of protein interfaces, IA and IB, we first create an interface vector. This vector is a directed vector from the centre of mass of each protein to the centre of mass of the interface calculated using the c-alpha atoms only. The two directed protein interface vectors, VA and VB, are aligned which allows an initial overlap of IA and IB interfaces. Since generating a large number of rotation states in 3D space is complex and time consuming, we have used Principal Component Analysis (PCA) which identifies the significant variations in the dataset for generating a smaller number of initial samples. Thus, the initial aligned interfaces using the interface vectors are projected on a 2D plane perpendicular to VA (IA' and IB' represents projected IA and IB). Using PCA for both IA' and IB' the first and second Principle Component vectors (PCa1 and PCa2 for IA', PCb1 and PCb2 for IB') are calculated. Initially IB is rotated so that PCa1 is aligned with PCb1 and consequently PCa2 and PCb2 are also aligned. Let's call this state of IA and IB as *initial alignment*. Once an *initial alignment* is created (outputted from the top red box in Figure 5.7) then ICP is used to optimally align IA and IB (green box in Figure 5.7). But as suggested by Besl & McKay (Besl & McKay 1992) several *initial alignments* are required to avoid getting stuck in local minima. Therefore, to generate the next new initial alignment state, the previous *initial alignment* IB is rotated by 10° around VA (small red box on left side of Figure 5.7). This process is repeated $n = 36$ times (covering 360° around the VA axis). The RMSDs from the 36 different runs of ICP are compared and the lowest one is selected as the final alignment of the interfaces (Final step in Figure 5.7). As discussed above the

association stage of ICP (green box in Figure 5.7) has also been improved which is discussed below.

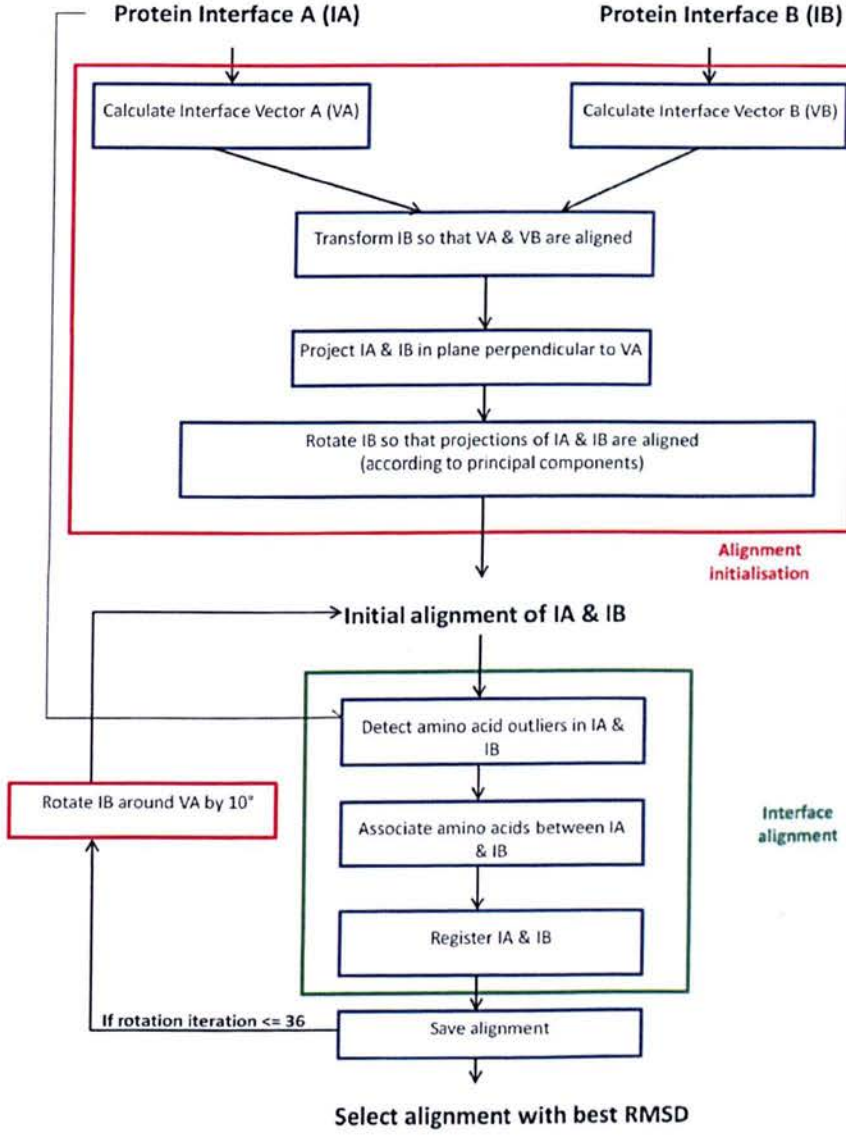


Figure 5.7: Protein Interface Alignment Pipeline. The inputs to this pipeline are two point clouds of protein interfaces (IA and IB) which need to be structurally aligned. Alignment initialisation (top red box): interface vectors (VA and VB are calculated for IA and IB, respectively). Then VB is aligned on VA and consequently IB is overlapped with IA. IA and IB are projected on a plane perpendicular to VA and PCAs are calculated. IB is rotated so that PCAs of IA and IB are aligned. The output of this staged in called *initial alignment*. To create several *initial alignments* (left red box) IB is rotated by 10° around VA and this process is repeated 36 times. Interface Alignment (green box): ICP is used to perform interface alignment using *initial alignment* of IA and IB. but prior this, outliers are detected and removed from the alignment stage. Point association is performed using the Hungarian algorithm. Then registration is performed. Once alignment is performed the RMSD is stored. Finally, following 36 *initial alignments* of IA and IB and 36 runs of interface alignment, the best alignment with the lowest RMSD among the 36 is kept as the final result.

5.3.2.2.2 Unique Association and Outlier Detection

For each *initial alignment*, the interfaces, IA and IB are aligned using ICP approach (The green box in Figure 5.7). However, the association step of the standard ICP does not provide a unique correspondence for each 3D point, i.e. one point on IA can be matched to several points on IB. In addition, ICP is not able to perform partial alignment. This means that if IA and IB should only partially overlap, ICP still attempts to associate all points of IA with IB, even 'outliers'. Those two limitations of ICP affect negatively alignment quality.

To deal with these issues we have used the Hungarian algorithm (Kuhn 1955; Munkres 1957) which was initially developed to solve unique assignment problem. The Hungarian algorithm finds the minimum cost of assigning jobs to a set of workers where there are as many jobs as workers. More formally, considering a matrix the rows of which correspond to workers and column to jobs, each cell (i.j) represents the cost of the i-th worker doing the j-th job. The Hungarian algorithm associates all workers to all jobs minimising the total cost. It also enforces that only one unique job is associated to each worker. Therefore, all jobs will be performed by different workers. In ICP, the rows and columns are each point in the point cloud of interface IA and IB and the costs correspond to the distances between each two points of the point cloud of IA and IB. A limitation of the standard Hungarian algorithm is that it does not support partial assignments: registration requires two point sets of the same size. To overcome this limitation, the Hungarian algorithm was extended for 'rectangular problems', i.e. where the input is a mxn matrix, m≠n. It finds k independent elements ($k = \min(n, m)$) that corresponds to the minimum cost (Bourgeois & Lassalle 1971). Figure 5.8.A gives a schematic view of how the Hungarian algorithm deals with the assignment problem. The numbers of blue and green points are identical. On the right side of Figure 5.8.A, standard ICP may associate one data point to several data points (red arrows) while using the Hungarian algorithm (left side) only unique associations are produced. Figure 5.8.B left image shows that the rectangular Hungarian algorithm can be applied to associate subset of the data points while standard ICP on the same data (right image) associates all points (red arrows). . However, in cases such as the one showed in Figure 5.8.A, where the tails of the cyan and green data points are considered as outliers, the rectangular Hungarian still tries to associate them because the two sets have a similar

number of points. Therefore, those outliers need to be detected before performing the Hungarian algorithm. A solution is to prevent their assignment in the *mxn* matrix, by assigning them a prohibitive cost. In that case, the Hungarian algorithm will leave them out because of their high impact on the overall cost function. An example of this case is shown in Figure 5.8.C. where the detected outlier points are not considered during the association stage. So far ICP with Hungarian algorithm have been only used for unique association of four atoms (Ellingson & Zhang 2012). Moreover, it does not deal with either rectangular Hungarian or outlier removal.

To benefit from the rectangular Hungarian algorithm, we created a *mxn* matrix where m and n corresponds to the number of points in IA and IB, respectively. The cells of this matrix represent the distances between each point pair from IA and IB.

In order to detect outliers before starting the Hungarian algorithm, for each point on IA the nearest neighbour distance to IB is calculated. Then, the standard deviation of these distances is calculated and any point with a distance above two standard deviations is considered as outliers. To avoid their assignment in the Hungarian algorithm they are assigned a distance of Infinity in the *mxn* matrix. These outliers are not removed from the whole process. They are estimated at each iteration of ICP, where they are not considered in the calculation of the rotation and translation matrix.

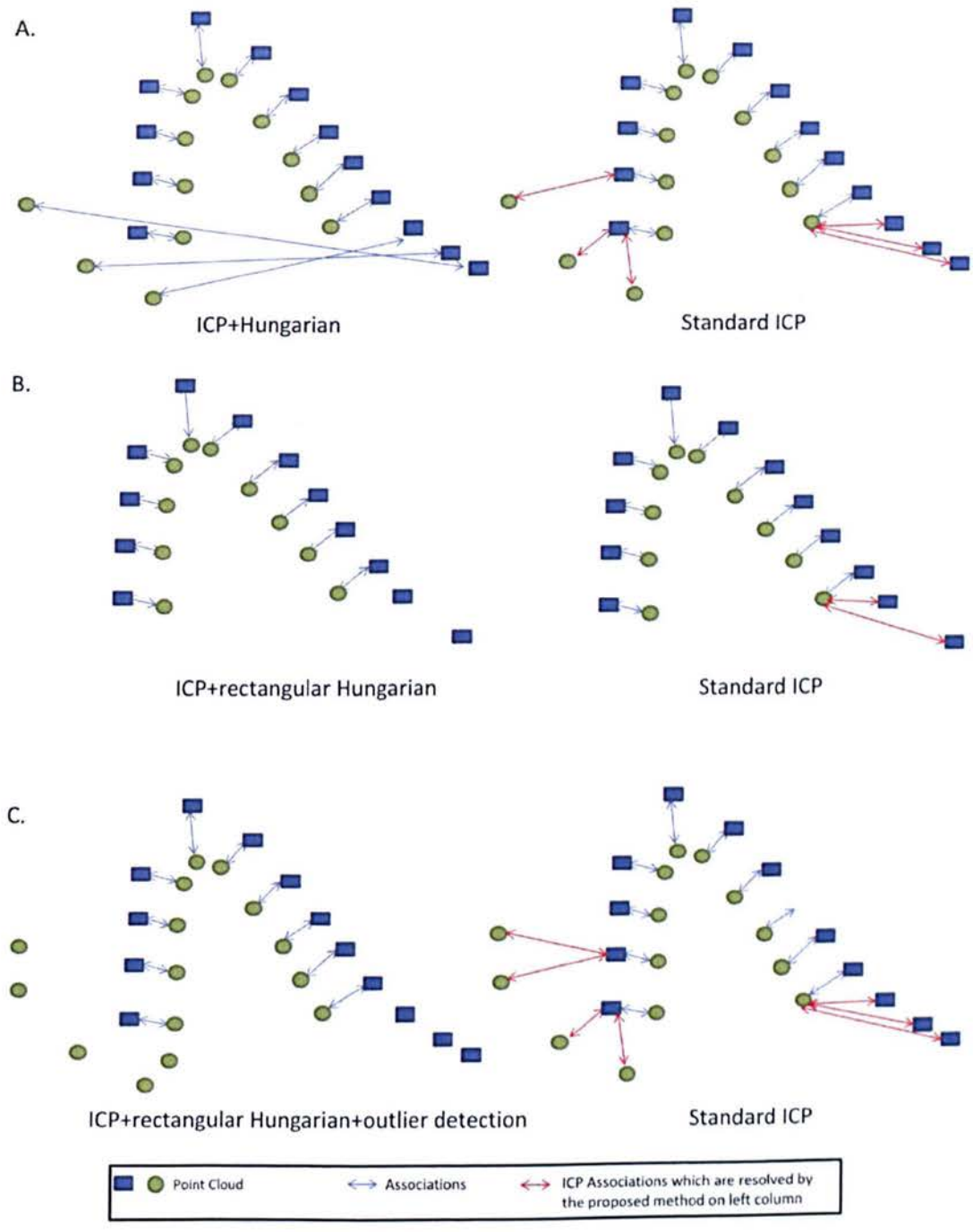


Figure 5.8: Hungarian Algorithm for Registration. Two point clouds points are shown in green circles and blue rectangles. A. The number of points is identical in the two point clouds. On the left, unique associations using ICP along with the Hungarian (ICP+Hungrian) is shown. On the right, the same association is generated by standard ICP where red arrows show that one point can be associated to several points. B. On the left, ICP+rectangular Hungarian allows unique association between point clouds with unequal numbers of points. On the right, standard ICP associates all the points. C. On the left, outliers in the example in A are detected and are left out from the association process, while on the right, standard ICP associates all data points even outliers (red arrows).

For ICP we have used an available Matlab package (Kjer & Wilm 2010) which was further edited to include the rectangular Hungarian algorithm (Cao 2008) . In this work ICP iteration was set to 100.

5.3.3 Detection of the Best Homologue

As discussed previously, to detect which A homologue is more likely to correspond to the real interface, Ai interfaces (here a1,a2 and a3) should be structurally compared against the interacting partners of Bj (here pb1, pb2 and pb3). In the example of Figure 5.4, there would be 9 combinations. In the previous section, we discussed how to generate the best 3D alignment of two interfaces with the lowest RMSD. While alignment comparisons between different configurations involving the same interfaces is relatively straight forward, comparisons between pairs of totally different aligned structures are more complex. For example in Figure 5.4, we need to use some metric allowing to express that the aligned pair of a1 and pb1 is a *better* alignment than a1 and pb2. Unfortunately, RMSD alone is not appropriate to judge the quality of alignments involving different sets of points. For example, a RMSD of zero can be achieved by aligning only two points. Balance between the number of aligned residues and RMSD can be express by a score that considers their ratio. Such score (Q) (Equation 5.8) has been used in previous studies (Krissinel 2012; Krissinel & Henrick 2004) for calculating the quality of protein structure alignments.

$$Q = \frac{N_{aligned}^2}{(1 + RMSD^2)N_A N_B} \qquad \textbf{Equation 5.8}$$

Where $N_{aligned}$, N_A and N_B represent number of aligned interfaces between interface A and B, number of points in interface A and number of points in interface B, respectively. But a drawback of this score is that it does not take into account the size of the interfaces. For example, aligning two interfaces of size 10 and two interfaces of size 20 with similar RMSD will give the same Q score. While in our algorithm, conservation between larger interfaces accounts for more binding site similarity. In addition, in the same example if RMSDs were different then again RMSD would have been the criteria for comparing the two sets.

Therefore, to overcome these limitations, we propose an adjusted score in Equation 5.9.

$$Q' = \frac{N_{aligned}^2}{(1 + RMSD)(N_A \cup N_B)} \qquad \textbf{Equation 5.9}$$

where $N_A \cup N_B$ is equal to $N_A + N_B - N_{aligned}$. To reduce the effect of RMSD and giving more weight to $N_{aligned}$, RMSD to the power of 1 is used. Going back to the example of two interfaces of size 10 and two interfaces of size 20, with similar RMSDs the interface with size 20 will be the winner. This is in agreement with the fact that larger binding sites are more likely to be biological (Tyagi et al. 2012). Also, when the RMSDs are different; removing the power 2 of RMSD balances the effect of RMSD based on the size of the dataset. Once the Q' score of all pairs of aligned interfaces is calculated, the highest score represents the best alignment. In the concept of binding site transitivity, the interface of homologue Ai (a1 in Figure 5.4) should correspond to Bi partner interface (pdb1 in Figure 5.4) in terms of shape and size to allow an interaction between Ai and Bi. In other words, two proteins interacting with the same protein at the same location, should show some degree of similarity in their interfaces. Therefore, before calculating the Q' score, the pairwise alignments which have shown poor structural similarity (RMSD> 4Å) or dissimilar interface sizes ($[\max(N_A, N_B)/ \min(N_A, N_B)] > 1.5$) are rejected. Eventually, based on Q' scores, ICPIP provides a *ranking list* of all the possible pairwise alignments, where the first rank corresponds to the ICPIP prediction of the best homologue called ICPIP hit.

5.4 Evaluation

5.4.1 Evaluation of Best Homologue Detection

ICP was used to detect the best homologue on DS93 dataset from Chapter 4. This dataset was categorised as 'homologous' by T-PIP and results showed that selection of best homologue in this category can improve the interface prediction. Since ICPIP is based on analysing the homologues of two query chains, out of 93 chains in DS93 only 80 chains (from 40 different targets) could be processed. This new subset is

called DS80 and is used in this study to evaluate prediction performance. DS80 was processed by ICPIP and for all the 80 targets their corresponding ICPIP ranking list was generated. The first experiment evaluates how well ICPIP can detect the best homologues. This is performed by using the relative rank of the best homologue of each target as extracted from its ICIPIP ranking list. As mentioned above best homologues can be detected by its best F1 score in comparison to the ground truth. Note that in order to obtain a meaningful comparison only target with more than 12 homologues were kept. The histogram in Figure 5.9 shows that the best homologue tends to be present in the top of the ranking lists. For example, for 21 targets the best homologue is positioned in the top 20% of ICIPIP ranking list. Actually, the frequency of finding the best homologue can be interpolated by a decreasing monotonic polynomial.

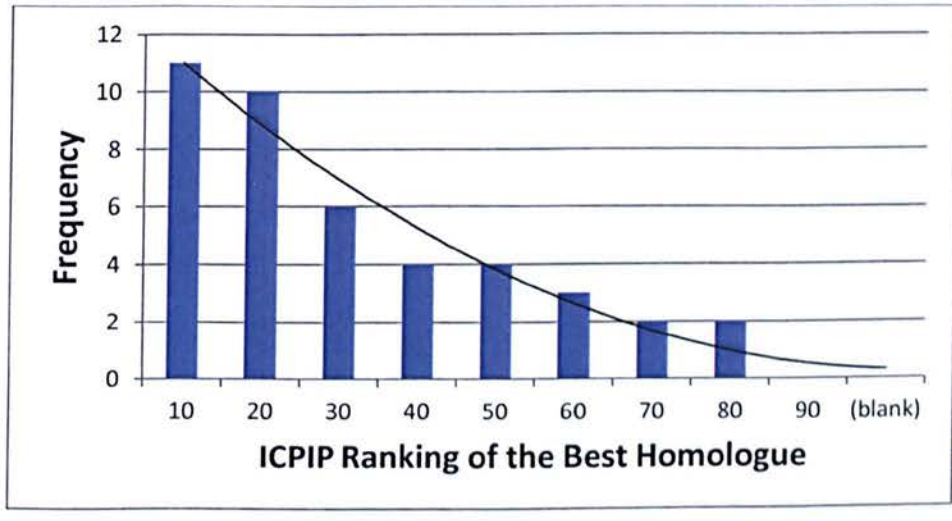


Figure 5.9: Histogram of the relative ICPIP rank of the best homologue in the ranking list. Each blue bar shows how many times ICIPIP ranks the best homologue within the specified interval (two successive values on the x axis). The first bar shows that for 11 targets ICIPIP placed the best homologue in the top 10 of its ranking list, whereas the second bar shows that 10 times it is placed between the 10^{th} and 20^{th} position. Those frequencies can be modelled by a decreasing monotonic polynomial.

In a second experiment, we evaluate the quality of the 'ICIPIP hit' by comparing the rank of the ICPIP hit in the ranking list of homologues based on ground truth. The results show that in 43% of the cases, ICIPIP hit is located in the top 5 based on ground truth ranking list. To visualise these results, Figure 5.10 shows F1 score of ICPIP hit in comparison to the best, worse and average F1 score of all the homologues for a specific target. In 70% of the cases ICPIP score is equal of above the average F1 score. In

addition, in 20 cases ICPIP hit is the best homologue. These results confirm that ICPIP can detect good homologues for a reasonable number of targets.

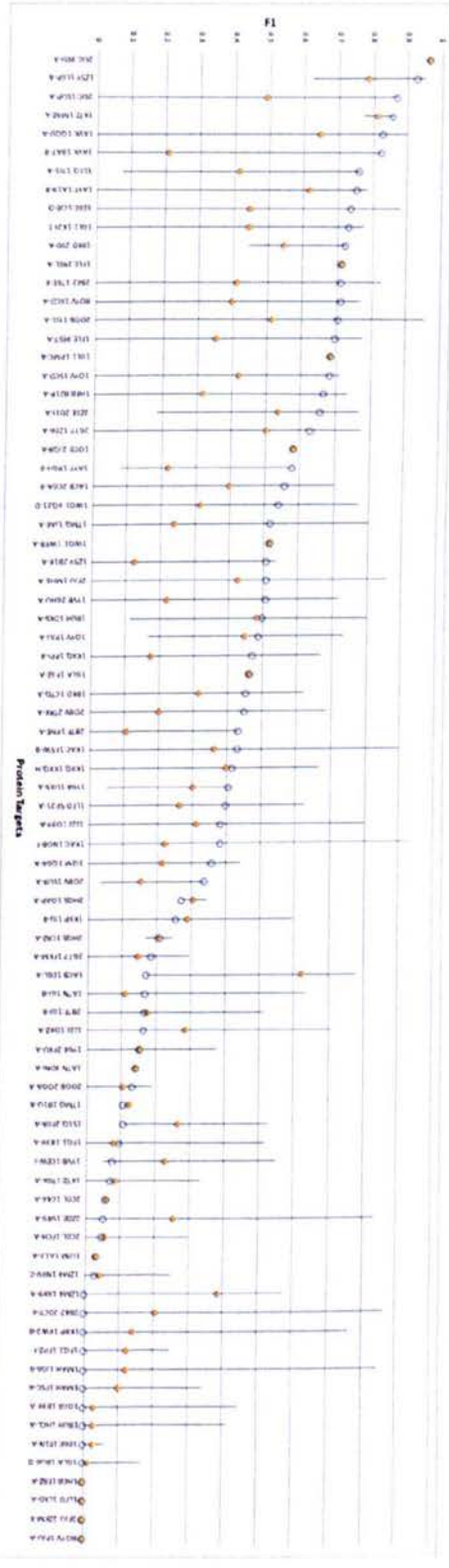


Figure 5.10: Interface F1 score of DS80 targets in respect to available homologues. Horizontal line connects the maximum and minimum F1 calculated for homologues of a given target. Average homologue F1 and ICPIP F1 are shown by yellow diamonds and blue circles, respectively.

5.4.2Performance Evaluation

Now that ICPIP has shown to perform well in detection of the best homologues, its performance was compared to other state-of-the-art methods. As described above to predict a QP interfaces, ICPIP hit interface is mapped on the QP. ICPIP was evaluated on DS56unbound (Table 5.1), DS120 (Table 5.2) and DS236 (Table 5.2). Since ICPIP can only be used for QP pairs from the 'homologous' category, which are 80, 40 and 17 chains of DS236, DS120 and DS56unbound, respectively. The remaining pairs are processed by T-PIP framework which also includes single chains in the 'homologous' category.

The result in Table 5.1 and Table 5.2 shows that although the use of ICPIP in the T-PIP framework performs better than other groups' methods, it has not improved the results obtained by T-PIP framework itself. T-PIP performance in precision, recall and F1 is ~2-5%, ~0.9-4.6% and ~ 1-2% above ICPIP. These results suggest that when ICPIP fails to correctly detect the best homologue, T-PIP prediction outperforms.

Table 5.1: Evaluation of interface prediction methods using the Ds56unbound dataset.

Predictor (*DS56unbound*)	**Precision**	**Recall**	**F1**	**Accuracy**	**MCC**
T-PIP	**53.8**	**48.5**	**49.6**	**84.0**	**41.1**
ICPIP	51.1	50.8	48.9	82.7	39.7
IBIS	48.2	29.3	34.4	82.5	27.9
PrISE	43.7	44.0	43.8	81.2	32.6
PredUs	43.3	53.6	47.9	73.2	30.4

Table 5.2: Comparison of interface predictors' performance on DS120 and DS236.

Predictor & dataset	Precision	Recall	F1	Accuracy	MCC
T-PIP DS120	**52.6**	**56.1**	**52.5**	**85.4**	**45.1**
ICPIP DS120	51.0	55.7	51.3	84.9	43.7
PredUs DS120	47.3	58.2	48.5	69.4	24.4
PrISE DS120	38.5	48.9	40.9	80.7	31.2
IBIS DS120	42.6	37.4	37.4	83.8	29.9
T-PIP DS236	**53.2**	**55.3**	**52.1**	**85.3**	**44.8**
ICPIP DS236	52.2	54.8	51.3	85.1	43.8
PrISE DS236	41.2	47.5	41.5	81.0	32.0
IBIS DS236	40.9	36.9	36.2	83.6	28.8

As in chapter 4, statistical significance of ICIPIP's results is first evaluated by providing standard deviations of predictors, Table 5.3. Apart from ICIPIP all the figures are taken from Table 4.6. ICIPIP standard deviation is very similar to T-PIP's which is in line with the fact that both methods provide predictions of similar quality. In addition, in Table 5.4, P-values are calculated for ICIPIP in comparison to other predictors. The statistically significance of ICIPIP performance is similar to what was shown for T-PIP in Table 4.7. Also, based on P-values in Table 5.4 the differences between ICIPIP and T-PIP performances are not statistically significant.

Table 5.3: Standard deviations of interface predictors' performance on DS120 and DS236. Numbers in the table display the standard deviation for every metric across the specified dataset.

Predictor & dataset	Precision	Recall	F1	Accuracy	MCC
T-PIP DS120	0.32	0.32	0.31	0.13	0.35
ICIPIP DS120	0.33	0.33	0.32	0.13	0.36
PredUs DS120	0.31	0.29	0.26	0.17	0.29
PrISE DS120	0.22	0.21	0.19	0.10	0.21
IBIS DS120	0.36	0.35	0.33	0.13	0.37
T-PIP DS236	0.32	0.32	0.30	0.12	0.34
ICIPIP DS236	0.33	0.33	0.31	0.13	0.35
PrISE DS236	0.23	0.22	0.20	0.11	0.22
IBIS DS236	0.36	0.34	0.32	0.12	0.36

Table 5.4: Significance of ICIPIP predictions: comparison with other predictors on DS120 and DS236. Numbers represent P-values. Those highlighted in red are not judged significant (P-values>0.05).

ICIPIP, Predictor & dataset	**Precision**	**Recall**	**F1**	**Accuracy**	**MCC**
ICIPIP, *T-PIP* DS120	0.65	0.93	0.76	0.78	0.74
ICIPIP, *PredUs* DS120	0.38	0.54	0.45	1.09E-13	9.96E-06
ICIPIP, *PrISE* DS120	0.0007	0.06	0.002	0.01	0.001
ICIPIP, *IBIS* DS120	0.07	5.47E-05	0.001	0.55	0.005
ICIPIP,*T-PIP* DS236	0.71	0.89	0.76	0.84	0.77
ICIPIP, *PrISE* DS236	3.61E-05	0.01	7.25E-05	0.0002	1.96E-05
ICIPIP, *IBIS* DS236	0.0005	2.46E-08	6.17E-07	0.19	8.84E-06

5.5 Discussion

The idea behind ICPIP was to predict protein interfaces by detection of the best homologue using binding site transitivity. The reason behind this idea was that the detection of best homologue results in a better interface prediction than T-PIP as shown in chapter 4 discussion. In Table 5.5, the same results are shown but only for DS80. We recall that 'Average homologues' is calculated by averaging all targets *average homologue* which is simply the average of all the scores among homologues of that target. T-PIP performance is much better than average but is below what the 'Best Homologues' approach would produce. Therefore, ICPIP was designed to achieve this optimal prediction by detecting the best homologue.

Table 5.5: Results based on selection of best homologues and average of homologues compared to T-PIP performance on DS80. Best homologues are selected by F1 score.

Predictor (*DS80*)	Precision	Recall	F1	Accuracy	MCC
Best Homologues	56.9	62.0	57.0	86.6	49.7
T-PIP	39.0	44.3	39.7	82.0	30.1
Average homologues	30.4	34.5	29.8	79.1	19.7

But as shown in the result section, ICPIP overall performance is below T-PIP, but at the same time ICPIP was able to detect a good quality homologue for 36% of the targets in DS236. Figure 5.11, shows the F1 score of ICPIP (blue circles) and T-PIP (red squares) for each target of DS80. These results confirm that two methods are orthogonal to each other, since their approach to interface prediction is different. T-PIP generates a censuses based on all the homologues while ICPIP focuses only on one homologue. Therefore, a meta method which can benefit from both predictions would improve the results.

To confirm this idea, in Table 5.6 we have evaluated the performance of the combined predictions of TPIP and ICPIP (ICPIP+TPIP). To combine these methods, for each target we have predicted the interfaces using both T-PIP and ICPIP. Then the best prediction based on the ground truth is taken as the result for that target. This combination method cannot be used in prediction since it requires the ground truth. But it displays that developing an intelligent meta predictor can significantly improve the T-PIP prediction by ~8% on F1 score. The F1 score of ICPIP+TPIP is 43% while the best that can be achieved is 57% (Table 5.5). Although, still there is a gap to the best that can be achieved, an improvement of 8% is a big step toward better interface prediction.

Based on Table 5.6, PrISE shows a better prediction results than T-PIP. However, Table 5.6 only displays a subsection of the whole T-PIP framework. We recall that Table 5.2 which displays the overall performance of T-PIP framework demonstrates the superiority of the T-PIP framework over PrISE. Note that T-PIP has other advantages over PrISE: first, T-PIP can predict interfaces for a QP chain whose structure is not available. Second, T-PIP prediction is faster than PrISE, since its structural alignment is performed once among homologues while PrISE requires searching in the repository of SEs, for every surface residue of the QP.

Table 5.6: Comparison of the combined interface prediction (ICPIP+TPIP) to each of the predictors separately. Note that: IBIS results are only for 74 chains since IBIS prediction is limited to availability of close homologues. PredUs performance is not displayed since results are not available on DS80 which is a subset of DBMK4.0.

Predictor (*DS80*)	Precision	Recall	F1	Accuracy	MCC
IBIS	36.0	32.1	31.7	83.0	24.1
ICPIP	36.2	42.9	37.2	81.3	27.5
T-PIP	39.0	44.3	39.7	82.0	30.1
PrISE	40.0	47.5	40.8	81.8	32.0
ICPIP+TPIP	42.8	47.8	43.0	83.5	34.6

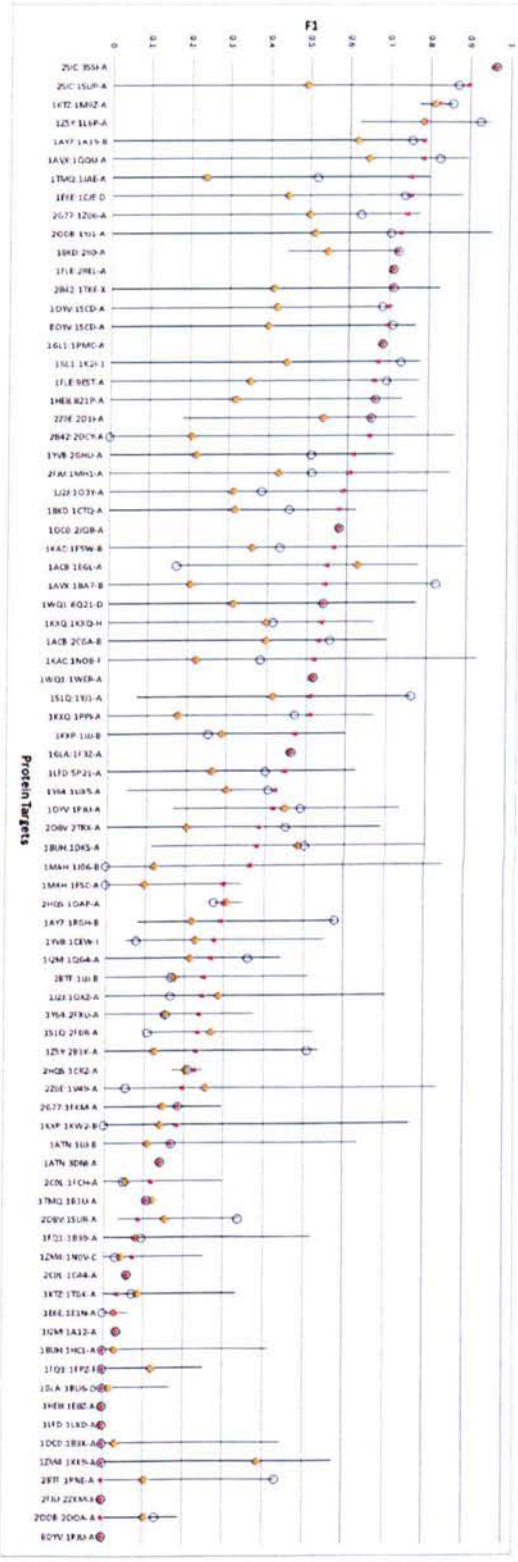


Figure 5.11: Interface F1 score of DS80 targets in respect to available homologues. Horizontal line connects the maximum and minimum F1 calculated for homologues of a given target. Average homologue F1, T-PIP and ICPIP F1 are shown by yellow diamonds, red square and blue circles, respectively.

Although a combined predictor is the optimal goal, currently both T-PIP and ICPIP frameworks (Table 5.1 and Table 5.2) have demonstrated to be the best predictors. Both methods have their own advantages and limitations and their use can be customised to the user goal. ICPIP covers fewer predictions than T-PIP, since it requires the homologues of two interacting chains. But at the same time ICPIP can produce the docked structure of the two QPs since it aligns the first QP chain over the interacting partners of the second QP chain homologue. Figure 5.12 displays an example of ICPIP docked model. For target 1S1Q:A-B (in blue), the chains in the unbound form 1YJ1-A and 2F0R-A are docked using ICPIP (in pink). The docked model is generated using the interface alignment between 1ZGU-B and 1UZX-B where the former is 1YJ1-A homologue and the latter is the interacting partner of 2F0R-A homologue. It should be noted that based on ICPIP alignment score 1ZGU-B had been selected as best homologue of 1YJ1-A. Since the RMSD between the actual and docked complexes is 0.98 Å, this shows that ICPIP can predict a near-native docked model. In this instance, ICPIP provides a more accurate model than Cluspro whose best docked model for this target has a RMSD of 17.5 Å. The output of a complex configuration by ICPIP is advantage benefit over current state-of-the-art methods, including T-PIP, which only specify if residues are part of the interface or not without considering their pairwise residue interactions.

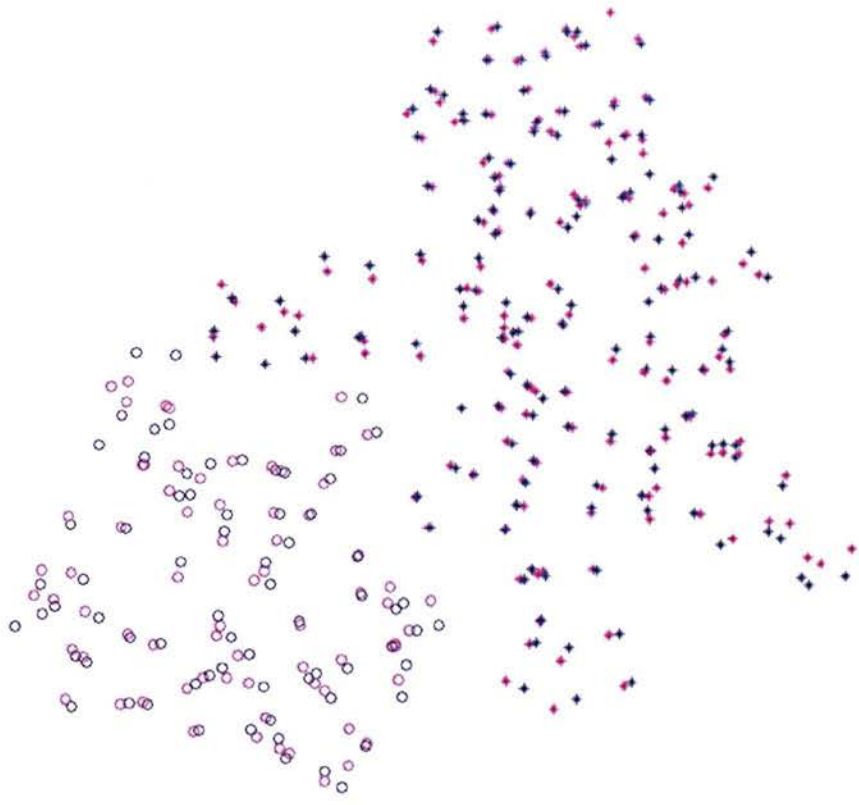


Figure 5.12: An example of ICPIP docked model. Each data point represents a C-alpha atom. The actual X-ray target 1S1Q:A-B is shown in blue where stars represent chain A and circles represent chain B. The docked ICPIP complex is shown in pink using the unbound chains 1YJ1-A (pink stars) and 2F0R-A (pink circles). This docked model is generated by the interface alignment of homologue of 1YJ1-A (1ZGU-B) and 2F0R-A homologue partner (1UZX-B). The RMSD between the docked and actual model is 0.98 Å.

Both ICPIP and T-PIP perform structural alignments but ICPIP has the risk of not finding the global maximum since it is based on local interface alignments. Apart from that, the best selected homologues for two QP are not always consistent. Consistency means that for example in Figure 5.4, since the alignment of a2 and pb1 suggests A1 as the best homologue for A; then B1 should be selected as the best homologue of B when comparing its binding site to interacting partners of A1 and A2. This is not always the case in ICPIP, since homologues can interacting with a diverse set of partners using the same binding site structure (Martin 2010; Tsai et al. 1996). Since ICPIP prediction is based on one homologue, a continue interface is predicted while T-PIP can predict any residues regardless of their overall location in 3D space.

Therefore, depending on the application of interface prediction one can take advantage of the differences between T-PIP and ICPIP.

5.6 Conclusion

In chapter 4, we demonstrated that T-PIP performance can be improved if its prediction was based on detection of best homologue. In this chapter, we have presented ICPIP, which aims at predicting interfaces of two QP chains through detection of the best homologous. This is achieved by interface structural comparison among homologue of the first QP chain with the homologues binding partners of the second QP chain and vice versa.

ICPIP uses ICP to perform interface alignments but due to its limitations, initial alignments stages have been introduced. Apart from this, Hungarian algorithm along with outlier detection has been integrated in ICP to account for both unique and partial assignment of data points.

Comparisons show that ICPIP is capable of improving T-PIP results in 36% of the cases, while the overall performance remains lower than T-PIP. Since these methods are orthogonal their combination has shown to improve T-PIP prediction by ~8%. Therefore, design of a meta-predictor combining T-PIP and ICPIP can significantly improve interface prediction. In addition, ICPIP interface alignment procedure may be enhanced by performing labelling data points with their physico-chemical properties.

6 Conclusion

This chapter concludes the dissertation. We first summarise our scientific contributions in section 6.1. Then, in section 6.2 we discuss the achievements and limitations of the proposed methods. In section 6.3, we propose a number of avenues for future research directions and finally closing remarks are provided in section 6.4.

6.1 Summary of Contributions

In this thesis, we have explored the field of protein interface prediction and detection of native-like 3D protein complexes by re-ranking their docked conformations.

First, chapter 2 provided an extensive literature review of methods in protein interface prediction. We then gave an overview of docking algorithms and discussed methods, in particular those using predicted interfaces, for re-ranking docked poses. This detailed analysis allowed us to highlight issues which had not been fully addressed by the research community. In this thesis, we have investigated ways of dealing with some of them.

Chapter 3 focused on creating 3D motifs which are binding site descriptors based on common structural pattern among homologues of a target protein. Although, 3D motifs provide biological insight of protein-protein interaction and was successfully applied to a real application as shown with the LDLR-HNP1 complex, their usage is limited. First, its dependence to literature studies prevents its application for high-throughput analysis, and, second, its creation requires homologous proteins which show

a homogeneous binding site pattern even on the interacting partner side. Finally, 3D motif creation is limited to the availability of close homologous complexes of the target protein.

To overcome these limitations, in chapter 4, we introduced T-PIP, a protein interface predictor which processes more predictions by using remote homologues. In addition, T-PIP differs from other state-of-the-art methods by considering diversity among interacting partners of remote homologues. The scoring strategy of T-PIP results in a performance higher than other state-of-the-art methods. However, we also discovered that a T-PIP prediction would generally not outperform a prediction made by using the actual interface of the best available homologue if it had been identified.

Therefore, in chapter 5, we proposed ICPIP which exploits the binding site transitivity concept to find the best homologue. This was achieved by structural interface comparison of the homologues of the first QP and the binding site of the homologues partner of the second QP. Comparisons with T-PIP showed that ICPIP is capable of improving T-PIP results in 36% of the cases.

We also investigated usage of predicted interface residues for re-ranking docked conformations. In chapter 4, we introduced PioDock which used T-PIP prediction to re-rank docked conformations, where higher rank was given to docked models showing more overlap with T-PIP prediction at their binding site. Comparison with other methods showed PioDock performs better on a standard benchmark.

6.2 Discussion

Despite our contribution to the state of the art, accurate prediction of interfaces and identification of near-native conformations remain challenging tasks. First, the actual binding site of the target protein does not always have a representative among homologous complexes. Second, when homologues binding sites refer to distinct potential sites it is a challenge to select the one corresponding to the actual interface. Third, docking software are still not able to produce native like models for every target. Finally, the performance of re-ranking using interface knowledge depends on the quality of predicted interfaces which is still far from being satisfying.

Although the combination of our two orthogonal interface predictors, T-PIP and ICPIP, has shown potential for improving predictions, the choice of methods is not

straight forward and depends on the application. First, ICPIP covers fewer predictions than T-PIP, since it requires the homologues of two interacting chains. Second, while T-PIP can predict interface from sequence of proteins, ICPIP requires the structure. Finally, using ICPIP, the best selected homologues for two QPs are not always consistent. On the other hand, unlike T-PIP which does not consider pairwise interactions between residues in its predictions, ICPIP predictions provide a description of the binding site environment. In addition, when ICPIP is able to detect the best homologue for a specific target, it performs better than T-PIP on that target.

Both methods rely on availability of homologues complexes. Although, we have shown that T-PIP covers more proteins in comparison to homologous template-based predictors by considering remote homologues, methods based on structural neighbours are applicable on a larger range of targets than T-PIP.

In addition, ICPIP detection of best homologue relies on a good interface alignment. Therefore, in cases where ICPIP is stuck in local minima it fails to detect the correct homologue.

In regards to re-ranking docking conformations, 3D motif has been used for a real application. Although its usage is limited, it provides constraint which allows short listing only a few complexes. This makes it useful for further wet-lab investigation. In addition, 3D motifs can be used to identify putative partners of proteins with known 3D structures. On the other hand, PioDock is a high-throughput method whose performance depends on the quality of interface predictions. Using PioDock along with ground truth interfaces, which are continuous in the 3D space (patch-like), improved ranking. Therefore, usage of ICPIP, which provides a continuous interface prediction (using one homologue) can improve ranking in comparison to T-PIP. In addition, ICPIP itself provides a docking conformation which requires further investigation.

Although we have made significant contributions to interface prediction by proposing T-PIP and ICPIP (chapter 4 and 5), and re-ranking docked conformations using 3Dmotif and PioDock (chapter 3 and 4), there is still scope for improvement. A few suggestions on how to further explore current methods are outlined in section 6.3.

6.3 Future Work

While T-PIP is able to cover more predictions than other homologous template-based predictors (introduced in section 2.3.2.1) by considering remote homologues, its coverage can be further expanded by exploring close and remote structural neighbours. T-PIP scoring can be improved by using structural alignments based on structural similarities among homologous/structural neighbours' binding sites, since similar interface architectures have been detected among proteins displaying different functions and global structures (Keskin & Nussinov 2007).

Moreover, we showed that knowledge of T-PIP predictions improves detection of near-native docked models. Also, we investigated that patch-like interface information can provide a more reliable ranking in comparison to non-continues interfaces. Therefore, T-PIP could be used along with clustering, similarly to the method used in JET interface predictor (Engelen et al. 2009), to provide a patch-like interface prediction. In consequence, this may result in better ranking of docked conformations.

Although ICPIP improves T-PIP prediction in cases where it detects the best homologue, its detection relies on finding an optimal interface alignment using the ICP algorithm. ICP could be improved by considering physico-chemical properties of residues while performing the alignments, since two proteins interacting with a third protein using a similar binding site will show similar interface properties.

In addition, ICPIP provides a template-based docking of models which requires further evaluation. This is an advantage over generating template-free docked poses which are computationally expensive.

Finally, artificial combination of ICP and ICPIP has resulted in promising prediction of protein interfaces. Therefore, developing an intelligent meta predictor is our main future objective. An initial approach is to develop an SVM-based classifier trained on different properties of ICP and ICPIP algorithms. This classifier will then be able to distinguish whether T-PIP or ICPIP prediction is more reliable for a given target.

6.4 Closing Remarks

This research has contributed to the fields of protein interface prediction and detection of native-like poses. Not only have we provided a mechanism for detailed analysis of interfaces which will benefit wet-lab investigations, but also, we have produced a tool for high-throughput interface analysis for the research community. Usage of this interface information may help many scientists in detection of the potential 3D structure of two protein chains allowing the investigation of drug design.

Finally, the presented contributions are intended to motivate future research in the area of interface prediction and detection of native-like 3D complexes; and hopefully, directing a more significant contribution to those fields.

References

Afonnikov, D.A., Oshchepkov, D.Y. & Kolchanov, N.A., 2001. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. *Bioinformatics*, 17(11), pp.1035–1046.

Ahmad, S. & Mizuguchi, K., 2011. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PloS one*, 6(12), p.e29104.

Aloy, P. et al., 2004. Structure-based assembly of protein complexes in yeast. *Science*, 303(5666), pp.2026–2029.

Aloy, P. et al., 2003. The relationship between sequence and interaction divergence in proteins. *Journal of molecular biology*, 332(5), pp.989–998.

Aloy, P. & Russell, R.B., 2002. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences*, 99(9), pp.5896–5901.

Altschul, S.F. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17), pp.3389–3402.

Amos-Binks, A. et al., 2011. Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences. *BMC bioinformatics*, 12(1), p.225.

André, I. et al., 2007. Prediction of the structure of symmetrical protein assemblies. *Proceedings of the National Academy of Sciences*, 104(45), pp.17656–17661.

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

Andreeva, A. et al., 2008. Data growth and its impact on the SCOP database: new developments. *Nucleic acids research*, 36(suppl 1), pp.D419–D425.

Andrusier, N. et al., 2008. Principles of flexible protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 73(2), pp.271–289.

Andrusier, N., Nussinov, R. & Wolfson, H.J., 2007. FireDock: fast interaction refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 69(1), pp.139–159.

Ashkenazy, H. et al., 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic acids research*, 38(suppl 2), pp.W529–W533.

Aytuna, A.S., Gursoy, A. & Keskin, O., 2005. Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics*, 21(12), pp.2850–2855.

Azuma, Y. et al., 1999. Model of the ran-RCC1 interaction using biochemical and docking experiments. *Journal of molecular biology*, 289(4), pp.1119–1130.

Bahadur, R.P. et al., 2003. Dissecting subunit interfaces in homodimeric proteins. *Proteins: Structure, Function, and Bioinformatics*, 53(3), pp.708–719.

Baldi, P. et al., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5), pp.412–424.

Bastard, K. et al., 2003. Docking macromolecules with flexible segments. *Journal of computational chemistry*, 24(15), pp.1910–1920.

Bastard, K., Prévost, C. & Zacharias, M., 2006. Accounting for loop flexibility during protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 62(4), pp.956–969.

Beglov, D. et al., 2009. Structural insights into recognition of β2-glycoprotein I by the lipoprotein receptors. *Proteins: Structure, Function, and Bioinformatics*, 77(4), pp.940–949.

Ben-Zeev, E. et al., 2003. Prediction of the structure of the complex between the 30S ribosomal subunit and colicin E3 via weighted-geometric docking. *Journal of Biomolecular Structure and Dynamics*, 20(5), pp.669–675.

Berchanski, A. & Eisenstein, M., 2003. Construction of molecular assemblies via docking: modeling of tetramers with D2 symmetry. *Proteins: Structure, Function, and Bioinformatics*, 53(4), pp.817–829.

Berman, H.M. et al., 2000. The protein data bank. *Nucleic acids research*, 28(1), pp.235–242.

Bernauer, J. et al., 2007. A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, 23(5), pp.555–562.

Bertolazzi, P., Guerra, C. & Liuzzi, G., 2010. A global optimization algorithm for protein surface alignment. *BMC bioinformatics*, 11(1), p.488.

Besl, P.J. & McKay, N.D., 1992. Method for registration of 3-D shapes. In *Robotics-DL tentative*. pp. 586–606.

Blacklow, S.C., 2007. Versatility in ligand recognition by LDL receptor family proteins: advances and frontiers. *Current opinion in structural biology*, 17(4), pp.419–426.

Blas, J.R., Segura, J. & Fernandez-Fuentes, N., 2012. Computational Tools and Databases for the Study and Characterization of Protein Interactions. In W. Cai, ed. *Protein-Protein Interactions - Computational and Experimental Tools*. In Tach, pp. 380–404.

Bogan, A.A. & Thorn, K.S., 1998. Anatomy of hot spots in protein interfaces. *Journal of molecular biology*, 280(1), pp.1–9.

Bonvin, A.M.J.J., 2006. Flexible protein-protein docking. *Current opinion in structural biology*, 16(2), pp.194–200.

Bordner, A.J. & Abagyan, R., 2005. Statistical analysis and prediction of protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 60(3), pp.353–366.

Bourgeois, F. & Lassalle, J.-C., 1971. An extension of the Munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, 14(12), pp.802–804.

Bourquard, T. et al., 2011. A collaborative filtering approach for protein-protein docking scoring functions. *PloS one*, 6(4), p.e18541.

Bradford, J.R. et al., 2006. Insights into protein-protein interfaces using a Bayesian network prediction method. *Journal of molecular biology*, 362(2), pp.365–386.

Bradford, J.R. & Westhead, D.R., 2005. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8), pp.1487–1494.

Bronowska, A., 2013. Entropy and quantum effects for in silico drug design. Available at: http://www.h-its.org/english/research/mbm/projects/structure-based_molecular_design/entropy_and_quantum_effects/entropy_and_quantum_effects.php.

Browne, F. et al., 2010. From experimental approaches to computational techniques: a review on the prediction of protein-protein interactions. *Advances in Artificial Intelligence*, 2010, p.7.

Brückner, A. et al., 2009. Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, 10(6), pp.2763–2788.

Brylinski, M. & Skolnick, J., 2008. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of sciences*, 105(1), pp.129–134.

Burgoyne, N.J. & Jackson, R.M., 2006. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, 22(11), pp.1335–1342.

Caffrey, D.R. et al., 2004. Are protein--protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13(1), pp.190–202.

Cao, Y., 2008. Munkres Assignment Algorithm. , p.MATLAB Central File Exchange.

Carl, N. et al., 2010. Protein-Protein Binding Site Prediction by Local Structural Alignment. *Journal of chemical information and modeling*, 50(10), pp.1906–1913.

Carl, N., Konc, J. & Janezic, D., 2008. Protein surface conservation in binding sites. *Journal of chemical information and modeling*, 48(6), pp.1279–1286.

Carter, P. et al., 2005. Protein-protein docking using 3D-Dock in rounds 3, 4, and 5 of CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.281–288.

Cazals, F., 2010. Revisiting the Voronoi description of protein-protein interfaces: Algorithms. In *Pattern Recognition in Bioinformatics*. Springer, pp. 419–430.

Chang, T.L. et al., 2005. Dual role of α-defensin-1 in anti-HIV-1 innate immunity. *Journal of Clinical Investigation*, 115(3), pp.765–773.

Chen, H. & Skolnick, J., 2008. M-TASSER: an algorithm for protein quaternary structure prediction. *Biophysical journal*, 94(3), pp.918–928.

Chen, H & Zhou, H.X., 2005. Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins: Structure, Function, and Bioinformatics*, 61(1), pp.21–35.

Chen, Huiling & Zhou, H.-X., 2005. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic acids research*, 33(10), pp.3193–3199.

Chen, P. & Li, J., 2010. Sequence-based identification of interface residues by an integrative profile combining hydrophobic and evolutionary information. *BMC bioinformatics*, 11(1), p.402.

Chen, P., Wong, L. & Li, J., 2012. Detection of Outlier Residues for Improving Interface Prediction in Protein Heterocomplexes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4), pp.1155–1165.

Chen, R., Mintseris, J., et al., 2003. A protein-protein docking benchmark. *Proteins: Structure, Function, and Bioinformatics*, 52(1), pp.88–91.

Chen, R., Li, L. & Weng, Z., 2003. ZDOCK: An initial-stage protein-docking algorithm. *Proteins: Structure, Function, and Bioinformatics*, 52(1), pp.80–87.

Chen, R. & Weng, Z., 2002. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins: Structure, Function, and Bioinformatics*, 47(3), pp.281–294.

Chen, X. wen & Jeong, J.C., 2009. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, 25(5), pp.585–591.

Chen, Y. et al., 2012. A novel method for prediction of protein interaction sites based on integrated RBF neural networks. *Computers in Biology and Medicine*, 42(4), pp.402–407.

Cheng, T.M.-K., Blundell, T.L. & Fernandez-Recio, J., 2007. pyDock: Electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 68(2), pp.503–515.

Chenna, R. et al., 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research*, 31(13), pp.3497–3500.

Chothia, C. & Janin, J., 1975. Principles of protein-protein recognition. *Nature*, 256(5520), pp.705–708.

Chowdhury, R. et al., 2013. Protein-Protein Docking with F2Dock 2.0 and GB-Rerank. *PloS one*, 8(3), p.e51307.

Chung, J.L., Wang, W. & Bourne, P.E., 2005. Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins: Structure, Function, and Bioinformatics*, 62(3), pp.630–640.

Clackson, T. & Wells, J.A., 1995. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196), pp.383–386.

Cohen, M.L., 1992. Epidemiology of drug resistance: implications for a post—antimicrobial era. *Science*, 257(5073), pp.1050–1055.

Cole, C. & Warwicker, J., 2002. Side-chain conformational entropy at protein-protein interfaces. *Protein Science*, 11(12), pp.2860–2870.

Comeau, S.R. et al., 2004. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, 20(1), pp.45–50.

Comeau, S.R. et al., 2007. ClusPro: performance in CAPRI rounds 6--11 and the new server. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.781–785.

Comeau, S.R. & Camacho, C.J., 2005. Predicting oligomeric assemblies: N-mers a primer. *Journal of structural biology*, 150(3), pp.233–244.

Connolly, M.L., 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science*, 221(4612), pp.709–713.

Conte, L. Lo, Chothia, C. & Janin, J., 1999. The atomic structure of protein-protein recognition sites. *Journal of molecular biology*, 285(5), pp.2177–2198.

Darnell, S.J., Page, D. & Mitchell, J.C., 2007. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins: Structure, Function, and Bioinformatics*, 68(4), pp.813–823.

Davis, F.P. et al., 2006. Protein complex compositions predicted by structural similarity. *Nucleic acids research*, 34(10), pp.2943–2952.

DeLano, W.L., 2002. Unraveling hot spots in binding interfaces: progress and challenges. *Current opinion in structural biology*, 12(1), pp.14–20.

Deng, L. et al., 2009. Prediction of protein-protein interaction sites using an ensemble method. *BMC bioinformatics*, 10(1), p.426.

Diamond, G. et al., 2009. The roles of antimicrobial peptides in innate host defense. *Current pharmaceutical design*, 15(21), p.2377.

Van Dijk, A.D.J., De Vries, S.J., et al., 2005. Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.232–238.

Van Dijk, A.D.J., Boelens, R. & Bonvin, A.M.J.J., 2005. Data-driven docking for the study of biomolecular complexes. *Febs Journal*, 272(2), pp.293–312.

Dirlam-Schatz, K.A. & Attie, A.D., 1998. Calcium induces a conformational change in the ligand binding domain of the low density lipoprotein receptor. *Journal of lipid research*, 39(2), pp.402–411.

Dominguez, C., Boelens, R. & Bonvin, A.M.J.J., 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7), pp.1731–1737.

Dong, Q. et al., 2007. Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC bioinformatics*, 8(1), p.147.

Douguet, D. et al., 2006. Dockground resource for studying protein-protein interfaces. *Bioinformatics*, 22(21), pp.2612–2618.

Duhovny, D., Nussinov, R. & Wolfson, H.J., 2002. Efficient unbound docking of rigid molecules. In *Algorithms in bioinformatics*. Springer, pp. 185–200.

Dundas, J. et al., 2006. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic acids research*, 34(suppl 2), pp.W116–W118.

Ellingson, L. & Zhang, J., 2011. An efficient algorithm for matching protein binding sites for protein function prediction. In *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. pp. 289–293.

Ellingson, L. & Zhang, J., 2012. Protein surface matching by combining local and global geometric information. *PloS one*, 7(7), p.e40540.

Emekli, U. et al., 2008. HingeProt: automated prediction of hinges in protein structures. *Proteins: Structure, Function, and Bioinformatics*, 70(4), pp.1219–1227.

Engelen, S. et al., 2009. Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling. *PLoS computational biology*, 5(1), p.e1000267.

Ericksen, B. et al., 2005. Antibacterial activity and specificity of the six human α-defensins. *Antimicrobial agents and chemotherapy*, 49(1), pp.269–275.

Esmaielbeiki, R. & Nebel, J.-C., 2012. Unbiased Protein Interface Prediction Based on Ligand Diversity Quantification. In *German Conference on Bioinformatics 2012*. pp. 119–130.

Esmaielbeiki, R., P. Naughton, D. & Nebel, J.-C., 2012. Structure prediction of LDLR-HNP1 complex based on docking enhanced by LDLR binding 3D motif. *Protein and Peptide Letters*, 19(4), p.458.

Esquivel-Rodríguez, J., Yang, Y.D. & Kihara, D., 2012. Multi-LZerD: Multiple protein docking for asymmetric complexes. *Proteins: Structure, Function, and Bioinformatics*, 80(7), pp.1818–1833.

Ezkurdia, I. et al., 2009. Progress and challenges in predicting protein-protein interaction sites. *Briefings in bioinformatics*, 10(3), pp.233–246.

Fahmy, A. & Wagner, G., 2002. TreeDock: A tool for protein docking based on minimizing van der Waals energies. *Journal of the American Chemical Society*, 124(7), pp.1241–1250.

Fariselli, P. et al., 2003. A neural network method to improve prediction of protein-protein interaction sites in heterocomplexes. In *Neural Networks for Signal Processing, 2003. NNSP'03. 2003 IEEE 13th Workshop on*. pp. 33–41.

Fariselli, P. et al., 2002. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry*, 269(5), pp.1356–1361.

Faure, G., Andreani, J. & Guerois, R., 2012. InterEvol database: exploring the structure and evolution of protein complex interfaces. *Nucleic acids research*, 40(D1), pp.D847–D856.

Feng, Y., Kloczkowski, A. & Jernigan, R.L., 2007. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins: Structure, Function, and Bioinformatics*, 68(1), pp.57–66.

Fernandez-Recio, J. et al., 2005. Optimal docking area: a new method for predicting protein-protein interaction sites. *PROTEINS: Structure, Function, and bioinformatics*, 58(1), pp.134–143.

Fernández-Recio, J., 2011. Prediction of protein binding sites and hot spots. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5), pp.680–698.

Fernández-Recio, J., Totrov, M. & Abagyan, R., 2003. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins: Structure, Function, and Bioinformatics*, 52(1), pp.113–117.

Fernández-Recio, J., Totrov, M. & Abagyan, R., 2004. Identification of protein--protein interaction sites from docking energy landscapes. *Journal of molecular biology*, 335(3), pp.843–865.

Fernández-Recio, J., Totrov, M. & Abagyan, R., 2002. Soft protein-protein docking in internal coordinates. *Protein Science*, 11(2), pp.280–291.

Fink, F. et al., 2011. PROCOS: Computational analysis of protein-protein complexes. *Journal of Computational Chemistry*, 32(12), pp.2575–2586.

Fisher, C., Beglova, N. & Blacklow, S.C., 2006. Structure of an LDLR-RAP complex reveals a general mode for ligand recognition by lipoprotein receptors. *Molecular cell*, 22(2), pp.277–283.

Fitzgerald, J.E. et al., 2007. Reduced Cβ statistical potentials can outperform all-atom potentials in decoy identification. *Protein science*, 16(10), pp.2123–2139.

Fleishman, S.J. et al., 2011. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. *Journal of molecular biology*, 414(2), pp.289–302.

Fleming, E. et al., 2008. Effect of lipid composition on buforin II structure and membrane entry. *Proteins: Structure, Function, and Bioinformatics*, 73(2), pp.480–491.

Frieden, T.R. et al., 1993. The emergence of drug-resistant tuberculosis in New York City. *New England Journal of Medicine*, 328(8), pp.521–526.

Fuentealba, R.A. et al., 2010. Low-density lipoprotein receptor-related protein 1 (LRP1) mediates neuronal Aβ42 uptake and lysosomal trafficking. *PLoS One*, 5(7), p.e11884.

Gabb, H.A., Jackson, R.M. & Sternberg, M.J.E., 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of molecular biology*, 272(1), pp.106–120.

Gaboriaud, C. et al., 2003. The crystal structure of the globular head of complement protein C1q provides a basis for its versatile recognition properties. *Journal of Biological Chemistry*, 278(47), pp.46974–46982.

Gallet, X. et al., 2000. A fast method to predict protein interaction sites from sequences. *Journal of molecular biology*, 302(4), pp.917–926.

Gamliel, R. et al., 2011. A library of protein surface patches discriminates between native structures and decoys generated by structure prediction servers. *BMC structural biology*, 11(1), p.20.

Gao, M. & Skolnick, J., 2010. Structural space of protein--protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, 107(52), pp.22517–22522.

Gardiner, E.J., Willett, P. & Artymiuk, P.J., 2001. Protein docking using a genetic algorithm. *Proteins: Structure, Function, and Bioinformatics*, 44(1), pp.44–56.

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

Ghoorah, A.W. et al., 2011. Spatial clustering of protein binding sites for template based protein docking. *Bioinformatics*, 27(20), pp.2820–2827.

Gibrat, J.-F., Madej, T. & Bryant, S.H., 1996. Surprising similarities in structure comparison. *Current opinion in structural biology*, 6(3), pp.377–385.

Glaser, F. et al., 2001. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 43(2), pp.89–102.

Gold, N.D. & Jackson, R.M., 2006. SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic acids research*, 34(suppl 1), pp.D231–D234.

Gong, S. et al., 2005. A protein domain interaction interface database: InterPare. *BMC bioinformatics*, 6(1), p.207.

Goodall, C., 1991. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.285–339.

Gottschalk, K.E., Neuvirth, H. & Schreiber, G., 2004. A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Engineering Design and Selection*, 17(2), pp.183–189.

Gower, J.C., 1975. Generalized procrustes analysis. *Psychometrika*, 40(1), pp.33–51.

Gray, J.J., 2006. High-resolution protein-protein docking. *Current opinion in structural biology*, 16(2), pp.183–193.

Gray, J.J. et al., 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of molecular biology*, 331(1), pp.281–300.

Grünberg, R., Leckner, J. & Nilges, M., 2004. Complementarity of structure ensembles in protein-protein binding. *Structure*, 12(12), pp.2125–2136.

Guerois, R., Nielsen, J.E. & Serrano, L., 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of molecular biology*, 320(2), pp.369–387.

Le Guilloux, V., Schmidtke, P. & Tuffery, P., 2009. Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, 10(1), p.168.

Günther, S. et al., 2007. Docking without docking: ISEARCH—prediction of interactions using known interfaces. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.839–844.

Guo, F. et al., 2012. Protein-protein binding site identification by enumerating the configurations. *BMC bioinformatics*, 13(1), p.158.

Guttman, M., Prieto, J.H., Croy, J.E., et al., 2010. Decoding of Lipoprotein- Receptor Interactions: Properties of Ligand Binding Modules Governing Interactions with Apolipoprotein E. *Biochemistry*, 49(6), pp.1207–1216.

Guttman, M., Prieto, J.H., Handel, T.M., et al., 2010. Structure of the minimal interface between ApoE and LRP. *Journal of molecular biology*, 398(2), pp.306–319.

Halperin, I. et al., 2002. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4), pp.409–443.

Halperin, I. et al., 2004. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure (London, England: 1993)*, 12(6), p.1027.

Han, J.-H. et al., 2006. Divergence of interdomain geometry in two-domain proteins. *Structure*, 14(5), pp.935–945.

Hart, T.N. & Read, R.J., 1992. A multiple-start Monte Carlo docking method. *Proteins: Structure, Function, and Bioinformatics*, 13(3), pp.206–222.

Heifetz, A. & Eisenstein, M., 2003. Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking. *Protein Engineering*, 16(3), pp.179–185.

Heifetz, A., Katchalski-Katzir, E. & Eisenstein, M., 2002. Electrostatics in protein-protein docking. *Protein Science*, 11(3), pp.571–587.

Henikoff, S. & Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22), pp.10915–10919.

Henrick, K. & Thornton, J.M., 1998. PQS: a protein quaternary structure file server. *Trends in biochemical sciences*, 23(9), p.358.

Hermes, L. & Buhmann, J.M., 2000. Feature selection for support vector machines. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on.* pp. 712–715.

Herz, J. & Bock, H.H., 2002. Lipoprotein receptors in the nervous system. *Annual review of biochemistry*, 71(1), pp.405–434.

Heuser, P. et al., 2005. Refinement of unbound protein docking studies using biological knowledge. *PROTEINS: Structure, Function, and Bioinformatics*, 61(4), pp.1059–1067.

Higa, R.H. & Tozzi, C.L., 2008. A simple and efficient method for predicting protein-protein interaction sites. *Genetics and Molecular Research*, 7(3), pp.898–909.

Higazi, A.A.-R. et al., 2000. The α-defensins stimulate proteoglycan-dependent catabolism of low-density lipoprotein by vascular cells: a new class of inflammatory apolipoprotein and a possible contributor to atherogenesis. *Blood*, 96(4), pp.1393–1398.

Hill, C.P. et al., 1991. Crystal structure of defensin HNP-3, an amphiphilic dimer: mechanisms of membrane permeabilization. *Science*, 251(5000), pp.1481–1485.

Holm, L. & Sander, C., 1993. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1), pp.123–138.

Hu, Z. et al., 2000. Conservation of polar residues as hot spots at protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 39(4), pp.331–342.

Huang, B. & Schroeder, M., 2008. Using protein binding site prediction to improve protein docking. *Gene*, 422(1), pp.14–21.

Huang, S.-Y. & Zou, X., 2008. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins: Structure, Function, and Bioinformatics*, 72(2), pp.557–579.

Hwang, H., Vreven, T., Pierce, B., et al., 2010. Performance of ZDOCK and ZRANK in CAPRI rounds 13--19. *Proteins: Structure, Function, and Bioinformatics*, 78(15), pp.3104–3110.

Hwang, H. et al., 2008. Protein--protein docking benchmark version 3.0. *Proteins: Structure, Function, and Bioinformatics*, 73(3), pp.705–709.

Hwang, H., Vreven, T., Janin, J., et al., 2010. Protein--protein docking benchmark version 4.0. *Proteins: Structure, Function, and Bioinformatics*, 78(15), pp.3111–3114.

Inbar, Y. et al., 2005. Prediction of multimolecular assemblies by multiple docking. *Journal of molecular biology*, 349(2), pp.435–447.

Jackson, R.M., Gabb, H.A. & Sternberg, M.J.E., 1998. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *Journal of molecular biology*, 276(1), pp.265–285.

Janin, J., 2005. Assessing predictions of protein--protein interaction: the CAPRI experiment. *Protein science*, 14(2), pp.278–283.

Janin, J. et al., 2003. CAPRI: a critical assessment of predicted interactions. *Proteins: Structure, Function, and Bioinformatics*, 52(1), pp.2–9.

Janin, J., 2013. Docking Predictions of Protein-Protein Interactions and Their Assessment: The CAPRI Experiment. In I. Roterman-Konieczna, ed. *Identification of Ligand Binding Site and Protein-Protein Interaction Area*. Springer Netherlands, pp. 87–104. Available at: http://dx.doi.org/10.1007/978-94-007-5285-6_5.

Janin, J., 2010. Protein--protein docking tested in blind predictions: the CAPRI experiment. *Molecular BioSystems*, 6(12), pp.2351–2362.

Janin, J. & Wodak, S., 2007. The third CAPRI assessment meeting Toronto, Canada, April 20--21, 2007. *Structure*, 15(7), pp.755–759.

Jensen, G.A. et al., 2006. Binding site structure of one LRP--RAP complex: Implications for a common ligand--receptor binding motif. *Journal of molecular biology*, 362(4), pp.700–716.

Jiang, F. & Kim, S.-H., 1991. "Soft docking": matching of molecular surface cubes. *Journal of molecular biology*, 219(1), pp.79–102.

Johansson, K.E. & Hamelryck, T., 2013. A simple probabilistic model of multibody interactions in proteins. *Proteins: Structure, Function, and Bioinformatics*.

Jones, G. et al., 1997. Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology*, 267(3), pp.727–748.

Jones, S. & Thornton, J.M., 1997a. Analysis of protein-protein interaction sites using surface patches. *Journal of molecular biology*, 272(1), pp.121–132.

Jones, S. & Thornton, J.M., 1997b. Prediction of protein-protein interaction sites using patch analysis. *Journal of molecular biology*, 272(1), pp.133–143.

Jordan, R.A. et al., 2012. Predicting protein-protein interface residues using local surface structural similarity. *BMC bioinformatics*, 13(1), p.41.

Kagan, B.L. et al., 1990. Antimicrobial defensin peptides form voltage-dependent ion-permeable channels in planar lipid bilayer membranes. *Proceedings of the National Academy of Sciences*, 87(1), pp.210–214.

Kanamori, E. et al., 2007. Docking of protein molecular surfaces with evolutionary trace analysis. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.832–838.

Karaca, E. et al., 2010. Building Macromolecular Assemblies by Information-driven Docking INTRODUCING THE HADDOCK MULTIBODY DOCKING SERVER. *Molecular & Cellular Proteomics*, 9(8), pp.1784–1794.

Kastritis, P.L. & Bonvin, A.M.J.J., 2010. Are Scoring Functions in Protein- Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *Journal of proteome research*, 9(5), pp.2216–2225.

Katchalski-Katzir, E. et al., 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences*, 89(6), pp.2195–2199.

Kawashima, S. & Kanehisa, M., 2000. AAindex: amino acid index database. *Nucleic acids research*, 28(1), p.374.

Keskin, O. et al., 2004. A new, structurally nonredundant, diverse data set of protein--protein interfaces and its implications. *Protein Science*, 13(4), pp.1043–1055.

Keskin, O. et al., 2005. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of molecular biology*, 345(5), pp.1281–1294.

Keskin, O. & Nussinov, R., 2007. Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, 15(3), pp.341–354.

Khashan, R., Zheng, W. & Tropsha, A., 2012. Scoring protein interaction decoys using exposed residues (SPIDER): A novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins: Structure, Function, and Bioinformatics*, 80(9), pp.2207–2217.

Kiel, C. et al., 2005. Recognizing and Defining True Ras Binding Domains II: In Silico Prediction Based on Homology Modelling and Energy Calculations. *Journal of molecular biology*, 348(3), pp.759–775.

Kim, E. et al., 2010. Predicting direct protein interactions from affinity purification mass spectrometry data. *Algorithms for Molecular Biology*, 5, p.34.

Kim, W.K. & Ison, J.C., 2005. Survey of the geometric association of domain--domain interfaces. *PROTEINS: Structure, Function, and Bioinformatics*, 61(4), pp.1075–1088.

Kjer, H.M. & Wilm, J., 2010. *Evaluation of surface registration algorithms for PET motion correction*. Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark.

Klotman, M.E. & Chang, T.L., 2006. Defensins in innate antiviral immunity. *Nature Reviews Immunology*, 6(6), pp.447–456.

Koike, A. & Takagi, T., 2004. Prediction of protein-protein interaction sites using support vector machines. *Protein Engineering Design and Selection*, 17(2), pp.165–173.

Konc, J. & Janezic, D., 2007. Protein-protein binding-sites prediction by protein surface structure conservation. *Journal of chemical information and modeling*, 47(3), pp.940–944.

Konc, J. & Janežič, D., 2010a. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, 26(9), pp.1160–1168.

Konc, J. & Janežič, D., 2010b. ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic acids research*, 38(suppl 2), pp.W436–W440.

Korkin, D. et al., 2006. Structural modeling of protein interactions by analogy: application to PSD-95. *PLoS computational biology*, 2(11), p.e153.

Korkin, D., Davis, F.P. & Sali, A., 2005. Localization of protein-binding sites within families of proteins. *Protein science*, 14(9), pp.2350–2360.

Kowalsman, N. & Eisenstein, M., 2009. Combining interface core and whole interface descriptors in postscan processing of protein-protein docking models. *Proteins: Structure, Function, and Bioinformatics*, 77(2), pp.297–318.

Kozakov, D. et al., 2010. Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. *Proteins: Structure, Function, and Bioinformatics*, 78(15), pp.3124–3130.

Kozakov, D. et al., 2006. PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins: Structure, Function, and Bioinformatics*, 65(2), pp.392–406.

Krishnamoorthy, B. & Tropsha, A., 2003. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12), pp.1540–1548.

Krissinel, E., 2012. Enhanced fold recognition using efficient short fragment clustering. *Journal of Molecular Biochemistry*, 1(2).

Krissinel, E. & Henrick, K., 2004. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D: Biological Crystallography*, 60(12), pp.2256–2268.

Kufareva, I. et al., 2007. PIER: protein interface recognition for structural proteomics. *Proteins: Structure, Function, and Bioinformatics*, 67(2), pp.400–417.

Kuhn, H.W., 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2), pp.83–97.

Kundrotas, P.J. et al., 2012. Templates are available to model nearly all complexes of structurally characterized proteins. *Proceedings of the National Academy of Sciences*, 109(24), pp.9438–9441.

Kundrotas, P.J. & Alexov, E., 2006. Predicting 3D structures of transient protein-protein complexes by homology. *Biochimica et biophysica acta*, 1764(9), p.1498.

Kundrotas, P.J., Lensink, M.F. & Alexov, E., 2008. Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles. *International journal of biological macromolecules*, 43(2), pp.198–208.

Kundrotas, P.J. & Vakser, I.A., 2010. Accuracy of protein-protein binding sites in high-throughput template-based modeling. *PLoS computational biology*, 6(4), p.e1000727.

Kuo, P., Makris, D. & Nebel, J.C., 2011. Integration of bottom-up/top-down approaches for 2D pose estimation using probabilistic Gaussian modelling. *Computer Vision and Image Understanding*, 115(2), pp.242–255.

Kuzu, G. et al., 2012. Constructing structural networks of signaling pathways on the proteome scale. *Current opinion in structural biology*, 22(3), pp.367–377.

Lamdan, Y. & Wolfson, H.J., 1988. Geometric hashing: A general and efficient model-based recognition scheme. In *ICCV*. pp. 238–249.

Launay, G. & Simonson, T., 2008. Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. *BMC bioinformatics*, 9(1), p.427.

Lawrence, M.C. & Colman, P.M., 1993. Shape complementarity at protein/protein interfaces. *Journal of molecular biology*, 234(4), pp.946–950.

Lazaridis, T., 2005. Implicit solvent simulations of peptide interactions with anionic lipid membranes. *Proteins: Structure, Function, and Bioinformatics*, 58(3), pp.518–527.

Leeuw, E. de et al., 2010. Functional interaction of human neutrophil peptide-1 with the cell wall precursor lipid II. *FEBS letters*, 584(8), pp.1543–1548.

Lensink, M.F., Méndez, R. & Wodak, S.J., 2007. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.704–718.

Lensink, M.F. & Wodak, S.J., 2010. Docking and scoring protein interactions: CAPRI 2009. *Proteins: Structure, Function, and Bioinformatics*, 78(15), pp.3073–3084.

Lesk, V.I. & Sternberg, M.J.E., 2008. 3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. *Bioinformatics*, 24(9), pp.1137–1144.

Li, B. & Kihara, D., 2012. Protein docking prediction using predicted protein-protein interface. *BMC bioinformatics*, 13(1), p.7.

Li, B.-Q. et al., 2012. Prediction of protein-Protein interaction sites by random forest algorithm with mRMR and IFS. *PloS one*, 7(8), p.e43927.

Li, J.-J. et al., 2006. Identifying protein--protein interfacial residues in heterocomplexes using residue conservation scores. *International journal of biological macromolecules*, 38(3), pp.241–247.

Li, L., Chen, R. & Weng, Z., 2003. RDOCK: Refinement of rigid-body protein docking predictions. *Proteins: Structure, Function, and Bioinformatics*, 53(3), pp.693–707.

Li, M.-H. et al., 2007. Protein-protein interaction site prediction based on conditional random fields. *Bioinformatics*, 23(5), pp.597–604.

Li, N., Sun, Z. & Jiang, F., 2008. Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC bioinformatics*, 9(1), p.553.

Li, X. & Liang, J., 2005. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. *Proteins: Structure, Function, and Bioinformatics*, 60(1), pp.46–65.

Liang, S. et al., 2009. Consensus scoring for enriching near-native structures from protein-protein docking decoys. *Proteins: Structure, Function, and Bioinformatics*, 75(2), pp.397–403.

Liang, S. et al., 2006. Protein binding site prediction using an empirical scoring function. *Nucleic acids research*, 34(13), pp.3698–3707.

Lichtarge, O., Bourne, H.R. & Cohen, F.E., 1996. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257(2), pp.342–358.

Lin, N. et al., 2004. Information assessment on predicting protein-protein interactions. *BMC bioinformatics*, 5(1), p.154.

Littler, S.J. & Hubbard, S.J., 2005. Conservation of orientation and sequence in protein domain-domain interactions. *Journal of molecular biology*, 345(5), pp.1265–1279.

Liu, B. et al., 2009. Prediction of protein binding sites in protein structures using hidden Markov support vector machine. *BMC bioinformatics*, 10(1), p.381.

Liu, S., Li, Q. & Lai, L., 2006. A combinatorial score to distinguish biological and nonbiological protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 64(1), pp.68–78.

Liu, S. & Vakser, I.A., 2011. DECK: Distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC bioinformatics*, 12(1), p.280.

Lo, S.L. et al., 2005. Effect of training datasets on support vector machine prediction of protein-protein interactions. *Proteomics*, 5(4), pp.876–884.

Lovell, S.C. & Robertson, D.L., 2010. An integrated view of molecular coevolution in protein-protein interactions. *Molecular biology and evolution*, 27(11), pp.2567–2575.

Lu, H., Lu, L. & Skolnick, J., 2003. Development of unified statistical potentials describing protein-protein interactions. *Biophysical journal*, 84(3), pp.1895–1901.

Lu, L. et al., 2003. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome. *Genome research*, 13(6a), pp.1146–1154.

Lu, L., Lu, H. & Skolnick, J., 2002. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins: Structure, Function, and Bioinformatics*, 49(3), pp.350–364.

Ma, B. et al., 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences*, 100(10), pp.5772–5777.

Madabushi, S. et al., 2004. Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. *Journal of Biological Chemistry*, 279(9), pp.8126–8132.

Madaoui, H. & Guerois, R., 2008. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proceedings of the National Academy of Sciences*, 105(22), pp.7708–7713.

Mandell, J.G. et al., 2001. Protein docking using continuum electrostatics and geometric fit. *Protein Engineering*, 14(2), pp.105–113.

Martin, J., 2010. Beauty is in the eye of the beholder: proteins can recognize binding sites of homologous proteins in more than one way. *PLoS computational biology*, 6(6), p.e1000821.

Martin, O. & Schomburg, D., 2008. Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 70(4), pp.1367–1378.

Mashiach, E., Nussinov, R. & Wolfson, H.J., 2010. FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins: Structure, Function, and Bioinformatics*, 78(6), pp.1503–1519.

Méndez, R. et al., 2005. Assessment of CAPRI predictions in rounds 3--5 shows progress in docking procedures. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.150–169.

Mintseris, J. et al., 2007. Integrating statistical pair potentials into protein complex prediction. *Proteins: Structure, Function, and Bioinformatics*, 69(3), pp.511–520.

Mintseris, J. et al., 2005. Protein--protein docking benchmark 2.0: an update. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.214–216.

Morris, G.M. et al., 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of computational chemistry*, 19(14), pp.1639–1662.

Mosca, R., Céol, A. & Aloy, P., 2012. Interactome3D: adding structural details to protein networks. *Nature methods*, 10(1), pp.47–53.

Movshovitz-Attias, D., London, N. & Schueler-Furman, O., 2010. On the use of structural templates for high-resolution docking. *Proteins: Structure, Function, and Bioinformatics*, 78(8), pp.1939–1949.

Mukherjee, S. & Zhang, Y., 2011. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure*, 19(7), pp.955–966.

Munkres, J., 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1), pp.32–38.

Murakami, Y. & Jones, S., 2006. SHARP2: protein-protein interaction predictions using patch analysis. *Bioinformatics*, 22(14), pp.1794–1795.

Murakami, Y. & Mizuguchi, K., 2010. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics*, 26(15), pp.1841–1848.

Murphy, J. et al., 2003. Combination of scoring functions improves discrimination in protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 53(4), pp.840–854.

Murzin, A.G. et al., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4), pp.536–540.

Nakashima, H. et al., 1993. Defensins inhibit HIV replication in vitro. *Aids*, 7(8), p.1129.

Nassar, T. et al., 2002. Human α-defensin regulates smooth muscle cell contraction: a role for low-density lipoprotein receptor-related protein/α2-macroglobulin receptor. *Blood*, 100(12), pp.4026–4032.

Nebel, J.-C., 2006. Generation of 3D templates of active sites of proteins with rigid prosthetic groups. *Bioinformatics*, 22(10), pp.1183–1189.

Nebel, J.-C., 2012. Proteomics and bioinformatics soon to resolve the human structural interactome. *Journal of Proteomics & Bioinformatics*, 5, pp.xi–xii.

Nebel, J.-C., Herzyk, P. & Gilbert, D.R., 2007. Automatic generation of 3D motifs for classification of protein binding sites. *BMC bioinformatics*, 8(1), p.321.

Negi, S.S. et al., 2007. InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics*, 23(24), pp.3397–3399.

Neuvirth, H. et al., 2004. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *Journal of molecular biology*, 338(1), p.181.

Neuvirth, H. et al., 2007. ProMateus—an open research approach to protein-binding sites analysis. *Nucleic acids research*, 35(suppl 2), pp.W543–W548.

Ngan, S.-C., Inouye, M.T. & Samudrala, R., 2006. A knowledge-based scoring function based on residue triplets for protein structure prediction. *Protein Engineering Design and Selection*, 19(5), pp.187–193.

Nguyen, M.N. & Rajapakse, J.C., 2006. Protein-protein interface residue prediction with SVM using evolutionary profiles and accessible surface areas. In *Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB'06. 2006 IEEE Symposium on.* pp. 1–5.

Nussinov, R. & Schreiber, G., 2010. *Computational protein-protein interactions* B. Raton, ed., CRC Press.

Ofran, Y. & Rost, B., 2003a. Analysing six types of protein-protein interfaces. *Journal of molecular biology*, 325(2), pp.377–387.

Ofran, Y. & Rost, B., 2007a. ISIS: interaction sites identified from sequence. *Bioinformatics*, 23(2), pp.e13–e16.

Ofran, Y. & Rost, B., 2003b. Predicted protein-protein interaction sites from local sequence information. *Febs Letters*, 544(1), pp.236–239.

Ofran, Y. & Rost, B., 2007b. Protein-protein interaction hotspots carved into sequences. *PLoS computational biology*, 3(7), p.e119.

Ogmen, U. et al., 2005. PRISM: protein interactions by structural matching. *Nucleic acids research*, 33(suppl 2), pp.W331–W336.

Oliva, R., Vangone, A. & Cavallo, L., 2013. Ranking multiple docking solutions based on the conservation of inter-residue contacts. *Proteins: Structure, Function, and Bioinformatics*, 81(9), pp.1571–1584.

Othersen, O.G. et al., 2012. Application of information theory to feature selection in protein docking. *Journal of molecular modeling*, 18(4), pp.1285–1297.

Palma, P.N. et al., 2000. BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 39(4), pp.372–384.

Pande, J., Szewczyk, M.M. & Grover, A.K., 2010. Phage display: concept, innovations, applications and future. *Biotechnology advances*, 28(6), pp.849–858.

Pazos, F. et al., 1997. Correlated mutations contain information about protein-protein interaction. *Journal of molecular biology*, 271(4), pp.511–523.

Petrey, D., Fischer, M. & Honig, B., 2009. Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proceedings of the National Academy of Sciences*, 106(41), pp.17377–17382.

Petrey, D. & Honig, B., 2003. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods in enzymology*, 374, pp.492–509.

Pettit, F.K. et al., 2007. HotPatch: a statistical a pproach to finding biologically relevant features on protein surfaces. *Journal of molecular biology*, 369(3), pp.863–879.

Pierce, B. & Weng, Z., 2007. ZRANK: reranking protein docking predictions with an optimized energy function. *PROTEINS: Structure, Function, and Bioinformatics*, 67(4), pp.1078–1086.

Pierce, B.G., Hourai, Y. & Weng, Z., 2011. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one*, 6(9), p.e24657.

Ponomarev, S.Y. & Audie, J., 2011. Computational prediction and analysis of the DR6--NAPP interaction. *Proteins: Structure, Function, and Bioinformatics*, 79(5), pp.1376–1395.

Pons, C. et al., 2010. Present and future challenges and limitations in protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 78(1), pp.95–108.

Porollo, A. & Meller, J., 2006. Prediction-based fingerprints of protein-protein interactions. *PROTEINS: Structure, Function, and Bioinformatics*, 66(3), pp.630–645.

Poupon, A., 2004. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Current opinion in structural biology*, 14(2), pp.233–241.

Qin, S. & Zhou, H.X., 2007a. A holistic approach to protein docking. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.743–749.

Qin, S. & Zhou, H.X., 2007b. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, 23(24), pp.3386–3387.

Qiu, Z. & Wang, X., 2012. Prediction of protein-protein interaction sites using patch-based residue characterization. *Journal of Theoretical Biology*, 293, pp.143–150.

Quinn, K. et al., 2008. Human neutrophil peptides: a novel potential mediator of inflammatory cardiovascular diseases. *American Journal of Physiology-Heart and Circulatory Physiology*, 295(5), pp.H1817–H1824.

Rausell, A. et al., 2010. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proceedings of the National Academy of Sciences*, 107(5), pp.1995–2000.

Ravikant, D.V.S. & Elber, R., 2010. PIE—Efficient filters and coarse grained potentials for unbound protein--protein docking. *Proteins: Structure, Function, and Bioinformatics*, 78(2), pp.400–419.

Reš, I., Mihalek, I. & Lichtarge, O., 2005. An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, 21(10), pp.2496–2501.

Ritchie, D.W., 2008. Recent progress and future directions in protein-protein docking. *Current Protein and Peptide Science*, 9(1), pp.1–15.

Ritchie, D.W. & Kemp, G.J.L., 2000. Protein docking using spherical polar Fourier correlations. *Proteins: Structure, Function, and Bioinformatics*, 39(2), pp.178–194.

Rudenko, G. et al., 2002. Structure of the LDL receptor extracellular domain at endosomal pH. *Science*, 298(5602), pp.2353–2358.

Rudenko, G. & Deisenhofer, J., 2003. The low-density lipoprotein receptor: ligands, debates and lore. *Current opinion in structural biology*, 13(6), pp.683–689.

Russell, R.B. et al., 2004. A structural perspective on protein-protein interactions. *Current opinion in structural biology*, 14(3), pp.313–324.

Russell, R.B., Sasieni, P.D. & Sternberg, M.J.E., 1998. Supersites within superfolds. Binding site similarity in the absence of homology. *Journal of molecular biology*, 282(4), pp.903–918.

Sali, A. et al., 2003. From words to literature in structural proteomics. *Nature*, 422(6928), pp.216–225.

Sánchez, I.E. et al., 2008. Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. *PLoS computational biology*, 4(4), p.e1000052.

Savojardo, C. et al., 2012. Machine-Learning Methods to Predict Protein Interaction Sites in Folded Proteins. In *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer, pp. 127–135.

Scheffzek, Klaus and Wittinghofer, A., 2001. Structural Views of the Ran GTPase Cycle. In P. Rush, Mark and D'Eustachio, ed. *The Small GTPase Ran*. pp. 177–201. Available at: http://dx.doi.org/10.1007/978-1-4615-1501-2_10.

Schisterman, E.F. et al., 2005. Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, 16(1), pp.73–81.

Schneider, J.J. et al., 2005. Human defensins. *Journal of molecular medicine*, 83(8), pp.587–595.

Schneider, S. & Zacharias, M., 2012. Scoring optimisation of unbound protein-protein docking including protein binding site predictions. *Journal of Molecular Recognition*, 25(1), pp.15–23.

Schneidman-Duhovny, D. et al., 2012. A method for integrative structure determination of protein-protein complexes. *Bioinformatics*, 28(24), pp.3282–3289.

Schneidman-Duhovny, D. et al., 2005a. Geometry-based flexible and symmetric protein docking. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.224–231.

Schneidman-Duhovny, D. et al., 2005b. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research*, 33(suppl 2), pp.W363–W367.

Schneidman-Duhovny, D. et al., 2003. Taking geometry to its edge: Fast unbound rigid (and hinge-bent) docking. *Proteins: Structure, Function, and Bioinformatics*, 52(1), pp.107–112.

Schrödinger, LLC, 2010. The {PyMOL} Molecular Graphics System, Version~1.3r1.

Schuck, P., 2007. *Protein interactions: biophysical approaches for the study of complex reversible systems*, Springer.

Schymkowitz, J. et al., 2005. The FoldX web server: an online force field. *Nucleic Acids Research*, 33(suppl 2), pp.W382–W388.

Segura, J., Jones, P.F. & Fernandez-Fuentes, N., 2012. A holistic in silico approach to predict functional sites in protein structures. *Bioinformatics*, 28(14), pp.1845–1850.

Segura, J., Jones, P.F. & Fernandez-Fuentes, N., 2011. Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi Diagrams. *BMC bioinformatics*, 12(1), p.352.

Shatsky, M., Nussinov, R. & Wolfson, H.J., 2004. A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*, 56(1), pp.143–156.

Sheinerman, F.B., Norel, R. & Honig, B., 2000. Electrostatic aspects of protein-protein interactions. *Current opinion in structural biology*, 10(2), pp.153–159.

Shen, M. & Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. *Protein science*, 15(11), pp.2507–2524.

Shen, Y. et al., 2007. Docking with PIPER and refinement with SDU in rounds 6--11 of CAPRI. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.734–742.

Shih, E.S.C. & Hwang, M.-J., 2013. A Critical Assessment of Information-guided Protein-Protein Docking Predictions. *Molecular & Cellular Proteomics*, 12(3), pp.679–686.

Shoemaker, B.A. et al., 2010. Inferred Biomolecular Interaction Server—a web server to analyze and predict protein interacting partners and binding sites. *Nucleic acids research*, 38(suppl 1), pp.D518–D524.

Shoemaker, B.A. & Panchenko, A.R., 2007. Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS computational biology*, 3(3), p.e42.

Shoemaker, B.A., Panchenko, A.R. & Bryant, S.H., 2006. Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein science*, 15(2), pp.352–361.

Shulman-Peleg, A. et al., 2008. MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions. *Nucleic acids research*, 36(suppl 2), pp.W260–W264.

Shulman-Peleg, A., Nussinov, R. & Wolfson, H.J., 2004. Recognition of functional sites in protein structures. *Journal of molecular biology*, 339(3), pp.607–633.

Šikić, M., Tomić, S. & Vlahoviček, K., 2009. Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. *PLoS computational biology*, 5(1), p.e1000278.

Sinha, R., Kundrotas, P.J. & Vakser, I.A., 2010. Docking by structural similarity at protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 78(15), pp.3235–3241.

Sippl, M.J., 1995. Knowledge-based potentials for proteins. *Current opinion in structural biology*, 5(2), pp.229–235.

Sivasubramanian, A. et al., 2006. Structural model of the mAb 806-EGFR complex using computational docking followed by computational and experimental mutagenesis. *Structure*, 14(3), pp.401–414.

Smith, G.R. & Sternberg, M.J.E., 2002. Prediction of protein-protein interactions by docking methods. *Current opinion in structural biology*, 12(1), pp.28–35.

Smith, G.R., Sternberg, M.J.E. & Bates, P.A., 2005. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *Journal of molecular biology*, 347(5), pp.1077–1101.

Sobolev, V. et al., 2005. SPACE: a suite of tools for protein structure prediction and analysis based on complementarity and environment. *Nucleic acids research*, 33(suppl 2), pp.W39–W43.

Del Sol Mesa, A., Pazos, F. & Valencia, A., 2003. Automatic methods for predicting functionally important residues. *Journal of molecular biology*, 326(4), pp.1289–1302.

Soman, S.S., Sivakumar, K.C. & Sreekumar, E., 2010. Molecular dynamics simulation studies and in vitro site directed mutagenesis of avian beta-defensin Apl_AvBD2. *BMC bioinformatics*, 11(Suppl 1), p.S7.

Stawiski, E.W., Gregoret, L.M. & Mandel-Gutfreund, Y., 2003. Annotating nucleic acid-binding function based on protein structure. *Journal of molecular biology*, 326(4), pp.1065–1079.

Strynadka, Ncj. et al., 1996. Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. *Nature Structural & Molecular Biology*, 3(3), pp.233–239.

Sumikoshi, K. et al., 2005. A fast protein-protein docking algorithm using series expansion in terms of spherical basis functions. *GENOME INFORMATICS SERIES*, 16(2), p.161.

Summa, C.M., Levitt, M. & DeGrado, W.F., 2005. An atomic environment potential for use in protein structure prediction. *Journal of molecular biology*, 352(4), pp.986–1001.

Suzek, B.E. et al., 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10), pp.1282–1288.

Talavera, D., Robertson, D.L. & Lovell, S.C., 2011. Characterization of protein-protein interaction interfaces from a single species. *PloS one*, 6(6), p.e21053.

Terashi, G. et al., 2007. The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.866–872.

Thompson, J.D., Higgins, D.G. & Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22), pp.4673–4680.

Tjong, H., Qin, S. & Zhou, H.-X., 2007. PI2PE: protein interface/interior prediction engine. *Nucleic acids research*, 35(suppl 2), pp.W357–W362.

Tjong, H. & Zhou, H.-X., 2007. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic acids research*, 35(5), pp.1465–1477.

Tobi, D. & Bahar, I., 2006. Optimal design of protein docking potentials: efficiency and limitations. *Proteins: Structure, Function, and Bioinformatics*, 62(4), pp.970–981.

Tobias, D.J., 2001. Electrostatics calculations: recent methodological advances and applications to membranes. *Current opinion in structural biology*, 11(2), pp.253–261.

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

Totrov, M. & Abagyan, R., 1994. Detailed ab initio prediction of lysozyme-antibody complex with 1.6Å accuracy. *Nature Structural & Molecular Biology*, 1(4), pp.259–263.

Tovchigrechko, A. & Vakser, I.A., 2005. Development and testing of an automated approach to protein docking. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.296–301.

Tress, M. et al., 2005. Scoring docking models with evolutionary information. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.275–280.

Tsai, C.-J. et al., 1996. Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Critical reviews in biochemistry and molecular biology*, 31(2), pp.127–152.

Tuncbag, N. et al., 2012. Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement. *Proteins: Structure, Function, and Bioinformatics*, 80(4), pp.1239–1249.

Tuncbag, N. et al., 2011. Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature protocols*, 6(9), pp.1341–1354.

Tyagi, M. et al., 2012. Homology Inference of Protein-Protein Interactions via Conserved Binding Sites. *PloS one*, 7(1), p.e28896.

Vakser, I.A., 2013. Low-resolution structural modeling of protein interactome. *Current opinion in structural biology*, 23(2), pp.198–205.

Vakser, I.A., 1996. Main-chain complementarity in protein-protein recognition. *Protein Engineering*, 9(9), pp.741–744.

Venkatesan, K. et al., 2008. An empirical framework for binary interactome mapping. *Nature methods*, 6(1), pp.83–90.

Verdaguer, N. et al., 2004. X-ray structure of a minor group human rhinovirus bound to a fragment of its cellular receptor protein. *Nature structural & molecular biology*, 11(5), pp.429–434.

Viswanath, S., Ravikant, D.V.S. & Elber, R., 2012. Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins: Structure, Function, and Bioinformatics*, 81(4), pp.592–606.

Vreven, T. et al., Evaluating template-based and template-free protein-protein complex structure prediction. *Brief Bioinform first published online July 1, 2013 doi:10.1093/bib/bbt047.*

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

De Vries, S.J. et al., 2010. Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins: Structure, Function, and Bioinformatics*, 78(15), pp.3242–3249.

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

De Vries, S.J. & Bonvin, A.M.J.J., 2008. How proteins get in touch: interface prediction in the study of biomolecular complexes. *Current Protein and Peptide Science*, 9(4), pp.394–406.

De Vries, S.J., Van Dijk, A.D.J. & Bonvin, A.M.J.J., 2006. WHISCY: What information does surface conservation yield? Application to data-driven docking. *Proteins: Structure, Function, and Bioinformatics*, 63(3), pp.479–489.

Wang, B. et al., 2006. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS letters*, 580(2), pp.380–384.

Wang, B., Chen, P. & Zhang, J., 2012. Protein interface residues prediction based on amino acid properties only. In *Bio-Inspired Computing and Applications*. Springer, pp. 448–452.

Wang, T., 2012. Iterative Closest Point algorithm-point cloud/mesh registration | My point cloud on WordPress.com. Available at: http://taylorwang.wordpress.com/2012/04/06/iterative-closest-point-algorithm-point-cloudmesh-registration/ [Accessed August 14, 2013].

Wei, G. et al., 2009. Through the looking glass, mechanistic insights from enantiomeric human defensins. *Journal of Biological Chemistry*, 284(42), pp.29180–29192.

Wei, G. et al., 2010. Trp-26 imparts functional versatility to human α-defensin HNP1. *Journal of Biological Chemistry*, 285(21), pp.16275–16285.

Van Wetering, S., Tjabringa, G.S. & Hiemstra, P.S., 2005. Interactions between neutrophil-derived antimicrobial peptides and airway epithelial cells. *Journal of leukocyte biology*, 77(4), pp.444–450.

Winter, C. et al., 2006. SCOPPI: a structural classification of protein-protein interfaces. *Nucleic acids research*, 34(suppl 1), pp.D310–D314.

Wodak, S.J. & Méndez, R., 2004. Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Current opinion in structural biology*, 14(2), pp.242–249.

Wolfson, H.J. et al., 2005. From structure to function: methods and applications. *Current Protein and Peptide Science*, 6(2), pp.171–183.

Xie, C. et al., 2005. Reconstruction of the conserved β-bulge in mammalian defensins using D-amino acids. *Journal of Biological Chemistry*, 280(38), pp.32921–32929.

Xu, D., Li, H. & Gu, T., 2007. Protein structure superposition by curve moment invariants and iterative closest point. In *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on*. pp. 25–28.

Xue, L.C. et al., DockRank: ranking docked conformations using partner-specific sequence homology based protein interface prediction. *Proteins: Structure, Function, and Bioinformatics, 2013 Jul 20. doi: 10.1002/prot.24370.*

Xue, L.C. et al., 2010. Ranking Docked Models of Protein-Protein Complexes Using Predicted Partner-Specific Protein-Protein Interfaces: A Preliminary Study. In *2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. pp. 441–445.

Xue, L.C., Dobbs, D. & Honavar, V., 2011. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*, 12(1), p.244.

Yan, C., Dobbs, D. & Honavar, V., 2004. A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, 20(suppl 1), pp.i371–i378.

Yan, Z. et al., 2013. Specificity and affinity quantification of protein-protein interactions. *Bioinformatics*, 29(9), pp.1127–1133.

Yang, A.-S. & Honig, B., 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of molecular biology*, 301(3), pp.665–678.

Zacharias, M., 2003. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science*, 12(6), pp.1271–1282.

Zellner, H. et al., 2012. Prescont: Predicting protein-protein interfaces utilizing four residue properties. *Proteins: Structure, Function, and Bioinformatics*, 80(1), pp.154–168.

Zhang, C. et al., 2004. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein science*, 13(2), pp.400–411.

Zhang, C. et al., 1997. Determination of atomic desolvation energies from the structures of crystallized proteins. *Journal of molecular biology*, 267(3), pp.707–726.

Zhang, C., Liu, S. & Zhou, Y., 2005. Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function. *Proteins: Structure, Function, and Bioinformatics*, 60(2), pp.314–318.

Zhang, Q.C. et al., 2011. PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Research*, 39(suppl 2), pp.W283–W287.

Zhang, Q.C. et al., 2013. PrePPI: a structure-informed database of protein--protein interactions. *Nucleic acids research*, 41(D1), pp.D828–D833.

Zhang, Q.C. et al., 2010. Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences*, 107(24), pp.10896–10901.

Zhang, Q.C. et al., 2012. Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature*, 490(7421), pp.556–560.

Zhang, Y., Lu, W. & Hong, M., 2010. The membrane-bound structure and topology of a human α-defensin indicate a dimer pore mechanism for membrane disruption. *Biochemistry*, 49(45), pp.9770–9782.

Zhao, N. et al., 2011. Feature-based classification of native and non-native protein-protein interactions: Comparing supervised and semi-supervised learning approaches. *Proteomics*, 11(22), pp.4321–4330.

Zhou, H.X. & Qin, S., 2007. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17), pp.2203–2209.

Zhou, H.X. & Shan, Y., 2001. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Structure, Function, and Bioinformatics*, 44(3), pp.336–343.