Theoretical and experimental restraints to drive the docking of protein-protein complexes

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Abbas Alameer BSc (Kuwait)

Department of Biochemistry

University of Leicester

March 2012

Abbas Alameer
Title: Theoretical and experimental restraints to drive the docking of protein-protein complexes.

Abstract

Biological processes are frequently driven by protein-protein interactions. The number of known protein interactions is much higher than the number of known protein complex structures. To bridge this gap, data-driven protein-protein docking utilizing experimental or theoretical restraints is applied. In this study the PROTIN_ID method for generating theoretical docking restraints is introduced. PROTIN_ID generates residue clusters on the protein surface based on sequence conservation. Compared to WHISCY and CCRXP, PROTIN_ID performs equally well or better. Furthermore, PROTIN_ID has user-friendly features such as the ability to improve the quality of sequence alignments, which improves its performance, and automatically utilizing up-to-date sequence data for experimentally determined proteins or homology models to generate theoretical restraints. A webserver version of PROTIN_ID was implemented for the academic community.

Statistical analyses of the conservation of interface residues using the latest version of Benchmark4.0 demonstrated that interface residues are more conserved than non-interface residues. The application of spatial clustering of residues is more efficient to exploit the conservation signal of interface residues, resulting in reliable predictions that are better than predictions generated by ‘non-clustering’ or at random.

Theoretical restraints derived from PROTIN_ID were applied to drive docking and compared to *ab initio* docking, demonstrating that data-driven docking was more successful. Combining theoretical and experimental restraints to drive docking was compared to experimental-data driven docking. It was shown that combined restraints-driven docking improved because of increased interface residue recall, demonstrating that consensus-data is possibly useful for improvement of docking performance.

Acknowledgements

I would like to express my immense gratitude to my supervisors Dr. Ralf Schmid and Prof. Mark Carr for giving me the opportunity to conduct this research, and for their valuable guidance, support, and encouragement during my PhD studies. I would also like to thank my committee members Dr. Mark Pfuhl and Dr. Cyril Dominguez for their valuable assistance and advice. I would also like to extend my thanks to Kuwait University for funding this studentship.

I would like to thank and acknowledge all those whose advice and suggestions assisted me during my research. In particular, I thank Dr. Vaclav Veverka, Dr. Lorna Waters, and Dr. Catherine Hall for their invaluable help in applying NMR data in protein docking. Also, I express my thanks to Dr. Wei-cheng Huang for providing me with helpful input and explanations about programming. I would like to offer my deepest thanks to Dr. Seyyed Imran Shah for his true friendship, kindness, and support during difficulties I encountered by reminding me that ‘knowledge gives life to the soul’. I would like to thank Dr. Imitaz Shafiq for his friendship during my studies. Also, I thank Dr. Philip Renshaw, Dr. Kirsty Lightbody, Dr. Ian Wilkinson and all present and former members of the Structural Biology group for their helpful support.

I sincerely thank my dear friend Seyyed Al-Mahdi and his great family for always being in my corner. I also express my deepest thanks to my friend Mohammad for encouraging me when I needed it most. I thank my in-laws as well for their support. My sincerest thanks go to my loving parents and family who have been with me all the way with their love and support. Finally, I would like to thank Maryam, the love of my life, for her love, kindness and support. In addition, my thanks go to my dear and lovely son Hassan for making me a proud father.

Abbas Alameer, Leicester, March 2012.

Abbreviations

ASA	Accessible surface area
CI	Confidence interval
CSP	Chemical shift perturbation
FP	False positive
FN	False negative
IPA	Intervector projection angles
MSA	Multiple sequence alignment
NOC	Number of correct models
RDC	Residual dipolar couplings
ROS	Rest of surface
TN	True negative
TP	True positive

List of Contents

Chapter 2, Preliminary Work

Chapter 3, Methods

Chapter 1

Introduction

1.1 Proteins

Proteins are composed of amino acid building blocks linked together through peptide bonds, forming a polypeptide chain known as the primary structure. The primary structure forms secondary structural elements, such as alpha helices and beta sheets via hydrogen bonds, which in turn interact to create the tertiary structure through protein folding. The tertiary structure may combine with another to form a quaternary structure (see Figure 1-1). The outcome is a protein with functional capabilities whose active site or interface, where biomolecular interaction occurs, is composed of amino acids that may be far apart on the primary structure-level but are close together because of the unique three-dimensional arrangement of the final tertiary or quaternary structures. Broadly, proteins are classified as globular proteins, which function in cellular activities (ex. enzymes) and fibrous proteins (ex. keratin) that assume a structural role. The holistic number of proteins produced from the genome of an organism is known as the proteome. Unless otherwise stated, “protein” will refer to globular proteins in this chapter.

1.2 Protein-protein interactions

Proteins compose the molecular machinery of cells and the association of a protein with others to form temporary or long-term functional complexes is fundamental in numerous biological processes. The region of specific binding between two or more proteins’ residues is known as the interface. Protein-protein interactions are diverse in functionality, highlighting their importance. For example, protein-protein interactions are involved in signalling in the bacterial phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS). The PTS system involves transfer of a phosphate derived from phosphoenolpyruvate to proteins of this pathway ultimately leading to phosphorylation of sugars coupled with their transfer through bacterial membranes

(Cornilescu *et al*., 2002; Wang *et al*., 2000). Protein interactions are also involved in enzymatic inhibitory activities such as inhibition of matrix metalloproteinases (MMPs) by tissue inhibitor of metalloproteinases (TIMP). MMPs are essential for breakdown of extracellular matrix constituents during embryogenesis, tissue regeneration, for example (Arumugam and Van Doren, 2003a; Gomis-Rüth *et al*., 1997). TIMPs regulate these enzymes by forming non-covalent inhibitory complexes (Williamson *et al*. 1997). A disturbance in this regulation favouring heightened activity of MMPs can result in pathophysiological conditions such as cardiovascular diseases like myocardial infarctions and aneurisms (Hovsepian *et al*., 2000). In addition, protein-protein interaction is important for the creation of multi-protein assemblies that perform specific tasks. For instance, DNA polymerases, DNA helicase, and DNA primase with other accessory proteins assemble into the replisome that undertakes the action of reproducing DNA during the DNA replication process, which is part of cell division (Perumal *et al*., 2011; Marians, 2008). Proteins are also involved in the degradation of others. As an example, in the ubiquitin proteasome pathway (UPP), ubiquitin protein molecules are linked via ubiquitin ligases to proteins. These ubiquitinated proteins are degraded into peptides by a multi-protein complex composed of multiple catalytic sites called the proteasome (Hershko *et al*., 1984; Hershko *et al*., 1983). Protein-protein interactions are important in immune responses. For example, T-cell activation is achieved through T-cell receptor interaction with antigens in complex with major histocompatibility complex proteins, leading to an immune response (Aleksic *et al*., 2010). The differing varieties of protein interactions are a part of an interaction network between different proteins and other biomolecules called the interactome. For example, in human cells an estimated number of 130,000 ± 32,000 binary interactions may occur, and there are 137,713 interactions currently known as presented in the BioGRID database, which is a repository for interaction data (Bonetta, 2010; Stark *et al*., 2010; Venkatesan *et al*., 2009). The number of binary interactions is higher than the estimated number of proteins (100,000) in a human cell (International Human Genome Sequencing Consortium, 2004). Indeed, the total protein-protein interactions may be as high as 375,000 based on an estimated 15 interactions per protein (Ramani *et al*., 2005). Protein interactions are undisputedly important and their disruption can lead to interruption of fundamental biological mechanisms, causing diseases.

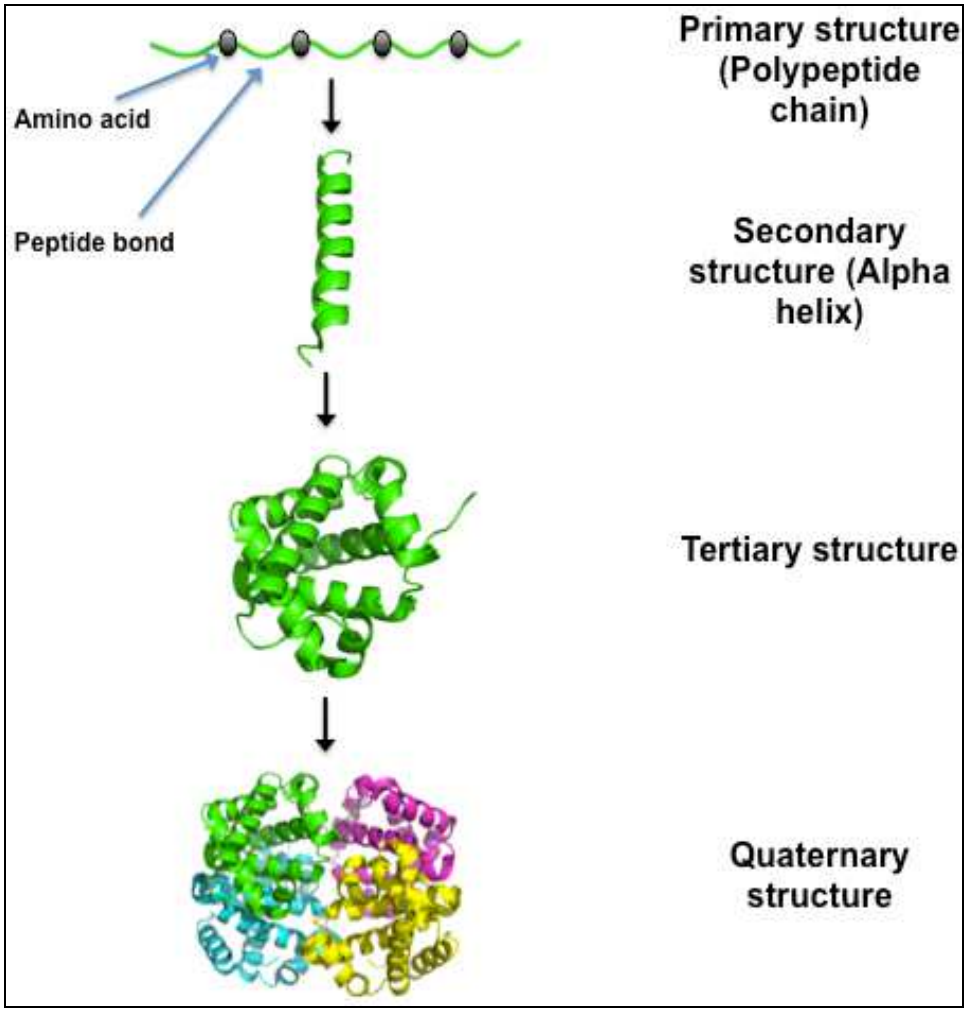

Figure 1-1: Examples of the primary, secondary, tertiary, and quaternary structures. The primary structure is composed of amino acids connected by peptide bonds, forming a polypeptide chain. Secondary structures (alpha helix) form through hydrogen bond interactions in the primary structure. This interaction of secondary structural elements leads to a tertiary structure, which may combine with others to form a quaternary structure. The example quaternary structure is human haemoglobin (PDB: 1MKO).

Therefore, it is vital to understand the different facets of proteins that deal with their interactions with others in order to comprehend protein complex formation and by extension biological processes (Keskin *et al*., 2004). This allows a wide-ranging understanding of the inner workings of the interactome as a whole (Spirin and Mirny, 2003). The sheer magnitude of differing proteins known to interact and those interactions yet to be discovered highlights the growing importance of understanding protein-protein interactions, which are important for all living things to function and exist (Alloy and Russell, 2004). For example, knowledge of protein-protein complex structure is important for drug screening and design, allowing targeting of those complexes linked to cancer to inhibit them (Lessene *et al*., 2013).

1.3 Classification of protein-protein interaction types

There are three main features that are employed to classify protein-protein interactions. These are based on protein complex composition, structural subsistence, and protein interaction lifetime (Ozbabacan *et al*., 2011). An overview of this classification of protein-protein interaction types will be discussed in this section. Figure 1-2 presents a summary of the classification of protein-protein interaction types.

1.3.1 Homo- and hetero-oligomers

This grouping of complexes is based on composition of the subunits of a complex. A protein complex where non-identical monomers compose it is referred to as a hetero-oligomer (Ozbabacan *et al*., 2011). A protein complex in which only identical subunits compose it is referred to as a homo-oligomer. Homo-oligomers can be subdivided further, if a homo-oligomer complex's subunits use the same interface for binding and have 2-fold structural symmetry, they are termed as being isologous in interaction, whereas heterologous interaction means that homo-oligomer subunits interact at different interfaces (Ozbabacan *et al*., 2011; Nooren and Thornton, 2003a; Goodsell and Olson, 2000; Monod *et al*., 1965).

1.3.2 Obligate and non-obligate interactions

An obligate protein complex is one where its bound constituents are unable to subsist as monomers (*i.e.* unstable) in unbound form. This means that they assume their final configuration upon protein complex formation to function (Ozbabacan *et al*., 2011; Nooren and Thornton, 2003a). The met repressor is an example of an obligate DNA binding protein, which inhibits DNA expression (Zhu *et al*., 2006; Rafferty *et al*., 1989). Non-obligate protein complex components can subsist autonomously from each other as stable monomers prior to protein complex formation (Ozbabacan *et al*., 2011; Nooren and Thornton, 2003a). Cyclin-dependent kinases are examples of non-obligate proteins essential for cell cycle regulation and which are anti-cancer drug targets (Shafiq *et al*., 2012).

1.3.3 Permanent and transient interactions

Based on protein interaction lifetime, there are permanent and transient interactions (Ozbabacan *et al*., 2011; Perkins *et al*., 2010; Nooren and Thornton, 2003a). Permanent complexes are stable in association and remain in complex, whereas transient complexes are transitory such that their components are able to bind and unbind. Obligate complexes are mainly permanent in interaction, while non-obligate complexes can be transient or permanent in interaction (Nooren and Thornton, 2003a). Transient interactions are further subdivided, based on their binding affinity (K_d; see section 1.4.5) and interaction lifetime, into strong and weak interactions. Strong transient interactions like the G-protein complex (αβγ) subunits remain stable with long lifetimes due to the binding of GDP (guanosine diphosphate), resulting in tight association of the complex subunits. This changes with the binding of GTP (guanosine triphosphate), triggering separation of the complex into Gα and Gβγ components (Nooren and Thornton, 2003a). Weak transient interactions cyclically engage and disengage in complex formation with short lifetimes (ex. seconds) (Ozbabacan *et al*., 2011; Perkins *et al*., 2010; Nooren and Thornton, 2003a, 2003b).

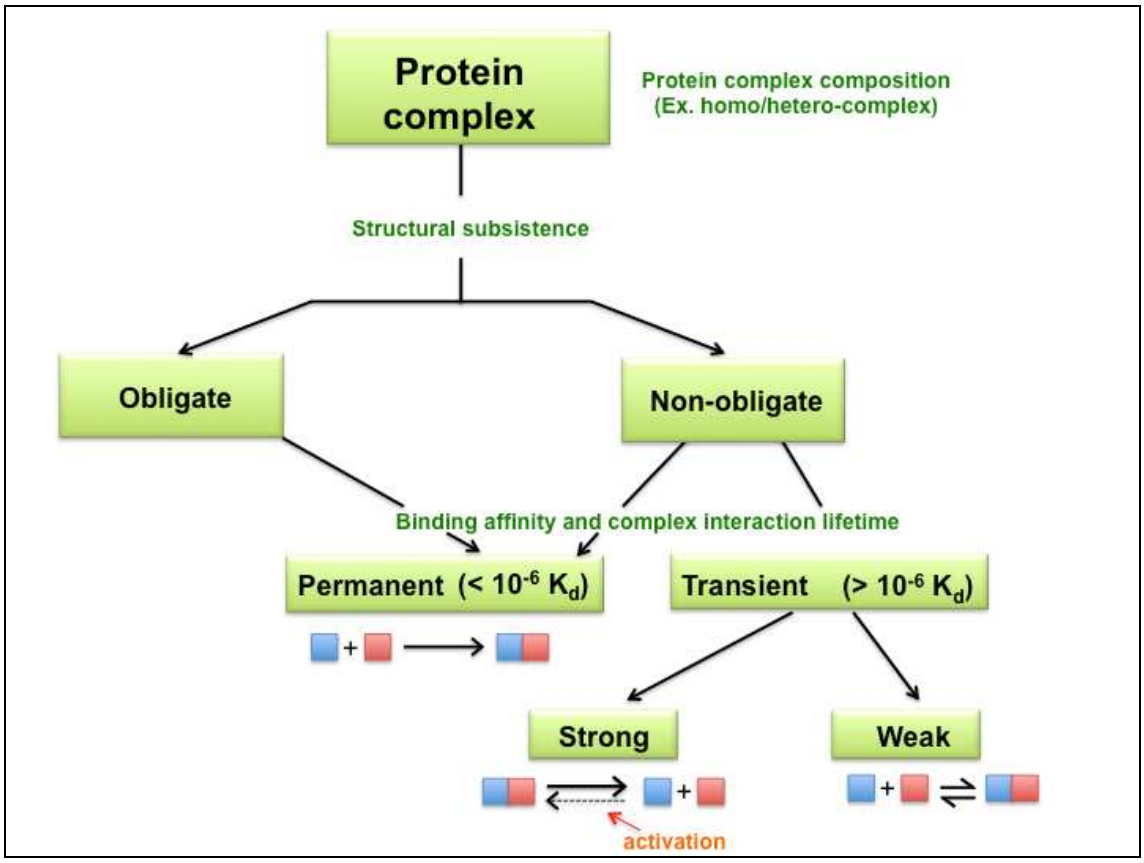

Figure 1-2: Classification of protein-protein interaction types according to protein complex composition, structural subsistence, and protein interaction lifetime. Adapted from Ozbabacan *et al*., 2011, pp. 2-3 and Perkins *et al*., 2010, p. 1234. In terms of composition, protein complexes with identical binding partners are called homo-oligomers, whilst complexes with different binding partners are known as hetero-oligomers. Proteins capable of existing as stable monomers upon association and dissociation are termed non-obligate. Oppositely, proteins whose association guarantees structural subsistence as monomers only and not the reverse are termed obligate. Interactions based on binding affinity (lifetime) are either permanent in which case they form a stable complex, or transient where they can unbind upon forming a complex. Transient complexes can be further sub-grouped as strong and weak transient interactions. Weak interactions bind and unbind with short lifetimes and, in contrast, strong interactions are stable in complex with longer lifetimes when an activating factor causes them to associate and leave their unbound state.

1.4 Characteristics of protein-protein interactions

There are differing features that characterize protein-protein interfaces upon protein complex formation and their interactions. In this section, an overview of these features will be presented.

1.4.1 Interface size

Interface sizes are defined by the buried surface area (BSA) measure, which determines the change in accessible surface area (ASA) in residues for proteins in their unbound and bound complexed states (Chothia and Janin, 1975). The BSA is measured as follows:

$$BSA = ASA_A + ASA_B - ASA_{AB} \qquad (1\text{-}1)$$

where ASA_{AB}, ASA_A, and ASA_B represent the accessible surface areas of the bound complex and unbound protein components A and B, respectively (Levy, 2010; Lo Conte, 1999). Non-obligate hetero-dimer protein-protein complexes have an interface area average of 1,910 $Å^2 \pm 760$ $Å^2$ (Dey *et al*., 2010). In general, an interface is a single patch when its interface area is < 2,000 $Å^2$ (Chakrabarti and Janin, 2002). In contrast, homo-dimers (mostly obligate with some non-obligate complexes) have a larger average interface area of 3,570 $Å^2 \pm 2{,}490$ $Å^2$ that bury more atoms, and these larger interfaces may consist of one or more patches (*i.e.* surface residues in structural proximity) (Dey *et al*., 2010; De *et al*., 2005; Bahadur *et al*., 2003). A recent study examined 42 non-obligate, weak homo-dimers and found that they had an interface area average of 1,620 $Å^2 \pm 480$ $Å^2$, which is similar to the protein-protein complexes (Dey *et al*., 2010). In general, the BSA measure is mainly directly related to the number of residues and their atoms that are buried upon protein complex formation (Dey *et al*., 2010). Similar proportions of main chain atom contributions to the BSA% are present for homo-dimers (17% ± 6) and hetero-dimer (19) protein-protein complexes (Dey *et al*., 2010). It has been suggested that larger interfaces of proteins (>2,000 $Å^2$) involve interactions of significant conformational change during protein complex formation,

whereas proteins with smaller interface sizes involve less conformational flexibility (*i.e.* rigid-body) during complex formation (Lo Conte *et al*., 1999). Obligate complex subunits cannot exist as monomers and are less ordered and only become ordered when forming complexes with each other. Therefore, since they have larger interfaces these types of proteins undergo major conformational change for them to become ordered upon protein complex formation (Janin, 2009).

1.4.2 Interface geometrical plane and complementarity

A feature of protein interface regions is that they are more planar than the rest of the surface of a protein (Wu *et al*., 2007; Murakami and Jones, 2006). Planarity is calculated by defining the least squares fit plane of interface atoms and determining the interface atoms' root mean square deviation from the plane (Chakrabarti and Janin, 2002). Compared to each other, non-obligate protein complex interfaces are more planar than obligate complex interfaces (Bera and Ray, 2009). Moreover, generally geometric (shape) complementarity is present in protein complex interfaces and this is due to close packing density of interface atoms of an interface (Bahadur *et al*., 2004). It can be determined by the shape correlation statistic that measures the fit between buried atoms of an interface of both complex proteins (Lawrence and Colman, 1993). It has been determined that, in general, the interface packing density is similar to a protein's interior packing density (Sonavane and Chakrabarti, 2008; Lo Conte, 1999). Of course there are exceptions to this. For example, electron transfer proteins have loose atomic packing for their interfaces and as a consequence geometric complementarity is less pronounced and this is most likely due to the nature of their interactions, which occur extremely rapidly and only generate short-term protein complexes, providing less emphasis on interface packing and geometric complementarity (Janin *et al*., 2007; Bahadur *et al*., 2004).

1.4.3 Interface secondary structural preferences

The secondary structural preference differs in protein complex interfaces for obligate homo-dimers (mostly obligate with few non-obligate complexes) and non-obligate hetero-complex interactions. Specifically, obligate homo-dimers interfaces have a

higher proportion of alpha helices than beta strands, whereas these secondary structural elements are almost similar in proportion in non-obligate hetero-complex interaction interfaces (Guharoy and Chakrabarti, 2007). Other structural elements such as loops, turns, and coils are present in higher proportion in non-obligate hetero-complex interfaces than obligate homo-dimer interfaces. Guharoy and Chakrabarti (2007) grouped alpha helices and beta strands under the category of regular structures, and loops, turns, and coils were grouped as non-regular structures. They observed a higher proportion for the regular group than the non-regular group in obligate homo-dimer interfaces. For the non-obligate hetero-complex interfaces this was not the case because both groups had similar proportions. However, the proportion of regular structures was found to increase with interface size (*i.e.* ΔASA) for non-obligate hetero-complex interfaces. Specifically, the alpha helices become longer in length with increased interface size. But for obligate homo-dimer complexes no change was observed in the proportion of secondary structural elements with increase in interface size. Examination of "pairing" of secondary structural elements across interfaces revealed interesting results. Pairing refers to the cross-interface interactions that occur between two proteins in complex. There were two pairing categories that were delineated. One category of cross-interface pairing was defined as being between regular secondary structural elements amongst themselves (*i.e.* intra-pairing). Another category was between non-regular structures pairing with either alpha helices or beta strands, or intra-pairing between non-regular structures themselves. It was found that for obligate homo-dimer complexes an approximately equal share of interactions occurred for both pairing categories, whereas non-obligate hetero-complexes favoured the latter category. For non-obligate hetero-complex interaction interfaces, bias to the latter category is due to their interfaces interchanging from exposed to buried and at the same time they must retain properties of a basic protein surface that allows their monomeric unbound states to remain stable in their natural soluble state (Guharoy and Chakrabarti, 2007; Bahadur *et al.*, 2004). This study highlighted that secondary structural preferences are dependent on protein-protein interaction type.

1.4.4 Interface regions and their physicochemical properties

It has been observed in protein interfaces that a slightly higher proportion of hydrophobic residues are found in homo-dimers interfaces (65% ± 7) that are mostly obligate with some non-obligate complexes compared to non-obligate hetero-dimer (58%) interfaces (Dey *et al*., 2010). In addition, polar and charged residues are slightly higher in proportion in non-obligate hetero-dimer interfaces (28% polar and 14% charged) than homo-dimers (22% ± 6 polar and 13% ± 6 charged), and both interaction types have roughly similar water molecule and hydrogen bond densities in their interfaces (Dey *et al*., 2010). Analysis of non-obligate weak homo-dimers indicates that their average contribution of non-polar and polar residues is 62% (± 8) and 24% (± 7), respectively, and this is similar to the general averages for homo-dimers. Furthermore, their proportion of charged residues is 13% (± 8), which is similar to non-obligate protein-protein complexes and the general average for homo-dimers (Dey *et al*., 2010). Chakrabarti and Janin (2002) proposed the core-rim model that divided an interface into core (inner) and rim (outer) regions, which contain buried residues and solvent accessible residues, respectively (see Figure 1-3A). They defined interface residues as protein residues (or their atoms) that lose >0.1 $Å^2$ ASA during protein complex formation based on equation 1-1. Core residues were defined as residues that have a minimum of one completely buried atom (zero ASA), while rim residues retain some solvent accessibility for all their atoms (>zero ASA). Based on this interface division, they examined the amino acid percentages for interface core and rim zones for non-obligate hetero-dimer complexes and later for homo-dimers composed of mostly obligate with some non-obligate interacting proteins (Dey *et al*., 2010; Bahadur *et al*., 2003). In terms of average amino acid percentage for non-obligate hetero-dimer complexes, the core region has a greater proportion of residues (55%) then the rim (45%). In homo-dimers complexes, the core's percentage contribution of interface residues is higher (59%) than the rim's (41%) percentage (Dey *et al*., 2010). Weak homo-dimers have identical percentages for their core and rim regions as the transient protein-protein complexes (Dey *et al*., 2010). The residue composition of the rim resembles the generic surface of a protein exposed to solvent (57% non-polar and 43% neutral polar/charged residue compositions of the BSA%) for non-obligate hetero-dimer

protein complexes. It is also similar for homo-dimers, but in it aliphatic residues are more prevalent relative to a protein's surface (Dey *et al*., 2010; Bahadur *et al*., 2003). In contrast, the core region is between a protein's interior and exterior surface in residue makeup (Levy, 2010). In non-obligate hetero-dimer complexes, aromatic residues enrich the core, while charged residues are low in presence with the exception of arginine (Chakrabarti and Janin, 2002). In homo-dimer protein complexes, aliphatic and aromatic residues makeup a major portion of core residues, providing a stronger hydrophobic character in these complexes than non-obligate hetero-dimer complexes due to their larger interface sizes (Levy, 2010; De *et al*., 2005; Bahadur *et al*., 2003). Also, in homo-dimers charged residues, excluding arginine, are reduced in the core region (Dey *et al*., 2010).

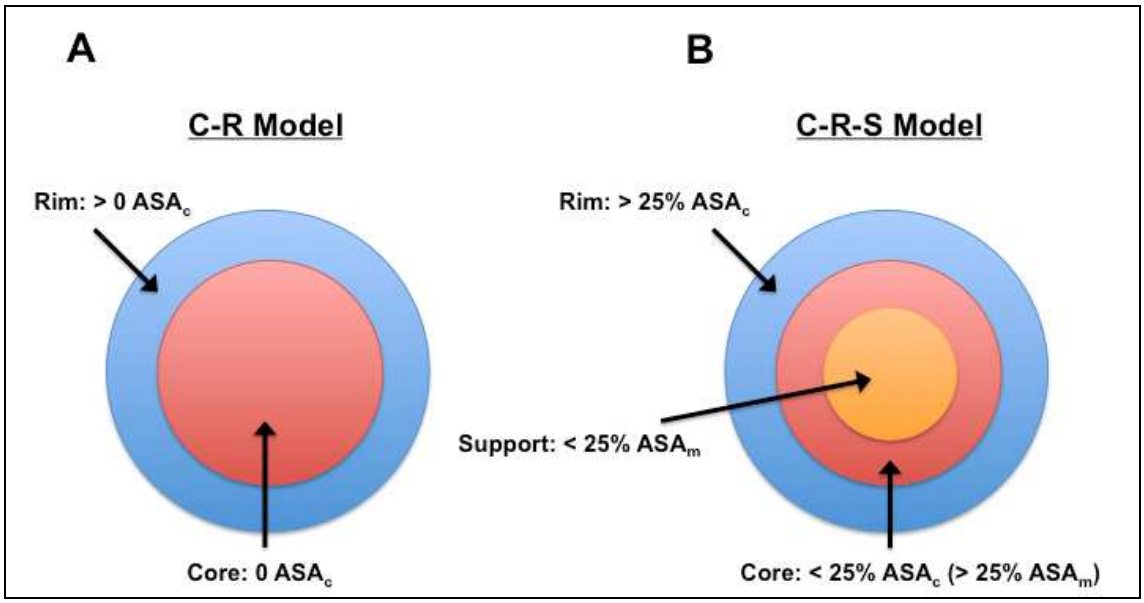

Figure 1-3: The delineation of an interface into regions based on change in buried surface area is presented based on two models. **A)** Core-rim (C-R) model. Core residues have a zero accessible surface area upon complexation for at least one residue atom (ASA_c), whereas rim residues have > 0 ASA_c. Adapted from Chakrabarti and Janin, 2002, p. 339 **B)** Core-rim-support (C-R-S) model. Core residues have > 25% accessible surface area in protein monomers (ASA_m) and upon complexation such residues have < 25% ASA_c. On the other hand, rim residues have > 25% ASA_c, whereas support residues have < 25% ASA_m and are buried further upon complexation. Adapted from Levy, 2010, p. 662. In general in terms of residue composition, the rim region is similar

to the generic surface of a protein, while core residues are midway between a protein's surface and interior. The support region is similar to a protein's interior.

Levy (2010) defines a third region, in addition to the rim and core regions of the interface, called support (see Figure 1-3B). Here, the regions of the interface (ASA > 0) are divided by a different accessible surface area (25% ASA) such that rim residues are exposed (> 0 ASA) in both unbound and complexed forms (> 25% ASA) of both complex proteins, while core residues are exposed (> 25% ASA) in a protein's unbound state, but become buried (< 25% ASA) upon complexation. The support residues are buried (< 25% ASA) in the monomer proteins and become even more so upon protein complexation (< 25% ASA). The rim's residues are still similar to a generic surface patch, while the core's residues are intermediary between a protein's surface and interior. The support, however, is similar to a protein's interior in residue composition (hydrophobic) due to its exclusive buried state upon unbound to bound protein transition during protein complexation. The residue makeup of the interface cores as delineated by Chakrabarti and Janin (2002) and Levy (2010) are similar. The rim as delineated by Levy (2010) is more similar to a protein surface, whereas the rim defined by Chakrabarti and Janin (2002) is slightly less similar to a protein surface in terms of amino acid frequencies. This slight difference is due to the differences of the ASA values used to partition an interface into its different regions. As a result in Levy's (2010) work, the proportions of rim, core, and support regions have analogous numbers of residues. Nevertheless, a single core residue buries almost double the surface area than both rim and support regions on average (Levy, 2010). But, in some cases where interfaces are small (< 1000 $Å^2$) like transient interfaces of approximately 800 $Å^2$, the core's presence is smaller relative to the interface rim; hence these interfaces are more polar in nature (Levy, 2010). Compared together holistically, the models' of partitioning the interface particularly for core and rim regions are similar in the information they convey about these regions most notably for the core region of the interface (Levy, 2010). In addition, the characteristics of the rim, core, and support were found to be regular in three species (*Homo sapiens*, *Saccharomyces cerevisiae*, and *Escherichia coli*), emphasizing the usefulness of this interface model (Levy, 2010).

1.4.5 The energetic terms of protein-protein interactions

The generation of a complex between two proteins (A and B) is described below:

$$A + B \underset{K_{off}}{\overset{K_{on}}{\rightleftharpoons}} AB \qquad (1\text{-}2)$$

where K_{on} and K_{off} are the association and dissociation rate constants, respectively. Following the mass action law, the ratio of these constants defines the equilibrium dissociation constant (K_d). K_d describes the binding affinity between two proteins (A and B) upon protein complex formation (AB) and is expressed in the following relationship:

$$K_d = \frac{K_{off}}{K_{on}} = \frac{[A][B]}{[A:B]} \qquad (1\text{-}3)$$

where [A], [B] and [A:B] are the concentration (Molarity) of the unbound proteins and bound complex. The tighter the binding affinity between complex proteins, the lower the K_d value. In addition, if the high concentrations of [A] and [B] are required to form [A:B], this reflects low binding affinity and hence a high K_d value.

The Gibbs binding free energy (ΔG) describes the thermodynamics of binding of protein monomers, depicting the affinity of the protein monomers for each other and the stability of the protein complex upon formation (Thirupathi *et al*., 2011; Janin, 1995). The thermodynamic terms that define the process of protein complex occurrence are defined in the equations below:

$$\Delta G_d = -RT \ln \frac{K_d}{c^\circ} \qquad (1\text{-}4)$$

$$\Delta G_d = \Delta H_d - T\Delta S_d \qquad (1\text{-}5)$$

where c° is the 1 standard reference concentration (1 $mol.l^{-1}$), R is the gas constant (8.314 JK^{-1} mol^{-1}), and T is the absolute temperature (Kelvin). By determining the K_d value, it is possible to calculate ΔG by using equation 1-4.

In equation 1-5, ΔH and ΔS refer to the change in enthalpy and entropy, respectively, which define the binding free energy (ΔG). The ΔH term is based on the interactions that form the protein-protein complex, such as electrostatic interactions, hydrogen bonds, and van der Waals interactions (Kastritis and Bonvin, 2013). Electrostatic interactions of the ΔH term involve dipole-dipole, charge-dipole, and charge-charge interactions. As an example, the interactions between glutamic acid and aspartic acid with lysine, arginine, or histidine represent instances of charge-charge interactions. These interactions generally contribute to stability in protein-protein interfaces (Xu *et al.*, 1997a; 1997b). The hydrogen bond is a dipole-dipole interaction, as it involves a polar-bonded hydrogen (ex. H-N or H-O hydrogen bond donors) interacting with an electronegative atom's non-bonded electron pair (ex. O or N hydrogen bond acceptors). Hydrogen bonds provide specificity in protein-protein interactions (Bissantz *et al*., 2010; Ponstingl *et al.*, 2005). This specificity of interaction is enabled, as hydrogen bonds follow stringent geometrical restrictions in a biological setting (Bissantz *et al*., 2010). Coupled with this, a hydrogen bond is weak bond, allowing it to swiftly come into being or break, thereby assisting in protein-protein interaction. van der Waals interactions occur through non-specific interactions that are a result of electron fluctuation of atoms, resulting in their irregular distribution, and the creation of induced dipoles for the atoms that interact with one another (Wood and Meyers, 1991). Although these are weak interactions, they are a contributing factor in protein-protein interaction specificity since many such interactions collectively occur during protein-protein binding (Kastritis and Bonvin, 2013).

The ΔS term defines the microstate dynamics of a thermodynamic system. For protein-protein interactions, the entropies of conformation (protein side-chain and main-chain), solvent, and association are an important component of the ΔS term (Brady and Sharp, 1997). The side-chain (ΔS^{side}) is the more prominent constituent of conformation entropy while in comparison the main-chain (ΔS^{main}) is limited in contribution to conformational entropy in protein complex formation (Stites, 1997). This may change upon the occurrence of limited protein folding during complex formation (Brady and Sharp, 1997). Prior to complex formation, the binding surface of a protein is exposed to water molecules. Upon protein-protein complexation, this changes and leads to a favourable (*i.e.* positive) solvent entropy (ΔS^{solv}). This occurs when water molecules

detach from a protein's binding surface and join the surrounding water molecules (Archakov *et al*., 2003). Furthermore, the hydrophobic effect (non-polar interactions), which is a non-specific interaction, results from entropic changes in the presence of water molecules. This is to reduce the effect of the "cage" ordering of water molecules in the presence of non-polar entities, which reduces entropy. Holistic aggregation of non-polar entities relative to water molecules reduces their "cage" ordering, diminishing the impact of entropic reduction (Bissantz *et al*., 2010). The hydrophobic effect is an important driving force for protein-protein complexation especially when the interface has non-polar and hydrophobic residues that become buried (ex. interface core) upon protein complex formation (Williams, 2011; Ponstingl *et al*., 2005; Tsai *et al*., 1997). With regards to the final entropic term, the association entropy (ΔS^{ass}), it decreases as translational and rotational restrictions occur upon the complexation of two proteins (Brady and Sharp, 1997). The sign of the overall entropy value determines if entropy drives protein-protein interactions. A positive ΔS drives the protein-protein interaction process, whereas a negative ΔS indicates that ΔH drives protein-protein interactions (Archakov *et al*., 2003).

1.5 An overview of hotspot residues

Protein interfaces are composed of residues with different physicochemical properties, and they facilitate protein-protein interactions. However, a specific subset of interface residues contributes to most of the binding free energy of a protein-protein interaction (Thorn and Bogan, 2001). These residues are known as hotspots. Clackson and Wells (1995) first used this terminology to describe such crucial residues for protein binding. In their work, the authors examined the complex of the human growth hormone (hGH) and the bound receptor protein (hGHbp). They applied alanine-scanning mutagenesis to mutate the complex's interface residues. Using this method, two tryptophan residues, forming part of a hydrophobic area of the interface, were found to provide most of the binding free energy (Clackson and Wells, 1995).

A hotspot is a residue that, upon mutation, causes a protein complex's binding free energy to change by ≥2kcal/mol (Bogan and Thorn, 1998). Tyrosine, tryptophan, and arginine residues have a common occurrence of being hotspots. Residues that contribute

less to the binding free energy of a complex are observed to encircle hotspots, and this configuration is called a hydrophobic O-ring. Their presence around hotspots occludes bulk solvent, which generally allows hotspots to maintain their contribution to overall binding free energy for a protein interaction (Bogan and Thorn, 1998). Specifically, this creates a localized low dielectric presence, heightening the impact of electrostatic and hydrogen bonding in the regions where bulk solvent are occluded (Li *et al*., 2005). Keskin *et al*., (2005) observed that hotspots were structurally conserved, while residues that encircle them were less structurally conserved in comparison. Moreover, these energetically important residues form clusters of hot regions in an interface, which are characterized by tight packing of hotspots as opposed to them being scattered across an interface (Keskin *et al*., 2005). Due to the importance of hotspots in binding free energy contribution, methods for predicting hotspots have been developed that utilise different properties (ex. evolutionary conservation data). For example, Ahmad *et al*., (2010), who observed that clusters of evolutionary conserved residues (CECRs) contained hotspot residues, developed a method to predict CECRs. This indicates that hotspots are evolutionary conserved (Guharoy and Chakrabarti, 2010). Likewise, Ofran and Rost (2007a) applied their ISIS method (discussed in section 1.8.1) to predict hotspot residues of a 30 protein complex dataset and demonstrated accurate predictions of hotspots. Tuncbang *et al*., (2009) developed an empirical method (HotPOINT) that incorporated residue conservation with other biophysical properties to predict hotspot residues in a protein complex, resulting in accurate predictions. The accurate prediction of hotspot residues in protein complexes facilitates their use in protein-protein docking studies.

1.6 *In silico* methods to predict protein-protein complexes: protein-protein docking

Given the sheer number (in the thousands) of predicted protein-protein interactions identified by high-throughput approaches like the yeast two-hybrid method, in comparison structurally solved protein-protein complexes by NMR and x-ray crystallography approaches, which are time-consuming, are lower in number. For example, the BioGRID database has about 490,000 known interactions and the PDB database has approximately 94,000 structures, including protein-protein complexes

(Berman *et al*., 2013; Stark *et al*., 2010). To bridge this gap, the application of protein-protein docking methods is a useful, inexpensive, and timesaving process of predicting and studying protein complexes as well as identifying biologically important interactions from those that are false positives obtained from high-throughput methods (Wass *et al*., 2011a; Lensink *et al*, 2007). Docking methods complement NMR and x-ray crystallography, providing important insights into protein-protein interactions (Goa *et al*., 2004). The objective of protein-protein docking, given two (or more) unbound proteins known to interact, is to predict their final protein complex configuration that is at the lowest free energy available to the system (Gray, 2006; Liang *et al*, 2006). For this to be achieved, a docking algorithm must sample repeatedly many structural binding poses between interacting proteins, using known experimentally determined unbound receptor and ligand protein models, or homology models if structures are lacking. Each predicted protein complex is scored during docking, filtering accurate predictions from erroneous ones in order to determine the most energetically minimal complex in free energy terms (Gray, 2006). In practice, however, the development of protein conformational sampling and binding free energy calculation that a docking method requires has been quite a lengthy process and is by no means complete, and the "docking problem" as it is known is an open problem under active research (Torchala *et al*., 2013). In this section, an overview of the steps in protein-protein docking will be presented.

1.6.1 Protein-protein docking: the sampling process

The initial stage of protein-protein docking is the sampling process. It involves the rapid generation of docked conformations of proteins while accounting for unbound to bound conformational flexibility of the interacting proteins in an attempt to produce putative complex models that are biologically meaningful. The first docking algorithm developed in 1978 performed docking on low-resolution (coarse-grained) structures where a residue is depicted as a sphere (Vakser, 2004; Wodak and Janin, 1978). It involved sampling through angular rotations coupled with translations (*i.e.* six degrees of freedom) of one protein's positional configuration located near its binding partner's active site surface to produce docked models (Wodak and Janin, 1978). This pioneering work demonstrated that docking two proteins using low-resolution representation was

possible. Further development of docking progressed to intermediary-resolution rigid-body docking methods such as fast Fourier transform (FFT) or geometric hashing techniques that depict interacting proteins as shapes and matches them accordingly (Fischer *et al*., 1995; Katchalski-Katzir *et al*., 1992). For example, FFT methods discretize proteins in three-dimensional space via a grid such that proteins are divided into interior, surface, and exterior sections, and subsequently perform docking of the two proteins rapidly by shape matching and complementarity through the overlap of surface regions between two proteins (Gabb *et al*, 1997; Katchalski-Katzir *et al*, 1992). Here, the proteins are kept rigid to perform six-dimensional sampling in translational and rotational space (Chen and Weng, 2002). Such docking methods simplify flexibility in protein interactions, for example, by means of introducing restricted structural flexibility in general and/or protein side-chain refinement (Jackson *et al*., 1998; Gabb *et al*, 1997).

Another approach is to permit sterical overlap between proteins during the sampling protocol, implicitly modelling flexibility (Fernández-Recio *et al*., 2002). Rigid-body docking in general performs well in predicting protein complexes whose protein constituents experience minimal conformational change when transitioning to their bound states, however, performance is poor for proteins that experience major conformational change during complexation (Janin, 2010; Ritchie, 2008). Recent development of high-resolution protein docking methods has enabled sampling using atomic-level representations of proteins, while having the capability to incorporate flexibility during the docking protocol (Wang *et al*., 2007; Dominguez *et al*., 2003; Gray *et al*, 2003). In high-resolution docking, the incorporation of flexibility can be achieved implicitly by using ensembles of protein conformations derived from NMR and molecular dynamics approaches, for example, or by using different experimentally-derived models of the same protein (Dominguez *et al*., 2004; Grünberg *et al*., 2004). This application of ensembles has been extended to rigid-body docking (Dominguez *et al*., 2003).

In principle, the use of an ensemble generated from an unbound protein presumes that the combined conformational snapshots for a protein cover a substantial portion of the actual bound conformational pose adopted by the protein in complex (Dobbins *et al*.,

2008; Grünberg *et al*., 2004). While helpful, it is possible that structural flexibility produced upon complexation cannot be generated through the ensemble model approach from unbound proteins that interact with each other, preventing the actual modelling of the final bound pose between interacting proteins. This may occur when an unbound protein's transition to it's bound configuration involves major surface transformation. Further means of incorporating conformational dynamics in docking in high-resolution docking methods involves the explicit inclusion of flexibility of protein side-chains and/or backbones through the use of molecular dynamics or Monte Carlo simulations (Dominguez *et al*., 2003; Gray *et al*., 2003).

For proteins that undergo large structural changes upon complexation, specific sampling protocols that model this flexibility have recently been developed (Karaca and Bonvin, 2010). For example, in the Flexible Multidomain Docking (FMD) data-driven approach of Karaca and Bonvin (2010), a flexible protein is split into sub-domains that are kept together by connectivity restraints during the FMD protocol. The sub-domains are docked as multi-bodies to the partner protein, which is followed by the introduction of explicit flexibility in the backbone and side chains of the interfacial regions of the docked models. The FMD approach successfully modelled a protein interaction where one protein constituent underwent backbone conformational changes of 19.5 Å when transitioning to its bound state (Karaca and Bonvin, 2010).

1.6.1.1 Data-driven sampling in protein-protein docking

There are two types of docking sampling strategies, *ab initio* and data-driven. The difference between the two is that the latter strategy limits sampling to specific regions on both proteins as dictated by data that guides a docking sampling algorithm. In contrast, the former strategy is unconstrained by data and may sample all possible poses between proteins, given adequate computational resources. The data used can be derived from experimental approaches (NMR) or theoretical approaches like protein interface predictors (see sections 1.7 and 1.10) (de Vries *et al*., 2006; van Dijk et al., 2005a). The composite data employed in sampling represents possible interfacial regions (ex. Chemical shift perturbations - CSPs) and provides orientational information (ex. Residual dipolar coupling - RDCs) of one protein to its partner, which is an

advantage in sampling because it vastly reduces the search space (van Dijk et al., 2005a). This is even more so significant for high-resolution docking where sampling is a computationally expensive process compared to intermediary and lower resolution sampling. While data-driven sampling is effective to steer docking, one side effect is the possibility the data used is wrong thereby "trapping" docking sampling in the wrong area of interest. A widely known and established high-resolution docking method that utilizes the data-driven strategy is the HADDOCK method (Dominguez *et al*, 2003). In HADDOCK, sampling data is converted into ambiguous interactions restraints (AIRs). These interaction restraints are referred to as ambiguous because interface residues identified from one protein are not known with which residues of the opposing partner protein they interact (Dominguez *et al*, 2003). HADDOCK randomly discards some restraints during its sampling protocol in case of some restraints being wrong. If wrong restraints are removed, this may improve docking.

1.6.2 Protein-protein docking: the scoring function

The prediction of a protein complex involves generating many models of different poses whether sampling is data-driven or not. From these models, it is anticipated that some models have similar structures to the native complex. Scoring functions are applied in order to rank each model in an attempt to distinguish biologically meaningful models from ones that are not (Moont *et al*., 1999). The application of a scoring function during docking is of two kinds. In docking, scoring functions may be applied directly in the sampling protocol influencing the generated models in which case they are termed as 'integrated', or they may be employed directly after sampling is completed where they are termed as 'edge' (Halperin *et al*., 2002). Regardless of their implementation in a docking method, the determination of the protein complex structure with the lowest binding free energy is the goal, assuming that the native complex corresponds to the lowest energy conformation (Yu *et al*., 2004).

Scoring functions employed in docking protocols can be knowledge-based functions or physical force field functions. The knowledge-based scoring functions are based on statistically deriving residue/atomic contact propensity data of experimentally solved protein-protein complexes and those from decoy protein complexes (Zhang *et al*.,

2005). This can assist in better discrimination from near-native complexes from decoys. The ClusPro docking method applies this type of scoring function using Atomic Contact Potential and electrostatic energy in its filtering stage after sampling is performed in order to select the best scoring models for later processing (Comeau *et al*., 2004). The force field scoring functions calculate the final energy of a docked complex, which is a weighted sum of the contribution of interaction terms derived from a molecular mechanics force field and also incorporate energy terms that evaluate the use of experimental/theoretical data in docking (Audie, 2009; Dominguez *et al*., 2003). An example of this type of scoring function (equation 1-6) is implemented in the HADDOCK docking method, which is applied in this study (de Vries *et al*., 2007; Dominguez *et al*, 2003).

$$E_{HADDOCK} = E_{vdW} + E_{elec} + E_{desolv} + E_{AIR} + E_{sani} + E_{vean} \qquad (1\text{-}6)$$

where E_{vdW} and E_{elec} represent the van der Waals and electrostatic energy terms, respectively. E_{desolv} is the desolvation energy. E_{AIR}, E_{sani}, and E_{vean} are pseudo-energy terms for the ambiguous interaction restraints, RDCs and intervector projection angles (IPAs) energies, respectively. They calculate the agreement between the generated models and the experimental data used in HADDOCK to guide docking, acting as a discriminator between near-native and incorrect models. The individual energetic terms are weighted to optimize the $E_{HADDOCK}$ score. The $E_{HADDOCK}$ score in various forms is applied during all stages of the HADDOCK docking protocol to select the best scoring models for each stage of the protocol.

1.6.3 The Critical Assessment of PRedicted Interactions (CAPRI)

CAPRI is a blind docking competition that assesses the capabilities of protein docking methods. A CAPRI prediction round is held upon the emergence of new experimentally determined protein-protein and protein-nucleic acid complexes (Lensink and Wodak, 2010). The starting protein structures of the ‘unknown’ CAPRI target complex are provided to each participating docking team for prediction of the final complex configuration. The CAPRI participants have no knowledge of the actual experimentally

determined protein complex. All generated docking solutions from participating docking teams are evaluated against the actual complex and classified as acceptable, medium, or high quality models (see section 3.8.2). Models not assigned to these classes are classified as incorrect (Méndez *et al*., 2003, 2005). Ideally, the starting protein structures used for docking would be in the unbound pose. However, if only a single unbound protein is available for a CAPRI target, a bound protein extracted from the CAPRI target complex is provided for docking, or in some cases when feasible, a homology model is used (Lensink and Wodak, 2010; Lensink *et al*., 2007; Méndez *et al*., 2003). The first CAPRI competition (2001-2002) was held for two rounds, involving 7 CAPRI targets and 19 docking teams overall (Janin *et al*., 2003). In this competition, acceptable to high quality models were generated for 5 CAPRI targets by 14 docking teams (Méndez *et al*., 2003). The second CAPRI competition (rounds 3-5; 2003-2004) involved more CAPRI targets (10) and docking teams (30). This competition was more successful than the first, as correct predictions were generated for all CAPRI targets (acceptable and higher models) by 20 docking teams (Méndez *et al*., 2005). Continuing the momentum of successes, rounds 6-12 of the third competition (2005-2007) involved 9 CAPRI targets and resulted in acceptable and medium models with only one high quality model in total for 8 CAPRI targets. In these rounds, 71 docking teams participated, but only 31 teams produced acceptable and above models. In addition, these rounds featured a scoring experiment. Here, 15 scoring teams scored models generated by the docking groups for 5 CAPRI targets and re-ranked them. The best 10 re-ranked models were submitted to evaluate their scoring methods' performance. The scoring groups identified only acceptable and medium models for 3 CAPRI targets from their submitted models (Lensink *et al*., 2007). Rounds 13-19 of the fourth CAPRI competition (2007-2009) performed blind docking experiments involving 14 CAPRI targets altogether (Lensink and Wodak, 2010). 76 docking groups participated in this competition. 51 docking groups generated acceptable quality models or above for 11 CAPRI targets. For the scoring function 'blind' test, 41 scoring groups participated and this was higher than previous participation. From their submitted models, the scoring groups identified acceptable and greater models for 7 CAPRI targets (Lensink and Wodak, 2010). The CAPRI competition has stimulated the development of protein docking methods through rigorous testing and assessment. This will help in docking methods' development to enable their deployment in high-throughput protein complex prediction at the proteomic scale. As a consequence, the

rapid generation of relevant protein complex models between putative proteins known to interact would be predicted for use in further experimental studies. The HADDOCK data-driven docking method, which has performed well in the CAPRI competition (de Vries *et al*., 2007), was used in this study to examine the effect of interface prediction data-driven docking vs. *ab initio* docking (see section 1.11).

1.7 *In silico* methods to predict protein-protein interfaces: protein interface predictors

The function of a protein complex at the molecular level can be understood by examining its interface. To realize this, it is essential to identify functional residues of the interface (Ofran and Rost, 2007b). Given two proteins known to interact, the goal is to predict their functional residues. Accordingly, protein interface predictors have been developed for this purpose. These protein interface predictors utilize known interface residue characteristics determined from analysis of interfacial and rest of the surface residues of proteins (Zhou and Qin, 2007). These *in silico* methods complement experimental approaches to characterize interface residues like site-directed mutagenesis or NMR CSP analysis (Fernández-Recio, 2011). In the context of protein-protein docking, interface predictors are useful as they can provide restraints to reduce sampling complexity to the region of interest and improve scoring of complex models (Ezkurdia *et al*., 2009).

1.7.1 Definitions of interface residues used for creating an interface residue dataset

Determining surface residues is necessary as this allows distinguishing between interface residues and the rest of the surface residues (ROS) of a protein's surface (see sections 3.2.1 and 5.3). Surface residues can be determined by defining the relative (percentage of accessible surface area) or absolute (accessible surface area) exposure to solvent of a residue (Wang *et al*., 2006; Chakrabarti and Janin, 2002). In either case, only interface residues that are above or equal to a certain threshold used for surface residues are retained, while those residues under the threshold are discarded from the final dataset of interface and ROS residues used for training a protein interface predictor

(see section 5.3) (Xue *et al*., 2011a).

It is necessary to define interface residues, for example, to characterize interface residue properties to be used as a basis to design, test, and benchmark interface predictors (see sections 3.2.1 and 5.3). Various definitions have been applied to interface residues in order to construct datasets. One definition is based on a distance between pairwise residues (*i.e.* alpha carbons) or any of their atoms of a protein complex. If the measured distance is smaller than or equal to a cut-off (ex. 5 Å or 1.2 nm), then residue pairs of the interacting proteins are counted as interface residues (Fariselli *et al*., 2002). Adjusting the distance cut-off to a higher value defines more residues as interface, and decreasing it results in fewer interface residues. An alternative interface definition is based upon the change in accessible surface area from unbound to bound protein states (see section 1.4.1) (Chakrabarti and Janin, 2002; Chothia and Janin, 1975). As with the distance-based definition, a cut-off is applied (ex. 4% ASA) to determine interface residues, and its adjustment controls the number of interface residues determined. A final definition considers the geometry of interfaces through the application of Voronoi diagrams (Pontius *et al*., 1996; Harpaz *et al*., 1994; Richards, 1974). The Voronoi diagram is the division of space around points (*i.e.* residues or residue atoms), leading to the creation of a polyhedron around each residue atom. Residues or residue atoms (points) from opposing proteins in complex that share the same Voronoi facet are part of the protein complex interface (Cazals *et al*., 2006; Valdar, 2002). Regardless of the definition used, similar results are produced in terms of interface size/area (Gong *et al*., 2005).

1.7.2 Interface residue predictive characteristics

Different features have been identified that characterize interface residues from other surface residues (ROS), and such features are applied as properties to predict protein interfaces. Interface residues are evolutionary conserved as opposed to non-interface residues due to functional/structural implications, indicating use as a predictive feature (see Chapter 5) (Chung *et al*, 2006; Bordner and Abagyan, 2005). The evolutionary conservation of residues is determined by comparing a protein sequence with other homologous sequences in a multiple sequence alignment (MSA; see section 3.4) (de

Vries *et al*., 2006). This is achieved using an evolutionary conservation score, which computes the amino acid variability in an MSA column while incorporating their physicochemical characteristics, for example, to determine the level of conservation of the MSA column (see section 3.4). Higher conservation implies functional significance (*i.e.* interface residue) due to evolutionary constraint and vice versa. In the context of the core-rim interface model, core residues are more conserved than rim residues (Guharoy and Chakrabarti, 2005). Also conserved residues cluster together spatially in interfaces (Guharoy and Chakrabarti, 2010). Hydrophobic amino acids along with aromatic/basic residues (tryptophan, tyrosine, and arginine), which are found in the core region of interfaces (see section 1.4.4), are clustered in interfaces, highlighting another predictive feature namely residue propensity (Guharoy and Chakrabarti, 2010; Chung *et al*, 2006; Liang *et al*, 2006; Bordner and Abagyan, 2005). Glycine also has a preferred propensity in conserved residue clusters although it is not enriched in an interface core (Guharoy and Chakrabarti, 2010). In addition, interface residues have higher solvent accessibility than ROS residues (Chen and Zhou, 2005). Moreover, interface residues are less probable to adopt different side-chain rotamers, which may be in preparation for loss of conformational entropy of side-chains upon protein complex formation (Liang *et al*, 2006; Cole and Warwicker, 2002). This predictive quality has been applied from x-ray crystallography B-factor data (Chung *et al*., 2006). Other features used for predictive purposes include secondary structure characteristics and interface shape, hydrophobicity, desolvation, and electrostatics (see section 1.4) (Burgoyne and Jackson, 2006; Hoskins *et al*., 2006; Bradford and Westhead, 2005).

1.7.3 Interface prediction approaches

Interface prediction methods are trained on datasets of determined interface and ROS residues, using a particular definition to determine interface residues (see section 1.7.1). Structural and sequence data pertaining to interface residues are incorporated as predictive features in an interface prediction method. In general, given a surface residue of unknown classification (*i.e.* interface or ROS), it is either classified as an interface or ROS residue by different methods (Bradford and Westhead, 2005). The classification can be done by interface prediction methods that, generally, can be divided into numerical value-based and probabilistic approaches (Zhou and Qin, 2007).

1.7.3.1 Numerical value-based approach

A numerical value-based approach is represented in a function taking into account sequence or structural-based predictive features of known interface residues (Zhou and Qin, 2007). This is represented as follows:

$$S_i = f(x_i, x_{j \in n_i}, c) \qquad (1\text{-}7)$$

where x_i is input data (*i.e.* predictive features like conservation) of an unknown residue *i* of interest. In addition, residue *i*'s neighbour list is considered, which is represented by spatially neighbouring residues' ($j \in n_i$) properties (x_j) in terms of their structural or primary sequence proximity (*i.e.* window) of the residue *i* (Capra and Singh, 2007). The group of coefficients, which is defined during training, is represented by c. The value of S_i determines whether residue *i* is classified as interface or ROS residue. A threshold for S may be set such that if $S_i > S$, then it is classed as an interface residue and vice versa. Numerical value-based approaches can be based on linear regression (Kufareva *et al*., 2007). For example, conservation data can be used as input data in a linear equation and compared to conservation of known interface residues (Li *et al*., 2006). Scoring functions, based on empirical energy functions, can also be applied, permitting different types of input data (*i.e.* interface discriminative features like conservation) to be used in scoring a residue of interest with later classification of that residue (see section 3.4). Finally, there are different machine learning-based approaches that have been applied. For instance, in support vector machines (SVM) non-linear mapping of input data of a training set is performed in high dimensional space, resulting in a hyper-plane that attempts to separate the input data points into two classes: interface and ROS residues (Zhou and Qin, 2007; Larrañaga *et al*., 2006). Subsequently, given a residue of interest, the objective is to assign it through SVM to the interface or ROS classes.

1.7.3.2 Probabilistic approach

The objective of a probabilistic approach is to determine the conditional probability for the residue of interest to be predicted as an interface residue, given a set of input data (Zhou and Qin, 2007). This is depicted as follows:

$$p(s|x_1,...,x_k) \qquad (1\text{-}8)$$

where **s** represents the residue of interest that can be either an interface or ROS residue. x_1 through x_k represent the input data (*i.e.* interface discriminative features) for **s**. If the obtained probability for a residue of interest is greater than a conditional probability cut-off, the residue is classified as interface and vice versa. The conditional probability cut-off is derived from the dataset of known interface and ROS residues that is used for training. There are different applications for this type of method. As an example, the naïve Bayesian method views the input data (x_1 through x_k) as being independent in order to calculate the conditional probability in the following manner:

$$p(s|x_1,...,x_k) = p(s)\prod_{l=1}^{k}\frac{p(x_l|s)}{p(x_l)} \qquad (1\text{-}9)$$

where *p(s)* represents the fraction of the class *s* (*i.e.* interface or ROS residues) of the training set composed of interface and ROS residues (Zhou and Qin, 2007). $p(x_l)$ is the probability density of the input data (x_l) in the entire dataset (Zhou and Qin, 2007). Finally, $p(x_l|s)$ is the likelihood probability of the data subset that it is of a specific class *s* (Zhou and Qin, 2007).

1.7.4 Prediction output of interface predictors

The interface predictors are grouped into patch and residue-based predictors (de Vries and Bonvin, 2008; de Vries *et al.*, 2006; Bradford and Westhead 2005). Patch-based methods divide a protein's surface into pre-defined patches of a given size and score them following a specific approach in an attempt to rank the patches in terms of

confidence such that the top scoring patch or group of defined patches are classified as the protein's interface (Bradford and Westhead, 2005). A residue-based method outputs a list of residues that have been predicted to be part of the interface. Such residues may be used as the final prediction although if mapped on a protein's surface there may be some isolated residues from some that are spatially close from one another (de Vries *et al*., 2006). Alternatively, residues of this approach are then clustered into patches that are re-ranked (Qiu and Wang, 2012). This eliminates isolated residues, which may be far apart from the main cohort of clustered residues (Guharoy and Chakrabarti, 2010; Ofran and Rost, 2007b). Also, applying clustering is supported by recent findings that conserved interface residues cluster spatially in close proximity on a protein's surface (Guharoy and Chakrabarti, 2010).

1.8 Historical description of protein-protein interface predictors

In this section, a variety of interface predictors in this field will be described, focusing on all aspects of their methodologies, the training and testing datasets used, and their reported performances. Thereafter, further discussions focusing on other themes relating to interface predictors will take place. These discussions will focus on advantages and disadvantages of the described interface predictors, their application in combination with protein-protein docking, and their limitations in the context of what the newly proposed interface predictor (PROTIN_ID) introduced in this work seeks to address, including in its application in combination with protein-protein docking.

1.8.1 Overview of interface prediction approaches

The first protein interface and functional residue prediction method developed in 1996, known as the Evolutionary Trace method (ET), detects evolutionary conserved surface residues using multiple sequence alignments (Lichtarge *et al*., 1996a; 1996b). In the ET method, the initial step is to generate a sequence identity dendrogram from a multiple sequence alignment. Sequences in a dendrogram are partitioned via partition identity cut-offs (PIC) and separated into sub-groups based on their identity to each other. These sub-group sequences each branch off from a dendrogram's node at a specific PIC. While

sharing sequence identities to each other, sequences in a sub-group may be similar in function as well. Based on the PIC selected, this allows scaling of functional resolution of entire sequences in a dendrogram through group clustering of all sequences. A higher PIC results in greater functional resolution and vice versa. The next step in this method is the construction of an evolutionary trace. This begins with the generation of a consensus sequence for each sub-group of sequences of a delineated partition, and invariant (conserved) residues of a consensus sequence are identified. The remainder of variant residues are designated as neutral. Following this, all consensus sequences for a given partition are aligned to identify: - 1) invariant residues conserved across all consensus sequences, 2) invariant residues that differ in residue type and designated as class-specific, and finally 3) variant residues (including gaps) for aligned positions, which are highlighted as neutral (Lichtarge *et al*., 1996a). The evolutionary trace data of conserved and class-specific residues are mapped onto the final three-dimensional structure of the query protein of interest to visualize the conservation data and interpret it through the application of residue clustering (Madabushi *et al*., 2002). An extension to the ET method incorporated the Shannon entropy score (see section 3.4.2) and treated gaps as the 21st amino acid (Mihalek *et al.*, 2004). The later iteration of the ET method outperformed the original implementation.

Jones and Thornton (1997a) developed, SHARPE^2, an interface residue predictor for prediction of protein-protein interfaces in homo-dimers, hetero-dimers, and antibody-antigen interactions. In their method they define surface patches for a given protein. A surface patch is composed of a seed residue and the number of neighbouring residues in close proximity to it. For homo-dimer interactions, surface patches of a certain size are determined based on a linear correlation between protein size and interface size, which are both defined in terms of their number of respective residues. In contrast hetero-dimer and antibody-antigen interactions' surface patch size for a given protein is based on the average size of the interface patch, and this is determined from their dataset. Each surface patch is scored according to predictive features applied, depending on the type of protein-protein interaction. The score obtained for each predictive feature is combined to create a composite score per surface patch. In complimentary work, Jones and Thornton (1997b) showed that interface properties differed, according to the type of protein-protein interaction. This is applied by varying groupings of predictive features used to score surface patches based on protein interaction type (Murakami and Jones,

2006; Jones and Thornton, 1997b). For homo-dimer interactions, these predictive features are accessible surface area, planarity, protrusion, hydrophobicity, solvation potential, and interface residue propensity. In hetero-dimer interactions, four features of the above predictive features are applied to score surface patches for the receptor protein known to interact with its smaller ligand protein; only hydrophobicity and solvation potential are not used. All six predictive features are applied to score the ligand protein of a hetero-dimer interaction's surface patches. In antigen interactions only the interface residue propensity feature is not used. When each feature contributing to the composite score scores highly, a surface patch is ranked high. In their method, the three highest scoring patches are taken as the final prediction. Overall, this method produced predictions of >70% specificity (see section 3.5.2) on a dataset of 59 complexes (homo- and hetero-complexes).

Another method developed by Landgraf *et al.* (2001) predicts functional residue surface patches of a protein and, like the ET method above, utilizes both three-dimensional structural and sequence data to generate predictions. This method utilizes structural data of a protein of interest and its sequence data (MSA). For the initial step, surface patches are generated the same way as the method of Jones and Thornton (1997a). Residues of a patch are removed in the global MSA and connected to form a regional MSA. The regional MSA represents the structural vicinity of residues that comprise a surface patch, and it is derived from the global alignment of the query sequence of interest with its related sequences. In the subsequent steps, global and regional sequence similarity matrices are created by the method, representing the global and regional MSAs. These matrices are compared to determine, if present, the degree of difference between them. This establishes if an individual seed surface residue and its structurally neighbouring residues are more conserved than the protein altogether. This procedure is iterated for all surface patches' seed residues and their neighbouring residues to determine their extent of conservation. The final output involves converting the associated structural conservation per seed residue scores into Z-scores, and these are mapped back onto the protein's surface. This method was tested on a dataset of 25 obligate and transient proteins, 6 proteins that bind to DNA (or RNA), and 15 proteins with catalytic sites.

A pioneering method used machine learning for the prediction of interface residues (Zhou and Shan, 2001). In their work Zhou and Shan (2001) employed PSI-BLAST position-specific sequence profile data and solvent accessibility data of spatially contiguous residues at the structure level to train a neural network for interface prediction. Their training set used consisted of 615 protein-protein complexes (552 homo-dimers and 63 hetero-dimers). Moreover, their testing set was composed of 129 complexes where 117 were hetero-dimers and the remaining were homo-dimers. Zhou and Shan's (2001) method, Protein-Protein Interaction Site Predictor (PPISP), showed specificity and sensitivity of 70% and 47%, respectively (see section 3.5.4). Later development of this predictor added consensus neural network prediction to alleviate over- and under-prediction problems (Chen and Zhou, 2005). Hence known as cons-PPISP, this later iteration of the method was trained (using the same predictive properties) on a much larger dataset (1156 protein chains: - 756 homo-dimers and 400 hetero-dimers) and tested on 100 (58 hetero-dimers and 42 homo-dimers) protein chains, resulting in much improvement interface residue prediction performance. This was exemplified by increased specificity (80%) and sensitivity (51%) values, when using the testing set. Further testing of cons-PPISP on independent validation datasets of 8 proteins determined by NMR and 68 transient proteins resulted in accurate predictions (NMR proteins: 69% specificity and 47% sensitivity; Transient proteins: 61.4% specificity and 38% sensitivity). Fariselli *et al.*, (2002) also used a neural network for prediction, training it on the same features used in the method of Zhou and Shan (2001). However, they used sequence profiles obtained from HSSP MSAs. In their method, the neural network is trained, using three-fold cross validation, on a dataset of 226 hetero-dimers and performed at 72% specificity and 56% sensitivity. Unlike the above authors, Ofran and Rost (2003a) used sequence data only and no structural data for interface prediction. They used protein interface composition by analysing their spatial proximity to each other at the sequence level, to train (using three-fold cross-validation) a neural network method. Although structural input is not utilized by their method, it is trained on interface residues determined through a distance cut-off of a set of 333 protein complexes predicted as transient. This dataset may contain obligate protein complexes due to its predicted nature. In comparison to this predicted dataset, an established transient protein dataset (Benchmark 4.0) exists with a lower number (176) of transient complexes and their unbound protein constituents (Hwang *et al.*, 2010). Ofran and Rost (2007b) further developed their method ISIS (Interaction Site

Identified from Sequence) by including more predictive properties in addition to the current property implemented in their method. They used secondary structure and solvent accessibility data predicted from sequence alone and evolutionary conservation data derived from PSI-BLAST sequence profile data as input to ISIS. The updated method of ISIS was trained (using cross-validation) on the original dataset and generated a minimum of one interface residue in its predictions for more than 90% of the dataset used, which corresponded to ~ 61% specificity at 20% sensitivity. The original implementation of ISIS managed to predict at 62% specificity at a low sensitivity (0.5%), and this performance enhancement of ISIS is chiefly from the addition of more predictive properties.

Neuvirth *et al.* (2004) developed the ProMate predictor, a Naïve Bayesian technique, which utilized various extensive descriptors drawing from both structure and sequence data of a training set of 57 transient proteins (via cross-validation) to predict interface residues. Specifically the method scans the surface of a protein, forming circles around a specific point at a certain radius. Neighbouring residues within a circle are analysed in terms of the subsequent parameters: (a) residue conservation derived from the PSI-BLAST position-specific sequence scoring matrix; (b) knowledge-based features (secondary structural composition, B-factor, and the presence of water molecules); (c) and physicochemical conformation (residue propensity and pairing, residue and atom type, and residues' sequence distance from one another). The surface "circles" are scored according to the above descriptors to determine their probability of being assigned as interface or not. ProMate produced accurate predictions of $\geq$ 50% specificity and $\geq$ 20% sensitivity. A further update to ProMate focused solely on re-optimization of the input parameters (via logistic regression) used for prediction, resulting in an increased number of correct predictions (*i.e.* proteins with specificities $\geq$ 50%) for 67% of the dataset for the optimized version of ProMate compared to 63% for the original ProMate (Neuvirth *et al*., 2007).

Koike and Takagi (2004) used a support vector machine (SVM), which is a machine learning technique, as an interface predictor (see section 1.7.3.1). Their method was trained using sequence residue neighbour profiles derived from a PSI-BLAST position-specific sequence scoring matrix and structural residue neighbour profiles based on the 10 closest residues' spatial distances to form a surface patch. In addition, other features

for training the SVM such as estimated interaction site ratio (interface residue to entire protein sequence), spatially contiguous residues' ASA, and planarity of surface patches were utilized. Using k-fold cross validation, their method was trained on 271 and 291 hetero- and homo-complexes, respectively, and produced a performance of 56.1% specificity and 44.6% sensitivity.

Keil *et al.*, (2004) developed a predictor, a neural network method, to predict functional residues. In their method, two topographical (cavity depth and surface topography index) and three physicochemical (electrostatic potential, hydrogen bond acceptor and donor densities, and lipophilicity) predictive properties were used to train their method for the purpose of scoring areas of a query protein's surface termed domains. Surface domain scores are mapped onto a protein's surface for visualization. They employed an extensive dataset composed of 7821 protein structures bound to proteins and peptides, ligands, and DNA and RNA that were partitioned into training and testing sets composed of 1,241,859 and 13,994 surface domains, respectively. They reported a sensitivity value of 76% when the top scoring surface domains were taken as final prediction for all bimolecular interactions of their dataset. For protein-protein interactions, 44% sensitivity was obtained when the top scoring surface domains were taken as final prediction

Chelliah *et al.*, (2004, 2006) developed the Crescendo method that employed environment specific substitution tables (ESST), which were used to discriminate between structural and functional constraints placed on residues in a given protein structure. Crescendo accepts as input a query protein structure along with a MSA containing the query protein's sequence with its associated homologous sequences, which may be obtained from known protein structures or not. The observed residue substitution pattern for each residue of an MSA column is compared to the expected residue substitution pattern, as determined from the ESST. The comparison can be quantified through a divergence score (*i.e.* Jensen-Shannon divergence) or a conservation score. The computed scores are converted to Z-scores and mapped onto the final three-dimensional structure of the query protein of interest where they are smoothed and contoured. High Z-scoring residues are clustered and all clusters are ranked by size. The Crescendo method was originally applied to representative proteins of 164 protein families to predict ligand binding sites and catalytic residues, producing

on average accurate performance (28% specificity and 40% sensitivity) (Chelliah *et al*., 2004). It was later applied for protein-protein interface prediction on a dataset of 20 proteins and achieved correct predictions (>50% specificity) for 85% of the dataset (Chelliah *et al*., 2006).

PPI-Pred is a method similar in concept to the pioneering SHARP^2 predictor, as it also analyses surface patches according to predictive properties (Bradford and Westhead, 2005). The six predictive properties used were conservation, solvent accessibility, hydrophobicity, interface residue propensity, electrostatic potential, surface shape index, and surface shape curvedness. It differs, however, in that it utilizes this data to train an SVM algorithm to differentiate between true interacting patches from ROS patches. PPI-Pred was trained, using leave-one-out cross validation, on a 180 protein dataset composed of obligate and transient interactions. PPI-Pred computes a confidence value for each patch and ranks them according to their scores. Like the SHARP^2 predictor, the top-three ranked patches are taken as the final prediction. A correct prediction was defined as a patch with >50% specificity and >20% sensitivity ranked in the top three patches. Based on this definition of success, PPI-Pred was able to obtain correct predictions for 76% of the dataset. Further testing on a mixed dataset of 47 proteins (obligates and transients) and a transient dataset of 57 proteins generated correct predictions for 72% and 53% of the datasets, respectively. In later work, Bradford and Westhead (2006) trained a naïve Bayesian classifier to predict protein interface patches using the same predictive properties applied for PPI-Pred. They used the same previous dataset, using leave-one-out cross validation for training, and the same criteria for success. This new predictor obtained 82% correct predictions (within top-three patches) for the dataset, performing better than PPI-Pred. This improvement in performance may be attributed to the lesser degree of data over-fitting by the naïve Bayesian classifier (Bradford and Westhead, 2006).

An SVM-based predictor used evolutionary conservation rates (analysed from a MSA) and the residue type frequencies for a seed surface residue and its closest 14 nearest neighbouring residues for training (Bordner and Abagyan, 2005). The residue type frequencies are determined from the residue type frequencies present in a MSA's columns that correspond to the 15 spatially nearest surface residues. This is iterated for all surface residues and their nearest-neighbours since all surface residues are

considered seed residues. The predictor was trained on 632 protein complexes (518 homo-dimers and 114 hetero-dimers), using five-fold cross-validation, which resulted in performance specificity and sensitivity values of 34% and 64%, respectively. In addition it was also tested on 43 transient hetero-dimers, producing accurate predictions (22% specificity and 67% sensitivity values).

An interesting method, developed by Hoskins *et al*., (2006), used secondary structural data to predict potentially interacting beta-strands of a query protein as an initial step for prediction of interacting residues. This method characterized beta-strands by their solvent accessibility, length and orientation (parallel or anti-parallel), strand type and localization (isolated or sheet; central or edge), beta-bulge occurrence, protective loops (PL) presence, and residue propensity to derive rules to classify beta-strands into interacting and non-interacting strands. The next step of this method is prediction at the residue-level, through scoring surface residues according to hydrophobicity and solvent accessibility that is based on secondary structure, residue type and atom type (main-chain; polar- and non-polar side-chain). If the residues are found in potentially interacting beta-strands in the first step, the PL presence is taken into account in the score too. The final step is the mapping of scored residues onto the three-dimensional structure of a protein followed by generating contours. This method was developed using a dataset of 467 proteins and tested on a dataset of 77 proteins, resulting in correct predictions (>50% specificity) for 79% of the interfaces of the testing dataset (Hoskins *et al*., 2006).

An interesting combination of an SVM with interface predictive properties augmented with structural conservation displayed enhanced prediction performance compared to not incorporating structural conservation in SVM training (Chung *et al*., 2006). Structural conservation data is mined from a multiple structure alignment (MSA^{struc}), which the predictor generates and scores via a structural conservation score. The structural conservation score accounts for distances between residues at an aligned position of an MSA^{struc} while weighting them by their mutability, as defined by a mutation data matrix. A novel step follows where each generated conservation score per MSA^{struc} position is additionally weighted by normalized B-factors obtained from the query structure. This results in rigid areas of residues in a query protein receiving higher structural conservation score values than flexible regions of residues of a query protein

that cause poor alignment of MSA^{struc} regions. Along with structural conservation, which distinguishes this method from those described thus far, Chung *et al*., (2006) trained their SVM driven method using PSI-BLAST position-specific sequence profile data and solvent accessibility data of spatially neighbouring residues at the structure level. Predicted residues are finally clustered on the query protein's structure. Chung *et al*., (2006) trained their method on a dataset of 274 hetero-complexes, using three-fold cross-validation. The method's specificity and sensitivity values were 50% and 67.3%, respectively. Without structural conservation data, their method generates a lower sensitivity (59.7%) at 50% specificity using the remaining predictive properties to train their SVM, highlighting the enhanced prediction performance derived from structural conservation data.

Wang *et al*., (2006) utilized sequence profile data from HSSP MSAs like Fariselli *et al.*, (2002), but also include evolutionary conservation data derived from phylogenetic tree analysis based on the methodology of the ET method. These predictive parameters are obtained for a seed residue and its structural residue neighbours, allowing an SVM to be trained on a dataset of 69 hetero-dimers. Leave-one-out cross validation analysis was applied, resulting in accurate performance (49.7% specificity, 66.3% sensitivity, 65.4% accuracy, and 0.297 MCC; see sections 3.5.1 and 3.5.7).

The PINUP (Protein Interface residue Prediction) interface residue predictor combines residue conservation, interface residue propensity, and an energy score to form a three-part empirical score function. It is the first predictor to utilize side-chain energy calculation to distinguish interface residues with higher side-chain energies than side-chains of ROS residues (Liang *et al*., 2006). The side-chain energy score is composed of knowledge-based and energy terms optimized by their respective weights. PSI-BLAST sequence profile data were used to compute the residue conservation term. The residue propensity term defines the contribution of residues to protein interfaces normalized by their contribution to the ROS of proteins and weighted by their accessible surface areas in interface and ROS areas. PINUP was trained on a set of 57 transient proteins, using leave-one-out cross validation, and produced 44.5% specificity and 42.2% sensitivity prediction performance values. PINUP was tester further on an independent dataset of 68 transient proteins, resulting in 29.4% specificity and 30.5% sensitivity values. The PINUP authors noted that the independent testing dataset had binding interfaces less

conserved than those in the training set, which may have contributed to the reduction in PINUP's prediction performance on the independent dataset.

de Vries *et al*., (2006) developed a method, WHISCY (WHat Information does Surface Conservation Yield?), that extracts evolutionary conservation data from a HSSP MSA through the use of a mutation data matrix-based evolutionary conservation score. All solvent accessible residues of a protein are assigned conservation scores weighted by the residues' interface propensities and mapped back to the protein surface. Following this, closely proximal surface residues are scored higher than isolated residues by a smoothing function. In the last step, high scoring surfaces residues are predicted as interface residues. WHISCY was developed using datasets of 57 (developmental) and 38 (testing) transient proteins, respectively. Its performance on the developmental dataset was 33% specificity and 30% sensitivity. The performance values of WHISCY evaluated on the test dataset were 40.8% specificity and 26.7% sensitivity.

Interface residues are solvent accessible in an unbound protein and buried upon binding. Methods designed for predicting solvent accessibility of residues from sequence alone predict interface residues as buried (Porollo and Meller, 2007). This observation has been exploited as a novel interface predictive 'fingerprint' in the SPPIDER predictor (Solvent accessibility-based Protein-Protein Interaction sites IDEntification and Recognition). The difference between solvent accessibility calculated from an unbound structure and predicted solvent accessibility from sequence alone is computed and used as an interface residue fingerprint (Porollo and Meller, 2007). This novel fingerprint and various other structural and sequence-related predictive features were applied. The sequence-based features are MSA-derived evolutionary conservation using the Shannon entropy and PSI-BLAST sequence profiles, and residue features (side-chain size, residue type and frequency, charge and hydrophobicity). The structure-based features employed were residue contact numbers and hydropathy constants. All predictive features were applied to train SPPIDER, which is a neural network-based method, via k-fold cross validation on a training set of 435 proteins (homo- and heterocomplexes). In addition a separate independent dataset of 149 (homo- and heterocomplexes) was used for validation. Overall, the authors found that SPPIDER achieved accurate prediction performance for the testing and (63.7% specificity and 60.3% sensitivity values) and training sets (67% specificity and 52.7% sensitivity values). In addition,

they showed that minus the novel solvent accessibility fingerprint, methods utilizing the other predictive features have less effective predictive performance (Matthews correlation coefficient- MCC of ~ 0.3) relative to SPPIDER (> 0.42 MCC), underscoring the usefulness of this novel fingerprint. Finally, further testing of SPPIDER on a dataset of transient proteins (43 interfaces) generated 47% specificity and 43% sensitivity.

HotPatch generates patches for high scoring residues through clustering based on spatial proximity and score (Pettit *et al*., 2007). The predictive structural properties used for scoring individual residues are concavity, surface roughness, electrostatic charge and potential. Predicted patches are assigned confidence values indicating their likelihood of functional importance. HotPatch, which is neural network-based, was trained (using jack-knifing) on a dataset of 618 varied protein interactions, including protein-protein interactions. The performance of HotPatch was ≥ 33% specificity.

Another predictor using structural predictive properties is InterProSurf (Negi and Braun, 2007). Interface residue propensity, defined as the contribution of a specific residue type in an interface normalized by the residue's contribution to the entire surface of a protein, and solvent accessibility, were used as discriminative parameters. Given a query protein, InterProSurf generates surface clusters or alternatively the neighbour density for each surface residue to form patches. Whichever surface partitioning method is applied, the clusters (or patches) produced are scored according to the predictive properties applied in the predictor and ranked by score. The predictor was trained on 72 protein complexes and independently validated on 21 protein complexes, achieving an accuracy value of ~70% (for both datasets) by using either surface partitioning technique.

Konc and Janežič (2007) used structural conservation data for prediction binding sites. In their method a query protein with a known partner protein is compared to related proteins through structural alignment to identify the most conserved surface region between the query protein and its related proteins. This conserved site is isolated from the first query protein and is compared to the second query protein's surface to find the best match on its surface using distance matrices and similarity of physical chemical features of the surface atoms. The best matching surface from the second query protein

is predicted as its binding site. The two surface sites isolated from both query chains are taken as the final prediction. This method was developed using a total of 8 proteins (6 obligate hetero-and homo-dimers and 2 transients). Overall, the method generated ~ 42% specificity and ~ 46.6% sensitivity values.

The PIER (Protein IntErface Recognition) predictor uses structural predictive discriminators (Kufareva *et al*., 2007). The structural properties are derived from statistical analyses of structures at atomic resolution to generate atom groups with significant predictive discrimination. This structural data is coupled with computed solvent accessibility for atom groups. PIER initially generates the solvent accessible surface of a query protein and extends it by a further 3Å. This is followed by the generation of uniformly spaced surface points on the protein surface. The local neighbourhood of solvent accessible atoms within a specific radius of a surface point are determined, creating surface patches. A surface patch is scored based on the structural discriminators utilized by PIER. Finally the scores are assigned to surface residues. Surface residues above a specific cut-off are predicted as interface residues. PIER was trained (using three-fold cross validation) on a dataset of 748 proteins composed of permanent and transient hetero- and homo-complexes, producing predictions at 60% specificity and 50% sensitivity values. Additionally, two independent validation datasets were used to gauge PIER's performance. The first dataset like the training set is mixed and composed of obligate and transient interactions. For this dataset the results were similar to those obtained using the training set (61.8% specificity and 50% sensitivity values). The second dataset was composed solely of transient interactions (340 interfaces obtained from 91 transient complexes). PIER generated predictions of reasonable accuracy (≥25% specificity at 50% sensitivity) for ~82% of the dataset (Kufareva *et al*., 2007).

1.9 Advantages and disadvantages of previous protein-protein interface predictors

Tables A-1 and A-2 of the Appendix summarize and include details regarding the advantages (in green) and drawbacks (in red) of the predictors discussed. The following categories are examined: - (i) the dataset essentials utilized for predictor development (tables A-1 and A-2); (ii) the training of predictors (Machine learning predictors only) and the application of benchmarking (table A-2); (iii) the use of structural and/or sequence data and the miscellaneous advantages and disadvantages of a predictor, including the availability of a predictor webserver or download (table A-2).

1.9.1 Protein dataset essentials: protein complex types' influence on predictor performance

Table A-1 shows the developmental (training/testing) and independent testing datasets used for the interface predictors with their performance measures for the specific dataset used. It can be seen that most interface predictors (15) are trained and validated on mixed datasets of different protein complex interaction types. These mixed datasets are composed mainly of obligate (or permanent) and transient complexes (see section 1.3). A few methods are trained and tested on transient complexes in comparison. The protein complex interaction types contained in those datasets differ in terms of their interface characteristics (Ofran and Rost, 2003b). For instance, interface size, hydrophobicity, and binding affinity are more prominent in obligate complexes than transient complexes (see section 1.4). In addition, obligate interfaces exhibit a strong evolutionary conservation signal and thus are more conserved than transient interfaces (Dey *et al*., 2010; Bradford and Westhead, 2006; Mintseris and Weng, 2005a; Caffrey *et al*., 2004). Moreover, obligate proteins cannot exist in the unbound form, as they are unstable. In the context of 'blindly' predicting unknown interfaces their 'prediction' as part of mixed datasets is a non-biologically relevant problem, because they only exist in the bound form in a complex. A more interesting and motivating challenge is to solve the complex of two unbound transient proteins through the aid of interface prediction. This scenario

would be relevant to a biologist intent on determining an unknown complex for proteins in the unbound form that are known to interact. Interestingly, computational methods have been developed to distinguish between obligate and transient interaction complexes, performing accurately in their predictions (Aziz *et al*., 2011; Zhu *et al*., 2006). These studies highlighted the differences between obligate and transient interfaces, which were utilized successfully in prediction for a biologically relevant problem in structural biology of distinguishing between obligate and transient interactions. As such, it would seem that obligate interactions are more suited for the protein complex type discrimination problem, which requires as input a protein complex for testing, rather than the interface prediction problem that requires monomeric (transient) proteins. Furthermore, interface predictors are increasingly utilized with protein-protein docking to predict the complex of two transient proteins known to interact (see section 1.10). Obligate proteins do not exist in the unbound form and hence generating interface prediction data for them and using them in protein-protein docking studies to dock obligate proteins is not biologically meaningful.

The differences of interface features results in much lower interface prediction difficulty for obligate interfaces than transient ones, influencing interface residue prediction performance. A consequence of this is that a interface predictor developed and validated using protein complex mixed datasets primarily consisting of obligate (or permanent) interfaces will perform better (higher specificity and sensitivity) in general than a predictor solely trained and validated on transient interfaces because these interfaces are easier to predict. Indeed, it can be seen that from the overall 15 predictors trained on mixed datasets, 6 of them were also independently validated using transient only datasets or mixed datasets and have specificities and sensitivities available (see table A-1). It can be observed that 5 predictors (cons-PPISP, PPI-Pred, SPPIDER, PIER, and Bradford and Westhead, 2006's predictor) perform better on mixed (developmental and/or independent) datasets then transient only datasets (developmental and/or independent) overall. An example of this is SPPIDER, which performs better in specificity (+20%) and sensitivity (+10%) for its mixed developmental dataset than its transient only independent dataset. The sixth predictor (Bordner and Abagyan (2005)) performed better for obligate complex types than transient ones. Although this predictor had similar sensitivities for both obligate and transient datasets, the specificity performance indicates much more accurate predictions for obligate proteins. In

summary, including obligate complexes in training and testing datasets boosts the performance of a predictor and this explains why predictors trained solely on transient proteins have lower reported performances (ex. PINUP and WHISCY) in comparison to those trained on mixed datasets (ex. SPPIDER).

1.9.2 Protein dataset essentials: bound vs. unbound transient proteins in datasets

Regarding dataset essentials, most predictors used datasets composed of bound transient protein-protein complexes where each protein chain of a complex is extracted. In contrast, an unbound transient protein refers to an independently determined protein monomer not in a complex with another protein. Bound transient proteins used for predictor development and testing may introduce potential bias by either influencing the interface residue predictive properties derived for prediction or artificially boosting prediction performance of a predictor when using bound models for testing and performance assessment. For example, a predictor that utilizes solvent accessibility data of bound interface residues may be using a biased property as it is linked to conformation change of a protein. It would be more ideal to utilize interface residue solvent accessibility data from unbound models to simulate a blind prediction setting, as interface residues are not in a buried and bound interacting pose. Another example of this is the use of B-factors obtained for interface residues as predictive discriminators for prediction. For the method of Chung *et al*. (2006) bound models are used to derive this data. Although interface residues have lower B-factors than ROS residues in the unbound form, in bound models their B-factor values are lower (Neuvirth *et al*., (2004). This may result in optimistic performance of a method using such data derived from bound models (and not unbound models) since interface residues have lower B-factors due to being in a ‘locked’ binding pose. Indeed, Porollo and Meller (2007) showed that incorporating B-factor data in SPPIDER and testing it on a dataset of 21 bound proteins resulted in better performance on the bound models (64.5% specificity and 37.5% sensitivity) in comparison to SPPIDER’s performance on their unbound protein counterparts (59.2% specificity and 31.4% sensitivity). Consequently, this possible bias of deriving B-factors data from bound models may influence the structural conservation predictive feature implemented in the method of Chung *et al*., (2006) where each conservation score assigned to a residue is normalized based on the ‘flexibility’ of the

region in which it is found. The effect is higher conservation scores for rigid residues than flexible ones, leading to optimistically reported performance values. An additional example is the use of bound model data in a prediction workflow, potentially introducing an unintended bias. This is manifested in the method of Konc and Janežič (2007), which was tested using a bound input protein ‘extracted’ from a known complex to predict the binding site of the input protein’s partner, which is biased developmental testing of their method. This can be avoided entirely through the use of an unbound protein model as input. Furthermore, during the first step of its prediction protocol (structural alignment between the query protein with its related proteins), a closer examination of the related proteins used in this step revealed that, for some instances, the related bound proteins were identical to the input protein and the protein partner for which predictions of their binding surfaces were sought. Such an error in this method could be avoided through use of a filtering heuristic prior to performing the first step in this method to remove identical ‘related’ (bound) proteins to the query protein.

Any potential bias introduced in the extracted ‘bound’ interface predictive features (ex. solvent accessibility or B-factor data) is accentuated by including spatial contiguity data of interface residue neighbours in the bound form for training a method. Interface residues are ‘closer’ during complexation than they would otherwise be in the unbound and uncomplexed form. This also highlights the linkage of residue spatial proximity with the conformational configuration of a protein. Similarly, biased predictive properties may indirectly affect other predictive properties not linked directly to conformational change of a protein used (ex. sequence profile data). For example, a predictor, which clusters its predicted residues to generate clusters, and is validated on bound proteins, may result in the enhancement of the effect of clustering and may artificially boost the effect of sequence profile data. This would be due to bound interface residues’ closer proximity to each other. For example, two highly conserved residues may be in closer spatial contiguity at the bound structural level, but not at the unbound structural level to be included in a cluster. Therefore a potential consequence is that a ‘bound cluster’ has a stronger conservation signal then the ‘unbound’ cluster, which may be lacking some of the conserved residues present in the ‘bound’ cluster.

Zhou and Shan (2005; 2001) tested their PPSIP predictor on the effect of unbound vs. bound change on a dataset of 35 unbound transient proteins and their bound counterparts. They did not find a difference between the 35 protein samples in overall specificity (bound 69% vs. unbound 70%) and concluded that their predictor performed as accurately on unbound models as bound models. In addition, cons-PPISP was also tested on Benchmark 1.0 bound and unbound models, producing slightly lower unbound specificity (61.4%) and sensitivity (38%) values relative to the bound models (63.6% specificity and 39% sensitivity) (Chen and Zhou, 2005; Chen *et al*., 2003a). Also, Bradford and Westhead (2005) tested PPI-Pred on a small dataset of 10 unbound models with their bound chain equivalents. They achieved success for 9 unbound models (> 50% sensitivity). However, all these analyses used protein datasets composed mainly of small-induced fit conformational changes during protein complexation and, in addition, focused on a specific prediction setting. For example, the 10 unbound model dataset used by Bradford and Westhead (2005) was composed mainly of the 'rigid-body' category (*i.e.* have minimal conformational change) as categorized by docking difficulty. This may have been the reason that the methods' performance was similar on bound and unbound models for transient proteins of their datasets due to lack of conformational differences on the dataset used by the cons-PPISP and PPI-Pred predictors. The analysis on unbound vs. bound transient models was not performed at a 'holistic' performance level, as offered through ROC analysis. Rather a specific prediction setting was used for the previous predictors. Using ROC analysis facilitates an overall representation of method performance at all possible predictor performance settings. The previous authors have not presented this in their papers when examining bound vs. unbound performance. Utilizing ROC analysis by these developers would have provided complete predictor performance without the potential bias of using only a single predictor setting (without the knowledge of predictor performance at alternative settings) for the specific testing of unbound vs. bound performance differences.

Later work by de Vries and Bonvin (2011a) examined predictor (for example WHISCY, SPPIDER, and PIER) performances on unbound vs. bound model prediction, using Benchmark 2.0 proteins (Mintseris *et al*., 2005b). ROC-like analysis (specificity vs. sensitivity plots) was performed to examine all possible scoring cut-offs for the

predictors. Even with small conformational changes, performance differed in favour of bound models for the majority of predictors. For example, performance was better overall for PIER on bound models. Additionally, at specific sensitivity ranges (approximately $\leq 45\%$), WHISCY performed better on bound models. Also, SPPIDER performed better on bound models than unbound models at approximately $\leq 20\%$ sensitivity. This differs from earlier work performed using SPIDDER on 21 bound vs. unbound models, which found only small differences in SPPIDER's performance on both datasets (Porollo and Meller, 2007). In general, the findings of de Vries and Bonvin (2011a) suggest that interface predictors perform better on bound models at specific or all sensitivity ranges. In their work, they pursue training and testing, using explicitly unbound models derived from Benchmark 2.0, to develop their consensus predictor, composed of individual predictors, for protein-protein docking (de Vries and Bonvin, 2011a).

Generally, interface predictive properties derived from bound models may introduce bias especially for the properties influenced by conformational changes. In addition using bound models may result in higher confidence in a predictor's performance during testing (Porollo and Meller, 2007). These disadvantages are avoided when using unbound models to derive predictive properties and for training and validation of a predictor. An examination of the predictors in Table A-1 shows that the majority were trained and tested on bound models. Only a few used unbound models. Additionally, predictors are increasingly utilized in docking (see below) of unbound proteins known to interact in an attempt to predict their final complexed configuration. In this regard, bound docking has minimal significance from a biological perspective, and recent development of a consensus predictor is aligned with the target of unbound (and blind) docking through generating 'unbound' docking constraints for data-driven docking (de Vries and Bonvin, 2011a).

1.9.3 Protein dataset essentials: Crystallization and antibody-antigen interactions

In some datasets, non-biological interactions caused by crystal packing interactions are not omitted. For instance, Zhou and Shan (2001) did not omit such complexes due to

the lower numbers of dimer interactions in the PDB to allow sufficient complexes for training their PPISP predictor. They used a filter of interface size (> 20 residues per protein) to select for large 'interfaces', arguing that such crystal packing binding sites are stabilized similarly to other dimer complexes in a soluble environment. The use of non-biologically relevant interactions was also present in the training and validation sets for their cons-PPISP predictor (Chen and Zhou, 2005). Crystal packing interactions are non-specific and do not have a biological function coupled with them such that they are not subjected to evolutionary pressures associated with biologically relevant interactions. These non-relevant biological complexes would not occur in a living organism (Zhu *et al*., 2006). While Zhou and Shan (2001) use interface size as a discriminator to filter some non-specific crystal interactions from their dataset, this sole criterion may not be enough, as non-relevant interactions may have large binding sites (Bahadur *et al*., 2004). In addition, crystal packing 'interfaces' are not conserved, having no different conservation as ROS residues, and this is not the case for biologically relevant interactions that have more conserved interfaces relative to their ROS residues (Dey *et al*., 2010). For a method like cons-PPISP that relies on conservation data as an interface predictive property among others, it is risky to include such complexes in their training and validation datasets. The use of non-biologically relevant complexes (homo-dimers) in training and testing a predictor is not justified and represents a disadvantage. As a consequence, attempts to remove such complexes from a dataset may involve using Swiss-Prot annotation checks as was done by Bordner and Abagyan (2005) (Magrane and UniProt Consortium, 2011).

Although the biological function of an antibody is interaction with its antigen, this is not reciprocated by antigens' biological functions to entail binding to antibodies. This makes their interfaces difficult to predict, as they do not follow the evolutionary model that assumes evolution of interfaces of two proteins that have overlapping biological functions (de Vries and Bonvin, 2008; Kufareva *et al*., 2007; Zhou and Qin, 2007). Antibody-antigen complexes mostly form through the complementary determining regions (CDRs) identifiable from their sequence variability relative to other sequence regions of an antibody in a multiple sequence alignment. Antigens, through antibody maturation, can have a number of epitopes for antibody interaction and these may overlap with other interface regions of other proteins. For predictors implementing evolutionary conservation data in their prediction protocol, this makes such interactions

unsuitable for prediction purposes. This is in contrast to other protein-protein complexes used in training and validation datasets, which have 'two-way' interactions coming from both binding partners and thus have an evolutionary model of overlapping biological functions, facilitating the use of evolutionary conservation data to identify their interfaces with other interface predictive properties. Therefore, the inclusion of antibody-antigen complexes in a dataset introduces a conflicting evolutionary model of interaction compared to that from other protein-protein complexes. Other authors have advised in the exclusion of antibody-antigen complexes from datasets (de Vries and Bonvin, 2008; Kufareva *et al*., 2007; Zhou and Qin, 2007; Liang *et al*., 2006). For example, the authors of PINUP and Crescendo excluded antibody-antigen complexes from their datasets, as their interactions were not developed under evolutionary pressure during long time spans, but under somatic cell mutations that occur swiftly, making such complexes not appropriate for prediction by their predictors since they utilize evolutionary conservation for prediction (Liang *et al*., 2006; Chelliah *et al*., 2004, 2006).

1.9.4 The training of predictors, their benchmarking to others, and their use of structural and/or sequence data

Like all predictors, machine learning-based predictors are trained on datasets of known interface and ROS residues. Since the proportion of ROS residues is greater than interface residues, the removal of some ROS residues from a dataset during cross-validation results in a decrease of FPs (*i.e.* ROS residues), biasedly affecting the accuracy measure. In a blind setting the interface and ROS residues are undetermined (Bordner and Abagyan, 2005). Only three predictors have this disadvantage; the majority of predictors do not reduce the number of ROS residues (see Table A-2).

It is important to benchmark a new predictor to existing predictors to ascertain its performance relative to theirs. This becomes important when a predictor aims to introduce, for instance, new algorithmic methodologies to improve prediction efficacy and/or applies novel interface predictive features combined with previous features to predict interface residues. A small number of predictors benchmark directly to other predictors, whereas a larger number do not benchmark directly, and with direct

benchmarking, all standard performance metrics can be used to gauge different aspects of each benchmarked predictors performance. In addition, the majority of predictors utilized interface predictive properties from both structure and sequence data, and eight other methods either are sequence- or structure-based. No given interface predictive property can unequivocally be used to absolutely predict interface residues amongst all the surface residues. However, utilizing multiple predictive properties from structure and sequence sources of data in a predictor provides a greater advantage than predictors that depend exclusively on either source of data.

Additionally, it is more advantageous to utilize the latest sequence data with structural data for a predictor. Some predictors do not utilize the latest sequence data. As an example, there are some predictors that utilize sequence data (*i.e.* MSAs) obtained from the HSSP database (Dodge *et al*., 1998). The HSSP database is not updated completely; instead only HSSP MSAs older than 6 months are updated in a weekly cycle (Joosten *et al*., 2010). Consequently, HSSP alignments under 6 months old, which do not contain the latest sequence data, are present in the HSSP database. This means a predictor may not have access to latest sequence data for a given prediction. Interestingly, a predictor that relies on HSSP alignments, and is applied to generate predictions for newly deposited structures in the PDB, would not be able to generate a prediction straight away. This is because newly deposited proteins do not have HSSP alignments immediately generated for them in the HSSP database.

1.9.5 The miscellaneous advantages and disadvantages of predictors and their availability

There were specific advantages or drawbacks unique to or shared by predictors (see Miscellaneous column of Table A-2). As shown in table A-2, there were a number of predictors that utilized manually curated datasets, avoiding the pitfalls of crystallization packing interactions (ex. PPI-Pred). Also, in some cases antibody-antigen complexes were excluded (ex. ProMate, WHISCY, and PIER), whilst in others they were present (ex. SHARPE^2). In summary, the use of manually curated datasets represents an advantage in relation to other predictors that were mainly developed using automatically generated datasets and in a specific instance using 77% of PDB-1999 (ex.

Keil *et al*., 2004). The size of the dataset for testing is also important. Using a small dataset would not be a robust means of testing a predictor's performance. For example, Konc and Janežič (2007) used 8 proteins to test their method. It would have been more appropriate to test their method using manually curated datasets like Benchmark 2.0 or those used by earlier predictors to systematically evaluate their predictor. Likewise, it is important to remove redundancy in a dataset to prevent bias in terms of artificially high confidence in prediction performance. A useful example is illustrated using the method of Fariselli *et al.* (2002) where originally a dataset of 452 proteins was used to validate it. This dataset was not checked for redundancy. When redundancy filtering was applied, only 59 chains were found to be non-redundant (Porollo and Meller, 2007). The testing of Fariselli *et al.* (2002)'s method on both datasets, using 10-fold cross validation, resulted in marked differences in performance. Specifically, using the redundant dataset this predictor yielded a 0.43 MCC. In contrast, a much lower MCC value of 0.28 was obtained when using the non-redundant dataset (Porollo and Meller, 2007). Thus, high redundancy in a dataset affects the performance evaluation of a predictor and should be filtered.

Generated predictions are compared, via a performance metric, to the actual number of interface and ROS residues, which are determined from experimentally available complexes (see section 3.5). In order to evaluate performance of a given predictor, most predictors use the actual interface and ROS residue numbers determined beforehand prior to prediction. This data would be used to ascertain their proportions in all predictions for proteins in a dataset. If ROS residues are present in predictions they are classed as false positives (FPs). Likewise all interface residues correctly predicted would be classed as true positives (TPs). Only two predictors do not adhere to this performance evaluation practice and convert some false positives into 'true positives' based on their proximity to true positives present in their predictions (ex. cons-PPISP and Chung *et al*., 2006). If such FP residues are in reality TPs then they should not be included initially as ROS residues prior to prediction to avoid introducing bias in performance assessment (*i.e.* increase the specificity value). The initial separation of surface residues into interface and ROS residues prior to prediction must remain fixed to enable unbiased performance evaluation. The conversion of some FPs into TPs, which follows after a prediction, is biased and boosts their predictor's specificity performance. An example of this was the NMR dataset used to independently validate

cons-PPISP; the specificity differed when accounting for TPs only (38.5%) and when accounting for TPs and FPs (converted to TPs) in close proximity (69%) at 47% sensitivity.

There are unique drawbacks to certain predictors. The methods of Fariselli *et al*., (2002) and Wang *et al*., (2006) both used permissive interface residues definitions (12 Å between alpha-Carbons of opposing protein chains). This resulted in 40% and 34.8% of their surface residues designated as interface residues for the former and latter predictors. This makes it easier to predict interface residues and improves predictor performance (Chen and Zhou, 2005). In addition, when this is coupled with a highly redundant dataset as used by Fariselli *et al*., (2002); the outcome is an optimistic performance of their predictor (72% specificity and 56% sensitivity).

Some predictors have notable advantageous features, for example the ProMate predictor is designed to be able to accept potentially new interface predictive properties as input (Neuvirth *et al*., 2007). Noteworthy predictive properties like SPPIDER's novel solvent accessibility fingerprint and the utilization of structural conservation in the predictor of Chung *et al*., (2006) demonstrate advantageous features unique to these predictors that improve the predictors' performance. All predictors developed were tested on datasets where an experimentally determined complex is known. This means the interface residues can be determined prior to testing. Of course most proteins bind to more than one partner and such complexes are not represented in a dataset for training and validation of a predictor. Therefore these unrepresented interfaces are considered ROS residues. In their valuable study, Porollo and Meller (2007) incorporate other known interface residues for each protein of their dataset, including the 'actual' interface of the specific experimentally determined complex of their dataset. This approach for developing a predictor has a further advantage in that potential false positives, which are in fact alternative interface residues, are avoided and assigned as TPs when correctly predicted. This prevents decreases in predictor specificity if they are not taken into account. Over half of the predictors are available to users as webservers (or downloads; see table A-2). This is beneficial and a user-friendly service. Also, it is useful for benchmarking of new predictors and fosters further development, as demonstrated in the recent progress of consensus prediction combined with protein-protein docking (de Vries and Bonvin, 2011a; Qin and Zhou, 2007a).

1.10 The use of interface predictors in combination with protein-protein docking

Interface residue predictors have been combined with protein-protein docking approaches to predict the binding poses of proteins known to interact. This is accomplished directly to drive docking sampling by reducing the sampling search space (*i.e.* front-end docking), or indirectly by re-ranking docked models via scoring (*i.e.* back-end docking). An early implementation of an interface residue predictor combined with back-end docking was achieved using an ET-based method (Aloy *et al*., 2001). Five enzyme-inhibitor complexes' unbound proteins were docked using FTDock (Gabb *et al*., 1997). The resulting docked poses were retained, if a protein's surface was within a close distance to a functional residue prediction obtained from the predictor for its partner protein, while all other complexes not fulfilling this criterion were disregarded. It was shown that the prediction-based distance constraint data filter improved the ranking of near-native models (ranked within top 25 models) compared to no filtering applied (ranked 87th and above). Specifically, four complexes resulted in near-native models found within the top-ten docking predictions. In the CAPRI competition, Ben-Zeev *et al*., (2003) utilized ET method prediction data in front-end docking to predict the best putative docking model of the HPR kinase-HPR complex using its unbound proteins. A model of acceptable quality was generated using MolFit (Heifetz *et al*., 2002). Zhu and Tytgat (2004) used the ET method to predict binding sites from Hsp90 and p23 proteins. They mapped this data to the best-ranked docked model for the two proteins generated using BiGGER (Palma *et al*., 2000). It was observed that they were part of the modelled complex's putative interface, suggesting a possibly putative biologically relevant complex (Zhu and Tytgat, 2004). Gottschalk *et al*., (2004) used ProMate prediction data as part of a scoring function that computes the tightness of fit for the potential interfacial sites on docked models. A benchmark of 21 enzyme-inhibitor complexes was used for docking unbound proteins using FTDock. Using their scoring function, a 77% success rate was achieved for complexes of their benchmark. van Dijk *et al*., (2005b) combined PPISP prediction data and experimental data for 8 CAPRI targets during the CAPRI competition to drive their docking using HADDOCK (Dominguez *et al*., 2003). Acceptable or above models were obtained for five targets. Specifically, 3 medium and 2 high quality models were generated. PPISP contributed

34% specificities and 32% sensitivities for the targets' predictions. Tress *et al*., (2005) applied prediction data from the method of Fariselli *et al*., (2002) with data derived from other experimental and theoretical sources, during the CAPRI competition, to identify near-native models. Docked models' putative interfaces were examined and filtered according to overlap to the prediction and experimental data. 7 CAPRI targets were docked using Hex and GRAMM, which resulted in acceptable models being produced for 4 CAPRI targets based on their filtering strategy. de Vries *et al*., (2006) applied WHISCY prediction data as restraints to drive docking of unbound proteins for 25 complexes obtained from Benchmark 2.0. Using this approach, successful results were generated for 48% of their dataset compared to *ab initio* docking using HADDOCK (0% success). Combining WHISCY and ProMate prediction data, the number of the successful cases increased to 64% of the dataset. Crescendo predictions have been applied for back-end docking (Chelliah *et al*., 2006). Docked solutions generated by pyDock were scored using distance restraints derived from Crescendo prediction data. This protocol produced near-native models ranked in the top-20 docking solutions for 80% of the 10 complex docking dataset (7 bound-unbound docking cases, 1 unbound-unbound cases, and 2 bound-bound cases). This protocol was found to be superior to the native pyDock energy score used for ranking models. Furthermore, Crescendo prediction data was also applied to drive front-end docking of four protein complexes (one was bound-unbound docking, while the rest were in the unbound form), using HADDOCK and generated near-native complexes ranked highly (1^{st} for three cases and 4^{th} for one case). Kanamori *et al*., (2007) developed a docking method that applies shape complementarity weighted by residue evolutionary conservation retrieved from the ET method to perform docking. This method was used to dock seven CAPRI targets and it produced near-native models for 5 targets. Qin and Zhou (2007b) combined biochemical data with prediction data from cons-PPISP predictions to rank docked models obtained from a mixture of bound-unbound or entirely unbound docking generated by ZDOCK for 24 CAPRI targets (Chen *et al*., 2003b). For 23 CAPRI cases, near-native models were generated using ZDOCK. An individual model is ranked according to the number of residues of its putative interface that match the number of residues predicted as interface in both binding partners. Using their back-end docking approach, they attained an improvement in ranking near-native models for 9 CAPRI cases compared to the native ZDOCK ranking. Martin and Schomburg (2008) utilized evolutionary conservation prediction data derived from an

ET-based method and combined it with other discriminative features for developing an SVM to differentiate and rank near-native models among decoys, which were generated by the ckordo docking method (Zimmermann, 2002). When applied to protein complexes of Benchmark 2.0, the SVM significantly ranked near-native models in the top-ten ranked models compared to ranking of docking solutions based on geometric fit.

In this work, HADDOCK was applied to study the effect of docking performance when interface prediction data was combined with experimental data and this was compared to standard experimental data-driven docking (see section 1.11 below).

1.11 Aims of this present study

This work seeks to address limitations in the current status of the field of interface prediction exclusively pertaining to unbound transient protein (hetero-complexes) interface prediction and its utilization in protein docking. The reasons that this complex type is studied were clarified in the discussion of the disadvantages and drawbacks of the previous predictors (see section 1.9). These limitations of the field are summarized in Appendix table A-3. To the best of my knowledge, this is the first study to address these limitations and solve them.

The initial limitations pertain to interface prediction through use of sequence data and clustering of prediction data: -

1) Multiple sequence alignments (MSAs) that are generated automatically may have alignment errors, which may diminish conservation signals because of the alignment error noise, and may also be out-of-date (*i.e.* HSSP). For the predictors that utilize sequence data from a multiple sequence alignment, no interface predictor has been applied to explicitly improve multiple sequence alignments prior to deriving the evolutionary conservation from them for the systematic prediction of transient hetero-complexes (see table A-3). It is predicted that explicitly reducing or eliminating sources of sequence alignment errors may result in better conservation signal retrieval and subsequently improve interface prediction. This hypothesis is tested in this work through the use of improved MSAs vs. automatically generated MSAs and the impact

they have on evolutionary conservation calculation for the interface prediction of transient proteins. Specifically it will be studied through applying a sequence editing heuristic protocol introduced in this work on the up-to-date sequence data derived from the UniRef90 database for the most recently published transient complex dataset (Benchmark 4.0). The sequence alignments generated through the sequence editing heuristic approach introduced in this study will be compared to sequence alignments generated automatically from the same data sources. Statistical hypothesis tests will be applied to test the significance of explicitly improving MSA data.

2) For the predictors that utilize structural data from an unbound transient input protein, no interface predictor has been applied to systematically address the impact of interface prediction data clustering using the three-dimensional coordinates of the input (unbound) transient protein. Clustering may improve the impact of interface residue accuracy. This important hypothesis is tested in this work through the use of clustering vs. no clustering of interface prediction data and their assessment on prediction performance. It is predicted that applying three-dimensional clustering of interface prediction data may cause elimination of ROS residues for the interface prediction of transient proteins in the unbound state, leading to improved interface prediction quality. This hypothesis will be studied by clustering prediction data derived from improved MSAs and its comparison to non-clustering. Statistical hypothesis tests will be applied to test the significance of clustering on interface prediction performance improvement. In addition, the application of clustering will be examined in the context of whether interface residues are more conserved than rest of surface residues for proteins of Benchmark 4.0 and this analysis will be discussed in the context of previous literature.

3) The sequence data editing and interface prediction data clustering heuristics will be integrated into a new predictor, PROTein INterface IDentification (PROTIN_ID), designed for the prediction of transient protein interfaces using evolutionary conservation data. This interface predictor will be developed using the Benchmark 4.0 dataset of transient proteins. In addition, PROTIN_ID will be benchmarked to the WHISCY predictor, developed using transient proteins, and the CCRXP predictor. Finally, a web-server implementation of PROTIN_ID will be designed with user-friendly features, allowing convenient access of the method to the biological community at large.

The other identified limitation relates to the application of interface predictors in combination with protein docking: -

4) The systematic application of interface prediction data-driven docking vs. *ab initio* docking and its evaluation using stringent CAPRI evaluation metrics will be performed to explore interface prediction data's impact on docking performance vs. *ab initio* docking. Statistical analysis will be applied to examine data-driven docking performance. This aim paves the way for the exploration of combining interface prediction data and NMR data to drive docking. It has been observed that interface predictors have been used in successful front-end or back-end docking of transient proteins (see section 1.10). However, interface prediction data has not been combined with NMR data (Residual dipolar coupling (RDC) and chemical shift perturbation (CSP) data). Currently, standard data-driven docking using CSP and RDC data from NMR is able to produce good quality models. If using interface prediction data has been shown to have a successful impact on docking performance, then it is hypothesised that combining this data with NMR data may improve front-end docking performance by increasing the number of correct docking solutions and/or their quality according to CAPRI metrics. This improvement may arise from greater interface accuracy and coverage caused by the 'consensus' combination of all data sources. To test this hypothesis, a dataset of protein complexes with known NMR data (RDC and CSP) will be used and unbound docking runs of standard experimental NMR data-driven docking will be performed and compared to consensus data-driven runs using HADDOCK. This will be followed by the application of statistical hypothesis tests to investigate the significance of consensus-data driven docking.

Chapter 2

Preliminary work

2.1 Introduction

This chapter describes the preliminary work performed in this study, and it serves as a foundation to other work described in this thesis. Developers of previous interface predictors have not systematically studied and analysed the impact of multiple sequence alignment quality and the impact of three-dimensional clustering on transient proteins (hetero-complexes) for their interface predictors (see table A-3). Specifically, the first analysis described in this chapter investigates the effect of explicitly improving MSAs by reducing or eliminating causes of alignment errors from sequence data obtained for transient proteins (see section 3.2) of the Benchmark 4.0 dataset (see section 1.11 point 1 for this hypothesis). This may improve conservation signal retrieval from an MSA and hence interface prediction. A comparison will be made to automatic MSAs not subjected to any improvement (*i.e.* controls) and derived from the same sequence data sources. The second analysis studies the effect of three-dimensional clustering vs. non-clustering (*i.e.* control) of interface prediction data on unbound transient proteins derived from Benchmark 4.0 (see section 1.11 point 2 for this hypothesis). Unbound transient protein interfaces are harder to predict in comparison to non-transient proteins in the context of blind prediction (see sections 1.9.1 and 1.9.2 and table A-1). This analysis will ascertain whether elimination of ROS residues is improved after clustering, leading to improved interface prediction quality. Statistical significance tests will be applied to test the two hypotheses.

2.2 Preliminary work: multiple sequence alignment optimization

In this section, the optimization of MSAs vs. automatic MSAs on the effect of conservation signal retrieval and consequently interface prediction is studied and discussed. The optimization of MSAs was performed using the sequence editing heuristic introduced in this work (see section 4.3.1). Briefly, this heuristic generates a pair-wise alignment (via the Needleman-Wunsch algorithm) between a Benchmark 4.0 query protein's sequence and each of its homologous sequences (*i.e.* hits) retrieved from the UniProt90 database (Hwang *et al.*, 2010; Suzek *et al.*, 2007; Needleman and Wunsch, 1970). The query sequence contains only the residues present in its tertiary structure, as such sequence regions of the hit sequences like N/C terminal overhangs present only in the hit sequence and absent in the query sequence, are removed. Likewise, hit sequence insertions are removed, which cause the query sequence to split. This is important as only sequence regions present in the query 'structural' sequence and reciprocated in the each sequence hit are used for calculating residue conservation scores and projected onto the query protein structure.

2.2.1 Case study: Tissue inhibitor of metalloproteinase 1 protein

The sequence editing heuristic was written as a stand-alone PERL script and was initially tested on the Tissue inhibitor of metalloproteinase 1 protein (TIMP-1; PDB: 1D2B) and its related homologous sequence data. The hit sequence data used was filtered to remove sequence fragments and redundant sequences, leaving only hit sequences with a high fraction of coverage to the query sequence prior to testing (see section 4.3.1 and figure 4-2). The same sequence data is used to generate refined (with the heuristic) and unrefined (without the heuristic) MSAs. When 'edited' hit sequences, including the TIMP-1 query sequence, are aligned, this results in an improved and structured MSA (refined). In comparison, when generating an MSA using 'unedited' hit sequences, this causes a more unstructured MSA (unrefined) (see figure 4-4). For the TIMP-1 example, the quality of an interface residue prediction is tested using both refined and unrefined TIMP-1 MSAs. Reducing or eliminating sources of sequence

alignment errors (N/C terminal overhangs and insertions) that cause misalignment in an MSA may result in better conservation signal retrieval. To evaluate their impact on interface residue prediction, both alignments were used to calculate evolutionary conservation scores for their MSA columns. Only MSA columns representing surface residues of the TIMP-1 protein were extracted to avoid false positive results caused by more conserved protein core residues (see section 3.2.1). Rank analysis, which computes the fraction of top-20 ranked interface residue columns, was performed (Wang and Samudrala, 2006). The top-20 hits score is set for the best 20 conserved surface residues since it is equal to the average size of a protein interface obtained from the Benchmark 4.0 dataset (see table 5-2). For TIMP-1, it was determined that the fraction of interface residues according to the top-20 hits score was higher for the refined MSA (0.35) in contrast to the unrefined (0.20) MSA. Using the top-20 hit score analysis, this initial result on TIMP-1 suggests that it is possible to improve ranking of interface residues using evolutionary conservation with a refined MSA. For meaningful statistical evaluation, this initial analysis was extended to include all proteins derived from intra-species protein complexes of Benchmark 4.0 (see section 3.2).

2.2.2 Refined vs. unrefined MSAs and their impact on interface residue prediction

For all intra-species interacting proteins of Benchmark 4.0, sequence data for each protein were obtained and filtered in the manner described previously for the TIMP-1 test case. Following this, each protein's sequence data (query and hits) was submitted to the editing heuristic protocol to generate refined MSAs. The sequence data was also used to generate unrefined MSAs. The top-20 hit score analysis was performed for each MSA of the dataset and an average for the whole dataset was obtained using the refined and unrefined MSAs. Improving an MSA may not result in better interface prediction of transient protein interfaces and this null hypothesis (H_o) states that there is no difference in interface prediction performance when using either unrefined or refined MSAs. Statistical analyses were performed to test this hypothesis and demonstrate (actual hypothesis, H_a) that refined MSAs improved transient protein interface residue prediction.

The Wilcoxon matched pairs test was applied to test the effect of refined alignments on

interface prediction against their unrefined counterparts. This test was performed using GraphPad Prism (version 5.00) at default settings (GraphPad Software). The fraction of interface residues as determined by the top-20 hit scores for refined and unrefined MSAs for each protein of the dataset were compared to determine if refined MSAs produced better interface residue predictions such that their top-20 hit score difference ($\Delta_{top\text{-}20}$) was statistically significant at the 5% significance level. In addition, the 95% confidence intervals (CI) for the upper and lower bound limits were calculated for the top-20 hit score difference between both MSA types, using the bootstrap analysis (using 1000 randomly selected samples to calculate sample means per bootstrap repetition), as implemented in STATA version 11 (StataCorp LP, 2009).

The difference in the top-20 hit score ($\Delta_{top\text{-}20}$) between refined and unrefined alignments was computed for the dataset (see table 2-1). $\Delta_{top\text{-}20}$ may be positive (*i.e.* $\Delta_{top\text{-}20} > 0$), which indicates that refined MSAs have greater enrichment of interface residues, or negative, indicating the opposite (*i.e.* $\Delta_{top\text{-}20} < 0$). A $\Delta_{top\text{-}20}$ of 0 indicates no improvement provided by either refined or unrefined alignments. It can be seen that on average $\Delta_{top\text{-}20} > 0$, indicating that refined alignments have more interface residues with a high ranking based on evolutionary conservation. This suggests that the conservation signal is better detected from these refined alignments than unrefined alignments and translates into better interface predictions. The sequence editing heuristic is having its intended effect by removing sources of misalignment errors like N/C terminal overhangs or insertions both of which arise from hit sequences. This finding (*i.e.* $\Delta_{top\text{-}20} > 0$) is supported by the result of the 95% CI analysis, which quantifies the precision of the $\Delta_{top\text{-}20}$ population mean (μ) from which the analysed dataset represents a sample, and depicts the 95% chance that the true population mean is between its upper and lower boundaries (see table 2-1). Most importantly from this analysis it is shown that the lower bound limit of the 95% CI (0.07) is greater than zero ($\Delta_{top\text{-}20} = 0$), essentially excluding the possibility that $\Delta_{top\text{-}20} < 0$ where the unrefined MSA dataset on the whole exhibits a stronger conservation signal and better interface residue prediction compared to the refined MSA dataset. This indicates the usefulness of the sequence editing heuristic protocol in diminishing the effect of misalignment-inducing errors in up-to-date sequence data, resulting in more structured MSAs and ultimately better interface residue predictions compared to an approach in which it is not applied.

Complementing the 95% CI is the p-value calculated using the Wilcoxon matched pair test. This statistical test evaluates the statistical significance of the null hypothesis that states that there is no difference in interface prediction performance using residue conservation from either unrefined or refined MSAs. As shown in table 2-1, it can be seen that the null hypothesis is rejected and indicates the difference in favour of the refined MSAs (*i.e.* $\Delta_{top\text{-}20} > 0$) is statistically significant (p-value <0.0001).

Table 2-1: Comparison of top-20 hit score averages for the refined and unrefined multiple sequence alignments (MSA). The standard deviations are indicated in parentheses. The fractional difference average indicates that the majority of the refined MSAs generate a stronger conservation signal, enriching the top-20 most conserved surface residues with more interface residues than the unrefined MSAs (*i.e.* $\Delta_{top\text{-}20} > 0$). The 95% Confidence Interval indicates the upper and lower bound range limits of $\Delta_{top\text{-}20}$. The Wilcoxon matched pairs test P value indicates the probability that $\Delta_{top\text{-}20} > 0$ is statistically significant at the 5% significance level (see table A-4 of Appendix for dataset examined).

	Editing heuristic: refined MSA	**Non-editing heuristic: unrefined MSA**	**Fractional difference ($\Delta_{top\text{-}20}$)**	**P value $\Delta_{top\text{-}20} > 0$**	**95% Confidence Interval ($\Delta_{top\text{-}20}$)**
Top-20 hit score	0.34 (0.22)	0.24 (0.20)	0.10 (0.17)	<0.0001[a]	0.07 – 0.13

[a] P value < 0.0001 indicates extreme significance.

In the dataset, there were some cases where $\Delta_{top\text{-}20} < 0$ (see table A-4). It was ascertained that for such cases ROS residues that did not generate a high conservation score in the unrefined MSAs had high conservation scores in refined MSAs and were ranked among the top-20 surface residues. The MSA regions of "displacing" ROS residues aligned better in refined MSAs compared to unrefined MSAs and hence generated higher conservation scores. This may be due to the displacing ROS residues being part of other binding interfaces (*i.e.* crypto-interface residues) with high conservation signals. These ROS residues scored higher than interface residues in these

cases and displaced them from the top-20 ranked residues for refined MSA cases, resulting in $\Delta_{top\text{-}20} < 0$ for these cases.

2.3 Preliminary work: the effect of clustering vs. non-clustering on interface prediction

In this section, the effect of three-dimensional clustering of interface prediction data is examined and compared to non-clustering of the same prediction data. The working hypothesis is that there is a potential elimination of ROS residues from interface prediction data using the clustering approach, leading to improved interface prediction quality (actual hypothesis, H_a). Alternatively there may be no difference between the two approaches (null hypothesis, H_o). Both scenarios were examined to ascertain the extent of the effect of the clustering approach on interface prediction reliability.

Clustering is useful as it can identify potential interface residues, which are spatially contiguous, and likely functionally important. Also, an important effect of clustering is the removal of isolated residues on a different part of a protein's surface from the predicted interface (Guharoy and Chakrabarti, 2010; Ofran and Rost, 2007b). Isolated residues can erroneously influence protein-protein docking sampling, if used as input with a potential functionally important cluster in a docking method, leading to biologically irrelevant docking solutions. The clustering protocol implemented is described in detail in chapter 4 (section 4.3.2). Briefly, the top-N most evolutionary conserved surface residues are extracted and their carbon-α distances relative to one another are determined and stored in an all-against-all distance matrix (see figure 4-5). Single-linkage clustering is applied to cluster top-N predicted residues within a carbon-α radial distance cut-off (default ≤ 7Å, see section 5.7). The largest cluster formed from the top-N surface residues is considered the final prediction (see figure 4-6). If two or more clusters are of the same size, they are sorted by the average cluster conservation. Using the clustering protocol the top-ranking cluster is taken as the final interface prediction. Compared to the clustering protocol, the non-clustering protocol assigns all top-N evolutionary conserved surface residues as the final interface prediction data.

2.3.1 Analysis of clustering vs. non-clustering protocols

The clustering and non-clustering protocols were implemented in a PERL script and were tested on a dataset of 123 proteins derived from Benchmark 4.0 (see section 3.2). The top-N evolutionary conserved surface residues were extracted from refined MSAs for each protein of the dataset. For each top-N residue extraction, the clustering protocol was applied to derive the top-ranking cluster as the final prediction. Likewise, the non-clustering approach predicted all top-N surface residues as interface residues. This top-N residue extraction procedure was iterated one surface residue at a time starting from the most conserved surface residue, incrementing the next conserved surface residue, until all possible top-N residues were extracted and inputted in both clustering and non-clustering approaches. Both approaches' interface prediction data for each top-N extraction cut-off point were used to generate receiver operator characteristic (ROC) curves (see figure 2-1). The areas under the curve were computed for false positive rate (FPR; see section 3.5.5) ranges of 1.0 ($AUC_{1.0}$) and 0.166 ($AUC_{0.166}$; this analysis will be explained below) for both clustering and non-clustering approaches. A numerical comparison of AUCs ($AUC_{cluster}$ and $AUC_{non\text{-}cluster}$) at both FPR ranges was performed to determine if a difference ($\Delta_{AUC} > 0$, H_a) is present and statistically significant or not ($\Delta_{AUC} = 0$, H_o) at the 5% significance level, using a standard AUC comparison statistical test (Sonego *et al*., 2007). In addition, the 95% CI analysis was applied to determine the upper and lower bound limits of the Δ_{AUC} for both FPR ranges to complement the AUC comparison statistical test. All analyses performed above were carried out using GraphPad Prism (version 5.00).

For $AUC_{1.0}$ analysis, the AUC values for both clustering and non-clustering approaches are indicated in table 2-2. Although $\Delta_{AUC1.0} > 0$ and is in favour of the clustering approach, this difference is not significant since it does not support the rejection of the null hypothesis, as indicated by the p-value (0.8552). This finding is supported by the 95% CI analysis, which shows that the 95% CI range extends from negative to positive values. As the lower bound limit is a negative value, this indicates that the non-clustering approach may also have a higher AUC than the clustering approach, or no difference exists between both approaches overall relating to interface prediction performance since $\Delta_{AUC} = 0$ is within the 95% CI range. This suggests the lack of evidence to support the observation that the $AUC_{1.0}$ of the clustering approach (*i.e.* Δ_{AUC}

0.00160) is significantly better in terms of interface prediction than the non-clustering approach.

Table 2-2: Comparison of AUCs at two FPR ranges (1.0 and 0.166) for the clustering and non-clustering approaches. The 95% Confidence Interval indicates the upper and lower bound range limits of Δ_{AUC} for both FPR ranges. The AUC comparison statistical test P value indicates the probability that $\Delta_{AUC} > 0$ is statistically significant at the 5% significance level.

	Clustering	**Non-clustering**	**AUC difference (Δ_{AUC})**	**P value ($\Delta_{AUC} > 0$)**	**95% Confidence Interval (Δ_{AUC})**
$AUC_{1.0}$	0.70030	0.69870	0.00160	0.8552	-0.01563 - 0.01883
$AUC_{0.166}$	0.04760	0.04230	0.00530	0.0098[a]	0.00127 - 0.00933

[a] P value 0.001 - 0.01 indicates a very significant result.

There are two drawbacks to using the above $AUC_{1.0}$ analysis, which compares both interface prediction approaches holistically. The first drawback is that for most of the ROC analysis, a large number of top-N (ex. top-100) conserved surface residue cut-offs are used, which are not useful to compare both interface prediction approaches. This is because of the nature of the single-linkage clustering approach, which is to maximize as much as possible the nearest neighbours when generating a cluster via a specific carbon-α radial distance cut-off. Therefore, having many residues to cluster creates a cascading effect, resulting in the majority of the top-N surface residues or all of them being clustered via the clustering approach. In this scenario there would not be a major difference between both approaches, as can be seen by the ROC curves being similar for the majority of high TPR (see section 3.5.4) and FPR ranges in figure 2-1. A potential best-ranked cluster of conserved residues based on a low top-N cut-off, which may contain genuine interface residues, may merge with other distant clusters over a protein's surface simply because of this cascading effect. In the context of using interface prediction data to drive protein-protein docking, having a high top-N residue

cluster, results in a high FPR with many ROS residues in the final prediction. This is a risky use of interface prediction data derived by clustering (and even non-clustering) for data-driven docking and hence the biological non-relevance of high top-N cut-offs-derived data application in the context of data-driven docking. In a docking scenario, predictions are relevant when a sufficient TPR coupled with a low FPR are retrieved as input docking sampling restraints. Therefore, a high (or all) TPR is not useful as it introduces many false positives that can misdirect docking sampling and reduce docking performance.

Given that the above $AUC_{1.0}$ analysis focuses on all TPR and FPR cut-offs, and that high cut-offs are not relevant in the context of protein-protein docking, the application of this analysis to a region of biological relevance in the context of protein-protein docking should be pursued instead of comparing the entire ROC curves, which is misleading. This is tied to the second drawback of the $AUC_{1.0}$ analysis in this circumstance. The $AUC_{1.0}$ analysis does not indicate if the two compared approaches' ROC curves are the same overall or differ in localized regions. It can be seen that indeed they are not the same (see figure 2-1). In fact, the clustering approach seems better than the non-clustering approach at the low FPR range (0 - 0.166) and this interestingly is the region most important for biological application in the framework of protein-protein docking. The difference between the approaches is because of the reduction of ROS residue noise using the clustering approach, which is more impactful for interface prediction relative to the non-clustering approach. Besides this relevant region, there are also other regions of difference between both approaches at higher FPR ranges, which are not relevant in the context of docking. Here, the non-clustering approach performed better and this is because the clustering approach identified less interface residues within these FPR ranges, causing reduction in interface prediction performance relative to the non-clustering approach.

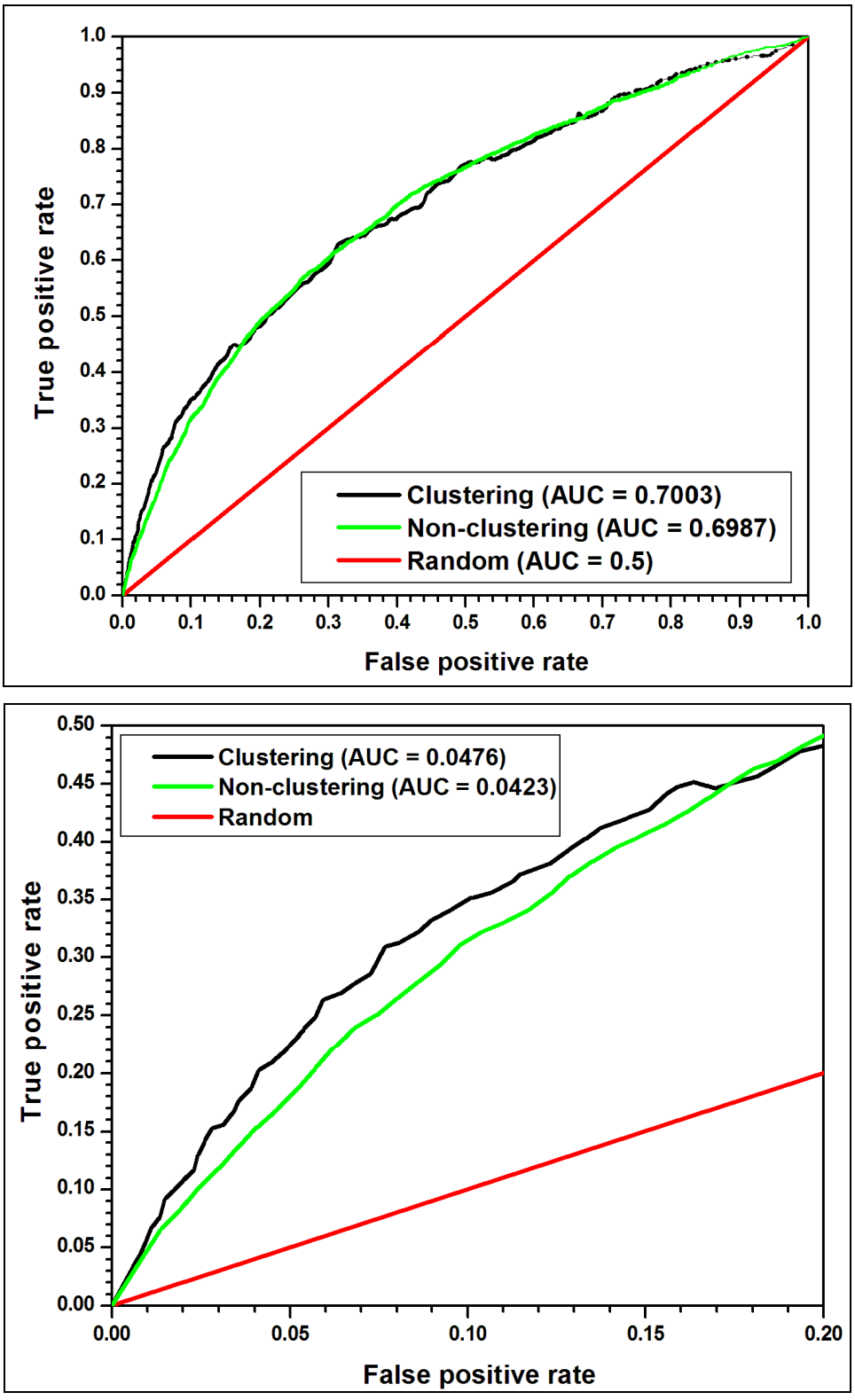

Figure 2-1: ROC curves comparing the clustering and non-clustering approaches for interface prediction. The AUCs are shown for both approaches at FPR ranges of 1.0 (top plot – $AUC_{1.0}$) and 0.166 (bottom plot – $AUC_{0.166}$), respectively.

The low FPR range (0 - 0.166) of the ROC curves was compared using $AUC_{0.166}$ analysis. For $AUC_{0.166}$ analysis, the AUC values for both clustering and non-clustering approaches indicate that $\Delta_{AUC0.166} > 0$, which is in favour of the clustering approach. This difference is significant and supports the rejection of the null hypothesis (p-value 0.0098), which states that no difference between these interface prediction approaches exists. In addition, this outcome is supported by the 95% CI analysis, which shows that the 95% CI lower bound limit is greater than zero, excluding the possibility that Δ_{AUC} is zero or lower. To summarize, there is a 95% chance that the 95% CI upper and lower bound limits support the observation that $\Delta_{AUC0.166} > 0$ at a low FPR range (*i.e.* biologically relevant range) upon expanding the size of the dataset. Using this inference, evidence exists to support the use of the clustering approach (*i.e.* H_a) since it significantly improves interface prediction quality at a biologically relevant low FPR range compared to the non-clustering approach. This demonstrates its usefulness in improving interface prediction quality of transient proteins for its utilization for protein-protein docking in comparison to lack of clustering.

2.4 Conclusion

The overall findings indicate that interface predictions can be significantly improved using explicitly refined multiple sequence alignments and three-dimensional clustering. These novel findings have not been systematically explored in previous work (see table A-3). Based on the above work, a new interface predictor to identify protein interfaces will be created (PROTIN_ID – PROTein INterface IDentification), which utilizes both state-of-the-art sequence editing and clustering heuristics. The default parameters of PROTIN_ID (top-20 conserved surface residues with clustering at 7 Angstroms) represent its operating point, which is within the significant 0 - 0.166 FPR range (see figure 2-1). The PROTIN_ID interface predictor is discussed in detail in chapter 4 and is benchmarked to other interface residue predictors (see chapter 6). Furthermore, the PROTIN_ID interface prediction data is used to compare data-driven docking to *ab initio* docking followed by its combination with experimental data to drive protein-protein docking in a novel docking study (see Chapter 7).

Chapter 3

Methods

3.1 Protein sequence database

The UniProt Reference Cluster dataset 90 (UniRef90) is composed of protein sequences derived from the Universal Protein Resource database (UniProt), which is a unified and comprehensive repository of protein sequences (Magrane and UniProt Consortium, 2011). UniRef90 contains protein sequences clustered at a 90% sequence identity threshold to reduce the presence of redundant sequences. The UniRef90 database was selected as it allows broad coverage of sequence space at 90% sequence identity resolution, hiding redundant sequences (*i.e.* > 90% sequence identity resolution) by grouping them into clusters that are represented by a single sequence (Suzek *et al*, 2007). This improves sequence search and retrieval speeds when a query sequence is searched against this database. This database is part of the PROTein INterface Identification predictor (PROTIN_ID) implementation. It is used to retrieve homologous sequences for protein chains derived from protein-protein complex structures of the Protein Docking Benchmark 4.0 dataset (see section 3.2).

3.2 Protein-protein complex database

The Protein-protein docking Benchmark 4.0 dataset is a manually curated dataset of non-redundant protein-protein complexes (Hwang *et al*, 2010). All protein complexes of this dataset form transient interactions. This dataset is primarily designed for testing the performance of protein-protein docking algorithms. However, it has found use outside the protein docking community as it is used for development of protein interface prediction algorithms (de Vries *et al*., 2006). There are a total of 176 experimentally determined protein complexes in this dataset. These are protein complexes that are in their bound form. In addition, the unbound forms of the proteins of each complex are included in the dataset. The 176 protein complexes were grouped into intra-species, inter-species, and antibody-antigen interactions. Overall, there are 73 intra-species, 76

inter-species, and 27 antibody-antigen complexes. A further six intra-species complexes mentioned in the Protein-protein docking Benchmark 4.0 paper (1J6T, 1O2F, 1P9D, 3EZA, 1UR6, and 1GGR) were included, raising the number to 79 intra-species complexes (Hwang *et al*, 2010). The unbound coordinates for the six complexes were located in the PDB database and included (1J6T: 1A3A and 1HDN; 1O2F: 1IBA and 2F3G; 1P9D: 1P9C and 1P98; 3EZA: 1HDN and 2KX9; 1UR6: 1E4U and 2ESK; and 1GGR: 1HDN and 2F3G). The intra-species complexes were further divided according to the number of complex protein constituents. This resulted in 63 binary and 16 multimeric intra-species complexes. One protein complex (1PXV) and the ligand protein of the 1ZHI complex were discarded from this dataset as the ligand and receptor proteins had very few homologs or no homologs retrieved from the UniRef90 database. As a result these proteins could not be used in the analyses of this study. The final dataset is therefore composed of 62 binary complexes that consist of 123 individual proteins (see appendix table A-5). The intra-species binary complexes and their corresponding unbound proteins were selected to create a dataset for this study (see section 5.2). All residues of the bound protein chains and their unbound counterparts were manually checked using the PyMol molecular visualization tool to ensure they agree in residue numbering (Delano, 2002). If there was disagreement, this was corrected. This is important for making sure interface and rest of surface (ROS) residues in the bound and unbound forms of a protein agree otherwise their subsequent analysis will be incorrect.

3.2.1 Determination of interface and "rest of surface" residues from the protein complex dataset

The dataset of 62 intra-species protein complexes was analysed to determine the number of interface residues for each individual protein complex and its unbound protein counterparts. This data is important as it is used to evaluate the performance of protein interface prediction algorithms, for example. Interface residues were determined based on the distance definition of the Critical Assessment of PRediction of Interactions (CAPRI) assessment established by the protein-protein docking community, which is used to measure the performance of protein docking algorithms. A residue is considered part of an interface if any of its atoms are $\leq$5 Å distance from the opposing protein's

residue atoms in a complex (Méndez *et al*, 2003). The contact script of the HADDOCK protein-protein docking suite was applied to calculate the distances of residue atoms of an interface that are ≤5 Å (Dominguez *et al*, 2003). A perl script was implemented to parse all atomic contacts of the contact script's output files to generate a list of interface residues of both interacting (bound and unbound) chains for each protein complex of the dataset. This provided a reference to calculate the total number of interface residues based on the distance definition in the entire dataset.

Solvent accessibility calculations were performed on all residues of the unbound proteins using Naccess (run via the PROTIN_ID method) where residues were defined as surface if their side-chain or main-chain atoms were ≥ 15% solvent accessibility (Hubbard and Thornton, 1993). This filters out core residues and generates surface residues. All unbound interface residues determined by the distance-based cut-off were examined to calculate how many are above the solvent accessibility cut-off. The interface residues for each complex's unbound proteins above ≥ 15% solvent accessibility were determined. This allowed a comparison between the interface sizes based on the distance and solvent accessibility criteria for each unbound protein of each complex for the entire dataset. A negligible loss of interface residues was determined using this solvent accessibility threshold. As a result of this, the 15% solvent accessibility was chosen as the default setting of the PROTIN_ID method for filtering out core residues prior to predicting interface residues. A perl script was implemented to determine the total number of surface residues, including interface residues (and rest of surface residues ROS) above the solvent accessibility threshold, and core residues for the unbound chains for each protein complex of the dataset.

3.3 Sequence retrieval and multiple sequence alignment generation

Homologous sequences of the unbound proteins' chains of the intra-species protein complex dataset were retrieved by the BLAST algorithm from the UniRef90 database. (Suzek *et al*, 2007; Altschul *et al*, 1990). Using the MUSCLE (version 3.8) multiple sequence alignment program (Edgar, 2004), all the unbound proteins' chains of the intra-species (binary) protein complex dataset were then aligned with their homologous counterparts. The BLAST search and MUSCLE alignment procedures were performed

within the PROTIN_ID method. PROTIN_ID also optimizes MSAs by dealing with sequence redundancy and sequence fragments prior to running MUSCLE (see Chapter 4).

3.4 Conservation score analysis of alignments

All multiple sequence alignments were scored with seven different conservation scores (described below) using the score conservation algorithm (default settings) as run from the PROTIN_ID program (Capra and Singh, 2007). A conservation window (default 3) in the score conservation algorithm is applied to incorporate the effect of the background conservation for each MSA column that is scored (see section 3.4.1). In addition, MSA columns with ≥ 30% gaps are disregarded and assigned a score -1000 by the score conservation algorithm since they are less likely to have functional significance (Capra and Singh, 2007). These scores were changed to 0.00 by PROTIN_ID to represent no conservation due to the presence of gaps. For MSA columns with < 30% gaps, a gap penalty is enforced. The score conservation algorithm implements position-based weights to weight all the sequences in an MSA based on the diversity of each MSA column. This prevents bias introduced from similar sequences that are overrepresented (Capra and Singh, 2007; Henikoff and Henikoff, 1994). All residues for each unbound chain of a protein complex of the dataset had their evolutionary conservation calculated. Using a perl script, the average conservation score and standard deviations were calculated from the parsed output files of the score conservation algorithm for each unbound chain's distance and solvent accessibility-based interface residues. This was also calculated for the rest of surface residues (ROS) for comparison purposes. This was done for all conservation scores in this study.

3.4.1 Window score heuristic:

Functionally important residues have neighboring residues (in sequence and structure) with higher conservation than average (Capra and Singh, 2007). In the score conservation algorithm, a window score heuristic feature is implemented that incorporates background conservation of residue neighbors of a residue of interest

within a sequence window when measuring conservation (Capra and Singh, 2007).

It is calculated as follows:

$$WindowScore = \lambda S_c + (1-\lambda)\frac{\sum_{i \in window} S_i}{|window|} \qquad (3\text{-}1)$$

where λ is a linear combination factor (default = 0.5) and S_c is the conservation score of the column of interest (*i.e.* foreground residue). S_i is the score of the neighbouring column (*i.e.* background residue). *window* refers to the total residue window on the left and right sides of the column of interest.

3.4.2 Shannon entropy score:

This is an early and widely used conservation measure (Valdar, 2002; Sander and Schneider, 1991; Shenkin *et al*., 1991; Shannon, 1948).

$$SE = -\sum_{\alpha}^{K} p_\alpha \log_2 p_\alpha \qquad (3\text{-}2)$$

where K and α represent the 20 amino acids and amino acid symbols, respectively. p_α represents the residue distribution in a multiple sequence alignment column. p_α is calculated as follows:-

$$p_\alpha = \frac{n_\alpha}{N} \qquad (3\text{-}3)$$

where n_α and N refer to number of amino acids of type α (*i.e.* symbol) and total number of amino acids in the alignment column, respectively.

3.4.3 Property entropy:

This score is a modified version of the above Shannon entropy (Valdar, 2002). The original Shannon entropy does not take into account amino acids' biochemical relationships and views all amino acids as just symbols. The property entropy remedies this by grouping amino acids according to physicochemical properties into 6 sets: aliphatic [AVLIMC], aromatic [FWYH], polar [STNQ], positive [KR], negative [DE],

special conformations [GP], and gaps (Capra and Singh, 2007; Valdar, 2002; Mirny and Shakhnovich, 1999). The distinct number of groups identified in an alignment column is then scored using this modified Shannon entropy score.

$$PE = \sum_{i}^{K} p_i \ln p_i \qquad (3\text{-}4)$$

where K represents 6 physicochemical sets i in which the 20 amino acids are partitioned. p_i represents the physicochemical set distribution in a multiple sequence alignment column over a given number of sequences (Manning *et al*, 2008; Valdar, 2002).

$$p_i = \frac{f_i}{N} \qquad (3\text{-}5)$$

where f_i is the stereochemical set's (i) frequency in an alignment column and N is the total number of sequences.

3.4.4 Property relative entropy:

This score is also a Shannon entropy variant and related to the Property entropy score in purpose such that it groups amino acids into physicochemical sets. It differs by using nine stereochemical amino acid sets and by having a normalizing term, which accounts for amino acid set frequencies of a multiple sequence alignment (Manning *et al*, 2008).

$$PRE = \sum_{i}^{K} p_i \ln\left(\frac{p_i}{\overline{p_i}}\right) \qquad (3\text{-}6)$$

where K represents 9 stereochemical sets i in which the 20 amino acids are partitioned (Manning *et al*, 2008; Valdar, 2002). The amino acid set are as follows:- [VLIM], [FWY], [ST], [NQ], [HKR], [DE], [AG], [P], [C] (Williamson, 1995). p_i represents the stereochemical set distribution in a multiple sequence alignment column over a given number of sequences. $\overline{p_i}$ is the average of p_i for the entire columns of an alignment.

$$p_i = \frac{f_i}{N} \qquad (3\text{-}7)$$

where f_i is the stereochemical set's (i) frequency in an alignment column and N is the total number of sequences.

3.4.5 *Relative entropy:*

$$RE = \sum_i p_i \log_2 \frac{p_i}{p_{ib}} \qquad (3\text{-}8)$$

where p_i represents the residue distribution in a multiple sequence alignment column. P_{ib} represents an amino acid background distribution derived from a dataset (default BLOSUM62 alignment data) of amino acid frequencies (Capra and Singh, 2007; Wang and Samudrala, 2006). Unlike previous scores, the utilization of a background distribution enhances the scores of residues that have low background frequencies when they are very common in an alignment column. Such residues are regarded as more likely to indicate functional significance. On the other hand, if a residue with a high background frequency is as prevalent with one that has a lower background frequency in an alignment column, it will be given a lower score than the other residue because of its higher background frequency.

3.4.6 *Jensen-Shannon divergence:*

$$JS = \lambda RE_{pc,r} + (1-\lambda) RE_{(q,r)} \qquad (3\text{-}9)$$

where p_c and q refer to the amino acid frequencies in an alignment column and background distribution of residues derived from a sequence dataset (default BLOSUM62), respectively (Capra and Singh, 2007). λ is a linear combination weighting factor (default 0.5) and *RE* is the Relative entropy. *r* is calculated as follows:-

$$r = \lambda p_c + (1-\lambda) q \qquad (3\text{-}10)$$

3.4.7 *Von Neumann entropy:*

$$VNE = -Tr(p \log p) \qquad (3\text{-}11)$$

where *p* is a density matrix normalized by its trace (Tr) (Capra and Singh, 2007; Caffrey *et al*, 2004). Initially, a matrix is constructed such that its off-diagonals are zero and only the diagonal values refer to the amino acid frequencies in an alignment column. These diagonal values are multiplied by the observed amino acid frequencies from a

similarity matrix, such as BLOSUM62, in order to create a density matrix (Caffrey *et al*, 2004). In doing so, physicochemical similarity is accounted for in this evolutionary conservation score.

3.4.8 Sum-of-pairs:

This score measures evolutionary conservation by pair-wise comparison of residues in an alignment column.

$$SP = \frac{1}{\sum_i^N \sum_{j>i}^N w_i \times w_j} \times \sum_i^N \sum_{j>i}^N w_i \times w_j \times S(C_i, C_j) \qquad (3\text{-}12)$$

where S is a similarity matrix used to compare the similarity of residues i and j of a column C. w_i and w_j are weighting factors for the ith and jth sequences (Capra and Singh, 2007; Valdar, 2002).

3.5 Statistical analysis of Interface prediction algorithms

In order to assess a method's performance in interface residue prediction, it is necessary to determine the number of interface and ROS residues. This was determined from the 62 protein-protein complex dataset used in this study. A surface residue (section 3.2.1) can either be an interface residue or not. This binary classification of residues by an interface predictor is compared to the known binary classification of surface residues of the dataset of this study. The predictions of interface prediction methods can have either a positive/or negative result, according to this binary classification applied to surface residues. A positive result is divided into true positives (TP) or true negatives (TN). A TP is a correctly predicted interface residue, whereas a TN is a ROS residue that has not been included in a prediction method's final prediction. In other words, it has been correctly discarded from the final result of an interface prediction method. A negative result is divided into false positives (FP) and false negatives (FN). An FP is a ROS residue that has been incorrectly predicted as an interface residue in the final prediction of an interface prediction method. A FN is an interface residue incorrectly discarded as a ROS residue. The relationship of TP, FP, TN, and FN residues is illustrated in the confusion matrix (table 3.1).

Table 3-1: A confusion matrix and the components that make up a positive or negative result. TP + FN refer to the total interface residues, whereas FP + TN refer to the total ROS residues.

Confusion Matrix		**Actual**	
		Positive	**Negative**
Predicted	**Predicted interface**	**TP**	**FP**
	Predicted ROS	**FN**	**TN**

The following standard performance metrics are derived from the confusion matrix (Fawcett, 2006).

3.5.1 Accuracy:

$$\frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (3\text{-}13)$$

The accuracy measure calculates the predicted TP and TN residue proportion out of the total positive and negative results of an interface prediction algorithm.

3.5.2 TP fraction (specificity):

$$\frac{(TP)}{(TP+FP)} \qquad (3\text{-}14)$$

The TP fraction calculates the percentage of TPs in the final prediction. A high fraction reflects more TPs and less FPs in the final prediction of an interface prediction algorithm.

3.5.3 FP fraction:

$$\frac{(FP)}{(TP+FP)} \qquad (3\text{-}15)$$

The FP fraction is the opposite of the TP fraction. It quantifies the fraction of FPs in a final prediction. A predictor that scores a low FP fraction generates low numbers of FPs in it final prediction results.

3.5.4 TP rate (sensitivity):

$$\frac{(TP)}{(TP+FN)} \qquad (3\text{-}16)$$

The TP rate quantifies the fraction of interface residues predicted in a final prediction from the total number of observed interface residues. A higher fraction represents a greater recall of interface residues.

3.5.5 FP rate:

$$\frac{(FP)}{(FP+TN)} \qquad (3\text{-}17)$$

The FP rate computes the fraction of ROS residues present in a final prediction from the total number of observed ROS residues. A lower fraction represents a lower recall of ROS residues.

3.5.6 F-measure:

$$\frac{2}{1/Specificity+1/Sensitivity} \qquad (3\text{-}18)$$

The F-measure quantifies the harmonic mean from the combination of the TP fraction (specificity) and the TP rate (sensitivity), which are weighted equally (Rennie, 2004; Van Rijsbergen, 1979). Hence, it measures the retrieval quality of an interface prediction method. A method that scores highly for both TP fraction and TP rate is awarded a high F-measure and vice versa. An F-measure of zero implies the lack of TPs in a final prediction and a score of one refers to total recall of observed interfaces with

the absence of FPs in the final prediction.

3.5.7 Matthew's correlation coefficient (MCC):

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \qquad (3\text{-}19)$$

The MCC score computes the correlation between the actual results of the confusion matrix (table 3-1) with the predicted results and scores 1 for completely correct predictions and -1 for the completely incorrect predictions (Murakami and Mizuguchi, 2010; Matthews, 1975). An MCC value of zero indicates prediction performed at random (Baldi *et al*, 2000).

A perl script was implemented to automate the parameterization of the PROTIN_ID prediction algorithm to maximize its performance in terms of interface residue prediction. This perl script ran PROTIN_ID at different settings and calculated its performance at each setting, according to the standard performance metrics defined above. Performance evaluation was also applied to the benchmarking of the performance of PROTIN_ID in comparison to WHISCY and CCRXP methods (section 3.7).

3.6 Statistical analysis of interface vs. ROS residues conservation

A statistical analysis was performed to determine whether interface residues are more conserved than ROS residues. A perl script was implemented to calculate the difference in conservation signal between the interface and ROS conservation for all proteins of the test dataset (ΔCons). This was done iteratively for average interface and ROS conservation as calculated for all seven conservation scores. ΔCons is calculated as follows: -

$$\Delta Cons = Interface_{AveCS} - ROS_{AveCS} \qquad (3\text{-}20)$$

where $Interface_{AveCS}$ and ROS_{AveCS} refer to the average interface conservation and average ROS conservation, respectively. The ΔCons value can be a positive, negative,

or zero difference. A positive difference indicates that interface residues are more conserved than the ROS residues of a protein. A negative difference is the opposite where ROS residues are more conserved than interface residues. If the difference is zero, this indicates that interface and ROS residues are equally conserved.

STATA version 11 (StataCorp LP, 2009) is a comprehensive statistical suite that is useful for the organization and analysis of data and its representation in convenient graphical output. STATA was used to determine if the ΔCons values of all dataset proteins follow the Normal distribution in order to analyse the ΔCons data with the appropriate statistical tests of STATA. This is important as the choice of a statistical test used for data analysis depends on the underlying probability distribution exhibited by the data. If the wrong statistical test is used for data analysis the subsequent statistical interpretation may be invalid (Park, 2008). To examine the ΔCons distribution, graphical methods, which compare a theoretical normal distribution with the ΔCons distribution, were applied. The graphical methods generated with STATA were histograms, Q-Q plots, and P-P plots. The example commands used for histograms and Q-Q plots (in that order) are given below.

- histogram difference, width(0.05) start(-0.3) frequency normal normopts(lwidth(thick)) ytitle(Frequency, size(large)) ylabel(0(5)30, labsize(large)) xtitle(Difference, size(large)) xlabel(-0.3(0.1)0.3, labsize(large)) xline(0, lwidth(thick)) title(Jensen-Shannon divergence, size(large))

- qnorm difference

In addition, numerical methods that examine skewness and kurtosis of the ΔCons variables (skewness-kurtosis test) and test for their normality (Shapiro-Wilks and Shapiro-Francia test) were performed in STATA. The standard commands for these tests are provided below.

- sktest difference
- swilk difference
- sfrancia difference

The Paired t-test was used to examine whether the ΔCons conservation difference in favour of interface residues over ROS residues was statistically significant. Furthermore, the 95% confidence intervals (CI) for the upper and lower bound limits of the ΔCons values were calculated as part of the Paired t-test results. Both tests presume that the data follow the normal distributions. An example command for this test performed using STATA is provided below.

- ttest asainterfacecsave == surfaceonlycsave

Further statistical tests were applied that assume no underlying probability distribution (ex. normal) as a precondition prior to data analysis. This was done for comparison with the Paired t-test and 95% CI analysis, which require as a perquisite normally distributed data, to determine if overlap in final statistical results exists. The bootstrap approach and the Wilcoxon matched pairs test were applied as equivalents to the CI analysis and Paired t-test, respectively. The Wilcoxon matched pairs test was performed using the biostatistical program GraphPad Prism version 5.00 at default settings for the ΔCons data. The bootstrap approach was performed using STATA for ΔCons data. For the bootstrap analysis, 1000 randomly selected samples were used to calculate the sample means per bootstrap repetition in order to calculate the CI for the ΔCons data. The example command is specified below.

- bootstrap m=r(mean), rep(1000) : summarize difference

3.7 Benchmarking of PROTIN_ID with WHISCY and CCRXP algorithms

The protein complexes selected for benchmarking analysis were based on the protein-protein complex datasets used by the authors of WHISCY and CCRXP (see section 6.3). The WHISCY interface prediction results were generated using default parameters for HSSP and UniRef90 (section 3.3) alignments of the proteins of the protein-protein complex dataset. The online webserver of WHISCY was used (http://nmr.chem.uu.nl/Software/whiscy/index.html). The output file that lists all residues by WHISCY score from the highest to lowest scores was parsed by a perl script and the residues that scored ≥0.18 where predicted as interface residues, following the

WHISCY authors in their inital study (de Vries *et al*, 2006). The predicted interface residues were evaluated by the statistical performance measures (see section 3.5). The CCRXP interface predictions were generated using default parameters using the webserver version of the method (http://ccrxp.netasa.org/). A perl script was implemented to parse the cluster output file generated by CCRXP and the largest cluster of residues (or most conserved cluster if the largest clusters are the same size) was predicted as interface residues (see section 6.2.2). The perl script further evaluated the predicted cluster using the previous performance metrics (see section 3.5). The PROTIN_ID interface predictions were performed using default parameters. Chapter 4 provides a full description of the PROTIN_ID method. A perl script was implemented to parse the output cluster file of PROTIN_ID (cluster_*filename*.dat) to retrieve the final cluster of predicted interface residues where it was assessed by the performance evaluation metrics.

3.8 Protein-protein docking driven by interface predictions

Protein-protein docking seeks to predict the protein complex of two proteins in their unbound pose that are known to interact with each other. To examine the effect of predicted interface data on the docking of protein-protein complexes and compare this to *ab initio* docking (*i.e.* without the use of any data), a docking dataset was created from the protein-protein complex dataset of this study. Protein-protein complexes were selected that had ≥ 10% TP rate (section 3.5.4) for both chains when prediction restraints were generated by PROTIN_ID. This resulted in a total of twenty-six complexes in a docking dataset, which consisted of their unbound forms for use in docking (see Chapter 7).

3.8.1 The HADDOCK protein-protein docking algorithm: Theoretical data-driven docking vs. ab initio docking

The High Ambiguity Driven protein-protein DOCKing (HADDOCK) algorithm was applied for the docking of the protein-protein docking dataset (de Vries and Bonvin, 2010; Dominguez *et al*, 2003). The HADDOCK algorithm was selected due to its

ability to use experimental or theoretical (*i.e.* predicted interface residues) data to drive protein-protein docking. The PROTIN_ID predicted interface residues were designated as active residues of the ambiguous interaction restraints (AIRs) and inputted into HADDOCK (version 2.0). The larger protein chain of the unbound input proteins was the receptor protein and the smaller chain was the ligand protein and this was reflected in the begin.par HADDOCK docking setup file. In the data-driven runs, default settings were used. For example in the run.cns parameter file of HADDOCK, the random removal AIRs (noecv = true) is set to 50% (ncvpart = 2) and this results in 50% AIRs removal per docking trial. This may result in the removal of possible false positives from the AIRs restraints that guide docking. The control docking runs (*ab initio* docking) were done using the centre of mass restraints (cmrest = true) while disabling the random removal of restraints parameter in run.cns (noecv = false). These centre of mass restraints enforce contact between the two unbound proteins during *ab initio* docking. Using default settings, 1000 complexes were generated in the rigid-body docking stage of HADDOCK for each run. Subsequently, the best 200 (default) complexes ranked according to the HADDOCK score were further subjected to simulated annealing and water refinements (Dominguez *et al*., 2003). The final water refined 200 protein-protein (docked) complexes generated by HADDOCK upon completion of a data-driven or *ab initio* docking run were analysed.

3.8.2 Analysis of predicted protein-protein docking complexes

The Critical Assessment of PRediction of Interactions (CAPRI) quality assessment of docked complexes was applied to analyse the protein-protein complexes generated by the docking runs (Lensink *et al*, 2007; Méndez *et al*, 2003). All predicted complex models for a particular protein complex were compared to the experimentally solved complex and were evaluated according to CAPRI criteria. The CAPRI criteria implemented are the fraction of native contacts (F_{nat}), ligand-rmsd (L-rmsd), and interface-rmsd (I-rmsd). F_{nat} is defined as the number of correct residue contacts in a predicted complex divided by the number of residue contacts in the actual experimental complex. A contact is determined from the experimentally solved complex, and it is defined as a residue pair of the protein interface whose atoms are ≤ 5 Å apart. The L-rmsd refers to the ligand (smaller protein) backbone (N, C_α, C, O atoms) rmsd

difference of the predicted and experimental complex ligand proteins, following the superimposition of both predicted and experimental complex receptor (larger) proteins. I-rmsd is all atom-atom contacts between residues of opposing proteins at ≤ 10 Å of the experimental complex's superimposed backbone (N, C_α, C, O atoms) over the same residues in the predicted complex. Both rmsd measures look at geometric fit between experimental and predicted complexes (Méndez *et al*, 2003).

The CAPRI criteria for ranking predictions are given below.

- **High quality predictions:** $F_{nat} \geq 0.5$ and l-rmsd ≤ 1.0 Å or i-rmsd ≤ 1.0 Å.
- **Medium quality predictions:** $F_{nat} \geq 0.3$ and l-rmsd $1.0 < x \leq 5.0$ Å or i-rmsd $1.0 < x \leq 2.0$ Å
- **Acceptable quality predictions:** $F_{nat} \geq 0.1$ and l-rmsd $5.0 < x \leq 10.0$ Å or i-rmsd $2.0 < x \leq 4.0$ Å
- **Incorrect predictions:** $F_{nat} < 0.1$ or (l-rmsd > 10.0 Å and i-rmsd ≤ 4.0 Å)

The CAPRI analysis (c shell) scripts for calculating the F_{nat}, L-rmsd, and I-rmsd were obtained from the HADDOCK authors. Both the F_{nat} and I-rmsd scripts were updated to perform CAPRI evaluation of the protein-protein models generated by docking. Only the L-rmsd script was re-written in perl for calculation of the L-rmsd values for predicted complexes. Both the L-rmsd and I-rmsd scripts utilize ProFit (version 3.1), which is a least squares fitting program, in the background for rmsd calculations (Martin, 1998). The input files for ProFit specifying the interface and ligand zones were generated for both I-rmsd and L–rmsd scripts, respectively. For each protein complex of the dataset under consideration, the number of correct models (acceptable and above) out of the final 200 refined models was determined for both data-driven and control docking runs. This data was analysed using the Fisher exact test using GraphPad Prism version 5.00 at default settings to determine if the difference in numbers of correct models between both runs was statistically significant.

3.9 Protein-protein docking driven by interface predictions and experimental data

The effects of using interface predictions and experimental data as restraints in protein-protein docking in order to improve docking performance was examined. The experimental data to be applied in docking were residual dipolar couplings (RDCs) and chemical shift perturbation (CSP) data. A search for binary protein-protein complexes with available unbound protein constituents, RDCs (derived from backbone amides), and CSP data was performed for all protein-protein complexes of the PDB database and the Biological Magnetic Resonance Data Bank (BMRB), which is a repository of NMR data derived from NMR analysis of proteins that are cross-linked to PDB entries (Ulrich *et al*, 2008; Bernstein *et al*, 1977). To this end, a dataset was created of protein-protein complexes with the required experimental data by applying a keyword-based search on the PDB database. Similar keyword-based search strategies have been applied in other studies (Choi *et al*, 2009; Nooren and Thornton, 2003b). Text searches were performed to get three lists of PDB entries associated with general keyword groups: - protein complexes, RDCs, and CSPs. The following keywords were used: “Protein Complex”, “Complex”, “Complexes”, “Chemical Shift”, “Chemical Shifts”, “Chemical Shift Perturbation”, “Chemical Shift Perturbations”, “CSP”, “CSPs”, “RDC”, “RDCs”, “Residual dipolar coupling”, and “Residual dipolar couplings”. This resulted in a list of PDB entries for each keyword of each group. A perl script was written to combine PDB entries for each keyword of a group and thus remove overlapping PDB entries. This resulted in non-overlapping PDB entries that were combined for each group. Following this, the PDB entries for the protein complex group where combined with those with the CSP group to retrieve overlapping entries. The same was done with the protein complex and RDCs groups. Therefore, two lists were created one for the protein complexes with associated CSP data and another for those protein complexes with associated RDC data in the PDB. These two lists were combined to create a final list of overlapping entries, which are protein complexes with CSP and RDC data. These were manually checked to retrieve only binary protein complexes for this study, which had CSP and RDC data in the BMRB (see section 7.2). For each protein complex of the dataset, the RDC and CSP data were retrieved from the BMRB. If only RDC data was present for a particular complex, the CSP data for that protein complex was retrieved from the literature. For all protein complexes their unbound constituents were found and these have been used as

input in HADDOCK docking runs.

3.9.1 RDC and CSP data preparation as restraints for protein-protein docking

To use RDCs in HADDOCK protein-protein docking, the PALES program was used to derive the axial (Da) and rhombic (R) components for each protein-protein complex's unbound protein constituent of the dataset (Zweckstetter, 2008; Zweckstetter and Bax, 2000). The standard PALES command is given below.

- Pales –bestFit –inD *file1*.tab –pdb *file2*.pdb –outD *file3*.tbl

The alignment tensor components and the RDC input file were used for each protein-protein complex run in HADDOCK (section 3.8.1). Furthermore, intervector projection angle (IPA) restraints generated using the RDC data were also used in HADDOCK runs. IPA restraints are not dependent on the alignment tensor (Meiler *et al*., 2000). Finally, CSP restraints for residues that displayed significant shifts (Bonvin, 2010) and were ≥15% in solvent accessibility were converted into AIRs active residues. Residues neighbouring to the designated (CSP) active residues and were ≥15% in solvent accessibility were included as passive residue AIRs restraints. The AIRs data were combined with the other forms of data (RDC and interface prediction data) in the docking runs. The following docking runs were performed:

- *Ab initio* run (control)
- Interface prediction run
- CSP run
- Interface/CSP run
- RDC/CSP run
- RDC, CSP, and interface prediction run

For each run, the final 200 water refined predicted protein-protein complexes were analysed using the previous CAPRI analysis scripts. This was followed by a statistical comparison of the difference in correct models generated for each run using the Fisher exact test of GraphPad Prism version 5.00 (at default settings). This is to determine if

the difference in numbers of correct models between the runs is statistically significant (see section 3.8.2). Using the same software, the Spearman r test was conducted to compare the correlation of TP fraction and TP rate overall to the number of correct models generated.

Chapter 4

The Protein Interface Identification method

4.1 Introduction

Proteins interact through specific residues on their surface, forming the interface. The identification of interface residues is important, as they can be applied as sampling restraints in protein-protein docking whose objective is to predict the complex of proteins known to interact (see section 1.6). To achieve this, the PROTein INterface IDentification (PROTIN_ID) method was implemented for the prediction of protein-protein interface residues (*i.e.* theoretical restraints) to be used to drive docking (see Chapter 7). This method is written in standard Perl 5 and uses the Bioperl toolkit, which provides functionality in biological data analysis (Ryu, 2009; Stajich *et al*, 2002; Wall *et al*, 2000). In order to generate theoretical restraints, the PROTIN_ID program requires as a minimum input a protein structure file in PDB format (Berman *et al*, 2000). In addition, an optional multiple sequence alignment (MSA) in CLUSTAL, HSSP, or FASTA file format may be inputted, if the default generation of an MSA is not preferred.

4.2 The rationale for the implementation of the PROTIN_ID method

The course of action followed in implementing this method was driven by previous work that has identified predictive features of interface residues (see section 1.7.2). Structural and sequence data were utilized as predictive properties for interface residues in this method. For example, in structural terms interface residues are solvent accessible and some residues cluster together in close proximity. Furthermore, interface residues are conserved due to evolutionary pressure and this data can be extracted from a multiple sequence alignment (MSA), taking into account the background environment of conserved residues, which is the spatial proximity of other residues varying in conservation in contiguous alignment columns (see section 3.4.1). Therefore, conserved interface residues cluster on a protein's surface, and being able to explore the extent of

conservation and clustering of conserved residues of an extensive dataset of intra-species proteins (see section 3.2), as a predictive feature of interface residues versus non-clustering, was a basis for developing this method (see section 5.7). And this goal was implemented in the context of whether interface residues are more conserved than rest of surface residues (see Chapter 5).

Sequence data in the form of MSAs often require manual editing before being data mined. This becomes tedious in the context of testing hundreds of proteins and their associated MSAs in a high-throughput setting for docking, and simply analysing an alignment generated automatically in a method without manually checking its sequences may introduce bias from the noise generated by alignment errors. This may introduce non-interface residues in a method's final prediction. When applied to docking, errors in theoretical restraints may misdirect the sampling. The implementation of an editing heuristic to 'check' the sequence data prior to MSA generation formed another basis for developing this method.

To recapitulate, the focus on conserved clusters of residues and elimination of factors from sequence data that may cause MSA issues affecting their prediction led to the development of PROTIN_ID that generates docking restraints for (unbound) proteins of the latest protein complex dataset (*i.e.* Benchmark4.0, see section 3.2). However, Ahmad, *et al*. (2010) later published a method, Clusters of Conserved Residues-XP (CCRXP), with a similar protocol, but with specific differences. The CCRXP method was designed to automate the computational tasks of generating conserved residue clusters, and provide the first publicly available method to achieve this (Ahmad *et al*., 2010). One difference between it and PROTIN_ID is that CCRXP does not process sequence data prior to MSA generation, and the final clusters of conserved residues in CCRXP method are not ranked further by a heuristic to create a final prediction as done in PROTIN_ID, but are annotated by structural features (ex. secondary structural composition). In addition, the CCRXP method was trained on a small dataset of protein-RNA and protein-protein complexes (25 proteins) to predict hotspot interface residues based on the premise that they form clusters of conserved residues. In contrast, PROTIN_ID was trained to cluster conserved interface residues based a bigger dataset of 123 intra-species proteins to be used in docking (see section 3.2). Since both methods seek to predict conserved residue clusters, PROTIN_ID was compared in interface

residue prediction performance to CCRXP and performed better than CCRXP (see Chapter 6).

4.3 Implementation of PROTein INterface Identification (PROTIN_ID)

The PROTIN_ID method requires a PDB coordinate file of an unbound protein of interest (*i.e.* query) to initiate its prediction of clusters of conserved interface residues. From the PDB file, the method extracts the query FASTA sequence from the file's ATOM coordinates. By default, chain 'A' of a PDB file is extracted unless the user specifies another PDB chain identifier. This preparatory step ensures the method has structural (PDB coordinate) and sequence data (query sequence) upon which to proceed with the next steps (below), utilizing predictive properties of interface residues, for the final prediction. An overview is of PROTIN_ID is shown in Figure 4-1, indicating the protocol's steps from start to finish.

4.3.1 Sequence data retrieval and processing in PROTIN_ID

PROTIN_ID performs a BLAST search to retrieve homologous sequences of the query protein sequence in the UniRef90 database at an e-value of $1x10^{-8}$ or as defined by the user (Altschul *et al*, 1990). Using this database greatly expands BLAST search speeds to retrieve sequence homologs (hit sequences) and allows better detection of sequences with distant relationships to the query sequence because redundant sequences (*i.e.* >90% sequence identity) are hidden (see section 3.1) (Suzek *et al*, 2007). UniRef90 also provides convenient access to hit sequence details like taxonomy ID and species name. The taxonomy ID is relevant for a subsequent data processing step in PROTIN_ID to filter out identical sequences of the same species that may be present. In PROTIN_ID, sequences in the generated MSA have their species names appended to their UniRef90 accession codes, making it easier for a user to relate sequences to each other, for example.

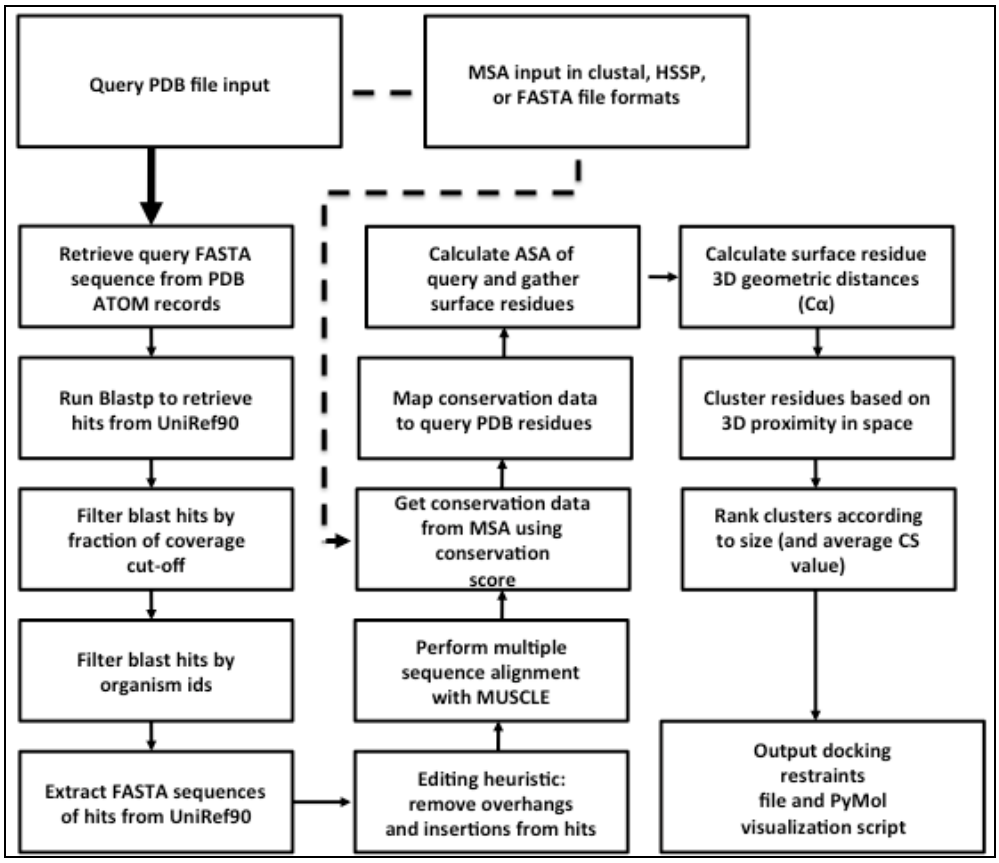

Figure 4-1: Overview of PROTIN_ID. The default protocol requires a PDB file as input. Along with this, an optional multiple sequence alignment (MSA) in CLUSTAL, HSSP, or FASTA format can be inputted by the user, which starts PROTIN_ID from the point indicated by the dashed arrow. The final output is a theoretical restraints file for data-driven docking. Also, two files for molecular visualization of clusters predicted and the conservation map of the protein of interest in PyMol are produced. PROTIN_ID has important features that have been implemented. The first feature involves minimizing noise in sequence data. This is achieved by filters to remove sequence fragments and redundant sequences, which are followed by the application of an editing heuristic to remove overhangs or insertions in the remaining sequences that occur when these sequences are aligned to the query sequence. The editing heuristic step minimizes or eliminates factors that cause multiple sequence alignment (MSA) errors and reduces the requirement of manual editing of an MSA. This is a desirable user-friendly feature useful in the context of high-throughput work to generate theoretical restraints from many proteins for application in protein-protein docking. The second feature is the use of structural data to apply spatial clustering of conserved surface residues. Residues that are spatially contiguous are more likely to be functionally significant, if conserved. Clustering is useful to remove isolated residues from being included in predictions.

Following the retrieval of sequence homologs, the hit sequences are filtered by the fraction of coverage (FC%). FC% is determined by comparing the pairwise alignment of a hit sequence to the query sequence taken from the BLAST report (see Figure 4-2). It is measured as follows:

$$FC\% = \frac{n_{hit}}{N_{query}} \times 100 \qquad (4\text{-}1)$$

where n_{hit} is the length of a hit sequence without gaps in the pairwise alignment (High-Scoring Segment Pair; HSP) between it and the query sequence and N_{query} is total length of the query sequence. HSPs are defined in the blast report. If more than one HSP is present for a hit and the query sequence, it is possible that they may overlap with each other in relation to the query sequence. In this instance, PROTIN_ID calls the tiling algorithm of Bioperl to retrieve the overall length of the hit sequence that is aligned with the query sequence. If 50% FC is set all hits that are ≥ 50% aligned to the query are retained; the rest of the hit sequences are discarded. This filter allows the user to filter out very short sequences or fragments caused from partial or incomplete experimental data from appearing in a multiple sequence alignment. If such short sequences are kept, gaps are inserted in areas of the missing sequence segments of the fragmented sequences. This can cause conservation scores to penalize for the presence of these 'artificial' gaps, which are assumed to be there because of biological significance, assigning a reduced score to the alignment columns enriched in them (see section 3.4).

Upon filtering by FC% another step may be applied by the user for further filtering of sequences. In this step, all hit sequences are grouped by their (NCBI) taxonomic ids in order to determine how many sequences are from the same species (Figure 4-2). The taxonomic IDs are retrieved from each sequence's UniRef90 webpage by PROTIN_ID. If a taxonomic ID number has more than one hit sequence associated with it, then only the sequence with the highest fraction of coverage is retained, while the rest are discarded. Conversely, if a taxonomic ID has only one sequence then it is retained by PROTIN_ID. This taxonomic ID filtering step eliminates sequence redundancy (*i.e.* identical sequences) in the same species to remove overrepresentation of identical

sequences of the same species. After applying the above filter(s) the final set of homologous hit sequences are defined. These are extracted from the UniRef90 database by the Fastacmd program when called by PROTIN_ID.

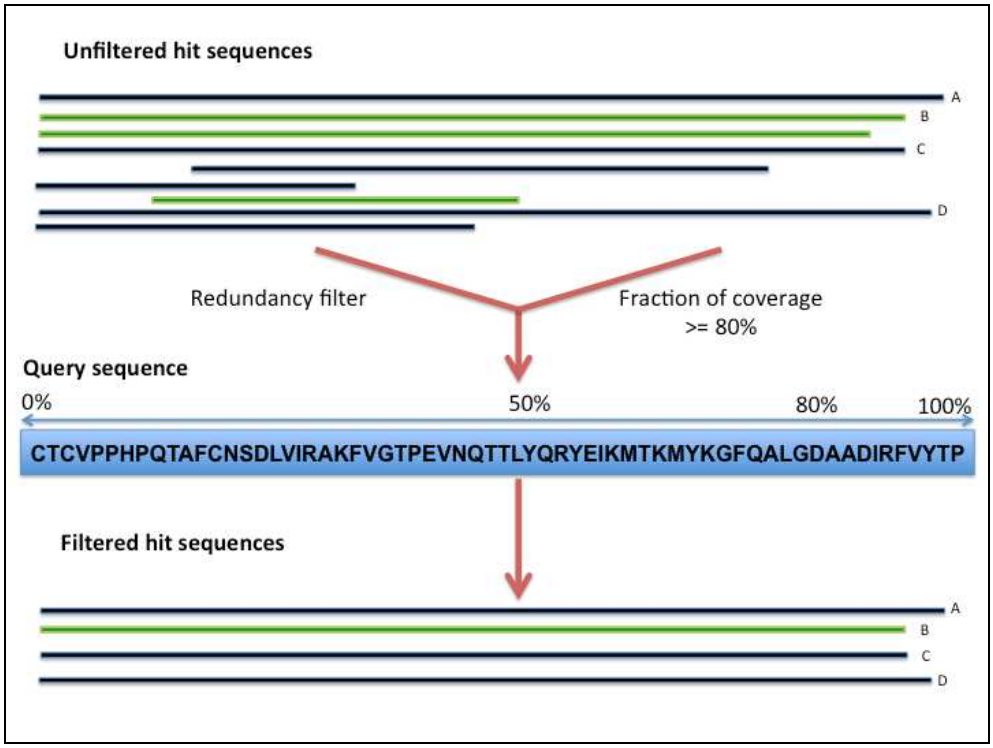

Figure 4-2: Overview of the fraction of coverage/redundancy filters that are implemented in PROTIN_ID. The plot shows hit sequence high-scoring segment pairs (HSPs) derived after pairwise alignments with the query (bold sequence). The hit sequences (dark blue) align with various regions of the query sequence. If a hit sequence covers ≥ 80% fraction of coverage (FC), for example, of the query sequence (centre), it is retained and if is under the cut-off it is discarded. In some instances, a hit sequence may have more than one HSP, which may overlap. In this case, their non-overlapping contribution to the overlap of the query is determined to calculate their FC%. Some hit sequences may be redundant (*i.e.* identical) sequences from the same species (green). In this case, redundancy filtering may be applied such that all identical sequences are removed, leaving only the intra-species sequence with the highest fraction of coverage (*i.e.* longest) to the query. In the end only the hit sequences A-D are retained. The above step is useful to remove many duplicate sequences and sequence fragments prior to multiple sequence alignment (MSA) generation, reducing the burden of repetitive manual editing of MSAs especially in the context of high-throughput generation of theoretical restraints for docking.

The PROTIN_ID editing heuristic is implemented, following the retrieval of all hit sequences via Fasatcmd (see Figure 4-3). First, pairwise global alignments between the query sequence and each hit sequence are created by the Needleman-Wunsch algorithm as implemented in the EMBOSS suite (Rice *et al*, 2000; Needleman and Wunsch, 1970). In this analysis, the query sequence is regarded as the canonical sequence and any deviations from the query sequence by the hits are corrected. The objective is to create a modified set of hit sequences that only align to the canonical sequence since protein functional prediction is mapped back to the known query 3-D structure ultimately and such information can only be gathered by hit sequence residues that align to the query sequence's residues. Only these residues have their evolutionary conservation measured. The hit residues not found in the query PDB structure are irrelevant for this purpose of mapping and are discarded even though their evolutionary conservation has been measured.

Often in manual editing of an alignment for input in an interface predictor method, the following steps are usually applied. If a hit sequence has overhangs that exceed the canonical sequence's N or C-termini, they are removed leaving only a hit sequence's areas, which align completely with the query sequence. Furthermore, if there are splits (*i.e.* gaps) in the canonical sequence observed in its pairwise alignment with a hit sequence, this is due to the hit sequence having additional residues (*i.e.* insertions) not found in the query sequence. Such insertions are removed from the hit sequence. These rules are automatically implemented by the editing heuristic. The outcome is an improved, structured MSA when the hits and the query are aligned in the next step by MUSCLE through the removal of the above factors. Figure 4-4 shows the difference between two MSAs generated by MUSCLE for the Tissue inhibitor of metalloproteinases 1 (TIMP-1) protein sequence. Both MSAs use the same number of sequences. One MSA has been generated using sequences edited by the novel editing heuristic (refined), while the other MSA uses the same, unedited sequences (unrefined).

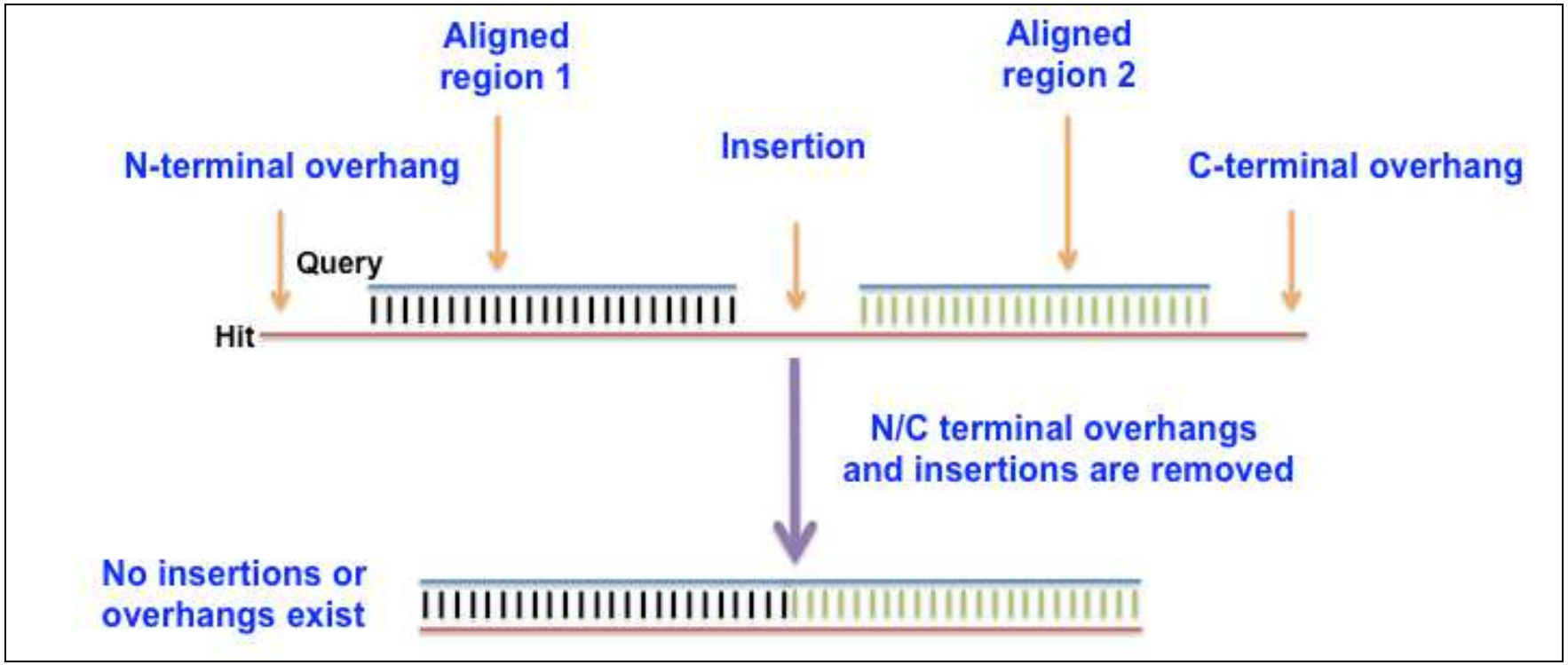

Figure 4-3: Overview of the sequence editing heuristic that is implemented in PROTIN_ID. A hit sequence is globally aligned to the query sequence by the Needleman-Wunsch algorithm. The query sequence is regarded as the canonical sequence and any N/C terminal overhangs beyond the query sequence or insertions that split the query by the hit sequence will be removed. The end result is hit regions that only align to the canonical sequence. This is important as these regions are used for computing conservation of the query. Hit residues not found in the query are non-essential because they do not exist in the query protein's 3D protein structure.

It can be seen for the unrefined MSA that there are misaligned residues (*i.e.* CVC), which is due to a long N-terminal overhang present in the hit sequence. This N-terminal overhang exceeds the length of the TIMP-1 query sequence, and is the longest sequence (*i.e.* to the left) in the holistic view of the unrefined alignment in Figure 4-4(A). In the refined MSA, this N-terminal overhang is removed from the hit sequence by the editing heuristic before MSA generation by MUSCLE, as it has no counterpart sequence region in the query sequence. This has resulted in the misaligned residues' positions likely aligned in the appropriate sequence columns that were originally occupied by residues (*i.e.* CLA) of the removed N-terminal region. In the context of high-throughput generation of theoretical restraints for docking such unstructured MSAs would require manual editing before input in an interface predictor. The aim of the sequence editing heuristic is to address this and lessen the need of manual editing of MSAs. The refined MSA is improved because the factors that caused unstructured alignments in the first place, namely N/C terminal overhangs and insertions that split the query (canonical) sequence have been eliminated. If more manual MSA editing is required at a later stage by the user, it will be more efficient to perform also. Both the refined and unrefined MSA were examined further to assess their impact on interface residue prediction performance of PROTIN_ID. Using the unrefined MSA as input, it was determined that no interface residues were predicted in the final prediction, whereas using the refined MSA resulted in interface residues being predicted in the final prediction at a 58% TP fraction (see 3.5.2). This is because more (7) interface residues were extracted for clustering based on the refined alignment in later steps of the protocol, which generated the largest (12 residue), top-ranking cluster (see section 4.3.2). Only four interface residues were extracted using the unrefined MSA, forming a smaller, five-residue cluster that was ranked second.

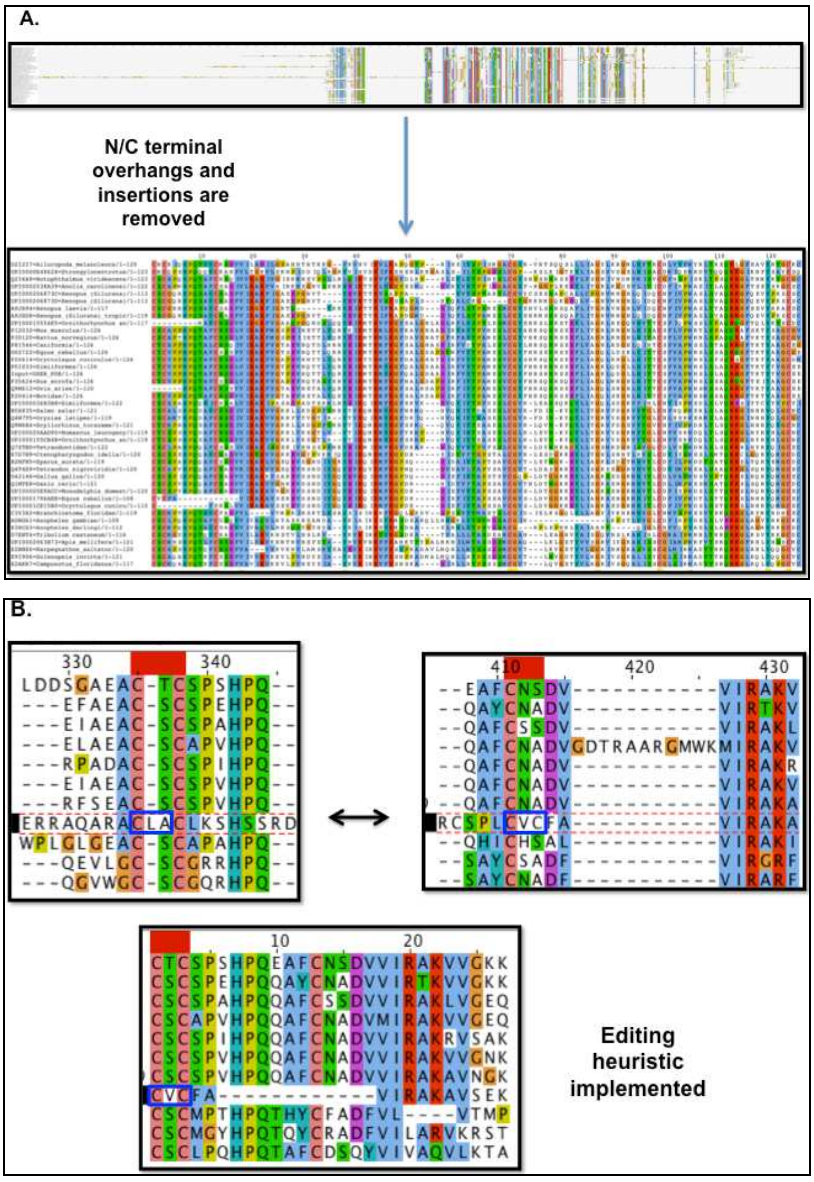

Figure 4-4: The comparison of the MSAs generated by MUSCLE of the same number of hit sequences for the Tissue inhibitor of metalloproteinase-1 sequence. **A)** The top MSA is when the editing heuristic is not applied (unrefined) and the bottom is when it is applied (refined). It can be seen that N/C terminal overhangs and insertions of longer hit sequences than the query cause an unstructured MSA, and this is not the case when the editing heuristic is applied **B)** There are three misaligned residues (CVC and CLA - blue outlines) in the top sections of the unrefined MSA indicated by the arrow. In the refined MSA, the three residues (CVC - blue outline) are likely aligned in the appropriate sequence columns compared to their original positions in the unrefined MSA. Using the refined MSA, PROTIN_ID predicts interface residues for its top prediction. With the unrefined MSA, no interface residues were predicted, indicating the editing heuristic's usefulness in improving PROTIN_ID prediction performance.

The cluster ranked first based on the unrefined MSA was composed entirely of eight non-interface residues. The four interface residues in the second-ranking cluster that were extracted based on the unrefined alignment overlapped with four out of seven interface residues extracted using the refined MSA. This leaves a difference of three residues that could not be extracted using the unrefined MSA. A comparison of these three residues' assigned conservation scores based on the two MSAs revealed that they score less in the unrefined MSA, decreasing their rank beyond the absolute residue extraction cut-off used in PROTIN_ID (see section 4.3.2). This indicates that using a structured MSA derived from 'edited' hit sequences via the editing heuristic can improve PROTIN_ID's prediction performance through extracting more interface residues clustered together, which can boost their cluster ranking.

As mentioned above, automatically edited hits are submitted along with the query protein's sequence to MUSCLE (version 3.8) to generate an MSA (Edgar, 2004). PROTIN_ID runs MUSCLE using default parameters to produce an output multiple sequence alignment in clustal format (see section 3.3). This is followed by conservation score analysis (see section 3.4). In this step, conservation scores are measured for the MSA columns by the score conservation algorithm using the Jensen-Shannon divergence score (default) or one of the other conservation scores at a three residue window to account for background conservation of the residue neighbours of a query residue of interest in the alignment column being scored (see section 3.4.1; Capra and Singh, 2007).

4.3.2 Structural data processing in PROTIN_ID

The conservation score values for each MSA column are mapped back to each query sequence's residues and are written in the residues' B-factor columns of the query PDB file in order to generate a conservation map to depict the residues' conservation signals on the query protein's surface. The "mapped" query PDB file is submitted to NACCESS by PROTIN_ID to calculate solvent accessible residues (Hubbard and Thornton, 1993). NACCESS calculates solvent accessibility by rolling a circular probe with a radius of 1.4 Å along a protein's van der Waals surface. The path undertaken by

the probe's centre for recursive slices of a protein's surface is known as the solvent accessible surface (Lee and Richards, 1971). Residues that are above a user-defined relative solvent accessible cut-off (default ≥15%, see section 5.3) are defined as surface and are retrieved by PROTIN_ID. The surface residues are ranked according to their conservation scores and the top N residues (default N = 20, see section 5.7) are extracted for further processing.

An option in PROTIN_ID exists to delete residues from the query protein where it is known that they are not part of the interface (*i.e.* using biological data). These surface residues may be important for other biological reasons (ex. a known cofactor binding site). Such residues may display strong conservation signals and mask the interface residues of interest by having higher conservation signals such that they may be part of the top-20 residues that are extracted and therefore the residue deletion list can eliminate these false positives from being part of the top-20 residues prior to clustering. This may allow other interface residues to take their place and be ranked within the top-20 conserved surface residues.

The next step involves clustering the extracted, top-20 conserved residues in order to determine clusters of conserved residues on the query protein's surface. To begin with, the top-20 extracted residues' carbon-α distances are measured from one another and the distances are recorded in an all-against-all distance matrix. The distance calculation between alpha carbons of two residues (Carbon α dis_{ij}) is defined as follows:

$$Carbon_{\alpha dis_{ij}} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \qquad (4\text{-}2)$$

where *x,y,* and *z* represent the α carbon coordinates of residues *i* and *j*. The result is the distance (Å) between two residues' α carbons. Data derived from this matrix is inputted for clustering using the OC clustering algorithm (Barton, 2002, 1993). The single-linkage clustering method, a form of hierarchical clustering, is used to cluster the residues at a carbon-α radial distance cut-off (Figure 4-5). In this method, residues ≤N Å radial cut-off (default ≤7 Å, see section 5.7) are clustered as neighbours from data derived from the all-against-all alpha carbon distance matrix until no residue nearest neighbours are found to any of the clustered residues comprising the current cluster (Ross, 1969; Johnson, 1967). This process is repeated to form a new cluster with the

remaining non-clustered residues until no more residues remain to be clustered out of the Top-20 extracted surface residues. PROTIN_ID applies a heuristic for sorting a cluster whereby it first sorts each cluster according to size (*i.e.* number of residues in a cluster) and calculates the average conservation of a cluster. The average conservation ($Average_{cons}$) is calculated as follows:

$$Average_{cons} = \frac{n_i}{N_j} \tag{4-3}$$

where n_i refers to the sum of the conservation of residues i of a cluster and N_j is the total number of residues of a cluster (*i.e.* cluster size). If some clusters are the same size, they are re-ranked according to their average conservation.

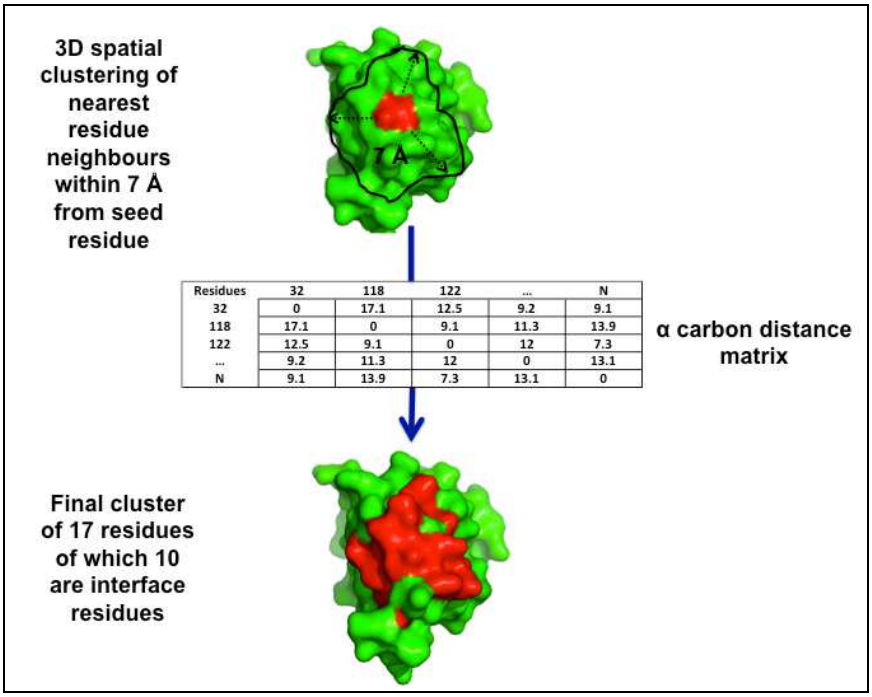

Figure 4-5: Overview of the clustering steps performed in PROTIN_ID. The above protein is transforming growth factor-beta receptor (TGF-beta receptor). The first seed residue of this protein has its distances from its alpha carbon to all other Top-20 extracted surface residues' alpha carbon atoms computed. This process is also performed for every other surface residue. The result is an all-against-all alpha carbon distance matrix. The next step begins with a seed residue to start a cluster. All nearest residue neighbours within 7 Å from it are clustered. This is repeated for the new residues of a cluster until no neighbours remain for all residues in a cluster. The final result is a cluster of 17 residues of which 10 are true interface residues.

The top-ranking cluster of conserved residues is predicted as being part of the protein interface by PROTIN_ID. Previous studies have demonstrated that interface residues are in close proximity to each other in three-dimensional structural space (Ahmad *et al*, 2010; Guharoy and Chakrabarti, 2010). It is hypothesized that interface residues would be part of the top-ranking cluster (see section 5.7). Furthermore, clustering allows the filtering out of lone residues on a protein surface and sorting clusters by size eliminates smaller clusters, which may be part of small molecule binding sites (Guharoy and Chakrabarti, 2010; Ofran and Rost, 2007b). In Figure 4-6, it can be seen that clustering of residues for two proteins adrenoxin and transforming growth factor beta receptor (TGF-beta receptor) results in the largest clusters with interface residues comprising them and the isolation of smaller lone clusters that consist entirely of non-interface residues. Moreover, the proteins' conservation maps, which display the level of conservation on their surfaces, indicate that the larger clusters are more conserved than the smaller ones. These maps can be a useful aid to the users of the PROTIN_ID method as they can be used to expand the top cluster if required by increasing the clustering radial cut-off.

4.3.3 PROTIN_ID theoretical restraints output for use in protein-protein docking

The first output file created by the PROTIN_ID method is a cluster of conserved interface residues prediction file that contains all clusters ranked by size and average conservation (see Figure 4-7). The top cluster prediction can be inputted as direct active residue restraints in the HADDOCK docking method (see section 3.8.1) for query proteins in order to guide docking sampling (Dominguez *et al*, 2003).

The second set of files created by PROTIN_ID are a PDB file (with conservation scores in the B-factor column) coupled with a PyMol script to generate the conservation map of the query protein, and another PDB file with its respective PyMol script to generate all clusters in different colours (sorted by size and average conservation) predicted for the protein as shown in Figure 4-6 (Delano, 1998). Both these sets of query protein PyMOL visualization files are meant to complement each other and guide the user into making an informed decision for later use of theoretical restraints in docking.

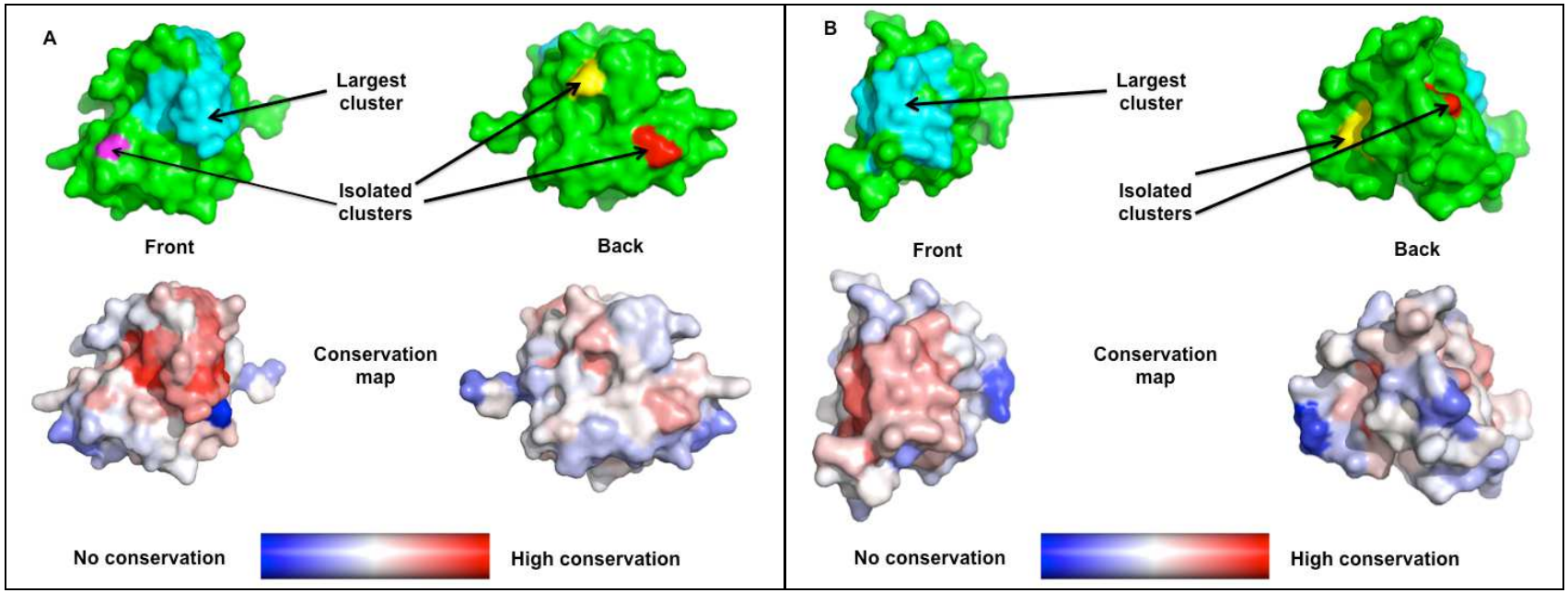

Figure 4-6: PROTIN_ID clustering and conservation map features. The effect of clustering in isolating lone clusters from the largest cluster is shown when the Top-20 conserved surface residues are clustered for two proteins. **A)** The Adrenoxin protein has 4 clusters predicted. Three clusters are small and isolated clusters (hot pink 1 residue, yellow 2 residues, and red 1 residue) in the front and back of the protein's structure with respect to the largest cluster (light blue). These small clusters are comprised of non-interface residues. The largest cluster contains 16 residues, including 12 interface residues. Based on its conservation map, it can be seen that the largest cluster is comprised of more conserved residues relative to the smaller clusters. **B)** The transforming growth factor-beta receptor has three clusters in total. Two clusters are small and isolated with regards to the largest one and contain non-interface residues (yellow 2 residues; red cluster 1 residue). The largest cluster comprises 17 residues, including 10 interface residues. Also, it is comprised of more conserved residues than the smaller clusters in the conservation map.

1M9Z.pdb
~~~~~~~~~~~~~~~~~~~~~~~~

Clustering 20 residues from 77 surface residues (20 extraction) using a 7 Angstrom(s) radial cutoff:
17, 0.73, Cluster_1 ==> 81 59 32 30 27 118 119 122 46 52 51 50 49 48 47 78 54
2, 0.73, Cluster_2 ==> 100 126
1, 0.73, Cluster_3 ==> 66

Figure 4-7: An example of PROTIN_ID's clusters of conserved interface residues prediction file for the transforming growth factor-beta receptor protein (TGF-beta receptor) is shown. In this file, the top prediction (*i.e.* Cluster_1) followed by the small and isolated clusters are indicated. The Top-20 residues are shown that have been extracted from 77 surface residues for this protein. The clustering of Top-20 residues was performed using a 7 Å radial cut-off. Each cluster is ranked by size (ex. 17 for the largest cluster) and then average conservation of a cluster (0.73), if two clusters are the same size. In this case, since all clusters have the same average conservation value, they are ranked by size.

The third file outputted by PROTIN_ID is a multiple sequence alignment of the query sequence and its homologous sequences (*i.e.* hits) as seen in Figure 4-4. A user may choose to manually re-edit the alignment or add other sequences to this alignment and then resubmit this file along with the query protein's PDB file to PROTIN_ID for re-prediction. PROTIN_ID allows a user to input a multiple sequence alignment derived from third party sources along with an input PDB file. In this case it initiates its protocol from the conservation score step to score the MSA and assign the residue conservation values to the query protein's residues. By default CLUSTAL, FASTA, and HSSP alignments are recognized. However, HSSP alignments from the HSSP alignment database are first converted into FASTA format using the program MView (version 1.52) prior to conservation score analysis (Brown *et al*, 1998; Dodge *et al*, 1998).

4.4 Conclusion

In general, most interface predictors utilize sequence and structural data to derive interface predictive features (see sections 1.8 and 1.9.4 and table A-2). Sequence data is represented in the form of a multiple sequence alignment, which is generated automatically. In some instances, manual re-editing of MSAs becomes necessary when they are unstructured, and this becomes a limiting factor when performed in a high-throughput setting of generating theoretical restraints for many proteins. For sequence data analysis, useful and user-friendly features were introduced to improve interface residue prediction performance by reducing the factors that cause MSA errors, which require manual re-editing by the user, and these new features distinguish PROTIN_ID from other interface predictors (see sections 1.8 and 1.11 and table A-3). These features introduced involve filtering and editing hit sequences prior to multiple sequence alignment generation. The filtering steps use the fraction of coverage of the hit to the query to remove short sequences, and a feature that removes redundant (*i.e.* identical) sequences from a sequence dataset to reduce overrepresentation of identical sequences in an MSA. In addition, the editing heuristic removes N/C-terminal overhangs and insertions in hit sequences that do not have corresponding residue counterparts in the query sequence of interest, resulting in better prediction results when comparing MSAs generated with (refined) and without (unrefined) the editing heuristic for the same number of sequences (see sections 2.2.2 and 4.3.1). Also, an implicit effect of the editing heuristic is that residues that are misaligned are likely aligned in the appropriate sequence columns in refined MSAs relative to their original positions in unrefined MSAs. This implicit effect is a result of the removal of alignment error-inducing factors (ex. long N-terminal overhangs), which cause residues to be aligned in the wrong regions, because other unlikely residues present in such overhangs are occupying their sequence columns.

Structural data is represented by a query protein's PDB coordinate file. The clustering of the top-20 conserved surface residues (≥15% ASA) in three-dimensional structural space using their alpha-carbon coordinates from a PDB file allows conserved residue clusters to be identified and sorted by size and average conservation when two clusters

have the same size. It also eliminates lone residues from the final prediction. A recently published method, CCRXP, has a similar protocol to PROTIN_ID, but lacks these practical and user-friendly features introduced in PROTIN_ID, namely the filtering and sequence editing heuristics, and the ranking of conserved residue clusters, which are useful to improve interface residue prediction (Ahmad *et al*., 2010). Also, if an experimentally solved protein structure is not available, it is possible to use homology models of proteins for prediction of interface residues using PROTIN_ID, which is not implemented in CCRXP. Furthermore, third-party MSAs may be used in PROTIN_ID, but this option is not present in CCRXP. The PROTIN_ID webserver has also been implemented for ease of use and convenient access for users in the generation of theoretical restraints for data-driven docking (Figure 4-8). In addition to this, a PROTIN_ID script that runs on Linux and UNIX operating systems is available to users that can be used in high-throughput prediction of clusters of conserved interface residues of unbound proteins.

In order to assess the prediction performance of the newly developed PROTIN_ID method, it will be systematically tested on a dataset of unbound proteins known to interact (Benchmark 4.0). The results for the unbound proteins will be compared to their known protein-protein complexes to examine PROTIN_ID's performance in interface residue prediction relative to random prediction (see Chapter 5). Furthermore, benchmarking of the new PROTIN_ID method's performance against other methods (ex. CCRXP) is useful to assess it strengths and highlight areas for further development of the method (see Chapter 6). Upon satisfactory optimization of PROTIN_ID's parameters and performance testing, the method will be applied to generate theoretical restraints to drive protein-protein docking to improve docking performance (see Chapter 7).

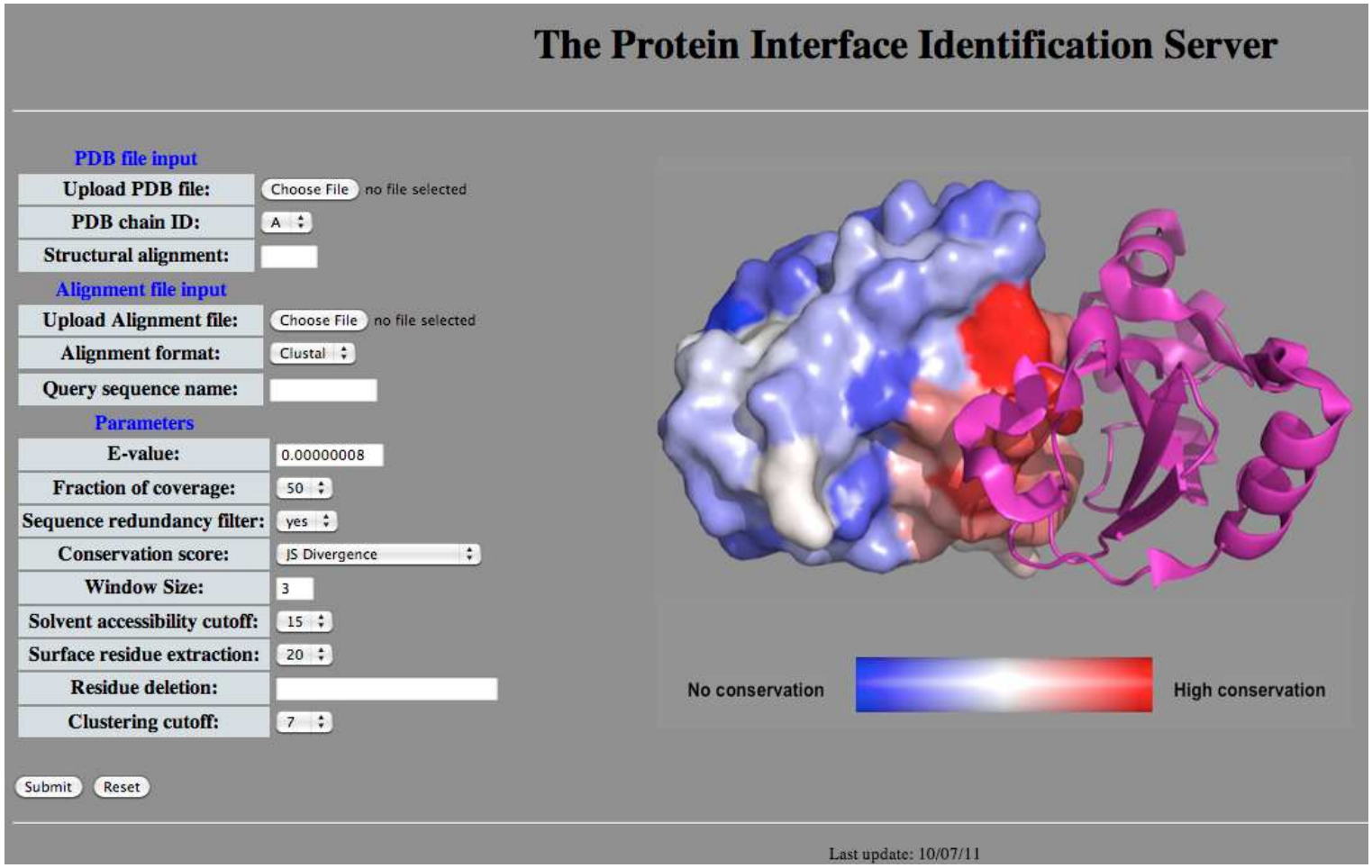

Figure 4-8: The input web interface of PROTIN_ID. The interface is divided into PDB file input, alignment file input, and parameters sections. The minimal input is a PDB file to upload to run the method. In addition a multiple sequence alignment can be uploaded in combination with a PDB file.

Chapter 5

Prediction of protein-protein complex interface residues

5.1 Introduction

Most cellular processes are driven by protein-protein interactions (see section 1.2). It is important to consider whether protein interface residues that drive these interactions are more conserved in comparison to the remainder of protein surface residues. Protein complexes and in particular intra-species protein complexes may have a tied development pattern in the evolutionary lifespan of an organism (actual hypothesis H_a). This may allow an evolutionary signal to be detected. For example, an enzyme-inhibitor interaction may have evolved into it current mature regulatory state where its interface residues are conserved because of functional constraint, and its interface is detectable because of this. On the other hand, it may well be possible that protein interfaces are no more conserved than the remainder of protein surface residues given that proteins can be promiscuous in nature. Therefore, conservation differences between a protein's interface and (non-interface) surface are no better than random, indicating an absence of an evolutionary signal (null hypothesis 'H_o'). To disprove the null hypothesis, statistical analyses of conservation signal differences between interface and rest of surface (ROS) residues of intra-species protein complexes were performed. The findings were discussed in the context of previous work (see section 5.5).

If interface residues emit a detectable conservation signal, it would be desirable to utilize it as a predictive property for such residues. For example, this would allow predicted interface residues to be applied as sampling restraints in protein-protein docking of two proteins known to interact (see section 1.6.1.1). However, the application of sequence conservation as an interface residue predictive feature is a debated topic. This issue is addressed in this chapter and an approach is explored where sequence conservation of surface residues is coupled with their spatial clustering (*i.e.* using known protein structures) to predict conserved interface residue clusters. This strategy is applied in the protein interface predictor, PROTIN_ID, which is used here.

Performance evaluation is done as a cluster is grown (to optimize PROTIN_ID) and is compared in terms of interface prediction reliability to a scenario where clustering is not applied, but residues that are conserved are predicted as interface residues. Both clustering and non-clustering are also compared in the framework of interface prediction at random to gauge their contributions in terms of interface prediction reliability.

5.2 Dataset for analysis of protein-protein interactions

The Benchmark 4.0 dataset is a manually curated dataset of protein-protein complexes (Hwang *et al*, 2010). There are a total of 176 structures of protein complexes in this dataset composed of intra-species, inter-species, and antibody-antigen interactions (see section 3.2). For each protein complex, its unbound (*i.e.* free form) protein constituents are available. Only intra-species binary complexes were selected for statistical analysis, which included their unbound proteins. This is because intra-species proteins evolve in the same organism in the context of an important biological function crucial for an organism's survival in which they are involved (ex. enzymatic inhibitory activity; Johnson *et al*., 2007). This allows the comparison of interface versus ROS residue conservation signals directly, which was done for binary interactions, and is useful for training of PROTIN_ID. All intra-species non-binary complexes (ex. three or more chains in a complex) were excluded from analysis because they may have overlapping interface residues, making an interface of interest difficult to predict by an interface predictor, if knowledge is not known in advance of other interfaces to exclude them from the final predictions. Because inter-species interactions are composed of proteins from different organisms they were excluded from analysis as a result. The antibody-antigen interactions were also excluded from analysis (see section 1.9.3).

5.3 Classification of residues of the dataset into interface and the rest of the surface

The number of interface residues, the rest of the surface residues (ROS), and core residues were calculated for the unbound proteins of intra-species binary complexes dataset based on the interface residues determined for the same proteins in their bound form (see section 3.2.1). This data is summarized in Table 5-1. Using the CAPRI interface distance criterion, overall 2,928 'bound' interface residues were yielded $\leq$ 5 Å in distance for the proteins in the bound pose of the dataset. Next the same proteins of the dataset in the unbound poses were examined to see if the same interface residues were present. It was found that 2,829 'unbound' interface residues were present in total in the dataset. There are 99 interface residues fewer in the total number of interface residues of the unbound proteins. This is because they are not found in those proteins' structures (*i.e.* in their PDB files). In the training of a protein interface predictor only unbound structures should be used, mimicking a blind prediction test (see section 3.5). Therefore, such interface residues found only in the bound proteins, but not in their unbound counterparts are discarded from training.

There are a total of 25,132 residues for the unbound proteins in the dataset. Using the solvent accessibility criterion ($\geq$ 15% accessible surface area %, ASA), they were divided into surface and core residues (see section 3.2.1). Accordingly, there are 17,256 surface residues and 7,876 core residues. The surface residues consisted of 2,498 interface residues that overlap with the distance criterion-derived unbound interface residues, and the remaining 14,759 residues are non-interface surface residues (rest of the surface, ROS). There are 331 interface residues < 15 % solvent accessibility cut-off and they are classed as core residues and hence are not included in further statistical analysis. The percentage of ASA interface residues out of the total surface residues is 14.5% for the unbound proteins dataset. As a result, there is a substantial bias towards ROS residues.

Table 5-1: Total interface, core, and rest of the surface residues of the unbound dataset of binary intra-species complexes. The bound proteins' total interface residues of the dataset are included for comparison.

Bound proteins total interface residues (Distance[a])	Unbound proteins total interface residues (ASA%[b])	Unbound proteins total ROS[c] residues (ASA%)	Unbound proteins total core residues (ASA%)
2,928	2,498	14,759	7,876

[a] CAPRI distance criterion (≤ 5 Å) to derive interface residues.

[b] Accessible surface area % criterion (≥15%) to derive interface residues.

[c] ROS = rest of the surface non-interface residues.

The average interface sizes were calculated for the two interface criteria for the unbound and bound proteins, and this is summarized in Table 5-2. The unbound ASA% average interface size calculated (20.31) differs by approximately three residues in comparison to the unbound interface average (23.00) determined by the distance-based criterion. This comparison is important because, in real-world terms, the ASA% is the only criterion that can be used to determine surface residues from the core residues of a protein. Therefore, this division into surface and core residues is an attempt to maximize the number of interface residues in the extracted surface residues. The ASA% is an interface residue predictive feature (see section 1.7.2). Using an ASA% threshold allows the discernment of how many interface residues above the threshold are determined on average as surface residues in the dataset. And the comparison of this ASA% interface residue average to the known unbound distance-based interface average size acts as a gauge to determine the suitability of the ASA% cut-off. Of course the lower the ASA% threshold the closer the ASA% interface residue average is to the distance-based average, which means more residues are designated as surface (and ROS). This threshold used (*i.e.* ≥ 15% ASA) indicates a negligible loss of interface residues (approximately three residues) and justifies using this specific solvent accessibility cut-off in the PROTIN_ID interface prediction method (see Chapter 4). Those 331 'lost' interface residues (found in 61 complexes studied) that are under the 15% solvent accessibility threshold had average solvent accessibilities of 5.76% (STDVs of 4.68) and 2.41% (STDVs of 3.96) for the side and main chains, respectively. This means the

ASA% threshold would have had to be lowered considerably just to extract those interface residues. Doing this would likely introduce many core residues as surface residues, which are conserved (Mintseris and Weng, 2005a). This would make it more difficult to assess the difference between interface and ROS conservation difference. It is worth mentioning that the same ASA% cut-off has been used in an earlier study for previous versions (1.0 and 2.0) of the protein-protein benchmark (de Vries *et al*, 2006; Mintseris *et al*., 2005b; Chen *et al*., 2003a).

Given that structural discrepancies can exist where residues are not found in an unbound chain relative to its bound chain counterpart, these missing residues can also be part of the interface. As mentioned previously, 99 interface residues are not found from the unbound chains compared to the bound chains. Comparing the averages of the bound vs. unbound interface based on the distance-based criterion indicates a minor difference of approximately one residue on average that is missing. This indicates that the loss of the 99 interface residues is negligible.

An important use of the calculated averages in table 5-2 is that they are relevant to calculate what is expected for interface prediction at random. This random prediction is important to know as it allows comparison to the prediction performance of an interface predictor method to determine if it performs better than random, is no better than random, or is worse than random prediction (see section 5.7).

Table 5-2: Calculated averages of the unbound/bound interfaces and unbound rest of surface (ROS) residues of the dataset of binary intra-species complexes. The solvent and distance-based definitions were used for the unbound/bound protein interfaces. The standard deviations of the average interface sizes are indicated in parentheses.

Unbound protein interface ave. (15% ASA)	**Unbound protein interface ave. (≤5 Å distance)**	**Bound protein interface ave. (≤5 Å distance)**	**Unbound protein ROS surface ave. (15% ASA)**
20.31 (±6.77)	23.00 (±8.09)	23.80 (±8.21)	119.99 (±80.44)

5.4 Analysis of protein interface residue conservation vs. ROS residue conservation

Conservation signals were calculated using seven different evolutionary conservation scores (available in PROTIN_ID) using multiple sequence alignment data for all dataset proteins (see sections 3.3 and 3.4). Using all conservation scores allows comparison of their results regarding interface residue versus ROS residues evolutionary conservation. The difference in conservation signal between interface and the rest of a protein's surface was calculated by subtracting the average ROS residue surface conservation signal from the average interface residue conservation signal (see section 3.6). The resulting signal difference values (ΔCons) can be a positive, negative, or zero value, which represents the presence of an interface conservation signal greater than the ROS, the presence of the non-interface surface residue signal greater than interface residues, or no detectable interface signal conservation because both the interface and the remainder of the surface residues are equally conserved and negate each others effects. Why are differences necessary to calculate? In a typical multiple sequence alignment, background noise may be introduced through difficult to align sequence regions, causing alignment errors, for example. This introduction of background noise equally affects the separately measured conservation signals of interface and the rest of the surface residues since they are derived from the same multiple sequence alignment. By calculating the ΔCons values, this noise does not affect the positive or negative direction of the ΔCons value of the interface conservation signal vis-à-vis the remainder of the non-interface surface residues. This effectively allows for statistical analysis to quantify the extent of statistical significance or lack thereof of ΔCons.

5.4.1 The probability distribution of empirical data (ΔCons)

Any statistical analysis assumes certain preconditions are satisfied upon its application. As such, knowing the probability distribution the empirical data (*i.e.* ΔCons) approximately models is important for applying suitable statistical tests (*i.e.* parametric or non-parametric) for valid data analysis to determine if interface conservation is more conserved than ROS residues (Park, 2008). Otherwise applying an incorrect statistical analysis that assumes data are normally distributed, for example, to data that does not

follow this statistical model will result in unreliable interpretations (Park, 2008). To apply the appropriate statistical test, the ΔCons data was tested for normality (see section 3.6). Accordingly, histogram distributions of ΔCons and their respective Q-Q plot diagnostic tests indicate that the underlying distribution relationship between interface and non-interface surface residues follows the Normal distribution (Figure 5-1). In the Q-Q plots, the quantiles of the actual data are plotted against those of the theoretical distribution (Normal distribution). In the Q-Q plot, the intersecting diagonal line represents the existence of perfect agreement between theoretical and actual data and implies that the actual data are normally distributed (Park, 2008). The plotted ΔCons empirical data residuals in all Q-Q plots have a linear pattern with respect to the intersecting diagonal lines. In addition, a normality test (Skewness-Kurtosis test) was applied to demonstrate that the data is normally distributed. This Skewness-Kurtosis test's null hypothesis is that the data are normally distributed. If the null hypothesis is disproven, then the data is not normally distributed. The p-values of this test show that no significant departure of normality has been observed, accepting the null hypothesis with regards to the data being normally distributed (Table 5-3).

Other normality tests (Shaprio-Wilk and Shapiro-Francia tests) were applied on the data and produced p-values in agreement with the Skewness-Kurtosis test, accepting the conclusion that the data are normally distributed. Two forms of normality tests (graphical and numerical) were conducted to determine the probability distribution the ΔCons data follow (*i.e.* normal distribution or not) as a precursor to the actual statistical analysis, which examines the significance of the conservation of interface residues to the ROS residues. All tests agree that the empirical data (ΔCons) calculated from all seven conservation scores are normally distributed, which means that parametric tests can be applied for the next step of analysis.

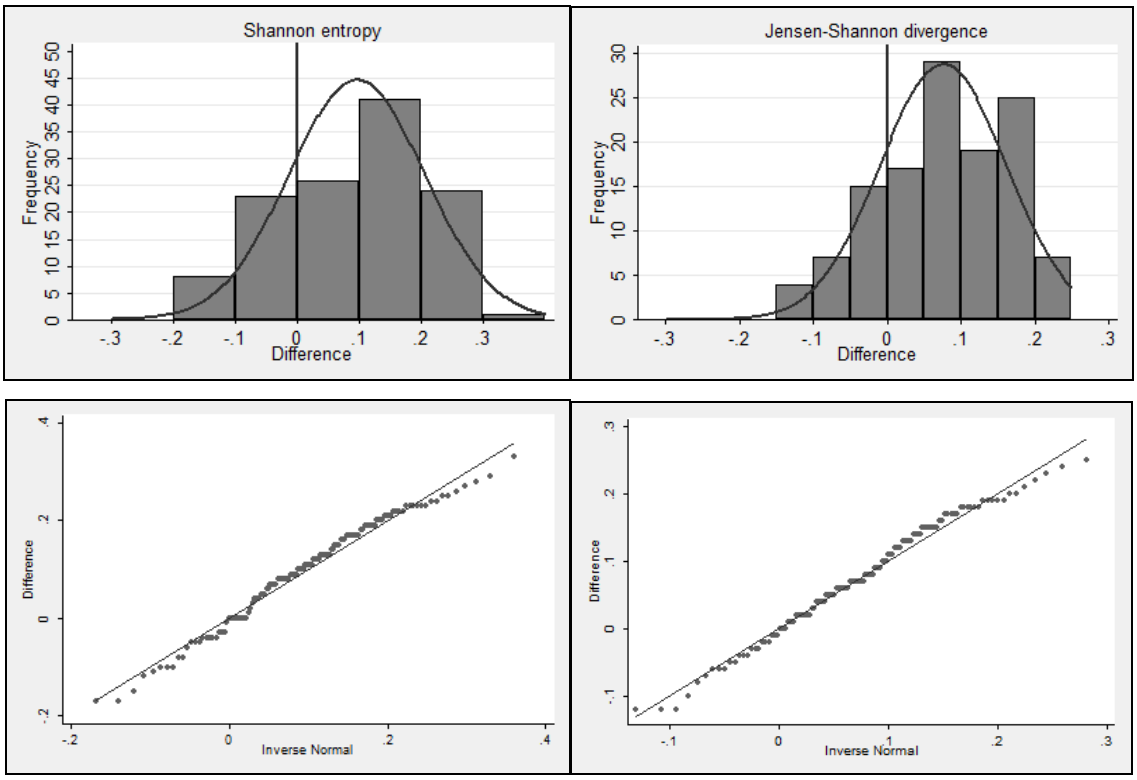

Figure 5-1: Histogram distributions of the differences (ΔCons, see equation 3-20) between average interface and rest of surface (ROS) residues of the dataset. The intersecting line at 0 marks no difference in conservation between the interface and the rest of a protein surface. ΔCons > 0 indicates an interface conservation signal. ΔCons < 0 indicates a non-interface surface residue signal. It can bee seen that ΔCons > 0 for the majority of the data distribution, indicating an interface conservation signal. The Shannon entropy and Jensen-Shannon divergence scores are shown with the Q-Q plots for each score below it. Q-Q plots depict the relationship between theoretical Normal distribution data (X-axis) and the actual experimental data (ΔCons, Y-axis). The intersecting diagonal line illustrates perfect agreement between empirical and theoretical data. As the residuals of the empirical data depict a linear pattern along the diagonal line, this suggests that the data follows the normal distribution.

Table 5-3: Comparison of conservation score averages of total surface ASA%, ROS ASA%, and interface residues ASA%. The standard deviations of the average conservation values are indicated in parentheses. The ΔCons > 0 shows that the majority of the interface conservation signals are better than the ROS ASA%. The Normality test P values were obtained from the Skewness-Kurtosis test. The null hypothesis (H_o) that indicates that the ΔCons values follow the normal distribution is not rejected for all scores. The 95% Confidence Interval shows the upper and lower bound range limits of ΔCons. Paired t test P values indicate the probabilities of the ROS residue conservation signal (ΔCons < 0), no detectable conservation signal between both groups of residues (ΔCons ≠ 0), and the interface residue conservation signal (ΔCons > 0).

Conservation scores	ASA[a] total surface	ASA ROS	ASA interface	Difference (ΔCons)	Difference above random (ΔCons%)	Normality test p value	95% Confidence Interval	P value ΔCons < 0	P value ΔCons ≠ 0	P value ΔCons > 0
Shannon entropy	0.53 (0.11)	0.52 (0.11)	0.61 (0.13)	0.10 (0.11)	79%	0.09	0.08 – 0.12	1.0[b]	< 0.001[c]	< 0.001
Property entropy	0.56 (0.11)	0.54 (0.11)	0.64 (0.14)	0.10 (0.12)	76%	0.18	0.08 – 0.13	1.0	< 0.001	< 0.001
Property relative entropy	1.59 (0.37)	1.53 (0.38)	1.85 (0.46)	0.33 (0.37)	79%	0.42	0.26 – 0.40	1.0	< 0.001	< 0.001
Relative entropy	0.52 (0.12)	0.50 (0.12)	0.60 (0.13)	0.10 (0.11)	80%	0.16	0.08 – 0.12	1.0	< 0.001	< 0.001
JS-divergence	0.5 (0.10)	0.49 (0.10)	0.57 (0.11)	0.08 (0.09)	75%	0.11	0.06 – 0.09	1.0	< 0.001	< 0.001
Von Neumann entropy	0.62 (0.11)	0.60 (0.11)	0.69 (0.13)	0.09 (0.10)	77%	0.11	0.07 – 0.11	1.0	< 0.001	< 0.001
Sum of pairs	2.00 (0.68)	1.87 (0.68)	2.53 (0.96)	0.66 (0.81)	77%	0.13	0.51 – 0.80	1.0	< 0.001	< 0.001

[a] ASA% = Accessible surface area percentage.
[b] P value 1.0 (> 0.05) indicates no significance.
[c] P value < 0.001 indicates extreme significance.

5.4.2 Statistical analysis of the empirical data (ΔCons)

For all the histogram distributions of ΔCons calculated for all evolutionary conservation scores, the results suggest that interface residues are more conserved than the remainder of a protein's surface where the majority of protein interfaces have ΔCons% > 0 (table 5-3). ΔCons% is the percentage of data greater than zero ΔCons. This indicates that for the majority of interfaces, their conservation signals are greater than the ROS residues (Figure 5-1). As the ΔCons data follows the Normal distribution, parametric statistical tests were applied to ascertain the significance of ΔCons (greater than zero) for all conservation scores used (see section 3.6). The confidence interval (CI), which describes the precision of the population mean (μ) of ΔCons, was calculated. The CI takes into account sample size and variability (standard deviation σ) when performing the analysis (Motulsky, 2007). The CI describes the 95% chance that the true population μ is defined within the CI range as delineated by the CI upper and lower bound limits. For example, for 95% of samples of the population (independently and randomly sampled) the CI reflects the probability that the true population mean is within the CI range and a 5% chance for the remaining 5% of the samples that it is beyond the CI range limits. Additionally, calculating the 95% CI allows the consideration of where the CI limit lower boundary falls in comparison to no detectable interface conservation signal (zero ΔCons). The interface conservation signal significance over the rest of a proteins surface (ROS) is demonstrated when the lower bound CI limit of ΔCons is greater than zero ΔCons. Indeed, for all evolutionary conservation scores applied it is observed that there is a 95% chance that all scores show lower bound CI ranges where ΔCons > 0. For example, the Jensen-Shannon divergence has a ΔCons mean of 0.08 with a lower bound CI limit 0.06 (ΔCons minimum) to an upper limit of 0.09 (ΔCons maximum), indicating a good precision of the score's ΔCons. All other scores are also in agreement, highlighting that there is an interface conservation signal to be exploited in the prediction of a protein-protein interface site since their ΔCons averages and ΔCons minimum are greater than zero (Table 5-3).

Taken as a whole, the dataset used for statistical analysis represents a sample of intra-species interacting proteins derived from a greater population of interacting proteins that

have evolved together in the same species. The CI provides the opportunity to make a broader inference about the interface conservation signal (ΔCons) by extrapolating the findings based on the sample to the population it is derived from. Since the calculated CI ranges for each score's ΔCons mean show good precision, indicating that there is a 95% chance that the population ΔCons mean (μ ΔCons) is similar to the sample ΔCons mean, there is a 95% probability that future extensions to the current dataset would have a 95% CI within the calculated ranges for all conservation scores. This means that if the current dataset were enlarged to include more intra-species interacting proteins, there is a 95% chance that the current CI results indicate that the ΔCons population mean is > 0, which shows that interface residues are more conserved than the rest of the surface residues.

To complement the CI analysis, the Paired t-test was used to assess the statistical significance of the ΔCons given that the null hypothesis's assumption is that there is no difference between interface and the rest of a protein's surface in terms of residue conservation (see section 3.6). The probabilities of actually observing the current ΔCons of all scores were computed. Table 5-3 summarizes the probabilities of ΔCons changes. It can be seen that for all evolutionary conservation scores the probabilities that ΔCons shifts in favor of interface or non-interface residue conservation (ΔCons ≠ 0) is extremely significant for all scores (p-value <0.001). This result is statistically significant and rejects the null hypothesis's assertion that no conservation signal is present between interface residues and other non-interface surface residues. However, the current result (ΔCons ≠ 0) only means that there is a conservation signal, but it does not elaborate if it is in favour of interface (actual hypothesis H_a) or ROS residues. Seeing that ΔCons ≠ 0 indicates the presence of a conservation signal, this ΔCons conservation signal could be a positive value (*i.e.* interface residues conservation signal, ΔCons > 0) or a negative value (*i.e.* ROS residues conservation signal, ΔCons < 0). The P values of negative ΔCons values are 1.0, which indicates no statistical significance to suggest that the conservation signal is present because ROS residues are more conserved than interface residues. This makes sense, as the lower bound CI limits of ΔCons were all greater than zero in the CI statistical analysis (*i.e.* no detectable conservation signal for ROS residues). This is confirmed for all conservation scores and therefore ROS residues do not have conservation signals that are more apparent than

protein interface residues. In contrast positive ΔCons P values are <0.001, indicating a statistically significant result and confirmed by the CI ranges computed for all scores in Table 5-3. All conservation scores form a consensus with regards to this. This accepts the actual hypothesis (H_a), and indicates a detectable conservation signal in favour of protein-protein interfaces when compared to non-interface surface residues.

It has been demonstrated that an interface conservation signal is present from analysis of the current dataset using parametric statistical tests that are based on the data being normally distributed. The statistical analysis does not support evidence to the contrary to accept the null hypothesis. For further confirmation of the results, the application of non-parametric tests that assume no underlying probability distribution (*i.e.* preconditions) for the data was performed to analyse the ΔCons significance for all evolutionary conservation scores (see section 3.6). The bootstrap approach and the Wilcoxon matched pairs test, for computing the CI and P values, respectively, both concur with the above results computed by the parametric tests. Overall, the evidence is in support of the actual hypothesis (H_a) in that interface residues are more conserved than ROS residues. The statistical data when looked at in the context of the interface size vs. non-interface surface residue size (see section 5.3) becomes more pronounced and illustrates the significance of the detectable conservation signal for protein interfaces given that they form 14.5% of the 17,256 unbound surface residues.

5.5 Interface vs. non-interface surface conservation: the views of others

Early work done by Grishin and Phillips (1994) focused on five enzyme oligomer proteins and sought to determine if interface residue conservation between subunits of oligomeric enzyme complexes is present and appreciable. They applied an identity score to examine the positional percentage identity of interface amino acids per sequence pair in a multiple sequence alignment compared to total sequence identity of the sequence pair. This allowed the analysis of the evolutionary rate difference of interfaces with respect to the rest of the surface. They concluded that interface residues are approximately 1.5 times more conserved than other surface residues, while enzyme

active sites and core residues showed the most appreciable conservation. This study bases its conclusions on a small dataset and on a reduced model of quantifying conservation based on comparing identities. However, accounting for physicochemical properties (see section 3.4.3, for example) of the interface amino acids may have improved their conservation signal (Valdar, 2002; Valdar, 2001). Nonetheless, based on the results of their study an interface conservation signal is still present waiting to be exploited. In another study, Valdar and Thornton (2001) analysed 6 homodimer proteins, using a more sophisticated score similar to the sum-of-pairs score (see section 3.4.8). They compared the conservation of interface residues vs. surface residues (that included interface residues) equal to the interface residues in number. They employ three ways: (1) interface residues vs. randomly selected surface residues, (2) interface residues vs. randomly selected structurally neighbouring surface residues and (3) interface residues vs. an almost circular patch of residues equalling the interface residues in number. Based on these thorough analyses, they concluded that interface residues are more conserved than the rest of the surface residues. Like Grishin and Phillips' (1994) study, the final conclusions are based on a small dataset. Prompted by the dataset size limitations of the previous studies, Caffrey, *et al*. (2004) used a considerably larger and diverse dataset to examine interface vs. non-interface surface residue conservation. Their dataset consisted of 64 proteins composed of 54 obligate complexes (42 homodimers and 12 heterodimers) and 10 transient complexes. Using the Von Neumann entropy score (see section 3.4.7) they conducted two analyses. The first analysis was comparing interface residues' conservation average to the rest of the surface's conservation average. They observed that interface residues were more conserved than the rest of the surface, producing a statistically significant result. Another analysis was comparing surface (*i.e.* non-interface and interface surface residues) residue patches' average residue conservation to that of the interface patches. Their results did not indicate statistical significance in support of interface conservation over the rest of the surface patches. Caffrey, *et al*. (2004) concluded that although the interface is more conserved than non-interface surface residues, because interface conservation vs. surface patches conservation was not significant in their patch analysis, evolutionary conservation as the only factor to predict interface residues is not sufficient. Burgoyne and Jackson (2006) took a different approach when analyzing interface conservation for 97 transient complexes. They divide a protein's surface into smaller sizes (clefts) and observed that interface conservation is not striking compared to surface residue clefts

when ranking clefts by conservation. Other studies also found that when compared to surface patches, interface patches did not display a significant conservation signal and that evolutionary conservation should be used with other features of protein interfaces and not by itself to predict interfaces (Capra and Singh, 2007; Reddy and Kaznessis, 2005). Mintseris and Weng (2005a) took a more cautious approach when analyzing interface conservation versus the rest of the surface, arguing that the estimation of conservation of non-interface surface residues is not an accurate estimate but an upper bound limit as the surface residues may contain other interface residues (*i.e.* crypto-interface residues). While acknowledging the possibility of crypto-interface residues being present in the rest of the protein's surface, Mintseris and Weng (2005a) compare conservation between core, interface, and surface/interface (*i.e.* non-interface surface) residues for 91 transient and 41 obligate complexes and show that the interface conservation is higher (statistically significant) than the surface/potential crypto-interface mixture but lower than core residues (Mintseris and Weng, 2005a). Bordner and Abagyan (2005), who show that interface residues are more conserved than the rest of the surface of proteins in the bound form for 518 homodimers, 157 hetero-dimers, and 862 multimers, also highlight the presence of crypto-interface residues classed as non-interface surface and their diminishing effect on prediction accuracy of transient heterodimer complexes. Choi, *et al*. (2009) in their study argued that multiple interfaces should be taken into account when comparing interface conservation vs. the rest of the surface. Their dataset consisted of 3844 protein complexes (bound form only) from which they isolated a total of 2646 interfaces (2344 and 302, obligate and transient interfaces, respectively). They demonstrated that interface conservation vs. the rest of the surface is improved and statistically significant when multiple interfaces are considered. This was more apparent for proteins where the rest of the surface was more conserved than a single interface and taking other multiple interfaces into account; the multiple interfaces were more conserved than the rest of the surface.

Some studies have much larger dataset sizes than the one used in this study and there are two reasons for this. Firstly, some datasets are composed of bound complexes only (Choi *et al*., 2009; Bordner and Abagyan, 2005), whereas the analysis of interface residue versus ROS residue conservation performed in this work was based on unbound proteins, which is important as these proteins will be used for testing of the

PROTIN_ID method prediction performance, simulating a realistic setting to predict interface residues (see section 5.6). Furthermore, the unbound proteins used in this work will be later applied in data-driven docking using PROTIN_ID's theoretical restraints to drive docking (see Chapter 7). Secondly, the dataset of this work is based on only intra-species interacting proteins, reducing the number of complexes in comparison to the other dataset of Mintseris and Weng (2005a).

To recap, previous studies compared average interface residue conservation (single or multiple interface models) versus the average rest of the surface residue conservation and showed that indeed the interface residues are more conserved, which is statistically significant. This is in agreement with the current analysis performed for interface and ROS residues in this work (see section 5.4.2). Therefore, there is a possibility to extract this signal by prediction. In some studies, they have implemented predefined protein surface patches of average interface size (Caffrey *et al*, 2004) or smaller sizes (clefts) (Burgoyne and Jackson, 2006) and did not yield an interface conservation signal that was significant when compared to other surface patches or clefts using conservation as a predictive factor. One question to ask is: are interface residues equally conserved relative to each other? If not, then only those residues exhibiting conservation may be used in interface prediction. If they are identified first, creating patches should then follow, which are examined to determine if a patch conservation signal is significant. Therefore, an approach should be tailored based on this observation. Herein, this is achieved through extracting a specific number of surface residues sorted by conservation, which are then clustered (*i.e.* structural window) according to proximity in three-dimensional space to create patches composed of clusters of conserved residues. This approach is implemented in the PROTIN_ID interface predictor and will be examined below.

5.6 Interface residues' conservation relative to one another

An analysis of interface residue conservation at different levels of conservation was carried out. Figure 5-2 shows the interface residue conservation of the entire dataset of interface residues compared to all ROS residues' conservation at different levels of conservation windows sorted from high to low conservation (calculated using the Jensen-Shannon divergence score, see section 3.4.6). Holistically, interface residues on average are more conserved than ROS residues. But it can be observed that individual interface residue conservation is spread out from low conservation to high conservation with other surface residues occupying the same window ranges. This non-homogeneity in interface residue conservation agrees with an earlier finding, indicating the differences of evolutionary pressure on interface residues (Guharoy and Chakrabarti, 2005). Relating this data to the core-rim interface model (see section 1.4.4), core residues are more conserved than rim residues (Guharoy and Chakrabarti, 2005). This implicitly indicates that conserved residues are also localized near to each other under the assumption they are core residues, according to the core-rim model.

It can also be seen that there are highly conserved ROS residues, which may be part of small molecule binding sites or other multiple interfaces. As such, if the goal is to retrieve all residues for the interface of interest directly, this may be hampered by a likely increase in the presence of conserved ROS residues (*i.e.* crypto-interfaces residues), which may influence patch analysis. In the context of generating theoretical (conservation) restraints in protein-protein docking, conserved ROS residues are not important for docking of two proteins since they play no role in the interaction of the proteins of interest, and if included, may have the undesirable effect of misdirecting data-driven docking sampling, producing incorrect binding poses in the final docked models (see section 1.6.1.1). Therefore these residues' presence should be minimized in a final interface prediction, while maximizing the presence of the number of residues of the interface of interest to ensure sufficient data is present to drive docking.

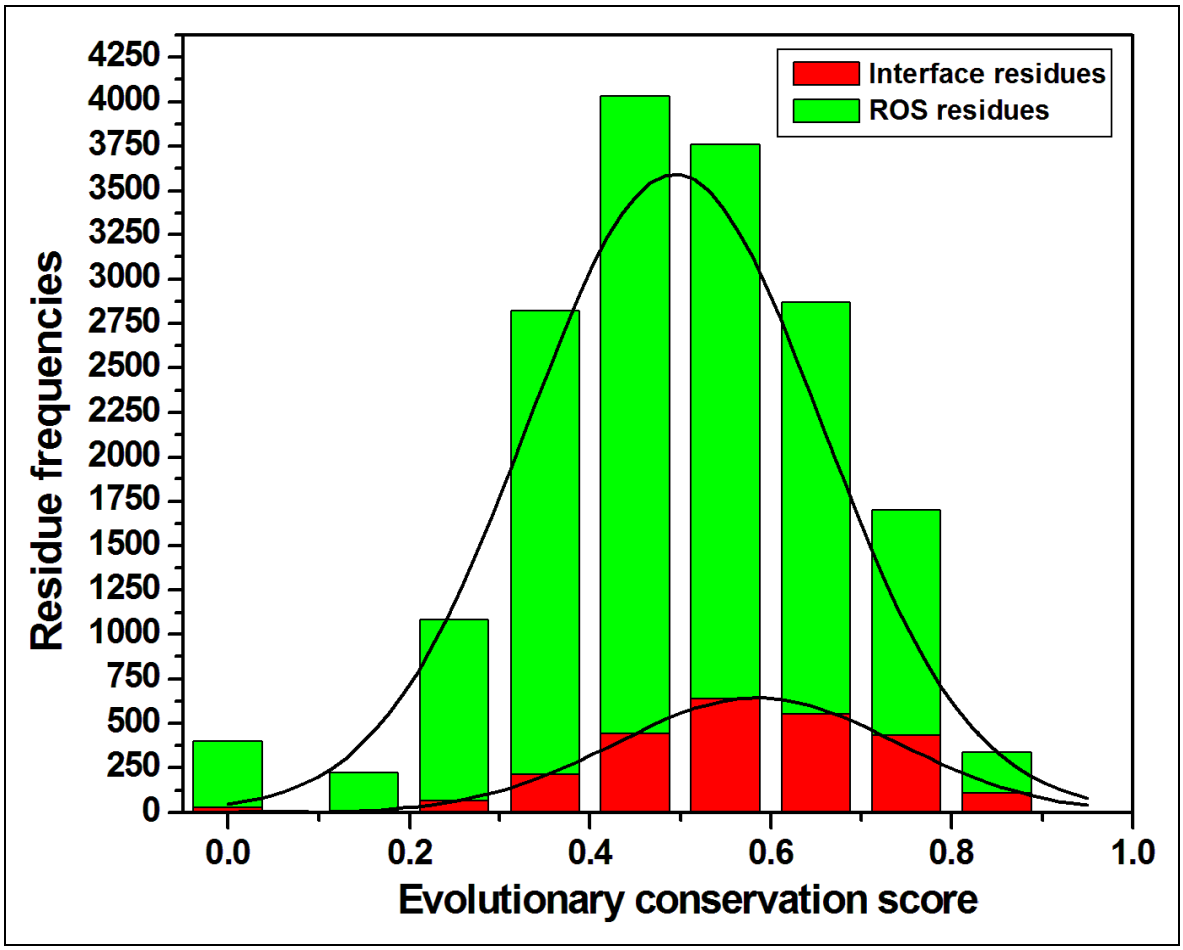

Figure 5-2: Histogram distributions of all interface residues (red) compared to rest of the surface residues (ROS, green) of the dataset, according to different conservation windows (calculations were performed using the Jensen-Shannon divergence score). Low and high conservation are 0 and 1, respectively. Holistically, interface residues on average are more conserved than rest of the surface residues. It can be seen that interface residues are spread out at different conservation windows, indicating non-homogeneity of conservation. This is the same for the rest of the surface residues. Attempting to predict all interface residues will increase background noise from the ROS residues. The ROS residues may be highly conserved because they are small molecule binding sites or crypto-interface residues.

5.7 Use of clustering to improve prediction of interface residues

Since interface residues differ in conservation, it is hypothesized that high ranking surface residues (sorted by conservation) will be comprised of 'N' conserved interface residues that are in close proximity in three-dimensional space (*i.e.* cluster of conserved residues) when visualized on a protein's surface. This means that if the top-N conserved surface residues are extracted, some of those residues may be residues of the interface of interest. Therefore, it may be possible to identify such residues since they could be in close proximity to each other through spatial clustering, which may also eliminate isolated residues to increase interface prediction reliability. An alternative to extracting top-N residues sorted by conservation is to use an absolute binary conservation cut-off and select residues above this cut-off; however, this approach was found to introduce many ROS residues, which may decrease prediction reliability through the increase of ROS residues in clusters. This is supported in a recent study, which has indicated that taking top-N residues instead of using absolute score cutoffs improves interface prediction reliability (de Vries and Bonvin, 2011a).

To test this hypothesis of taking top-N residues, the PROTIN_ID method was used to generate clusters of top-20 extracted surface residues sorted by conservation for all unbound proteins of the dataset (see Chapter 4 for description of PROTIN_ID). The top-20 residues extracted is equal to the average size of the (unbound) interface (see table 5-2). As stated earlier, the use of unbound proteins is necessary to ensure a realistic setting, especially for later use of theoretical restraints to drive docking (see Chapter 7). If bound proteins were used in this analysis, interface residues may possibly be in closer proximity because of bound pose conformational changes, introducing bias and may artificially enhance the effect of spatial clustering (see section 1.9.2). Indeed, a recent study suggested that using bound proteins for training resulted in better performance of methods in general when compared to using unbound proteins in training (de Vries and Bonvin, 2011a).

Starting with a seed residue from the Top-20 residues, a cluster was systematically increased in size to examine the extent of interface residues (True positives, TPs) that

cluster and to minimize the presence of ROS residues (False positives, FPs) in each unbound protein's final cluster (see section 3.5). More than one cluster of varying size may be generated for a single protein and clusters were ranked according to the cluster size first (*i.e.* number of residues in a cluster) and then average residue conservation of a cluster if two or more clusters were equal in size to re-rank them. The top ranked cluster is taken as the final interface prediction. In order to examine the effectiveness of clustering on the dataset as a whole, the average TPs (Interface residues, 20) and FPs (ROS residues, 120) of the entire dataset (see table 5-2) are needed for comparison to the average cluster generated by PROTIN_ID for the entire dataset. This average cluster size generated by PROTIN_ID was statistically analyzed by performance measures to examine the effect of clustering (see section 3.5). Table 5-4 displays the average cluster sizes when extracting the top-20 surface residues and clustering them at incremental cut-offs (from 4 - 8 Å) with the statistical analysis of each cluster. On average, when the top-20 surface residues are extracted, about 7 interface and 13 ROS residues are extracted. The next step is to maximize the ratio of TP:FP such that it is better than the random ratio of TP:FP (Kufareva, *et al*, 2007).

Random TP to FP ratio is derived from the average interface size (20.31) compared to the average ROS size (119.99) fractions' of the total surface (Table 5-2). Therefore the minimum (normalized) ratio is one interface residue to six ROS residues. If 100 residues are sampled from N population of surface residues (*i.e.* 17,256), then it is expected that 14.5 residues are interface residues and the remaining 85.5 are ROS residues. The TP (specificity) and FP fractions of random prediction for interface and ROS residues are approximately 14.5% and 85.5%, respectively. It can be seen that as the clusters are grown more TPs are included as are FPs in the growing cluster. To analyze the significance of each grown cluster, the TP fraction (specificity) and FP fractions are computed. The TP fraction quantifies the fraction of correctly predicted TPs in a cluster as it is grown by 1 Å (see equation 3.5.2). Also, the FP fraction computes the fraction of FPs present in a cluster (see equation 3.5.3). It can be seen that all TP fractions for the differing radial cut-offs are greater than random (14.5%) with an average fraction of 43.6%. This is also the same for the average FP fraction of 56.4% being smaller than 85.5%. This indicates that, for example, for 4.42 interface residues at the cluster radial cut-off of 7 Å an equivalent of 5.76 FP residues are clustered. This value is lower than the random value of 23.61 FP residues for every 4 TP residues.

Table 5-4: The extraction of the top-20 conserved surface residues by PROTIN_ID. The top-20 residues contain about 7 interface residues (true positives) with the remaining 13 residues as the rest of the surface (false positives). Clustering residues at increasing distance cut-offs increases the TP count. The FP count is also increased but minimized, compared to the minimum ratio of TPs to FPs (2:12), when selecting 14 surface residues at random. Clustering demonstrates that a TP signal can be exploited at improved ratios when comparing the TP and FP counts. TP and FP fractions indicate the cluster TP and FP percentages. TP and FP rates indicate how much actual TPs and FPs are present from the pool of TPs (20) and FPs (120) (table 5-3). Accuracy reports how successfully correct TPs are predicted in the cluster and successfully correct FPs have not been predicted in the cluster from the total 120 ROS residues. The F-measure and MCC measure increase as more TP residues are clustered; indicating that the effect of clustering is better than selecting the same number of residues of a cluster at random, where random is 0 for both measures. The average cluster conservation indicates that the clusters are conserved.

Radial cut-off (Å)	Extracted interface	TP count	FP count	Cluster size (TP+FP)	TP fraction (Specificity)	FP fraction	TP rate (Sensitivity)	FP rate	Accuracy	F-measure	MCC[a]	CS average
4	6.72	2.02	2.37	4.39	0.46	0.54	0.1	0.02	0.85	0.16	0.16	0.71
5	6.72	2.27	2.91	5.18	0.44	0.56	0.11	0.02	0.85	0.18	0.16	0.71
6	6.72	3.44	4.46	7.89	0.44	0.56	0.17	0.04	0.85	0.25	0.2	0.7
7	6.72	4.42	5.76	10.19	0.43	0.57	0.22	0.05	0.85	0.29	0.23	0.7
8	6.72	5.15	7.47	12.62	0.41	0.59	0.25	0.06	0.84	0.31	0.24	0.7

[a] MCC = Matthews correlation coefficient

Figure 5-3 displays the TP and FP counts for the top-20 most conserved residues when beginning clustering from 3 - 23 Å and creating clusters at 1 Å increments until all 20 surface residues are clustered in the final cluster at 23 Å. The clusters are compared to what is expected at random and it is shown that clustering achieves better TP retrieval than random. Furthermore, the TP and FP average counts are shown for the entire dataset when the top-N conserved residues are extracted, starting from 2 – 20 residues and incrementing by one residue each time. Each absolute residue extraction values are taken as the final prediction without applying spatial clustering of the residues. When compared to the spatial clustering approach, it can be seen that clustering minimizes the FP count unlike the conservation-only extraction approach, improving prediction reliability (*i.e.* TP fraction), with the best clustering cutoff at 7 Å that is furthest away from the conservation-only data. If all top-20 residues are clustered or conservation-only extraction is performed, they are still better than random, however, this introduces more FPs, which when used in data-drive docking studies may result in incorrect docking solutions because of sampling in the wrong area of interest. This may be caused if some residues are on different 'sides' of a protein. This is where clustering is most useful in removing such lone residues through FP reduction (see Figure 4-6 and section 2.3).

Both the TP and FP rate refer to the number of interface/ROS residues present in a patch out of the total observed interface/ROS residues (see equations 3.5.4 and 3.5.5). It can be seen that TP rate increases as cluster size is grown and more TP residues are incorporated in the cluster, ranging from 10% - 25% of the observed number of average interface residues (*i.e.* 20.31 interface residues). Because the aim is to retrieve the highest conserved interface residues since it was shown that interface residues are non-homogeneous in conservation (see section 5.6), and attempting to retrieve them all would increase the number of FPs in a cluster (see Figure 5-2), it can be seen that the spatial clustering approach reduces the fraction of FPs in a cluster, which is reflected in the FP rates from 4-8 Å. These values range from 2 - 6% of the observed number of FPs (*i.e.* 119.99).

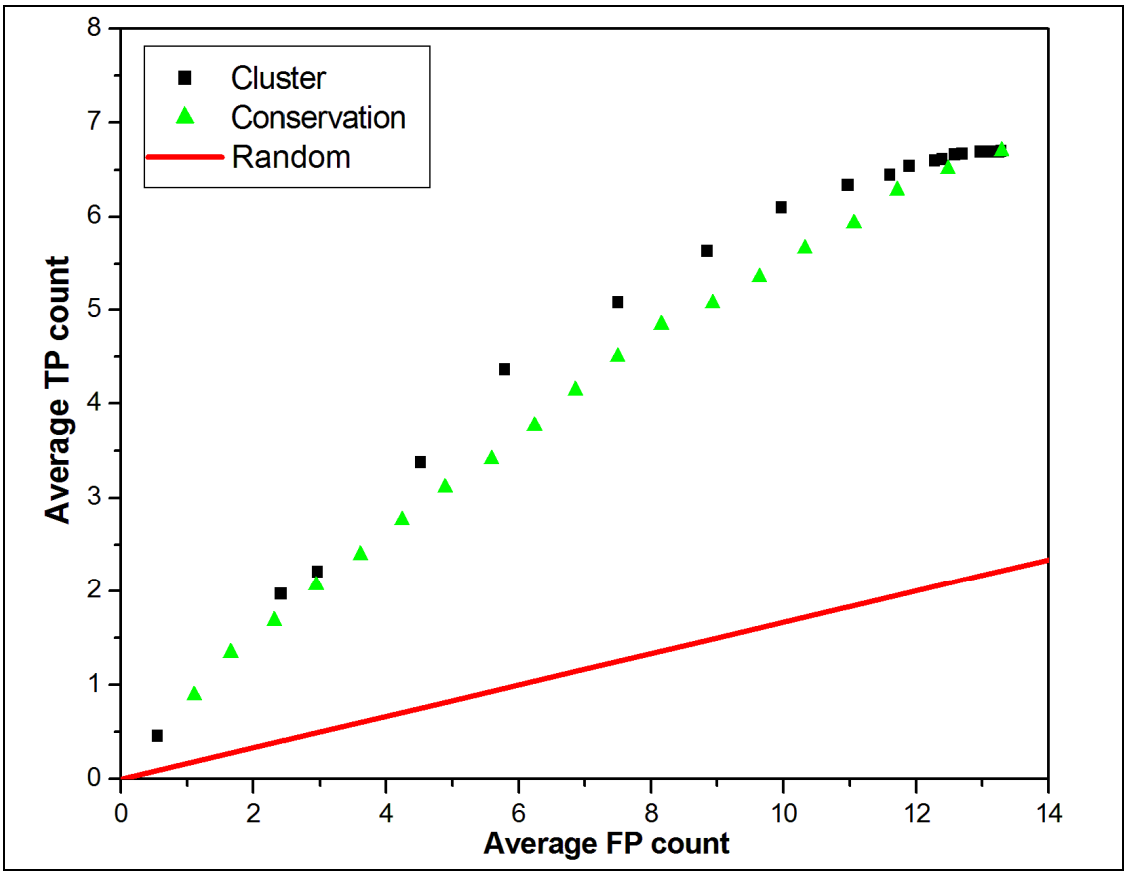

Figure 5-3: Clustering of the top-20 most conserved surface residues extracted for all proteins of the dataset. The average interface (True positives TP) and ROS (False positives FP) residues in the extracted Top-20 residues for all proteins are about 7 and 13, respectively. Clustering (black) is initiated with a 3Å radial cut-off and incremented by 1 Å at a time to form a new cluster. All extracted TPs and FPs are clustered at 23 Å. Clusters are compared to the top-N conserved residue extraction, starting from 2 – 20 residues, incrementing by one residue each time (green). Absolute residue extraction values are taken as the final prediction without the application of clustering. Both approaches are compared to random prediction (red). It can be seen that the effect of clustering improves the retrieval of TPs compared to random prediction and minimizes the effect of FPs when compared to conservation-only extraction, making it a useful approach for generating theoretical restraints for data-driven docking.

The accuracy measure quantifies the proportion of TPs and TNs (True negatives) correctly predicted (see equation 3.5.1) (Xue *et al*, 2011a). It can be seen that the accuracy is stable at all clustering cutoffs with an average value of 85% for the clusters. The accuracy measure's value should be taken as an approximation of performance of the clustering as the accuracy measure assigns the same weight to both correct TPs and non-prediction of FPs, resulting in optimistically high values due to the biased ratio of interface to ROS residues (de Vries *et al*, 2008). The accuracy measure has been featured for completeness. The F-measure computes the harmonic mean of the TP fraction (specificity) and the TP rate (sensitivity) (Van Rijsbergen, 1979) (see equation 3.5.6). A zero F-measure score indicates that no TPs are added to a cluster. It can be seen that the F-measure increases as the TP fraction and TP rate increase during clustering, thus signifying the presence of TPs in a cluster and highlighting the impact of the clustering approach for predicting interface residues. Because there is a skewed ratio of interface to ROS residues, a balanced evaluation is required for this imbalanced phenomenon and this is provided by the Matthews correlation coefficient (MCC) (Ezkurdia *et al*, 2009; Carugo, 2007; Baldi *et al*, 2000; Matthews, 1975) (see equation 3.5.7). The MCC score for the random prediction of TPs is 0. It can be seen that MCC increases as a cluster is grown from 0.16 – 0.24, indicating that predicting interface residues is better than random prediction.

It can be seen that prediction using the strategy of clustering conserved residues produces results better than random prediction. This is because not all interface residues are equally conserved, but those that are conserved are clustered. This observation is in agreement with previous studies. For example, Landgraf, *et al*. (2001) demonstrate the effect of clustering in predicting functional residues that are conserved. They analyzed a mixed dataset of protein-protein complexes (13 transient and 12 obligate complexes) and other protein-non-protein complexes; hence the approach they use is general-purpose in the type of functional residue predicted. A study by Guharoy and Chakrabarti (2010) used larger datasets of transient proteins in their bound forms (204 protein-protein complexes and Benchmark3.0), showing that conserved interface residues are clustered together (Hwang *et al*., 2008). They also demonstrate this for obligate interactions, using a separate dataset. In this chapter, a dataset of unbound, transient complexes was analyzed to identify clusters of conserved residues without the knowledge of bound pose conformational changes, and this is larger than the dataset

used by Landgraf, *et al*. (2001) (*i.e.* 13 transient complexes). This is important as it is necessary for later protein-protein docking of these unbound proteins using theoretical restraints to guide docking (see Chapter 7). Clustering has been applied to cluster around clusters. For example, Chung, *et al*. (2006) apply clustering in a support vector machine (see section 1.7.3.1) to eliminate lone residues from their final prediction of interface residues and also expand clustering around a minimum of three predicted interface residues (*i.e.* seed residues) to include other non-predicted residues around them. In essence, this approach uses spatial clustering to grow clusters of non-predicted residues around predicted residues. However, this strategy resulted in a marginal decrease in TP fraction and a marginal increase in TP rate after clustering when compared to before clustering.

5.8 Conclusion

Intra-species proteins evolve in the same organism and are thus subject to evolutionary pressure to maintain their interactions in the context of an important biological process. This hypothesis (actual hypothesis, H_a) was tested in this work through comparison of interface and ROS residue conservation, which is an important and debated topic. In this chapter, a dataset of intra-species interacting proteins was derived from Benchmark4.0 to examine this phenomenon (Hwang *et al*., 2010). Using intra-species interacting proteins, predictive interface features were applied to identify interface residues for training of the PROTIN_ID method (ex. solvent accessibility, conservation residue extraction, and spatial clustering), and this was examined in the context of whether interface residues are more conserved than ROS residues or not. It was observed that a 15% solvent accessibility cutoff to determine surface residues from core residues was suitable to identify interface residues, which had been identified using a distance-based definition, that were above the cutoff (*i.e.* solvent accessible). There were some interface residues that were under this cutoff with low average solvent accessibilities. As such, lowering the 15% cutoff to extract those interface residues would be risky as it would introduce conserved core residues, which may mask surface interface residues' conservation signals. This finding is important as it allows application of this cutoff (15% solvent accessibility) as the default cutoff in the PROTIN_ID method for extracting surface residues of unbound proteins for any future predictions. Ultimately, it was observed that interface residues are a minority of overall

surface residues (14.5%), allowing the comparison of interface and ROS residue conservation.

Analyzing interface conservation and ROS conservation revealed that there exists an interface conservation signal that is statistically significant, supporting the actual hypothesis (H_a). When attempting to predict interface residues, other studies divided a protein's surface into patches or clefts, and reported that conservation of interfaces is not significant by itself for prediction of interface residues (see section 5.5). It was shown that interface residues are non-homogenous in conservation and extracting the top-20 most conserved residues and clustering them using the PROTIN_ID method results in the best ranked cluster of predicted interface residues, which is better than random prediction. This was confirmed by the clustering performance measures, particularly the Matthews correlation coefficient. Compared to conservation-only extraction, the usefulness of clustering is in the reduction of ROS residues, improving prediction reliability. It was found that 7 Å clustering cutoff starting from a seed residue had the best reduction of ROS residues in the final cluster when compared to the conservation-only prediction approach. Based on these observations, the top-20 residues and 7 Å clustering cutoffs are included as final default values in the PROTIN_ID method.

It was revealed in this work that an interface conservation signal exists in the statistical analysis of the dataset, which can be exploited to retrieve interface residues via spatial clustering. Based on this, parameterization of the PROTIN_ID method using interface residue predictive features to predict such residues was achieved. This allows the use of this method in performance benchmarking with other methods (see Chapter 6). In addition, the generation of theoretical restraints of the method is useful to utilize in data-driven docking to compare its performance to *ab initio* docking (see Chapter 7). Furthermore, this allows the combination of theoretical restraints with NMR data (chemical shift perturbation data, CSP, and residual dipolar couplings, RDCs) to examine where the CSP restraints overlap with the theoretical restraints in the data they provide, and where they provide non-overlapping data to boost the interface residue recall. In combination with RDC orientational restraints, this consensus-data can be useful to examine the improvement in docking performance compared to standard CSP/RDC-driven docking (see Chapter 7).

Chapter 6

Benchmarking of the PROTIN_ID Method

6.1 Introduction

Many interface predictors have been developed with differing features and aims (see section 1.8 and tables A-1 to A-3). This chapter examines the performance of the PROTein INterface IDentification predictor (PROTIN_ID, see Chapter 4) compared to the other recently published interface residue predictors Clusters of Conserved Residues-XP (CCRXP) (Ahmad *et al*, 2010) and WHat Information does Surface Conservation Yield? (WHISCY) (de Vries *et al*, 2006). The rationale to select these two methods for comparison was as follows. CCRXP is similar in design to PROTIN_ID and it also seeks to predict conserved residue clusters that are part of protein-protein interfaces, specifically hot spot residues (Ahmad *et al*, 2010). As such, both predictors have overlapping aims and it was for this reason that CCRXP was selected for benchmarking performance comparison with PROTIN_ID. WHISCY is primarily designed in generating docking restraints to be used in the HADDOCK docking method. This aim is similar to PROTIN_ID's and it is for this reason that WHISCY has been selected for benchmarking. Metrics to measure the performance of each predictor have been implemented and the results have been analysed. In this analysis the dataset, interface definition, and performance metrics are standardized, thus allowing a meaningful comparison between the predictors. Such a comparison of PROTIN_ID with CCRXP and WHISCY is important as it highlights the strengths and weaknesses, which can assist in predictor development and their improvement (Aniba *et al*, 2010). Just comparing each predictors' reported performance in the literature with PROTIN_ID's performance can be misleading, since the predictors have been tested on different datasets, use different interface definitions, and use a limited number of performance metrics for self-evaluation (Ezkurdia *et al*, 2009). All predictors have been compared to each other using their default settings when using their own multiple sequence alignments (MSAs) (see section 3.7). Furthermore, since both PROTIN_ID and WHISCY accept external MSAs, both use the other predictor's default MSA as input to examine their performance.

6.2 Description of CCRXP and WHISCY predictors

In this section an overview of both WHISCY and CCRXP methods' protocols will be discussed. Both methods utilize structural and sequence data to derive interface residue predictive features to be used in their predictions (see section 1.7.2).

6.2.1 Overview of the WHISCY predictor

The WHISCY predictor by default uses a HSSP multiple sequence alignment (MSA) as a basis to calculate the conservation of a query protein when its PDB file is inputted in the method. A HSSP alignment is created for each protein deposited in the PDB database, using sequence homologues, and an alignment is stored in the HSSP database (Dodge *et al*., 1998). WHISCY can also accept as input a user-provided multiple sequence alignment. A score derived from a Dayhoff matrix (Dayhoff *et al*, 1978) for each surface residue of the query protein is determined through pair-wise alignment of homologous sequences to the query sequence. This is performed while taking into consideration the sequence distance (*i.e.* divergence) of the hit from the query through a maximum likelihood tool. WHISCY also weights all hit sequences to eliminate bias introduced by redundant sequences (*i.e.* identical or overrepresented in a specific species) present in a HSSP alignment prior to calculating conservation. The sequence weight is determined by ranking all hit sequences by their distance and calculating the weight of each hit sequence as half its sequence distance difference from the next sequence ranked below it (de Vries *et al*., 2006). Next, each residue's conservation score is changed to a p-value, which is divided by the residue's interface propensity (de Vries *et al*, 2006). The p-value is then changed back to a conservation score. As such, residues more likely to be part of an interface are scored higher than those that are not. Using the query protein's structure, mapped surface residues that form patches and those that are spread out on the query protein's surface have their scores weighted according to distance from each other by a smoothing function. Closely proximal residues score higher than those that are not since interface residues are more likely to be clustered together (de Vries *et al*, 2006). Finally all scored surface residues are sorted by conservation and all residues above a conservation score cut-off are predicted as

interface residues.

6.2.2 Overview of the CCRXP predictor

The CCRXP predictor requires the PDB file of the query sequence and uses the query FASTA sequence extracted from the PDB file to perform a BLAST search against the UniRef90 database (Suzek *et al*, 2007; Altschul *et al*, 1990). The top 50 sequences from the BLAST report are retrieved from UniRef90 to generate an MSA using the ClustalW MSA program (Larkin *et al*, 2007). Following this, the MSA is then scored by the Sum-of-pairs evolutionary conservation measure (see section 3.4.8) of the Scorecons server to calculate conservation (Valdar, 2002). This evolutionary conservation score takes into account sequence redundancy by weighting sequences. This is accomplished by computing the average genetic distance, which is the weighting factor, of a sequence to all other sequences (Valdar, 2002). The next step in CCRXP involves extracting all query residues above a conservation score cut-off and computing their geometric distances of their atoms from each other to generate clusters of conserved residues. All generated clusters are annotated by structural properties such as secondary structure composition, average solvent accessibility (via the DSSP program) and evolutionary conservation, and cluster size (Kabsch and Sander, 1983). Other properties like packing densities and dipole moments are also calculated for all clusters.

The final clusters generated by CCRXP are not processed further to recommend which cluster is the final prediction. However, the CCRXP authors recommend that the important features to identify in the final output of the predictor are cluster size and high cluster conservation, and in their work, they demonstrate that functionally important residues (hotspots) were located in large clusters (Ahmad *et al*, 2010). Therefore the cluster sorting heuristic implemented in PROTIN_ID (see Chapter 4) has been applied to sort CCRXP clusters according to size and then by average conservation if two clusters are equal in size; the top-ranking cluster is selected as the final prediction. Since both predictors are similar in design and outcomes, this allows an appropriate comparison to be made between interface residue prediction performance of CCRXP and PROTIN_ID.

6.3 The selected test dataset for benchmarking

In this study, the protein-protein complex datasets used by the authors of the WHISCY and CCRXP predictors were selected for performance analysis. The WHISCY dataset was derived from Benchmark versions 1.0 and 2.0 (Mintseris *et al*, 2005b; Chen *et al.*, 2003a). The CCRXP dataset was taken from a recent study (Tuncbag *et al*, 2009). Only intra-species protein complexes that consisted of two interacting chains were used for testing, following the criteria established in this work (see section 5.2). Overall, 26 proteins (13 protein complexes) were used to compare all the predictors' interface residue prediction performance.

6.4 Comparison of interface predictions of the predictors on the test dataset

The average number of interface residues (20.3), rest of surface (ROS) (107.4), and total surface residues (127.7) was determined from the proteins of the dataset (Table 6-1). The interface and ROS averages derived from all proteins of the dataset represent the random ratio of true positives (interface residues; TP) to false positives (ROS; FP) that a predictor is expected to surpass to predict better than random. The minimum (normalized) TP to FP ratio is 1.0 interface residue to 5.3 ROS residues. This indicates that if 100 surface residues were sampled randomly from the total population of surface residues, 15.9 will be interface residues while the rest are ROS residues (84.1). The TP (specificity) and FP fractions at random prediction are 15.9% and 84.1%, respectively. Naturally, the performance of any benchmarked predictor under consideration should be better than random prediction.

It is also important to consider the prediction performance depending on the objectives a predictor is aiming for (de Vries and Bonvin, 2008). On the one hand, a predictor may be interested in only predicting a few number of interface residues while minimizing false positives in the final prediction. This would ensure a final prediction enriched in true positives (TPs). However, on the other hand, a predictor may aim to predict as many interface residues as feasible, which comes at the expense of introducing more false positives in the final prediction. A predictor that fulfils its designed goals does not necessarily mean that it produces high scores for all standard benchmark performance

metrics. This is because performance metrics, while greatly informative in their own right, are designed to highlight different aspects of predictor performance. A predictor may be rated low by one performance measure (ex. TP rate) designed for over prediction of TPs, while still fulfilling its aims that it is designed for, and be rated high by another performance measure (ex. TP fraction) that focuses on the integrity of the prediction (*i.e.* minimize FPs). Consequently the statistical measure to be used for performance assessment ultimately depends on the purpose a predictor is designed for and seeks to fulfil (de Vries and Bonvin, 2008). The PROTIN_ID predictor has been developed to generate protein-protein docking restraints. As such, it has been optimized to predict sufficient interface residues, while minimizing the number of false positives. Certainly it is possible to incorporate more interface residues in the final prediction in PROTIN_ID but this introduces more false positives, which may affect the generation of reliable protein complex models when the prediction data are used as restraints in protein-protein docking. In this light, the most appropriate performance measures when comparing PROTIN_ID to the other predictors are TP fraction (specificity), FP fraction, FP rate, and to some extent TP rate (sensitivity) (see section 3.5). The TP rate measures the amount of interface residues recalled from the total pool of observed interface residues for all proteins of the dataset. Therefore with PROTIN_ID it relates to the number of interface residues that can be predicted at a minimal expense of false positives (FPs). A predictor that seeks to maximize interface residues in its final prediction will aim for a high TP rate. Both TP fraction and TP rate of a method can be combined in one measure, the F-measure. A standard measure like Matthew's correlation coefficient (MCC) while undoubtedly important, generally reports high values for predictors that favour over prediction (de Vries and Bonvin, 2008). A predictor like PROTIN_ID will achieve lower values if compared to a predictor designed for over prediction of TPs even if it produces perfect predictions (*i.e.* 100% TP fraction) while being assessed by these measures. All described measures are applied in this study for completeness (see section 3.5).

Table 6-1: The comparison of interface prediction performance using standard performance metrics is indicated for the PROTIN_ID, CCRXP, and WHISCY interface residue predictors when run at their default settings. The average interface, rest of surface (ROS), and total surface residues of the dataset and their standard deviations in parentheses are indicated.

Predictor	Interface	ROS	Surface Residues	TP count	FP count	Total count (TP+FP)	TP fraction (Specificity)	FP fraction	TP rate (Sensitivity)	FP rate	Accuracy	F-measure	MCC[a]
PROTIN_ID	**20.31** (±6.96)	**107.38** (±66.62)	**127.69** (±69.23)	4.42	5.69	10.12	0.44	0.56	0.22	0.05	0.83	0.29	0.22
CCRXP				5.23	18.35	23.58	0.22	0.78	0.26	0.17	0.74	0.24	0.08
WHISCY				5.27	7.50	12.77	0.41	0.59	0.26	0.07	0.82	0.32	0.23

[a] MCC = Matthews correlation coefficient

6.4.1 Benchmarking the predictors using the TP and FP fractions

The TP fraction results of the predictors for the entire dataset are shown in table 6-1 (see Tables A-6 to A-10 of the Appendix for individual proteins' results). The PROTIN_ID, WHISCY, and CCRXP predictors achieve 44%, 41%, and 22% TP fractions, respectively. This measure is important to compare the integrity (*i.e.* minimization of FPs) of the final predictions for all predictors, if reliability of predictions is most essential especially when applied as restraints in protein-protein docking. As expected, all predictors' TP fractions are greater than the random TP fraction (15.9%). This was also observed for the predictors' FP fractions where they were all lower than the 84.1% random FP fraction. PROTIN_ID achieves a marginally better TP fraction when compared to WHISCY. This is because the TP/FP count difference on average between the methods is minor since both are designed to generate docking restraints and aim to generate a satisfactory number of TPs in the final prediction without over prediction in order to minimize the number of FPs (de Vries *et al*, 2006).

Both methods' TP fractions are significantly better than CCRXP's TP fraction. The TP fraction of PROTIN_ID (and WHISCY) is two times better, which is relevant considering both predictors apply a similar approach for prediction. The reason why CCRXP has the lowest TP fraction is due to the average final prediction (23.58 residues), which is higher than the average final predictions for the other predictors WHISCY (12.77) and PROTIN_ID (10.12). This suggests CCRXP favours over prediction. Out of these 23.58 residues, a higher number are FP residues (18.35). This increased CCRXP's FP fraction (78%) and resulted in the lowest TP fraction. In contrast WHISCY, which generates 12.77 residues on average for a prediction, has a similar TP count (5.27) to CCRXP, but has a significantly lower FP count (7.50).

6.4.2 Benchmarking the predictors using the TP and FP rates and accuracy measures

Both CCRXP and WHISCY have the same TP rates of 26%, whereas PROTIN_ID has a lower value of 22% (see section 3.5.4). This difference in TP rates is marginal, as it is due to an approximately one interface residue difference between the average TP count

of PROTIN_ID with the other predictors' TP counts. PROTIN_ID has the lowest FP rate (5%), which is a marginal difference compared to WHISCY, because it has the lowest number of FPs from the total number of ROS residues (107.38) introduced in its final prediction. This analysis of the two measures depends on the design goals of a predictor. Both PROTIN_ID and WHISCY are designed for generating protein-protein docking restraints and aim to strike a suitable balance between TPs and FPs in their predictions. They produce similar results when comparing TP and FP rates, indicating that PROTIN_ID is as efficient and competitive as WHISCY.

The CCRXP predictor also has a TP rate of 26% that is higher than PROTIN_ID's (24%) and this is also due to an approximately one residue difference in their final TP counts. Interestingly, CCRXP's over prediction has not resulted in a TP rate that is substantially higher than the rest of the predictors but it has only resulted in the highest FP rate (17%), which is 3.4 times more than PROTIN_ID's FP rate (5%). Overall, PROTIN_ID and WHISCY achieve similar performance based on the TP and FP rate metrics since they overlap in their defined goals. In contrast, CCRXP incorporates more FPs in its prediction (*i.e.* FP rate) with a minor increase in TPs (*i.e.* TP rate), resulting in a decrease in prediction reliability.

The PROTIN_ID, WHISCY, and CCRXP predictors' interface prediction data were used to generate receiver operator characteristic (ROC) curves and their respective areas under the curves (AUC) at certain false positive rate (FPR) ranges (see table 6-2 and figure 6-1). This analysis examines the prediction performance of the predictors in an unbiased and independent manner. This is made possible as all prediction cut-offs, including the default predictor cut-off, are considered in this analysis, allowing the comparison of the performance of the predictors by their AUCs (via a standard AUC comparison statistical test and 95% CI analysis). The CCRXP interface predictor source code could not be obtained from its author. Therefore, all prediction cut-offs implemented in the CCRXP webserver version were used in this analysis. The relationship between the TP and FP rates of all predictors are compared to each other in ROC analysis (Fawcett, 2006). The objective of this comparison is to show the TP rate performance of a predictor while gauging its capability of reducing the FP rate (Figure 6-1). An ideal situation would be when a predictor generated predictions where all interface residues were correctly predicted without any ROS residues incorporated in

the prediction.

The AUC values (*i.e.* $AUC_{1.0}$) for all predictors at an FPR of 1.0 are indicated in table 6-2. WHISCY's AUC (0.6864) is marginally higher than PROTIN_ID's AUC (0.6828), however, this difference is not significant, as indicated by the p-value (0.8515). This result is supported by the 95% CI analysis, which shows that the 95% CI range extends from negative to positive values for the $\Delta_{AUC1.0}$ between these predictors (see table 6-2). Because the lower bound limit of the 95% CI is negative, this indicates that PROTIN_ID may also have a higher AUC than WHISCY, or that their AUCs' difference is zero ($\Delta_{AUC} = 0$), as it is also amongst the calculated 95% CI range. In addition, it can be seen that localized regions (FPRs 0.084-0.428, and 0.494-0.859) of the ROC curves for the PROTIN_ID and WHISCY predictors differ from one another (see figure 6-1 and table 6-2). For these localized AUC comparisons, the differences were not significant (see table 6-2). This is also indicated in the 95% CI analysis, which follows the same outcome as the 95% CI analysis at an FPR of 1.0.

In summary, the holistic and localized (FPR: 0.084-0.428) analyses for the AUCs suggest the absence of evidence to support the observation that WHISCY's marginally better interface prediction performance is significantly better than PROTIN_ID's performance on the same protein-protein complex dataset. This indicates that PROTIN_ID is as competitive and useful as the WHISCY predictor.

Compared to CCRXP, both methods have higher AUC values than CCRXP (0.5151), which is closest to random prediction (0.5), and this difference is significant (p-value <0.0001). The 95% CI analyses for both predictors compared to CCRXP have lower bound limit values greater than zero ($\Delta AUC = 0$), eliminating the likelihood that CCRXP can perform better than PROTIN_ID or WHISCY. The decrease in prediction reliability of CCRXP is caused by the increase of FPs (over prediction) in its predictions, which has resulted in similar prediction reliability to a random interface predictor.

Table 6-2: Comparison of area under the curves (AUC) for the PROTIN_ID, WHISCY, and CCRXP interface predictors. AUCs for PROTIN_ID and WHISCY are compared at three FPR ranges (1.0, 0.084-0.428, and 0.494-0.859). Both interface predictors are compared to CCRXP at $FPR_{1.0}$. The 95% Confidence Interval indicates the upper and lower bound range limits of AUC differences (Δ_{AUC}) for the FPR ranges. The AUC comparison statistical test P value indicates the probability that Δ_{AUC} is statistically significant at the 5% significance level.

Interface predictor	**PROTIN_ID**	**WHISCY**	**CCRXP**
$AUC_{1.0}$	0.6828	0.6864	0.5151
$AUC_{0.084-0.428}$	0.1655	0.1829	N/A
$AUC_{0.494-0.859}$	0.3202	0.3072	N/A
PROTIN_ID vs. WHISCY	**Δ_{AUC}**	**P value (Δ_{AUC})**	**95% Confidence Interval (Δ_{AUC})**
$FPR_{1.0}$	0.0036	0.8515	-0.0342 - 0.0414
$FPR_{0.084-0.428}$	0.0174	0.1219	-0.0047 - 0.0395
$FPR_{0.494-0.859}$	0.0130	0.4181	-0.0185 - 0.0445
PROTIN_ID vs. CCRXP ($FPR_{1.0}$)	**$\Delta_{AUC1.0}$**	**P value (Δ_{AUC})**	**95% Confidence Interval (Δ_{AUC})**
	0.1677	<0.0001[a]	0.1296 - 0.2058
WHISCY vs. CCRXP ($FPR_{1.0}$)	**$\Delta_{AUC1.0}$**	**P value (Δ_{AUC})**	**95% Confidence Interval (Δ_{AUC})**
	0.1713	<0.0001	0.1332 - 0.2094

[a] P value < 0.0001 indicates extreme significance.

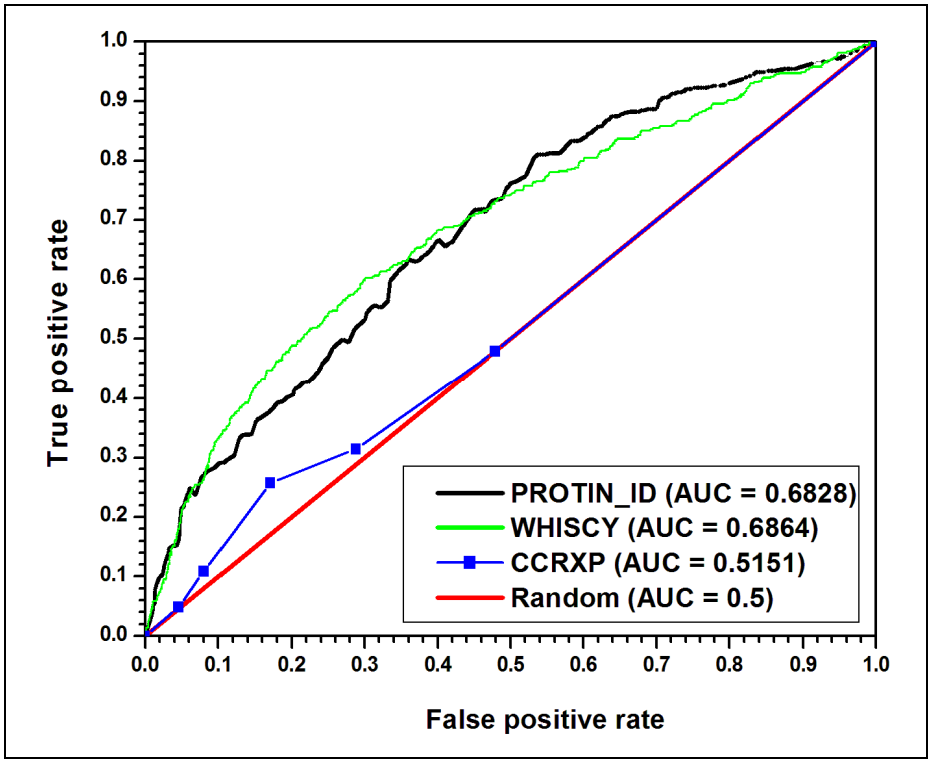

Figure 6-1: ROC curves comparing the PROTIN_ID, WHISCY, and CCRXP protein interface predictors. The $AUC_{1.0}$ values at an FPR range of 1.0 are shown for all interface predictors in parentheses. TP rate represents the recall of interface residues, while FP rate signifies the number of FP positives (ROS residues or noise) from the total number of observed ROS residues incorporated in the prediction of a predictor. An ideal predictor seeks to maximize TP rate and minimize FP rate. The intersecting red line (y = x) represents random prediction such that the TP rate equals the FP rate. Any curves above the random diagonal line represent prediction better than random, whereas curves below the line are considered predictions worse than random.

The fraction of correctly predicted TPs and TNs as described by the accuracy measure (see section 3.5.1) is highest for PROTIN_ID (83%) followed by WHISCY (82%), and CCRXP (74%). As discussed earlier (see section 5.7), the accuracy weights both correct predictions of TPs and TNs equally and, due to the skewed ratio of interface to ROS residues, results in an optimistically high value. Even so, it can be seen that CCRXP's accuracy is lower than the other predictors' accuracies due to the increased number of

FPs incorporated in its predictions. PROTIN_ID and WHISCY perform similarly due to their reduced number of FPs both fulfilling their intended design goals of generating protein-protein docking restraints.

6.4.3 Benchmarking the predictors using the F-measure and Matthews correlation coefficient

The F-measure describes the harmonic mean by taking into account the TP fraction (specificity) and TP rate (sensitivity) measures, which are weighted equally, of a predictor (see section 3.5.6). The WHISCY predictor has a higher F-measure (0.32) when compared to PROTIN_ID (0.29) and this is mainly because of the 4% difference between PROTIN_ID and WHISCY's TP rates, which is approximately one extra TP residue on average in WHISCY's TP count. This has translated in a value that marginally boosted WHISCY's F-measure compared to PROTIN_ID's F-measure. The F-measure of CCRXP is the lowest at 0.24. This is due to the predictor having the lowest TP fraction, caused by an increase of FP residues in its final prediction, which has been the factor that has reduced this measure even though the predictor has the same TP rate as WHISCY.

The final measure in this analysis is the Matthew's correlation coefficient and it provides a holistic interpretation of the performance of the predictors when compared to the F-measure since it takes into account the true negative data (*i.e.* ROS residues not predicted as interface) and other components that make up the positive and negative classes of a confusion matrix (see section 3.5.7 and Table 3-1 of Chapter 3). It can be seen that overall PROTIN_ID has a similar MCC (0.22) to WHISCY (0.23). Their MCC difference is minor as both methods seek to minimize FPs in their predictions. In contrast, CCRXP has the lowest MCC of 0.08 and is closest to random prediction (*i.e.* 0) due to a high number of FPs in its prediction.

6.5 The performance differences of the interface predictors on different complexes

The general performance of the interface predictors at the dataset level has been described. Herein, the performances of the predictors will be examined at the individual protein complex level. For example, there were certain proteins where one predictor performed well on such cases (*i.e.* >0 MCC), whereas other predictors did not (*i.e.* ≤0 MCC). There are eleven examples where this is the case (see appendix tables A6 to A-8). For instance, there are five cases where both PROTIN_ID and CCRXP extracted correctly predicted interface residues (TPs), but generated unsuccessful predictions because the correctly predicted interface residues could not be clustered (1BRS (CCRXP), 7CEI (PROTIN_ID) receptor proteins and 1FIN (CCRXP), 1E6E (CCRXP), and 2PCC (both predictors) ligand proteins). The TP residues extracted for these proteins were further apart than the cluster radial distance cut-offs used by both predictors to cluster them together in the final cluster and possibly generate successful predictions, as defined by the MCC performance metric (*i.e.* >0 MCC). An example of this is 1BRS complex's receptor protein, it was found that CCRXP extracted 7 TP residues that could not be clustered in one cluster and possibly lead to a successful interface prediction because the residues were not in close proximity to each other. A possible means to improve PROTIN_ID to deal with this scenario would be to select more than one cluster for its final prediction. To be effective and useful, such an implementation to improve PROTIN_ID would need to be tested via ROC analysis against the current top-ranking cluster implementation to ascertain if its AUC is significantly different.

Two examples (1EWY receptor and 1FQ1 ligand proteins) specifically failed for CCRXP due to poor quality MSA alignment columns, resulting in low conservation scores for MSA columns of interface residues. This prevented such TP residues from being extracted by CCRXP. In other protein cases (1E6E receptor protein (both predictors) and 1SPB ligand protein (PROTIN_ID)), it was ascertained that there were better aligned ROS residue MSA columns compared to the MSA interface residue columns. For these proteins, this resulted in high ROS residue conservation scores by PROTIN_ID and CCRXP and caused their final predictions to contain a majority of ROS residues (CCRXP - 1E6E receptor protein and PROTIN_ID - 1SPB ligand protein)

or entirely ROS residues (PROTIN_ID - 1E6E receptor protein). For the 1FIN complex's receptor protein, PROTIN_ID predicted a top ranked cluster of ROS residues. In the top-20 extracted surface residues for this example only three residues were TPs. Two TPs formed lone residue clusters and the remaining TP residue was part of a small two-residue cluster. Therefore, the few interface residues extracted by PROTIN_ID could not make a substantial impact on its prediction performance. Upon closer inspection of the 1FIN receptor protein's MSA, it was revealed that the interface residue MSA columns were composed of different amino acids, increasing the residue diversity of such MSA columns. This resulted in low conservation scores as calculated by the Jensen-Shannon divergence score applied in PROTIN_ID. Because of this, a few interface residues were part of the top-20 extracted residues for this protein case.

Regarding the WHISCY predictor, it was unsuccessful for 1BXI complex's ligand protein (-0.15 MCC). The steps in which the final score is predicted in WHISCY are more than one, unlike the other predictors where a conservation score is directly applied to score MSA columns. The effects of these steps (ex. sequence distance calculation) are difficult to disentangle. For 1BXI complex's ligand protein, it is possible that the interface residues did not predict with high WHISCY conservation scores simply because they did not appreciably differ in behaviour from the WHISCY evolutionary matrix used to predict them.

Finally, in contrast to the unsuccessful examples mentioned above, the successful performances of PROTIN_ID, WHISCY, and CCRXP on the individual proteins of the dataset (ex. 1WQ1 complex's ligand protein) were due to generally well-aligned interface residue columns in their MSAs. This resulted in high conservation scores being computed for such interface residues and led to significant (*i.e.* >0 MCC) and successful predictions for these protein cases by these predictors. There were proteins where all the predictors were unsuccessful in predicting their interfaces (*i.e.* <0 MCC). This is because conserved alternative binding sites were being predicted for these proteins.

Although PROTIN_ID and WHISCY overall perform equally well, for individual protein cases it was seen that they do perform differently in successful predictions, or even both fail to predict certain protein cases of the dataset. Where PROTIN_ID is successful and WHISCY is not (and vice versa) or where both are not successful, it would prove useful to combine them to create a compound conservation score predictor. Both these predictors apply conservation scores that differ in computation of conservation based on the commonness (or prevalence) of a residue. For example, arginine is more frequently occurring than tryptophan. As such, columns of conserved arginine or tryptophan score differently using the WHISCY conservation score or the Jensen-Shannon divergence score applied in PROTIN_ID. PROTIN_ID would score the conserved tryptophan column higher than that of the arginine column, because it has a lower background frequency (*i.e.* less common), and regards it as more strikingly conserved (see section 3.4.5). WHISCY takes the opposite route and scores the arginine column higher simply because it has more substitutable amino acid alternatives based on physico-chemically property than a tryptophan. Therefore, a highly conserved MSA column of arginine residues suggests evolutionary constraint that does not tolerate even physico-chemical substitution to similar alternative residues (Valdar, 2002). Potential cases where an interface is composed of common and uncommon residues may prove to be more successfully predicted when WHISCY and PROTIN_ID are combined.

In addition further improvements of both predictors would require addition of more interface predictive features and this would improve the prediction performance of both predictors for cases for which they already achieved success, and possibly for those where both produced unsuccessful predictions. Moreover, there are examples where both WHISCY and PROTIN_ID agree on the surface region of a protein such that their predictions overlap in terms of correctly predicted interface residues. For example, for the 1E6E ligand protein PROTIN_ID and WHISCY correctly predict 12 and 16 interface residues. The 12 residues predicted by PROTIN_ID were predicted by WHISCY. In blind protein prediction scenarios, this form of prediction strategy would be useful to assign confidence to a protein surface region (or subset region) where both predictors agree in prediction. It suggests the predictors are likely correct in predicting the true interface when they agree than scenarios where they disagree.

6.5.1 The underlying factors that caused PROTIN_ID and CCRXP performance differences

Based on the above results, it is important to determine the factor that contributed to an improved performance by PROTIN_ID compared to CCRXP. The main difference between both methods' performance lies in the number of average FP residues in their final predictions. CCRXP incorporated more FPs, which was reflected in its performance evaluation, and this was examined further. Table 6-2 shows a comparison of the total number of residues predicted as interface residues by both predictors in their final predictions for all dataset proteins. CCRXP predicts more residues (613) than PROTIN_ID (263). When broken down into their respective actual TPs and FPs, CRRXP has 21 TPs more than PROTIN_ID, but this comes at the cost of 329 additional FPs.

Table 6-2: The comparison of total TPs and FPs for all proteins of the dataset as predicted by the CCRXP and PROTIN_ID predictors when run at default settings.

Class	CCRXP	PROTIN_ID
True positives (TP)	136	115
False positives (FP)	477	148
Total	613	263

The dissimilarity in FPs is due to the implementation of the residue extraction steps in both methods prior to the final clustering step. In CCRXP, conserved residues above an absolute binary conservation score cut-off are extracted prior to the final clustering step, whereas the PROTIN_ID predictor extracted the top-N (default 20) surface residues and then clustered them. Since CCRXP extracts all residues above a binary conservation score cut-off, this introduces many ROS residues in its predictions (see section 5.7). As such, many conserved ROS residues in close proximity to each other will be extracted and clustered, increasing cluster size as a result. As an example, the receptor protein of the 1GLA complex is composed of a total of 497 residues, and the prediction generated by CCRXP resulted in 39% of these residues (*i.e.* 194) clustered in the final prediction,

which were all FP residues. This indicates the impact of the binary conservation score-based extraction step on prediction performance. Another example is the 1WQ1 complex's receptor protein for which CCRXP produced a prediction of 58 residues composed of 19 TPs and 38 FPs. In CCRXP, results like these bias the proportion of predicted residues in favour of FP residues even though TPs are predicted, resulting in decreased overall performance. For 1GLA's protein, both PROTIN_ID and WHISCY also failed to predict any TPs but, as they extract fewer surface residues, they reduce the effect of noise introduced from conserved ROS residues. As a consequence, these failed results consisting of only FPs (PROTIN_ID 4 FPs, WHISCY 7 FPs) do not have the same impact on these methods' overall performance as for CCRXP since they extract fewer FPs. The failure of these predictors on the 1GLA protein may be likely due to alternative binding sites, which exhibited a stronger conservation signal. For 1WQ1's protein, PROTIN_ID (11 TPs from 16 total residues) and WHISCY (6 TPs from 9 total residues) produced fewer TPs than CCRXP, however, their predictions are more reliable due to fewer FPs in their predictions than CCRXP.

In PROTIN_ID the stringent extraction cut-off reduces the number of ROS FPs in the overall final prediction (see section 5.7), which explains dissimilarity in performance between PROTIN_ID and CCRXP.

6.5.2 Performance analysis of WHISCY and PROTIN_ID using their respective MSAs as input for each other

The HSSP and UniRef90 alignments used by WHISCY and PROTIN_ID were reciprocated as input into each predictor. It was observed that the effect of this on prediction performance is more noticeable on PROTIN_ID than for WHISCY (table 6-3). PROTIN_ID has a reduced TP count (3.69) and increased FP count (6.5) when HSSP alignments are used even though the size of the final prediction is approximately 10 residues, which is the same as when the default UniRef90 alignments are used. This has caused a decrease in TP fraction and an increase in FP fraction compared to its default setup. Furthermore, this is evidenced in its TP and FP rates, which have decreased and increased, respectively. Finally, this has had a decreasing effect on the F- and MCC measures where both decrease, but a minor effect on the accuracy measure.

Table 6-3: The comparison of interface prediction performance is indicated for the PROTIN_ID-HSSP and WHISCY-UniRef90 interface predictors. The default MSAs for these two predictors is reciprocated as input into each other.

Method	TP count	FP count	Total count	TP frac.	FP frac.	TP rate	FP rate	Acc.[a]	F-measure	MCC[b]
PROTIN_ID	3.69	6.5	10.19	0.36	0.64	0.18	0.06	0.82	0.24	0.16
WHISCY	5.31	8.31	13.62	0.39	0.61	0.26	0.08	0.82	0.31	0.22

[a] Acc. = Accuracy

[b] MCC = Matthews correlation coefficient

The WHISCY predictor's performance was minor compared to its default setup. Only its FP count (8.31) has increased compared to its default setup's FP count; there is a negligible difference between the TP counts. Even with the increase in FP count, the impact on the other performance measures for WHISCY is minor with slight differences in TP/FP fractions and the remaining other measures have slight differences also. This suggests this drop in PROTIN_ID's performance scores stems from the alignments used. As such, the structure of HSSP alignments (ex. presence of sequence fragments) may be the differentiating factor in prediction performance for PROTIN_ID. To explore this, a closer inspection of the UniRef90 and HSSP alignments revealed that in most alignments sequence fragments, causing the presence of gapped columnar regions, and duplicate and overrepresented sequences were more apparent than in the UniRef90-based alignments. This indicates that HSSP alignments require further editing to be improved. This shift in alignment quality has had a greater impact on overall PROTIN_ID performance than on WHISCY's performance. With HSSP alignments, PROTIN_ID is not able to use its built-in features designed to improve the structure of an alignment such as the fraction of coverage measure, sequence redundancy filter, and the sequence editing heuristic prior to conservation analysis, which has impacted on its performance (see section 4.3.1). Since PROTIN_ID uses a conservation score (default is

Jensen-Shannon divergence, see section 3.4.6) that calculates evolutionary conservation based on the analysis of a column, the presence of artificial (*i.e.* not biologically important) gaps from sequence fragments introduces gap penalties during column-based conservation analysis, or in cases where a column has greater than 30% gaps, its conservation analysis is not performed since it is likely not functionally significant (Capra and Singh, 2007). This has affected the final conservation score of an alignment column(s) and influenced final predictions. Furthermore, since a conservation window is applied (see section 3.4.1) that takes into account the background conservation with respect to the current column analysed the presence of such gaps bearing no biological significance would also affect the final conservation score of the column(s) of interest in PROTIN_ID.

A comparison of total TP/FP count for PROTIN_ID when UniRef90 and HSSP alignments were used as input showed that there were 96 and 115 TPs for HSSP and UniRef90 alignments, respectively (Table 6-4). There were 19 residues fewer in the HSSP TP count and 21 extra FP residues where PROTIN_ID ‘s FP count increased from 148 to 169 FP residues when HSSP alignments were used. With regards to WHISCY, its performance did not dip significantly when using UniRef90 alignments. WHISCY was more robust in performance, which may be due to its design in scoring conservation. As mentioned previously (see section 6.2.1), WHISCY defines pair-wise alignments of hit sequences to the query sequence and calculates the conservation for residue pairs individually for each sequence, taking into account each hit’s sequence distance from the query, which is done independently from other residue pairs Also, background conservation in the form of residue smoothing is implemented after alignment analysis where it is reliant on the proximity of residues to each other in the three dimensional structure of a query protein (unlike the conservation windows). All pairwise scores for a given residue of the query sequence determined are summed to determine the final score for that query residue’s position. This is different in comparison to the columnar conservation score used in PROTIN_ID. Therefore, its means of calculating conservation may not be sensitive to alignment issues in HSSP alignments as PROTIN_ID was. This allowed WHISCY to retain robustness even with alignments where there were bad regions present (de Vries *et al*, 2006). In addition, WHISCY used all of its built-in functionality to ensure maximal analysis and optimal predictions according to its design for any alignment. These overall aspects of WHISCY allowed it

to derive relevant prediction data from both HSSP and UniRef90 MSAs and produce similar prediction performance overall, while avoiding the errors present in the HSSP alignments. This is not the case for PROTIN_ID and highlights the need to implement the use of all of its features even when an external alignment is used (future work). For example, for the ligand protein of the 1FIN complex, the difference in general in the UniRef90 and HSSP MSAs is that the latter has short sequence fragments where N or C termini are missing. PROTIN_ID generated 100% TP fraction (5 of 5 TPs) for the former MSA, whilst for the latter MSA it produced only FPs in its prediction (0 of 5 TPs), indicating the impact of alignment quality on prediction because of the presence of artificial gaps in the HSSP MSA resulting in the enforcement of gap penalties in PROTIN_ID. In contrast, WHISCY produced the same number of TPs for both MSAs, but had more FPs in its final prediction using the former MSA (12 of 25 HSSP, 12 of 31 UniRef90). Where both HSSP and UniRef90 MSAs were of similar quality, PROTIN_ID produced identical results (7 of 8 TPs) as illustrated for the 1BXI complex's ligand protein. Unlike the previous MSA, no sequence fragments were present in the same abundance, allowing PROTIN_ID to produce the same result. For the same protein, using either HSSP or UniRef90 MSAs, WHISCY did not predict a result (*i.e.* all surface residues were under WHISCY score cut-off - UniRef90) or produced only FPs (0 of 3 TPs - HSSP).

This analysis of PROTIN_ID with external alignments highlights the strengths of PROTIN_ID's current default protocol where it generates refined UniRef90 MSAs based on new features to implicitly improve MSAs. Applying the same concepts to improve alignments as implemented in the default PROTIN_ID protocol to HSSP alignments should allow PROTIN_ID to display a similar performance as its default setting that was found to perform as competitively to WHISCY. The findings indicate the importance of benchmarking in its ability to highlight areas for further improvement in future development of PROTIN_ID.

Table 6-4: The comparison of total TPs and FPs for all proteins of the dataset as predicted by PROTIN_ID using the UniRef90 (default) and HSSP alignments as input.

Class	PROTIN_ID-UniRef90	PROTIN_ID-HSSP
True positives (TP)	115	96
False positives (FP)	148	169
Total	263	265

6.6 Conclusion

The comparison of PROTIN_ID was performed with other predictors (CCRXP and WHISCY) for evaluating its performance and highlighting areas for further development. Performance comparison of the methods was conducted using a unified dataset, interface definition, and standard performance metrics. The selected methods for benchmarking overlapped with PROTIN_ID in design or aim. CCRXP was selected for comparison as it is similar in design and also predicts clusters of conserved residues. Like PROTIN_ID, WHISCY is designed to generate theoretical restraints to be used in data-driven docking and was selected for comparison to PROTIN_ID for this reason.

From a user's viewpoint, the reliability of a method's predictions in the context of the aim that a method seeks to fulfil is an important factor. This influences the choice of the method by the user. The two performance measures important in this framework are the TP fraction (specificity) and TP rate (sensitivity). According to the TP fraction metric, PROTIN_ID (0.44) performs marginally better than WHISCY (0.41), and both methods perform significantly better than CCRXP (0.22). This is because CCRXP incorporates more false positive residues (on average) in its predictions than the other methods. Based on the TP rate metric, CCRXP (0.26) and WHISCY (0.26) perform marginally better than PROTIN_ID (0.22). Other performance evaluation metrics employed in this study (ex. accuracy, F-measure, MCC, and ROC curve analysis with AUC comparisons) also indicated that the performances of PROTIN_ID and WHISCY are not significantly different. At the individual protein case level, performance differed such that one could aim for an enhanced predictive performance based on a compound conservation score approach. Furthermore, additional interface predictive features would be useful to

enhance their predictive capabilities. Finally, a useful prediction approach would be the assignment of high confidence to blind predictions when both predictors overlap in prediction. This has a potential to improve prediction success in such scenarios.

The performance evaluation metrics (including ROC curve analysis with AUC comparisons) scored CCRXP lower due to over prediction. It has the highest FP rate score (0.17) compared to WHISCY (0.07) and PROTIN_ID (0.05). This is due to CCRXP's binary conservation score cut-off extraction step, which extracted many ROS residues, causing more to be clustered and integrated in its prediction results. This resulted in PROTIN_ID and WHISCY being significantly better than CCRXP. Besides the performance difference between them, PROTIN_ID has additional user-friendly features that distinguish it from CCRXP, which are convenient to a user interested in the generation of clusters of conserved residues for later application in data-driven docking (see Chapter 4 and section 4.4).

In comparison to WHISCY, PROTIN_ID provides access to the latest sequence data from the UniRef90 database in order to construct up-to-date MSAs for predicting protein interface residues. PROTIN_ID's local sequence database is updated with each new release (biweekly) of UniRef90 (Suzek *et al*., 2007). In contrast, WHISCY relies on the HSSP database for alignments, which is not updated regularly in its entirety. Instead only HSSP files that have not been updated for more than 6 months are updated (Joosten *et al*, 2010). Therefore, HSSP alignments that are out-of-date (<6 months old) are present in the database. WHISCY in its default design utilizes available PDB structures and their HSSP MSAs to generate predictions. As a consequence, newly deposited protein structures in the PDB do not immediately have a HSSP alignment generated for it. In terms of generating theoretical restraints for docking, retrieval of the latest sequence data is essential for a user. Of course, it is possible to supply a third-party alignment to WHISCY but this will involve gathering homologous sequences manually that are needed for alignment with a query sequence of interest to create an MSA, and this MSA may require manual editing afterwards. These steps are already automated in PROTIN_ID in a user-friendly manner (see Chapter 4). With PROTIN_ID it is possible to generate an alignment for newly released structures and even homology models automatically. Homology models can be used in WHISCY, but only by manually creating custom MSAs. PROTIN_ID's default performance compared to it

using HSSP alignments revealed that PROTIN_ID's performance decreases. This is due to the poorer alignment quality of HSSP MSAs. The future development of PROTIN_ID will include processing any third-party MSAs in the same manner as it generates refined UniRef90 alignments in its current protocol. This is to avoid diminishment of prediction performance caused by any errors in third-party MSAs.

In this work, it was shown that PROTIN_ID is as competitive to WHISCY and CCRXP, and useful for generation of theoretical restraints to be applied in data-driven docking (see Chapter 7).

Chapter 7

The docking of protein-protein complexes using theoretical and experimental restraints

7.1 Introduction

This chapter examines the effect of using theoretical restraints (*i.e.* interface predictions) generated by the newly developed PROTIN_ID algorithm (see Chapter 4) to guide protein-protein docking of binary interacting proteins. HADDOCK is designed to accept theoretical and experimental restraints to guide docking of two or more proteins. This feature is ideal for docking using theoretical restraints. The performance of data-driven docking using theoretical restraints was compared to *ab initio* docking also performed in HADDOCK. The aim is to test the performance (*i.e.* number of correct models produced) of guided docking compared to *ab initio* docking when using theoretical data. The ultimate goal is using residual dipolar couplings (RDCs) and chemical shift perturbation data (CSP), both obtained from NMR experiments, in combination with theoretical data to drive docking. Both CSP and theoretical data seek to map the interface of two interacting proteins, suggesting complementary to each other when combined in a protein-protein docking context, especially if some interface residues do not display significant CSP signals (van Dijk, *et al*., 2005a). Therefore, the extent of overlap or lack thereof between these types of data was examined in the context of interface coverage. RDCs provide orientational information between two interacting proteins and their usefulness has been demonstrated in HADDOCK.

It is possible to generate biologically useful protein-protein complex models when using RDCs and CSPs in docking on their own (van Dijk, *et al*., 2005a). The addition of theoretical restraints to CSP and RDC data in docking was examined to determine if there is an improvement over CSP/RDC data docking. Different combinations of data-driven docking were performed and compared to each other to demonstrate the effectiveness of using all forms of restraints on docking performance. To the best of my knowledge, this study is the first to examine the impact on docking performance of

using of theoretical, CSP, and RDC docking restraints in combination.

7.2 Datasets used for protein-protein docking

The protein-protein docking dataset$_1$ used is derived from Benchmark 4.0 (see sections 3.2 and 3.8). A protein complex was included in the docking dataset$_1$ when both of its unbound protein chains predicted ≥ 10% TP rate (see section 3.5.4) when analysed by PROTIN_ID, irrespective of the number of false positives present in the final prediction (*i.e.* theoretical restraints). This was to ensure that a significant number of true positives composed the theoretical restraints to be used to drive docking with HADDOCK, according to the MCC measure (see section 3.5.7). A total of 24 complexes out of the intra-species dataset (61 complexes) fulfilled this criterion (≥ 10% TP rate) and were included in the docking dataset$_1$. All protein chains under < 10% TP rate are assumed to lack sufficient theoretical restraints data to generate successful results in data-driven docking. They were not included in the docking dataset$_1$ for testing. However, they are included in the final statistical analysis, which summarizes the performance for data-driven docking (*i.e.* conservative estimate taking into account these 'unsuccessful' cases) and is compared to *ab initio* docking. As such, the whole dataset of complexes taken from Benchamark4.0 (see section 3.2) is accounted for both data-driven and *ab initio* docking.

A second dataset$_2$ was also created comprising protein complexes that had RDC and CSP data associated with them. The protein complexes were identified following a keyword-based search strategy (see section 3.9). Briefly, protein complexes were identified when they were linked with terms representing RDCs and CSPs. It was found that two protein complexes of the docking dataset$_1$ (1GGR and 1J6T) satisfied this condition. A further two protein complexes not part of Benchmark 4.0 were found (1OO9 and 2L0T) that satisfied this criterion. It was ascertained that RDC data for the 2L0T protein was incorrect and hence is not included in the analysis. 2L0T was still included along with 1OO9 in the protein-protein docking dataset$_1$ as both provide useful theoretical restraints data (*i.e.* ≥ TP 10% rate for both unbound chains) to examine. This also increased the main docking dataset$_1$ to 26 complexes. Overall, dataset$_2$ is composed of three complexes for testing.

7.3 Data-driven and *ab initio* docking with HADDOCK

The theoretical restraints generated by PROTIN_ID for each unbound protein per complex were put into HADDOCK as ambiguous interaction restraints (see section 3.8). A corresponding *ab initio* run (control) for each complex using the same unbound proteins was performed using center of mass restraints. The final water refined solutions for the data-driven and *ab initio* runs were analysed and compared to each other in terms of the number of correct models and their energy ranking. Although the sampling process is set to produce 1000 models (default) in HADDOCK, it has been recently recommended that for *ab initio* runs a minimum of 5000 solutions should be produced (de Vries *et al*., 2010). However, for data-driven runs, using data generated by bioinformatics interface predictors, 1000 solutions are satisfactory (de Vries *et al*., 2010). For computational efficiency, all runs were performed using default settings. However, to test for bias, docking runs (*ab initio* and data-driven) where performed for three test cases setting the number of solutions generated at the rigid-body docking stage to 5000 (see section 7.3.3).

7.3.1 Analysis of correct models using CAPRI and Fraction of native contacts criteria

For all docking runs, two criteria were followed to evaluate the predicted protein complexes. The first is the CAPRI criteria (Critical Assessment of Prediction of Interactions), which combines (using Boolean expressions) different features of assessing a predicted model's geometric fit to a known experimental model (Lensink *et al*., 2007; Méndez *et al*., 2003; Wodak *et al*., 2003). This is achieved by determining the fraction of native contacts (F_{nat}) between complex proteins, the orientation of a ligand protein of a predicted complex with its counterpart of a known complex (L-rmsd), and the fit between the interface regions of predicted and known complexes (I-rmsd) (see section 3.8.2). Using CAPRI criteria, a correct model can be classified into one of three groupings (acceptable, medium, and high) depending on the similarity of the correctly docked model to the known experimentally determined complex. The second evaluation criterion uses solely the F_{nat} measure and has been applied in previous work (Bourquard

et al., 2011). This is a less stringent evaluation and as such all correctly docked models ($\geq$ 0.1 F_{nat}) are regarded as being near-native and grouped into one of three classes (acceptable, medium, and high) based on the number of native contacts detected (see section 3.8.2). The F_{nat} criterion is directly related to theoretical restraints used in docking, provided that the restraints have interface residues to guide docking in the right direction. Therefore, the more interface residues in the theoretical restraints, the better the chance of improving F_{nat}. In some scenarios, predicted models generated have actual native contacts, but may be classified as incorrect in CAPRI due to global orientational errors of their complex components. Such near-native models may still provide biologically valuable information of interest to the scientific community and disregarding them can be wasteful (Bourquard *et al.*, 2011). This provides a justification in using this less stringent evaluation for analysis given that a possibility exists of biologically relevant near-native complexes being produced during docking, but are classified as incorrect by CAPRI analysis. Both evaluation criteria assess a docking algorithm's sampling performance, allowing the examination of the effect of addition of theoretical restraints in HADDOCK and comparing this with unrestrained (*ab initio*) docking. The objective is to determine if there are significantly more correct models in the data-driven runs compared to *ab initio* runs (actual hypothesis, H_a). It may be possible that both run types produce no significant difference in correct models (null hypothesis, H_o).

7.3.2 General comparison of data-driven versus ab initio docking with HADDOCK

The results of data-driven versus *ab initio* docking for CAPRI evaluation are shown in table 7-1. For CAPRI analysis, there are 17 and 9 complexes from the $dataset_1$ that produced correct models for the data-driven and *ab initio* docking runs in the final 200 refined solutions, respectively. Of the 17 cases for data-driven docking, 10 protein complexes generated a statistically significant higher number of correct models, which represents 38% of the total $dataset_1$ and 59% of the 17 complexes. In contrast, only one protein complex (1Z0K) for the *ab initio* runs produced significantly better results out of the 9 cases, which is 4% of the docking $dataset_1$ and 11% of the 9 complexes. This is the only *ab initio* result that differed from the actual hypothesis where it produced a significant number of correct models than the data-driven run. In the 1Z0K data-driven

run, the restraints produced 92 models with F_{nats} greater than the CAPRI threshold (0.1 F_{nat}). These models' proteins are in the wrong orientation and have high I- and L-rmsds, which are beyond the CAPRI minimum thresholds. This prevented the data-driven run from being more successful than the *ab initio* run. However, there are more cases that are successful for the data-driven (38%) runs than *ab initio* runs (4%) of the docking dataset$_1$. The dataset$_1$ represents a sample of the whole intra-species complexes derived from Benchmark4.0 dataset. As some cases were excluded, a conservative estimate that accounts for those excluded cases that would likely produce no meaningful results (*i.e.* < 10% TP rate), and thus is a summary statistic, indicates that 16% cases are successful for data-driven runs when compared to the same cases in the *ab initio* runs. For the *ab initio* runs of dataset$_1$, they represent a sample of the 63 complexes and the results based on dataset$_1$ are extrapolated for all complexes and scaled accordingly to calculate the summary statistic, resulting in 4% cases that are successful when compared to the same cases in the data-driven runs. For the remaining cases of the whole dataset based on the combined summary statistics, both data-driven and *ab initio* docking did not produce a significant difference in correct models (80%). Overall, the success rate of data-driven docking (16%) is four times higher than *ab initio* docking (4%) based on CAPRI criteria. This supports the actual hypothesis indicating that data-driven using theoretical restraints generates significantly more correct models compared to *ab initio* docking of the entire dataset.

For dataset$_1$, only acceptable and medium quality models were generated for both types of docking; no high models were produced. Medium quality models occurred in 50% of protein complexes displaying statistical significance for data-driven docking runs, whereas the *ab initio* run which produced only 1 medium model was not statistically significant run (1SYX). For both run types the first correct model was ranked within the top 50 generated models for 7 out of 10 data-driven runs and 1 out of 9 *ab initio* runs that showed statistical significance. Comparing runs for which a significant number of correct models were generated to runs producing non-significant results, the first best-ranked correct model (lowest energy correct model) is not ranked lower for statistically significant runs than non-significant runs.

Table 7-1: CAPRI analysis of data-driven vs. *ab initio* docking. The number of correct models out of 200 is shown. Correct models are grouped as either acceptable (*) or medium (**). No high models (***) were produced. The best ranked correct model and its CAPRI grouping is indicated. The P values < 0.05 (Fisher exact test) are indicated. The TP fractions and TP rates for receptor (R) and ligand (L) proteins of the theoretical restraints used for the data-driven runs are indicated.

Complex	*Ab initio*	Data	*Ab initio*		Data		*Ab initio*	Data	P value	TP fraction		TP rate	
			**	*	**	*	Best rank	Best rank		R	L	R	L
2HRK	0	33	0	0	8	25	-	4**	<0.0001	0.55	0.88	0.43	0.47
1XD3	2	33	0	2	1	32	124*	19*	<0.0001	0.69	0.60	0.50	0.45
1O2F	1	30	0	1	0	30	15*	6*	<0.0001	0.33	0.61	0.15	0.85
1WQ1	0	24	0	0	0	24	-	6*	<0.0001	0.69	1	0.41	0.31
1OO9	0	32	0	0	2	30	-	59*	<0.0001	0.75	0.62	0.50	0.40
2L0T	1	22	0	1	7	15	115*	14*	<0.0001	0.70	0.67	0.64	0.63
3CPH	0	21	0	0	0	21	-	1*	<0.0001	0.60	0.71	0.50	0.67
2OOB	2	20	0	2	18	2	66*	79**	<0.0001	0.33	0.47	0.63	0.67
2O3B	0	9	0	0	0	9	-	32*	0.0036	0.63	0.50	0.24	0.44
1EWY	0	7	0	0	0	7	-	22*	0.0148	0.40	0.30	0.22	0.19
1Z5Y	0	4	0	0	0	4	-	18*	0.1231	0.71	0.44	0.36	0.25
2J7P	0	3	0	0	0	3	-	34*	0.1231	0.50	0.30	0.12	0.11
1JK9	0	4	0	0	0	4	-	114*	0.1231	0.76	0.67	0.57	0.22
1P9D	0	3	0	0	0	3	-	65*	0.2481	0.72	0.76	0.72	0.65
2NZ8	0	2	0	0	0	2	-	36*	0.4987	0.60	0.83	0.27	0.29
1J6T	2	1	0	2	0	1	113*	76*	1	0.43	0.64	0.18	0.69
1Z0K	8	1	0	8	0	1	13*	112*	0.0036	0.86	0.45	0.27	0.56
1GGR	1	0	0	1	0	0	41*	-	1	0.33	0.64	0.11	0.60
1SYX	2	0	1	1	0	0	1**	-	0.4987	0.33	0.45	0.17	0.64
1UR6	1	0	0	1	0	0	21*	-	1	0.40	0.40	0.27	0.21
2J0T	0	1	0	0	0	1	-	175*	1	0.78	0.62	0.33	0.57
1GRN	0	0	0	0	0	0	-	-	1	0.40	0.82	0.17	0.43
1BKD	0	0	0	0	0	0	-	-	1	0.67	1	0.19	0.19
2OT3	0	0	0	0	0	0	-	-	1	0.75	1	0.26	0.19
1FQ1	0	0	0	0	0	0	-	-	1	0.29	0.64	0.13	0.37
1F6M	0	0	0	0	0	0	-	-	1	0.83	0.60	0.23	0.41

Based on F_{nat} evaluation, the number of cases that produced correct models for both run types increases (Table 7-2). This is because of the less stringent evaluation. This is observed in the number of correct models (NOCs) produced for both types of runs. 25 out of 26 test cases produced correct models when theoretical restraints were used. For *ab initio* docking, 17 complexes produced correct models. As expected, a greater number of test cases produced correct models than with CAPRI evaluation. This is also reflected in the number of statistically significant cases, which have increased to 18 cases (69% of $dataset_1$ and 72% of the 25 complexes). No *ab initio* runs produced statistically significant runs, including 1Z0K, which was successful under CAPRI criteria. This is because the analysis to determine significance or not is a comparison of the correct models produced of both run types. Based on F_{nat}, the data-driven 1Z0K run produced 92 correct models to the 1Z0K *ab initio* run's 15, which was the reason for this. There was little increase in near-native models (7) to supplement the 8 CAPRI correct models generated by the 1Z0K *ab initio* run. This is precisely due to the absence of restraints to guide docking. Some (eight) runs that were not significant under CAPRI evaluation, however, were under F_{nat} evaluation (ex. 1GRN), signifying that the theoretical restraints used are having their intended effect in driving docking and producing a significant enrichment of near-native models. They were unsuccessful in CAPRI because of orientational errors based on L-rmsd and I-rmsd used in CAPRI.

Like with docking evaluated by CAPRI criteria, summary statistics were calculated for both run types based on the F_{nat} measure. A conservative estimate for data-driven runs indicates that 29% of cases are successful when compared to the same cases in the *ab initio* runs. The *ab initio* docking did not produce cases that were successful compared to data-driven docking on $dataset_1$. When extrapolated (based on results of $dataset_1$) to represent all 63 complexes, the summary statistic indicates 0% cases that are successful when compared to the same cases in the data-driven runs. This is because the analysis for significant results (for CAPRI or F_{nat} measures) is dependent on the NOCs produced by both run types that are compared to ascertain if significance exists. Only one case was successful under CAPRI criteria for *ab initio* docking (1ZOK, 8 correct models compared to one model in data-driven docking), but the same case did not a produce significant NOCs (15 for *ab initio* to 92 for data-driven docking). It is for this reason that 0% of *ab initio* cases produce no significant results when extrapolated to all complexes based on the F_{nat} criterion. Besides the significant cases, there are 71%

remaining cases of the whole dataset that do not produce a significant difference in correct models for both run types. Overall, the success rate of data-driven docking is better than *ab initio* docking based on the F_{nat} measure. This confirms the results shown using CAPRI criteria and supports the actual hypothesis such that theoretical data-driven docking generates significantly more correct models vs. *ab initio* docking.

More high, medium, and acceptable quality near-native models were generated for restraints-driven runs than their *ab initio* counterparts in $dataset_1$. These near-native models were found in all significant runs of data-driven docking. 61% of the significant runs produced medium quality near-native models and only 17% of those runs had high models. Ranking of the first correct near-native model was within the top 50 for all significant runs (100%), which is an improvement over ranking under CAPRI evaluation, because of the less stringent criterion.

7.3.3 Docking runs using 5000 rigid body models in HADDOCK

Docking runs (*ab initio* and data-driven) were performed for three test cases (2L0T, 1Z0K, and 1F6M) setting the number of solutions generated at the HADDOCK rigid-body docking stage to 5000 to examine if an increase in models affects the results significantly. 2L0T and 1Z0K were selected because they produced significant numbers of models (under CAPRI criteria) for data-driven and *ab initio* runs using 1000 models (default), respectively. 1F6M was neutral since both run types failed to generate any correct models for it (under CAPRI criteria). For 1F6M, both run types did indeed not produce any correct models. The data-driven 1Z0K run produced one acceptable model and this is the same result as the default run (1000 models). An interesting result was the *ab initio* 1Z0K run, which did not produce any correct models according to CAPRI (or F_{nat} criteria). This was because all the top-200 refined models were incorrect. The 2L0T run produced the opposite result in that the data-driven run produced more correct models (34), when compared to 22 correct models when default settings are run. However, the difference of 12 models is not significant (p-value = 0.1124). The *ab initio* 2L0T run produced no correct models. In summary, raising the number of rigid-body structures to 5000 (from 1000) produced similar results in data-driven docking. In general, this is also the same for the *ab initio* runs, indicating that increasing the number of models to 5000 (from 1000) for HADDOCK did not affect the results significantly.

Table 7-2: F_{nat} analysis of data-driven vs. *ab initio* docking. The number of correct models (≥ 0.1 F_{nat}) out of 200 is shown. Correct models are grouped as either acceptable (*), medium (**), and high (***). The best ranked correct model and its F_{nat} grouping is indicated. The P values < 0.05 (Fisher exact test) are indicated. The TP fractions and TP rates for receptor (R) and ligand (L) proteins of the theoretical restraints used for the data-driven runs are indicated.

Complex	*Ab initio*	Data	*Ab initio*			Data			*Ab initio*	Data	P value	TP fraction		TP rate	
			***	**	*	***	**	*	Best rank	Best rank		R	L	R	L
2HRK	1	153	0	0	1	4	21	128	180*	1*	<0.0001	0.55	0.88	0.43	0.47
1XD3	8	119	0	1	7	0	7	112	12*	2*	<0.0001	0.69	0.60	0.50	0.45
1O2F	1	72	0	0	1	0	7	65	15*	6*	<0.0001	0.33	0.61	0.15	0.85
1WQ1	0	48	0	0	0	0	0	48	-	2*	<0.0001	0.69	1	0.41	0.31
1OO9	2	136	0	0	2	0	19	117	4*	5*	<0.0001	0.75	0.62	0.50	0.40
2L0T	2	119	0	0	2	6	16	97	83*	1*	<0.0001	0.70	0.67	0.64	0.63
3CPH	0	154	0	0	0	0	11	143	-	1**	<0.0001	0.60	0.71	0.50	0.67
2OOB	4	54	0	1	3	18	1	35	14*	3*	<0.0001	0.33	0.47	0.63	0.67
2O3B	4	101	0	0	4	0	8	93	3*	1*	<0.0001	0.63	0.50	0.24	0.44
1EWY	0	12	0	0	0	0	0	12	-	19*	0.0004	0.40	0.30	0.22	0.19
1Z5Y	0	107	0	0	0	0	2	150	-	3*	<0.0001	0.71	0.44	0.36	0.25
2J7P	0	26	0	0	0	0	0	26	-	1*	<0.0001	0.50	0.30	0.12	0.11
1JK9	1	106	0	0	1	0	2	104	115*	3*	<0.0001	0.76	0.67	0.57	0.22
1P9D	0	63	0	0	0	0	3	60	-	2*	<0.0001	0.72	0.76	0.72	0.65
2NZ8	0	4	0	0	0	0	0	4	-	36*	0.1231	0.60	0.83	0.27	0.29
1J6T	8	14	0	0	8	0	0	14	29*	12*	0.2726	0.43	0.64	0.18	0.69
1Z0K	15	92	0	4	11	0	0	92	13**	1*	<0.0001	0.86	0.45	0.27	0.56
1GGR	4	10	0	1	3	0	0	10	18*	5*	0.1719	0.33	0.64	0.11	0.60
1SYX	2	4	1	1	0	0	0	4	1***	85*	0.6851	0.33	0.45	0.17	0.64
1UR6	3	8	0	0	3	0	0	8	21*	43*	0.2201	0.40	0.40	0.27	0.21
2J0T	1	149	0	0	1	0	0	149	15*	1*	<0.0001	0.78	0.62	0.33	0.57
1GRN	0	45	0	0	0	0	0	45	-	1*	<0.0001	0.40	0.82	0.17	0.43
1BKD	0	27	0	0	0	0	0	27	-	3*	<0.0001	0.67	1	0.19	0.19
2OT3	1	2	0	0	2	0	0	1	51*	3*	1	0.75	1	0.26	0.19
1FQ1	0	1	0	0	0	0	0	1	-	23*	1	0.29	0.64	0.13	0.37
1F6M	0	0	0	0	0	0	0	0	-	-	1	0.83	0.60	0.23	0.41

7.3.4 Examination of protein type and protein docking difficulty of the docking dataset

The docking dataset$_1$ was examined in the context of protein-protein interaction type and docking difficulty as categorized in Benchmark 4.0 and their relation to the number significant cases producing correct models in the data-driven runs (Hwang *et al*., 2010; Hwang *et al*., 2008; Mintseris *et al*., 2005b; Chen *et al*., 2003a). The relevant protein-protein interaction types of this study are enzymes/inhibitors or substrates, and 'other' interactions, whereas antibody/antigen interactions were excluded because of difficulty of generating theoretical restraints for them (see section 5.2). All protein complexes are classified according to docking difficulty based on structural changes during protein interaction (Hwang *et al*., 2010). The protein complexes included in the docking dataset$_1$, which were not part of the original benchmark 4.0 (see sections 3.2 and 7.2), were classified according to the same criteria applied to generate Benchmark 4.0.

Table 7-3 summarizes the percentage of statistically significant complexes for the protein interaction types under the different docking difficulty classes for the docking dataset. It can be seen that in general the percentage of statistically significant cases combined (*i.e.* Total) for both enzyme and 'other' interaction categories decreases as docking difficulty increases according to CAPRI or F_{nat} evaluation criteria. In addition, there is a clear improvement in the significance percentage in each docking difficulty category of the F_{nat} evaluation when compared to their CAPRI equivalents. For example, for the hard docking difficulty class of the 'other' proteins under CAPRI criteria, none of the protein complexes generated a statistically significant number of correct models, whereas in the F_{nat} evaluation for the same difficulty class there is a 67% improvement. A similar type of improvement is also demonstrated for the combined difficulty classes of the 'other' protein complexes when evaluated by the F_{nat} criterion (76%) compared to CAPRI criteria (35%). Because of the disparity in the total number of enzymes and other protein complexes, the percentages of statistical significance differences are misleading when compared. However, a greater improvement in the increase in significance percentage from CAPRI to F_{nat} evaluation is observed for 'other' protein complexes than for the enzyme complexes. The inclusion of F_{nat} evaluation has demonstrated that the seemingly unsuccessful (not significant) docking cases of varying docking difficulties analysed by CAPRI criteria, especially the difficult class, are indeed successful and significant in that biologically meaningful

complexes can be produced. This can been seen in the combined totals of each docking difficulty class of both protein interaction types from CAPRI to F_{nat} evaluation.

Table 7-3: The percentage of statistically significant enzyme (enzyme/inhibitor or substrate) and other complexes classed according to docking difficulty (rigid, medium, and hard) when analysed according to CAPRI and F_{nat} criteria in the docking $dataset_1$. The number of significant protein complexes out of the total complexes is indicated in parentheses. N/A (not applicable) indicates the lack of medium docking difficulty cases in the docking $dataset_1$ for the enzyme category.

Type	**Enzyme**	**Other**	**Total**	**Enzyme**	**Other**	**Total**
Difficulty	**CAPRI**			**F_{nat}**		
Rigid	40% (2/5)	60% (3/5)	50% (5/10)	60% (3/5)	100% (5/5)	80% (8/10)
Medium	N/A	33% (3/9)	33% (3/9)	N/A	67% (6/9)	67% (6/9)
Hard	50% (2/4)	0% (0/3)	29% (2/7)	50% (2/4)	67% (2/3)	57% (4/7)
Total	44% (4/9)	35% (6/17)	38% (10/26)	56% (5/9)	76% (13/17)	69% (18/26)

7.3.5 Comparison between protein docking's production of correct models and theoretical restraints prediction quality

The number of correct models (NOCs) produced by docking based on CAPRI and F_{nat} evaluation for each protein complex was compared to interface prediction quality as measured by TP fraction and TP rates (see sections 3.5.2 and 3.5.4) to determine their relationship with each other (see tables 7-1 and 7-2). Each protein complex had a single TP fraction and TP rate calculated by combining the data for its receptor and ligand protein components. The objective is to determine the extant of correlation existing between the NOCs and either protein complex TP fraction or TP rate to determine the magnitude of positive, negative, or lack of correlation between them at the protein complex level. Spearman's rank correlation, which is a non-parametric analysis, was applied since it assumes no underlying probability distribution for the data (Motulsky, 2007). This analysis determines whether an increase in TP fraction or TP rate of the

PROTIN_ID theoretical restraints used for docking influences the NOCs produced during docking. It may be that both interface prediction quality measures have no significant relationship with the NOCs, which is the null hypothesis (H_o). Table 7-4 summarizes the results of this analysis.

Table 7-4: The correlation between protein complex TP fraction and TP rate each compared with the number of correct models (NOCs) generated for the protein complexes of the dataset$_1$. The 95% Confidence Interval (CI) shows the upper and lower bound range limits of the Spearman's rank correlation coefficient (Spearman r). P values indicate the probabilities of the likelihood of a positive correlation relationship with the number of correct models produced and either TP fraction or TP rate.

Evaluation type	TP fraction and NOCs			TP rate and NOCs		
	Spearman r	95% CI	P value	Spearman r	95% CI	P value
CAPRI	0.16	-0.26 – 0.52	0.45	0.44	0.05 – 0.71	0.02[a]
F_{nat}	0.32	-0.09 – 0.64	0.11	0.47	0.09 – 0.73	0.01[a]

[a] P value (< 0.05) indicates statistical significance.

It can be seen that for CAPRI and F_{nat} evaluation, the correlation coefficients produced for both TP fraction and rate association with the NOCs are positive values. The correlation coefficient values are higher for the TP rate than for the TP fraction in CAPRI and F_{nat} evaluation. To examine the significance of all the correlation coefficient values, the 95% confidence interval (CI), which quantifies the precision of the determined correlation coefficient, giving the 95% probability that a correlation coefficient value is within the calculated upper and lower bound limits, was calculated for both TP fraction and TP rate. Furthermore, the probability that NOCs and TP fraction or TP rates do not have any correlation and that positive correlation coefficient values observed occurred by chance is calculated by p values. Both CI and p value analyses are complementary. For TP fraction of CAPRI and F_{nat} evaluations, the CI ranges are from negative to positive values. Because the lower bound limits of the CI are negative values, this suggests that their association does not indicate a tendency of the NOCs and TP fraction to increase together. Also, the 0 correlation coefficient value, which represents no correlation, is within the calculated range of CI values. In addition,

both p values (CAPRI p = 0.45, and F_{nat} p = 0.11) support the CI data and indicate lack of statistical significance, suggesting the absence of evidence that the positive correlations are genuine, but due to chance. On the other hand, the TP rates in CAPRI and F_{nat} evaluation both have upper and lower bound limits of the CI that are positive values, excluding the possibility of no correlation (*i.e.* Spearman r = 0) and negative correlation. The CI analysis suggests a tendency for NOCs and TP rate to increase together and is unlikely due to chance, which is also supported by p values that are statistically significant. From the data, it appears TP rate has a more positive correlation than TP fraction on protein docking's ability to generate NOCs. This suggests that theoretical restraints used in docking are more influential if they have a high TP rate. However, simply taking all the surface residues of two unbound proteins as theoretical restraints in docking would fulfil the criterion of having a high TP rate by generating 100% TP rates. But such docking runs would sample the entire surfaces of both input proteins instead of localized surface regions thereby generating similar results as *ab initio* runs. It is possible that the generation of NOCs may more likely be a qualified combination of TP rate with a reduction in false positives as indicated by the TP fractions (table 7-1).

In general, there are 18 protein complexes that have statistically significant NOCs and high TP rates and fractions when both statistically significant results of CAPRI and F_{nat} evaluation are pooled. For example, the protein complexes 1OO9 and 1P9D both have high TP rates and fractions for their unbound protein complex receptor (>50% TP rate and >70% TP fraction) and ligand (>40% TP rate and >60% TP fraction) proteins. 1OO9 has high NOCs for both CAPRI (32) and F_{nat} (136) evaluation while 1P9D has high NOCs (63) for the F_{nat} criterion. Figure 7-1 shows the L-rmsd, I-rmsd, and F_{nat} for both 1OO9 and 1P9D protein complexes when data-driven and *ab initio* docking are compared. It can be seen that the effect of theoretical restraints localizes protein docking sampling to specific binding poses as shown by the clouds of models produced for F_{nat}, L-rmsd, and I-rmsd. In general, more than one distinct cloud is formed in data-driven docking and this may be due to the manipulation of the restraints in HADDOCK where 50% of restraints are randomly removed during the initial rigid body docking stage in HADDOCK. This heuristic is performed to remove the presence of false positive residues. It may not always succeed in doing this and may remove more true positives, which may have contributed to the cloud of models (ex. L-rmsd and I-rmsd)

beyond the acceptable thresholds defined for correct models (see section 3.8). Also theoretical restraints (as ambiguous interaction restraints) do not provide orientational information of two input proteins, but localized sampling information (Dominguez *et al*., 2003).

Consequently, the formation of different clouds is likely due to random restraint removal and orientational errors where one protein is rotated incorrectly with respect to its partner. Such orientational errors are captured by L-rmsd and I-rmsd measures in CAPRI evaluation. In comparison to data-driven runs, *ab initio* docking does not have any restrictions on docking sampling and, as can be seen, models produced are more dispersed as a consequence. For the F_{nat}, most *ab initio* models are localized at 0 F_{nat}, reflecting the greater surface area beyond the true interface of both input proteins that HADDOCK is sampling. This is also evident in both rmsd evaluations with a formation of a dispersed cloud, indicating as with F_{nat}, the non-specific effect of *ab initio* sampling (Figure 7-1).

In terms of energy ranking, most data-driven models are lower in energy than the ones produced by *ab initio* docking. However, the HADDOCK score is not able to discriminate efficiently in ranking between correct and incorrect models produced through data-driven docking. At least one correct model is obtained in the top 200 final refined (out of 1000 rigid body models) HADDOCK solutions in the majority of cases, indicating the effectiveness of the HADDOCK score in selecting correct models from many incorrect ones. Nevertheless, the ideal goal would be to have such models consistently ranked as the best (*i.e.* lowest energy) such that the better quality correct models (ex. high or medium) would be ranked higher than those of lesser quality (ex. acceptable) and then followed by incorrect models. This has not been consistent in recent CAPRI evaluation of docking programs and scoring functions (Lensink *et al*. 2010; Lensink *et al*., 2007). Reliable prediction of binding free energies using docking scoring functions is at its nascent stages (Fleishman *et al*., 2011; Melquiond *et al*., 2011; Kastritis and Bonvin, 2010). However, latest developments such as a recently released binding affinity benchmark, or the application of a docking (and by extension its scoring function) to distinguish true interacting proteins from non-interacting proteins will provide insights into the development and application of improved scoring functions (Kastritis *et al*., 2011; Wass *et al*., 2011b).

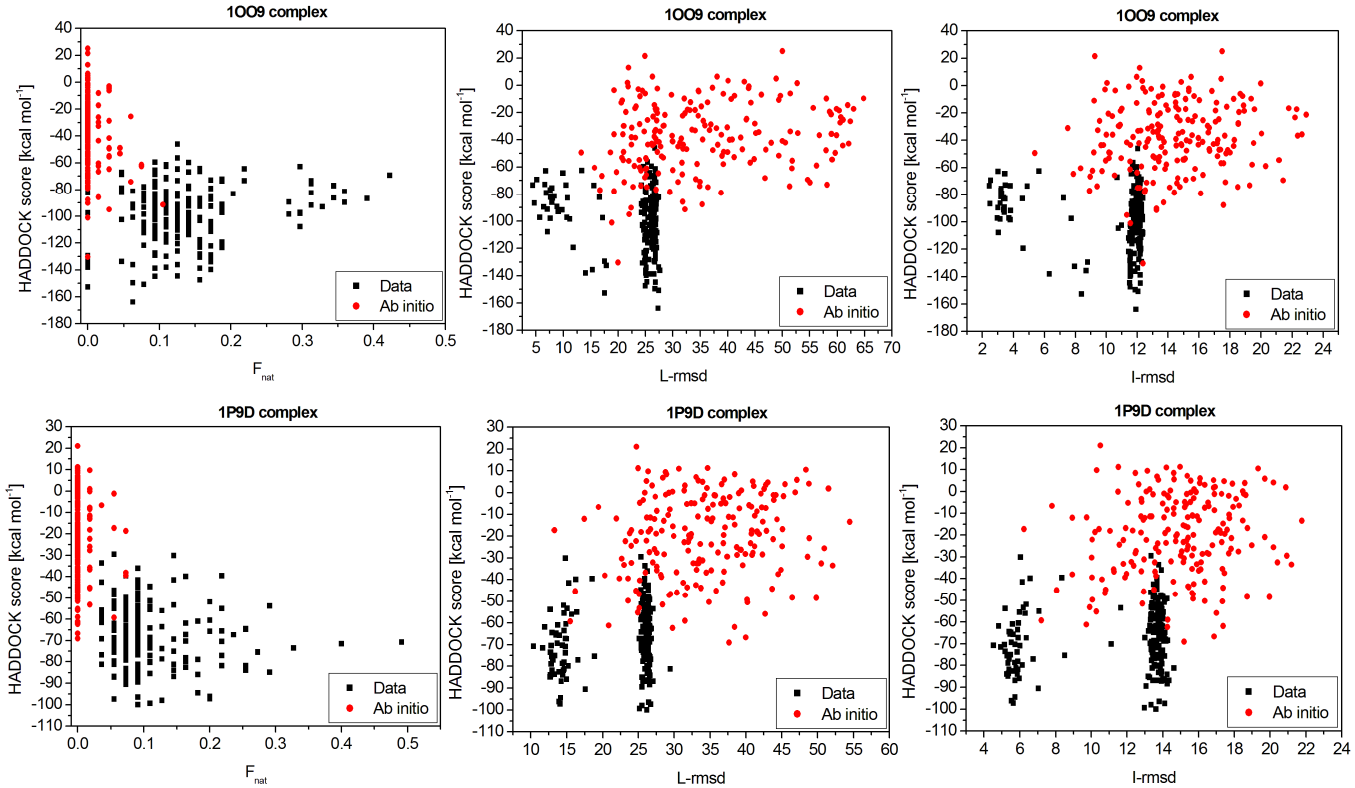

Figure 7-1: HADDOCK score versus F_{nat}, L-rmsd, and I-rmsd evaluators for data-driven (black) and *ab initio* (red) docked models generated for the 1OO9 and 1P9D complexes. All models are compared to experimentally solved 1OO9 and 1P9D complexes to derive the rmsds and F_{nat}. The combination of all three evaluators is what defines CAPRI criteria, whereas the evaluation of the presence of near-native models utilizes the F_{nat} criterion solely. It can be seen that the use of theoretical data to drive docking generates correct binding poses in comparison to *ab initio* docking.

It can be seen in Figure 7-1 that some incorrect models are ranked better (*i.e.* lower energy) than correct models. Even amongst the pool of correct models, their HADDOCK score ranking is varied. For instance, in the 1OO9 docking results the best ranking correct models (CAPRI or F_{nat} evaluation) are not the best in terms of binding pose quality (see tables 7-1 and 7-2). In Figure 7-2(A, B) the best-ranked 1OO9 correct (CAPRI) and *ab initio* models' binding poses in relation to the known experimentally determined 1OO9 complex are compared. The F_{nat}, L-rmsd, and I-rmsd values are 0.29, 7.1 Å, and 3.0 Å, respectively for the data-driven model, which is classified as acceptable under CAPRI criteria. The *ab initio* model in comparison has F_{nat}, L-rmsd, and I-rmsd values violating the evaluation criteria thresholds. This is illustrated by the ligand interface of the *ab initio* model being in the opposite direction of the true binding interface of the reference protein complex. Some near-native models classified as incorrect by CAPRI due to orientation errors in L-rmsd and I-rmsd may still provide biologically useful information. Figure 7-2(C) displays an example of a superimposed near-native model that does not satisfy CAPRI criteria (*i.e.* incorrect) compared with the known 1P9D protein complex and the best ranked *ab initio* model (Figure 7-2D). The near-native model produced by data-driven docking has 0.25 F_{nat}, whereas the *ab initio* model has none. It is considered incorrect due to orientational errors as measured by L-rmsd and I-rmsd. In spite of this, the number of correct native contacts (*i.e.* interface resides with correct intermolecular interactions) predicted in this model provides a good starting point for further investigative research.

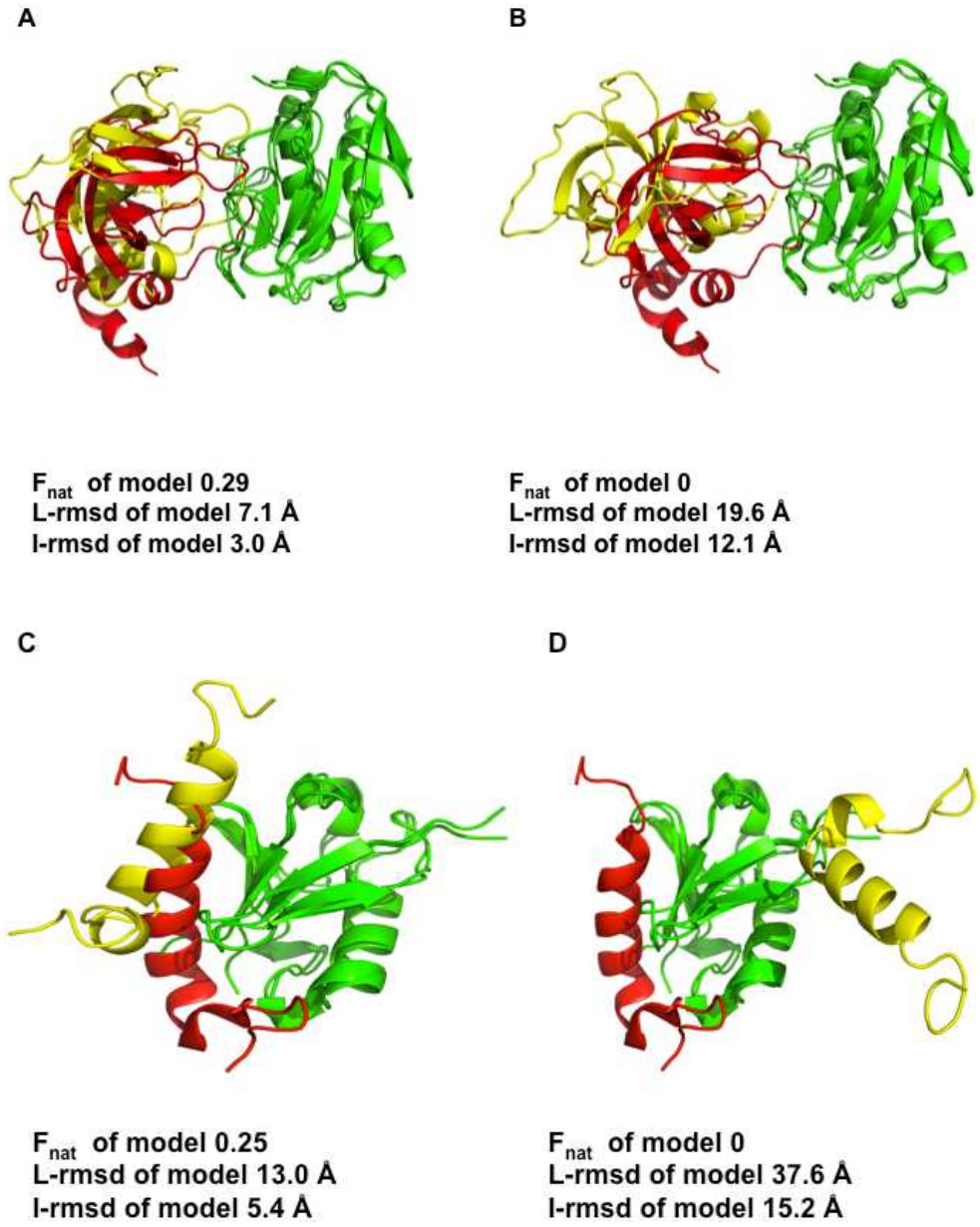

Figure 7-2: The comparison of the data-driven and *ab initio* models of 1OO9 (A & B) and 1P9D (C & D) with their experimentally determined protein complexes. The receptor proteins (green) of all models are superimposed on the receptor (green) of the experimental complex to illustrate the binding poses of the ligand proteins (yellow) of all models with respect to the ligand protein of the experimental complex (red). **A)** Best ranked data-driven docked model. **B)** Best ranked *ab initio* docked model. **C)** Near-native docked model. **D)** *ab initio* docked model.

7.3.6 The failure of certain protein docking cases

Only 8 protein complexes did not produce significant NOC for data-driven docking in relation to *ab initio* docking according to F_{nat} criteria. There are possible reasons for their failure. Firstly, there are examples for which a substantial number of near-native complexes were produced, but were not regarded as statistically significant due to a high number of near-natives complexes produced in their *ab initio* comparison runs that undercut their significance (ex. 1J6T, 1GGR, and 1UR6). Secondly, other protein complexes produced a few near-native models for both run types (2NZ8, 1SYX, 2OT3 and 1FQ1) and hence were not statistically significant. Thirdly, only 1F6M produced no near-native models for data-driven and *ab initio* runs. A closer examination of the theoretical restraints data used for docking revealed that there are native contacts correctly predicted for both unbound receptor and ligand proteins. It was observed that 136 of 200 models were produced with F_{nat} values spanning 0.017 – 0.089. To determine if acceptable or better models were not selected based on low energy scores in HADDOCK's water refinement stage, the rigid-body stage's models (1000) were examined. The same F_{nat} value range was found for 590 of 1000 models, excluding this possibility. The total possible true interfacial residue-residue pairs F_{nat} that can occur based on the theoretical restraints was determined to be 0.164, which is above the F_{nat} minimum threshold. In theory HADDOCK had minimum data to produce acceptable models based solely on the F_{nat} criterion, its failure is likely due to the small number of residue pairs derived from the theoretical restraints data combined with HADDOCK randomly removing some of those restraints. This may have resulted in the restraints not being sufficient for HADDOCK to effectively sample and predict correct models for the 1F6M complex (see Figure 7-3).

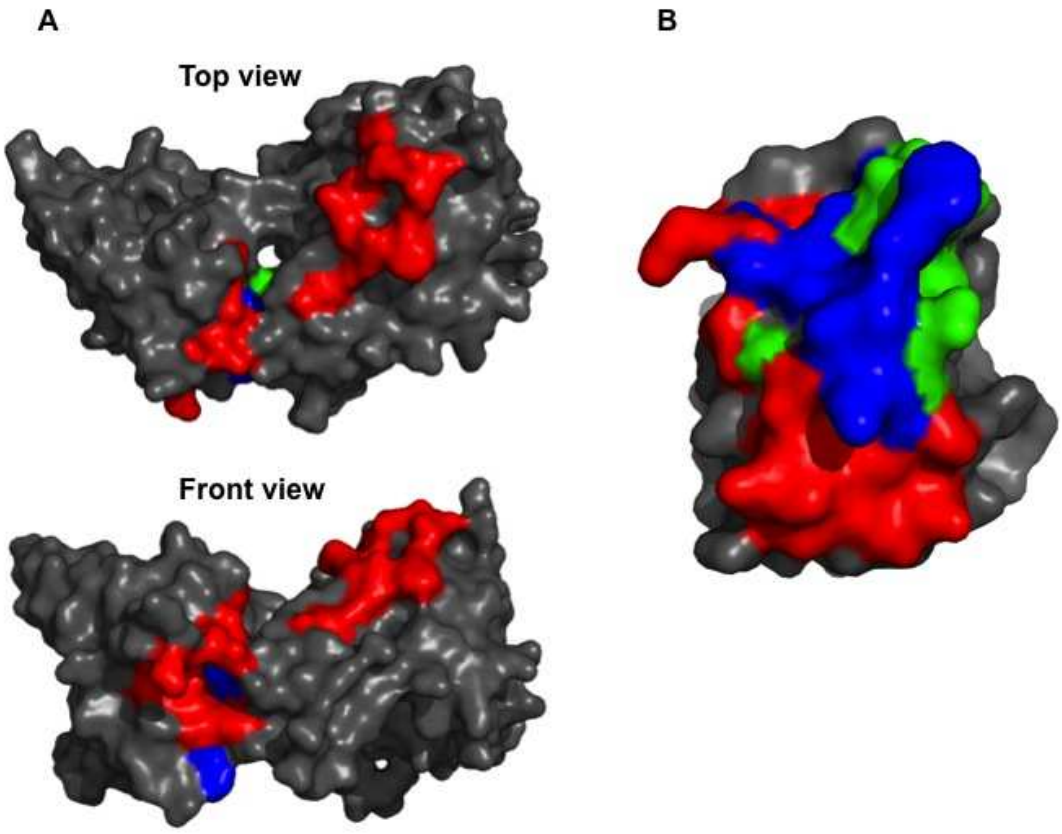

Figure 7-3: Theoretical data mapped onto the surface of 1F6M complex protein and compared to the real 1F6M interface. The actual interface (red), true positive interface predictions (blue), and false positive interface predictions (green) are shown. HADDOCK was unsuccessful in predicting correct models for this complex because of a small number of possible native contact residue pairs (F_{nat} 0.164) derived from the prediction data combined with HADDOCK randomly removing some of those restraints. **A)** Thioredoxin reductase (receptor protein; PDB: 1CL0). **B)** Thioredoxin-1 (ligand protein; PDB: 2TIR).

7.4 The use of experimental and theoretical data to improve protein-protein docking performance

Theoretical data have been applied in protein-protein docking as a means to improve docking sampling and scoring of docked protein complex models using front- or back-end approaches (see section 1.10). In this work, systematic front-end application of conservation data to drive docking has indicated that in general data-driven docking using theoretical restraints outperforms *ab initio* runs, generating statistically significant results. Likewise, the use of prediction data to re-rank docked models has demonstrated encouraging results in recent work (Huang and Schroeder, 2008). In addition, in a related approach re-ranking with structurally derived 'conservation' data of protein complexes of interest originating from their known homologous protein complex counterparts has been shown to be successful when compared to a docking approach's ranking of models (Xue *et al*., 2011b). Relatedly using experimental data such as CSP and RDC data to drive protein docking for front-end docking application has been demonstrated successfully (van Dijk, *et al*., 2005a; Clore and Schwieters, 2003; Dominguez *et al*., 2003; McCoy and Wyss, 2002). Moreover, back-end application using CSP data solely or in combination with RDC data has achieved considerably effective ranking of docked models (Stratmann *et al*., 2011; Montalvao *et al*., 2008; Dobrodumov and Groenborn, 2003; Morelli *et al*., 2001). It is possible that theoretical and CSP data complement one another by mapping potential interface regions of two interacting proteins to aid protein docking algorithms in predicting intermolecular interactions of two partner proteins. Coupling their use with RDC orientational restraint data in protein docking simulations also allows the relative orientation of two protein partners with respect to each other to be enforced, decreasing orientational errors that may otherwise arise in there absence. Ultimately, the examination of whether it is possible to improve docking performance using consensus-data derived from experimental and theoretical sources to drive docking when compared to CSP/RDC-driven docking is the target. The case studies described below represent the examination of consensus-data derived from the merger of CSP, RDC, and theoretical data in front-end application to protein docking.

7.4.1 Application of RDCs, CSPs, and theoretical restraints for docking of the 1OO9 protein complex

1OO9 is a complex formed by the interaction of matrix metalloproteinase (MMP-3) and its partner known as tissue inhibitor of metalloproteinase (TIMP-1). MMP-3 is a zinc endopeptidase involved in extracellular matrix protein breakdown during embryogenesis and tissue regeneration (Arumugam and Van Doren, 2003a). MMP-3 is controlled by TIMP-1 and the interruption of this regulation results in, as an example, arthritis and cancer (Gomis-Rüth *et al*., 1997).

CSP data obtained for both proteins were used as ambiguous interaction restraints (AIRs) for docking along with RDC data as described in section 3.9.1 (Arumugam *et al*., 2003b; Arumugam *et al*., 1998). The mapping of the CSP and theoretical data on the MMP-3 and TIMP-1 protein surfaces is indicated in Figure 7-4(A, B). Furthermore, the TP fractions and TP rates for CSP, theoretical, and CSP/theoretical (*i.e.* consensus data) restraints for all docking runs performed are shown in Table 7-5. Treating the CSP data using standard performance metrics to evaluate interface predictors, allows a comparison between CSP and theoretical data to gauge their contributions in terms of true positive (*i.e.* interface residue) recall and precision. It must be highlighted that passive residues used with the CSP restraints were considered as CSP data even though they have insignificant CSPs but are in close proximity to active residues that have been identified as having significant CSPs (Dominguez *et al*., 2003). This is because passive residues are part of the CSP docking restraints and as such the TP fraction and TP rate analysis of the CSP data takes into account passive residues. In Figure 7-4 (A, B) it can be seen that there is overlap between the theoretical and CSP data for both proteins of the complex. This overlap includes active and passive residues of the CSP data. This indicates that experimentally identified (active) residues with significant CSPs are conserved. In addition, residues (passive) with less significant CSPs but are in close proximity to active residues are also conserved. Therefore, the knowledge of passive CSP residues' conservation from theoretical data allows their "promotion" to active residues, which was done in the consensus-data driven runs (section 3.9.1).

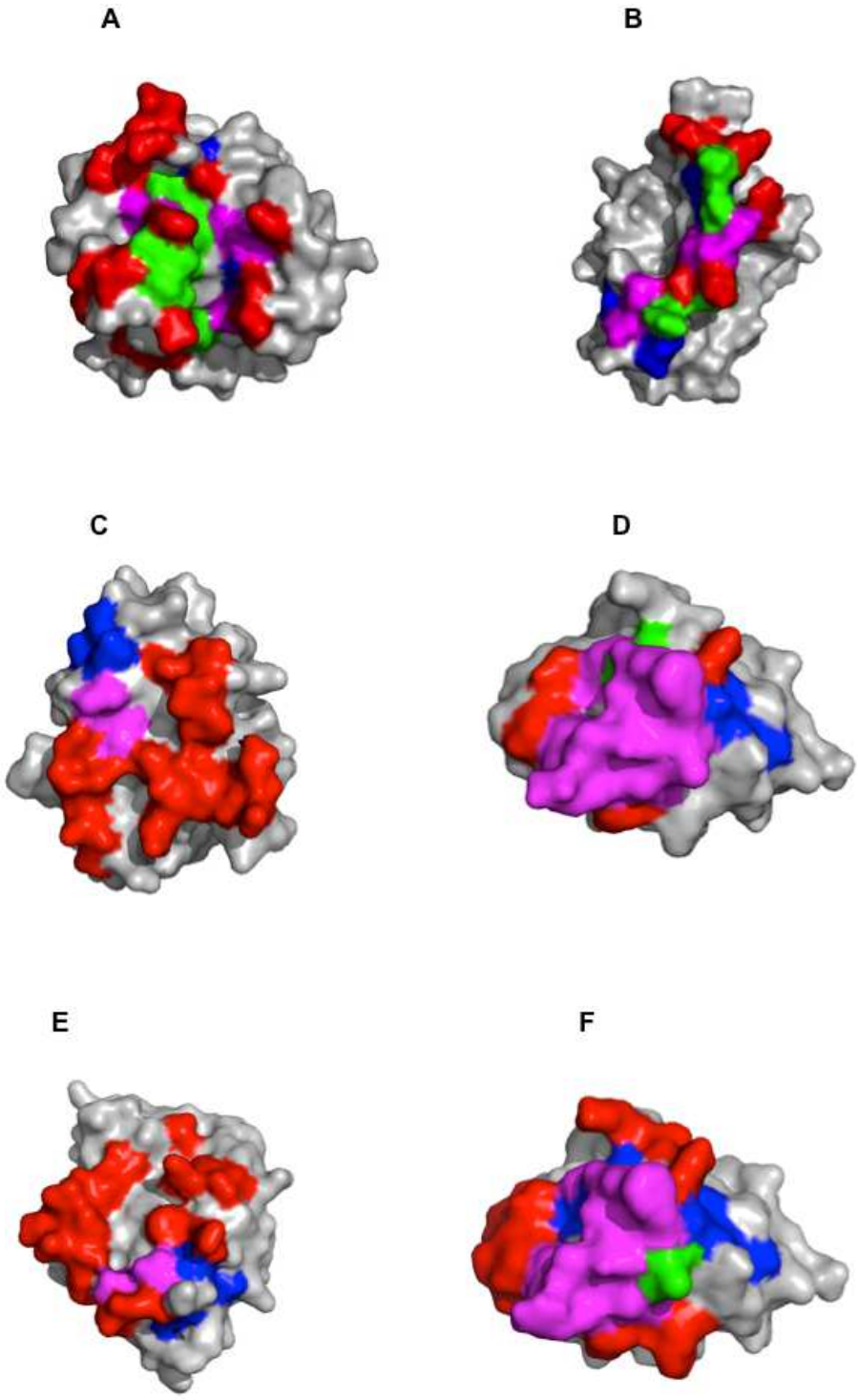

Figure 7-4: CSP and theoretical data mapped onto the surfaces of 1OO9 (A & B), 1J6T (C & D), and 1GGR (E & F) complex proteins. Non-overlapping CSP (active and passive) and theoretical data are coloured red and blue, respectively. Theoretical data overlapping with CSP-active and CSP-passive are coloured magenta and green, respectively. **A)** MMP-3 **B)** TIMP-1 **C)** $E2A^{Mtl}$ **D)** HPr **E)** $E2A^{Glc}$ **F)** HPr

Table 7-5: Comparison of CAPRI and F_{nat} analysis of consensus-data (ALL = CSPs, RDCs, and theoretical restraints) and experimental data-driven docking (CSPs/RDCs restraints) for 1OO9, 1J6T, and 1GGR complexes. Various combinations of the theoretical and experimental data data-driven runs are included along with the *ab initio* runs for comparison. The number of correct models (NOCs) out of 200 is shown. They are grouped as acceptable (*), medium (**), or high (***). The best ranked correct model and its CAPRI or F_{nat} grouping is indicated. The statistical significance is indicated by bold NOCs values for data-driven runs vs. *ab initio* runs. An italicized 'ALL' docking run value indicates statistical significance when compared to CSPs/RDCs docking run. The TP fractions and TP rates of receptor (R) and ligand (L) proteins derived from the restraints used to drive docking are indicated for all runs.

Complex	CAPRI [b]					F_{nat}					TP fraction		TP rate	
	NOCs	***	**	*	Best rank	NOCs	***	**	*	Best rank	R	L	R	L
1OO9														
Ab initio	0	0	0	0	-	2	0	0	2	4*	-	-	-	-
TH	**32** [a]	0	2	30	59*	**136**	0	19	117	5*	0.75	0.62	0.50	0.40
CSPs	**47**	0	0	47	16*	**106**	0	5	101	4*	0.54	0.71	0.61	0.60
CSPs/TH	**35**	0	0	35	4*	**111**	0	9	102	4*	0.54	0.67	0.65	0.70
CSPs/RDCs	**116**	0	4	112	1*	**140**	0	43	97	1*	0.54	0.71	0.61	0.60
ALL	***157***	0	7	150	2*	***171***	2	87	82	2*	0.54	0.67	0.65	0.70
1J6T														
Ab initio	2	0	0	2	113*	8	0	0	8	29*	-	-	-	-
TH	1	0	0	1	76*	14	0	0	14	12*	0.43	0.64	0.18	0.69
CSPs	7	0	2	5	5**	**142**	2	4	136	1*	0.79	0.67	0.71	0.77
CSPs/TH	**35**	0	25	10	3**	**162**	25	9	128	2*	0.67	0.61	0.76	0.85
CSPs/RDCs	**137**	0	87	50	11**	**143**	22	74	47	5*	0.79	0.67	0.71	0.77
ALL	***168***	0	83	85	1**	***176***	44	108	24	1***	0.67	0.61	0.76	0.85
1GGR														
Ab initio	1	0	0	1	41*	4	0	1	3	18*	-	-	-	-
TH	0	0	0	0	-	10	0	0	10	5*	0.33	0.64	0.11	0.60
CSPs	**69**	0	48	21	1**	**161**	36	29	96	1***	0.79	0.68	0.65	0.87
CSPs/TH	**30**	0	21	9	1**	**158**	18	4	136	1***	0.65	0.58	0.65	0.93
CSPs/RDCs	**183**	11	99	73	1**	**185**	77	76	32	1***	0.79	0.68	0.65	0.87
ALL	**180**	9	49	122	1***	**180**	51	91	38	1***	0.65	0.58	0.65	0.93

[a] P value < 0.05 indicates statistical significance (Fisher exact test).

[b] No *** models were found according to CAPRI criteria.

For MMP3 and TIMP-1, there are non-overlapping theoretical restraint residues that are part of the interface, indicating that theoretical restraints can provide additional true positives when included with CSP data in consensus-data docking. This is important to demonstrate because if all theoretical restraints overlap with CSP data and/or provide only non-overlapping false positive data they do not offer additionally relevant data for the protein docking sampling process. Consequently, the combination of theoretical restraints and CSP data results in an increase in TP rate to 0.70 and 0.65 for both ligand (TIMP-1) and receptor (MMP-3) proteins, respectively, suggesting their practical use with CSP data in consensus-data driven docking. This is a significant achievement considering that the CSP data have higher TP rates than the theoretical data, meaning that the experimental data identified higher numbers of interface residues than the theoretical data. The TP fractions for both proteins are decreased, which is not surprising, because of inclusion of false positives (*i.e.* ROS residues) caused from combining the two sets of data.

As discussed earlier (section 7.3.5), TP rates have a higher correlation with NOCs than TP fractions provided that the TP fractions are not too low. In this case, the TP fractions of the consensus data are over 50% for ligand and receptor consensus restraints, which is high specificity. All data-driven runs produced a statistically significant number of correct models in comparison to the *ab initio* run for CAPRI and F_{nat} evaluation. Docking using theoretical (TH), CSP, and combining CSP/TH data (*i.e.* consensus-data) was implemented to examine docking performance minus orientational restraints in comparison with the runs using them (CSP/RDC and consensus-data/RDC). The NOCs are generally similar for the TH, CSP, and CSP/TH runs, according to CAPRI evaluation with differences that are not statistically significant. However, the use of RDCs increased the NOCs produced with the consensus-data/RDC (‘ALL’) run having a statistically significant result compared to the standard CSP/RDC run (Table 7-5). Furthermore, more models are converged and populate the first cluster for RDC incorporating runs than for the rest of the runs that do not use RDC data (Figure 7-5). The mean I-rmsd of the first cluster of the ‘ALL’ run is 3.4 Å (STDV of 0.85 Å), which is lower than the CSP/RDC run’s first cluster I-rmsd mean of 4.2 Å (STDV of 1.43 Å). The remaining runs have similar means for their first clusters to the CSP/RDC run (ex. CSP and CSP/TH runs) or the ‘ALL’ run (TH run). However, these runs have smaller clusters containing correct models. There are distant clusters with high I-rmsds and low

energies, and this is especially notable for data-driven runs that do not use RDC data. The reason these distant clusters have similar (or better) HADDOCK scores for some of their models than correct models of the first clusters is due to several charged patches on the MMP-3 and TIMP-1 proteins, preventing the HADDOCK score from discriminating between correct and incorrect models that may be different by, for example, 180° rotations of their ligand proteins from the correct models' ligand proteins. And this may explain the absence of a clear overall positive linear correlation between energy and I-rmsd of correct and incorrect models in all 1OO9 runs.

Chiefly, acceptable models are produced for all runs. The 'ALL' run has the highest number of medium models (7). The ranking of the first correct model for CAPRI evaluation improves when RDC data is used. Consensus data docking only generated one correct model in the top ten ranked docked models, whereas adding RDCs with consensus data produced 8 correct models in the top ten models. This is higher than standard CSP/RDC, which generated 6 correct models in the top ten models. Docking using TH or CSP restraints do not have ranked models in the top-ten HADDOCK score-ranked models.

As expected, the RDC incorporating CSP data-driven run shows that the application of orientational restraints for both 1OO9 partners clearly influences the NOCs generated in docking and ranking of correct models. This is particularly significant as CAPRI criteria take into account the L and I-rmsd for model assessment of orientational errors, showing that more models are satisfying the criteria hence the higher NOCs. The effect of using TH data with CSP and RDC restraints in providing additional true positives is pronounced as it boosts the NOCs and improves ranking of correct models even though multiple charged patches are present on both ligand and receptor proteins.

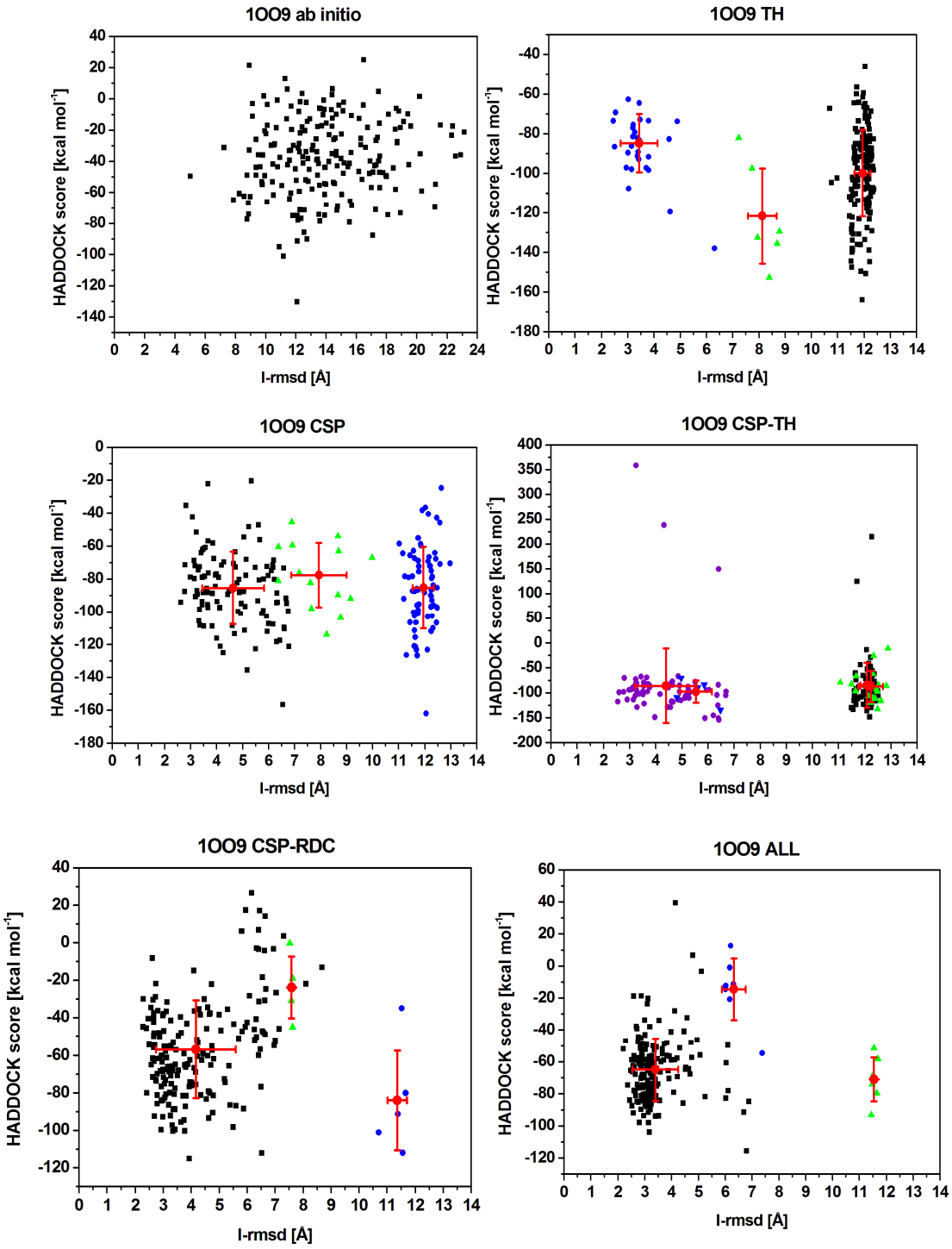

Figure 7-5: HADDOCK score versus I-rmsd for consensus-data (ALL = CSPs, RDCs, and theoretical restraints) and experimental data-driven docking (CSPs/RDCs restraints) for the 1OO9 complex. Various combinations of the theoretical and experimental data-driven runs, including *ab initio* docking are included for comparison. All models are compared to the experimentally solved protein complex to derive the I-rmsd. Structural clusters are coloured differently. *Ab initio* models are not clustered due to diversity of poses, resulting in small structural clusters. Red points and crosses indicate the cluster means and standard deviations for the HADDOCK score and I-rmsds, respectively.

The F_{nat} evaluation of the runs indicates that the consensus data/RDC performs the best with statistically significant results when compared to the CSP/RDC run. The 'ALL' run is also better than the runs, which lack RDCs in docking. The TH, CSP, and CSP/TH runs amongst themselves show a greater difference in NOCs where the TH run generated the highest number of statistically significant NOCs than the CSP and CSP/TH runs based on the F_{nat} criterion. This is interesting considering that the TH run's restraint data has the lowest TP rates for both ligand and receptor proteins than the other runs. To investigate this further, the rigid-body stage models (1000) were examined to ascertain the NOCs. These were found to be similar in number for the runs with differences that were not significant (TH 228, CSP 235, and CSP/TH 242 NOCs). Thus, the TH run was successful in the F_{nat} evaluation because the HADDOCK score ranked more NOCs in the top-200 models for it than the other runs. In terms of model quality, RDC-less data-driven runs produced more acceptable (*) models with less medium (**) models compared with the RDC-implementing runs that enrich both acceptable and medium categories. The 'ALL' run produced the most medium quality models and the only high quality models (2), underscoring the effect of using consensus-data/RDC to improve docking performance. Because the F_{nat} criterion is less stringent in comparison to CAPRI criteria, the inclusion of more correct near-native models has an impact on the ranking of the first correct near-native model. For all runs, the first correct model (*) is ranked amongst the top ten HADDOCK score-sorted models. The 'All' data-driven run produced 9 correct near-native models ranked in the top ten, which was higher than the CSP/RDC run which produced 7 correct models. All RDC-less runs produced 4 correct models in the top ten ranked solutions. The *ab initio* run produced only 1 correct model ranked in the top ten. The capability of consensus-data/RDC docking in enriching in correct models in the top-ten ranked solutions is likely due to an increased number of near-native models produced during docking sampling.

7.4.2 Application of RDCs, CSPs, and theoretical restraints for docking of the 1J6T protein complex

The 1J6T complex is created by the interaction of cytoplasmic domain A of the mannitol-specific phosphotransferase enzyme II ($E2A^{Mtl}$) and the histidine-containing phosphocarrier protein (HPr). This protein complex is a part of the bacterial

phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS) that phosphorylates carbohydrates while transferring them through bacterial cell membranes (Cornilescu *et al*., 2002).

The CSP and RDC data were obtained and applied for docking (section 3.9.1) of this complex (Clore and Schwieters, 2003; Cornilescu *et al*., 2002). The mapping of CSP and TH data on both receptor ($E2A^{Mtl}$) and ligand (HPr) proteins indicate that there is a contribution of true positives from the TH data when combined with CSP data (Figure 7-4C, D). Furthermore, it can be seen that conservation of experimentally identified active residues and their neighbouring passive residues (HPr ligand protein) of CSP data is present, as overlap exists with TH data. As stated previously, passive residues are designated active if conserved (section 3.9.1). The quantitative contribution of TH data shows an increase in the TP rate for both ligand (0.85) and receptor (0.76) proteins when it is combined with CSP data (Table 7-5). The TP fractions for both proteins decrease, but are still acceptable since they indicate >60% specificity for both proteins. In comparison to the *ab initio* run, only the RDC-incorporating and consensus data-driven (CSP/TH) runs produced statistically significant runs according to CAPRI criteria (Table 7-5). RDC data improved the results appreciably with the consensus data/RDC run producing the most NOCs (168) when compared to the standard CSP/RDC run (137), and this difference in NOCs is a statistically significant result. Figure 7-6 indicates the I-rmsds of the docked models compared to the reference 1J6T complex as a function of the HADDOCK score for all runs. Both RDC-incorporating runs produced the most correct models that converge in their first clusters than the other data-driven runs, producing the same I-rmsd cluster mean of 2.4 Å (STDVs of 0.57 Å (‘ALL’) and 0.56 Å (CSP/RDC)), but they differ in average energies due to differences in the ambiguous interface restraint (AIR) energy term of the HADDOCK score, which reflects AIR violations. For the CAPRI criteria, the TH run did not produce enough NOCs because the majority of 200 complexes generated in TH docking had F_{nat} values under 0.085 and orientational restraint errors as indicated by L and I-rmsd values, which where above the CAPRI cut-offs. Interestingly, the CSP run, which had higher TP fractions and TP rates than the TH run also did not produce a significant result for CAPRI criteria due to orientational errors in its generated models. Both CSP and TH runs’ first cluster had I-rmsd means of 4.6 Å and 5.9 Å, respectively (STDVs of 1.03 Å and 1.29 Å, respectively). It seems that higher TP rates were needed to generate

significant results as indicated in the CSP/TH run, which generated a mean I-rmsd for the first cluster of 4.1 Å (STDV of 0.83). For the TH, CSP, and CSP/TH runs, it can be seen that the majority of models cluster at I-rmsd values >9 Å.

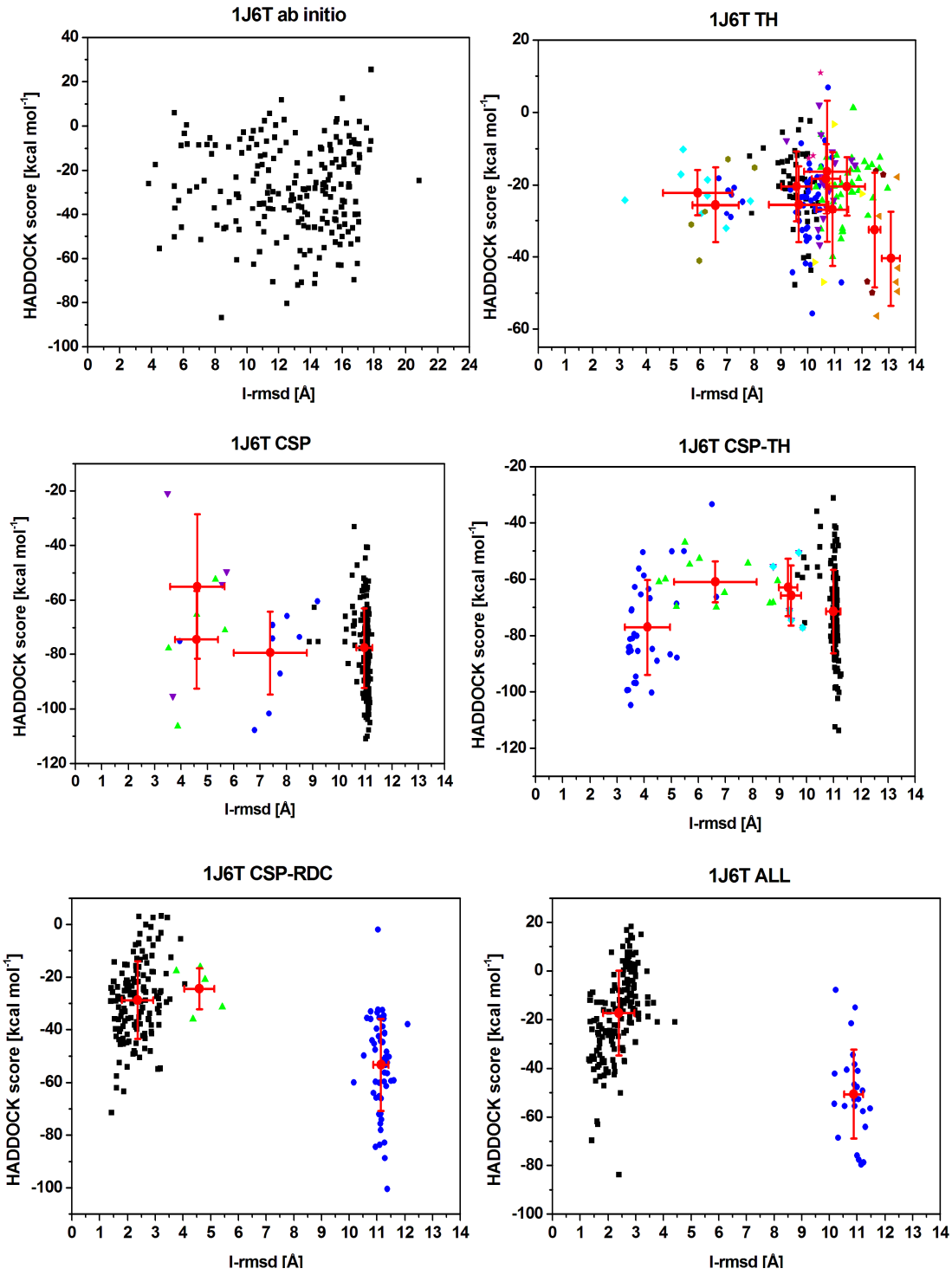

Figure 7-6: HADDOCK score versus I-rmsd for consensus-data (ALL = CSPs, RDCs, and theoretical restraints) and experimental data-driven docking (CSPs/RDCs restraints) for the 1J6T complex. See the legend of Figure 7-5 for further details.

These distant clusters have arisen because of the nature of the ambiguous interaction restraints used in docking, whether derived from CSP or TH data, which do not provide orientational input to docking sampling, and this explains the enrichment of incorrect models in those clusters where ligand proteins are rotated incorrectly with respect to the true ligand's binding pose to the receptor protein, especially for the TH and CSP runs. These clusters exist in the RDC-incorporating runs with similar incorrect orientational poses for their ligands (180°) as the models produced in the CSP and CSP/TH runs, however, they are smaller in size, indicating the effect of the RDC data in enriching the first clusters with more NOCs and minimizing these orientational errors. The average cluster energies for these clusters differ between RDC-incorporating and RDC-less runs because the HADDOCK score incorporated additional RDC-energy terms for the former runs.

In general, the correct and incorrect models differ by 180° rotation of their ligands even when RDC data is used. With RDC data, there are four possible 180° orientations of proteins in a complex with respect to the axes of an alignment tensor, which is four-fold degeneracy, and combining this data with ambiguous interaction restraints data (ex. CSP data) usually identifies the protein-protein orientation out of the possible four that agrees with both sets of data. For this protein complex, the combination of RDC and ambiguous interaction restraints data has resulted in two-fold degeneracy reduction, causing the majority of models to adopt two ligand poses that differ by 180° rotations, and this can be removed through the enforcement of restraints to favour the orientation most compatible with the CSP data (Clore and Schwieters, 2003). It can also be seen that the distant clusters have better HADDOCK scores as correct models for all runs. This is due to an extended charged patch on the receptor ($E2A^{Mtl}$), which includes the interface, where models localize on it, preventing HADDOCK discriminating the correct models from incorrect ones.

The correct models for the consensus-data/RDC ('ALL') run represent a balanced proportion of acceptable (83) and medium (85) quality models based on CAPRI assessment, which constitutes the highest number of NOCs produced altogether when compared to the other runs. Although the RDC-incorporating runs generated the most NOCs, because of inconsistency of discrimination by the HADDOCK score between correct and incorrect model ranking, the ranking of the correct models in the top ten

solutions was affected. For example, no correct models were ranked in the top ten for standard CSP/RDC docking. The first correct model is ranked 11th and is of medium quality for this run. For consensus data/RDC docking, only 4 correct models were ranked in the top ten where the 1st ranked model is of medium (**) quality. The CSP/TH generated 4 correct models, while the CSP run produced 1 correct model in their top ten ranked solutions. The correct solutions of the TH and *ab initio* runs were not ranked in the top ten, instead they are ranked 76th and 113th, respectively.

For F_{nat} evaluation, all runs except the TH data run produced statistically significant results compared to the *ab initio* run. The reason for the failure of the TH run was explained previously (see section 7.3.6). The consensus data/RDC run produced more near-native models (176) versus conventional CSP/RDC docking (143), which is a statistically significant result. In addition, it produces more near-native models than the rest of the runs that do not use RDCs, although the difference between its result and the CSP/TH run's result (162) is not statistically significant. Comparing the RDC-less data-driven runs's results to each, the CSP/TH run's result (162) is statistically significant compared to the CSP and TH runs. This may be due to the higher TP rates for the docking restraints used in the CSP/TH run. In terms of model quality, the consensus data/RDC run generated the highest proportion of high (44) and medium (108) F_{nat} models along with 24 acceptable models and this is a superior result when contrasted to the standard CSP/RDC docking. It is interesting to note that whilst CSP/TH docking produced a statistically significant number of near-native models compared to RDC/CSP docking and a non-significant difference compared to consensus/RDC docking, this achievement is not reflected fully in model quality. Most models for CSP/TH runs are enriched in the acceptable (128) F_{nat} category with only a few models (9) of medium quality. Although 25 models are rated as high for CSP/TH docking, however, the RDC-incorporating runs produced more medium quality models, and an almost double number of high models were produced by the consensus data/RDC run, highlighting the effect of orientational restraint data's influence on docking performance. The ranking of top ten models is improved under F_{nat} evaluation. The 'ALL' run produced more (6) near-native models ranked in the top ten solutions than the CSP/RDC run (1), where a high near-native model is ranked first for the 'ALL' run. Only the CSP/TH (7) and CSP (8) runs produced near-native models in top ten solutions from the non-RDC integrating runs. While these runs produced more near-native

models in the top ten then the 'ALL' run and in one case (CSP) ranked as 1st a near-native model, the 'ALL' has the best quality model ranked in first place by HADDOCK. The results based on CAPRI and F_{nat} assessment of the $E2A^{Mtl}$-HPr docking demonstrate that consensus/RDC data driven docking enriches in the number of correct models compared to standard CSP/RDC docking.

7.4.3 Application of RDCs, CSPs, and theoretical restraints for docking of the 1GGR protein complex

The 1GGR complex is an interaction involving glucose-specific phosphotransferase enzyme IIA ($E2A^{Glc}$) and the histidine-containing phosphocarrier protein (HPr). Like the 1J6T complex (section 7.4.2), this protein complex participates in the bacterial phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS) that phosphorylates carbohydrates and transfers them through bacterial cell membranes (Wang *et al*., 2000). Both protein complexes share the same ligand protein (HPr), however, their receptor proteins differ in primary sequence and tertiary structure (Clore and Schwieters, 2003).

The CSP and RDC data were applied for protein docking of this complex as discussed in section 3.9.1 (de Vries and Bonvin, 2011b; Clore and Schwieters, 2003). The mapping of CSP and TH data is indicated for both ligand (HPr) and receptor ($E2A^{Glc}$) proteins in Figure 7-4 (E, F). It can be seen that TH data contribute non-overlapping true positive residues to the ligand protein, increasing the TP rate to 0.93 for this protein combined with a TP fraction of 0.58 (Table 7-5). No additional non-overlapping true positive residues are provided for the receptor protein by the TH data, keeping the TP rate at 0.65. Instead all non-overlapping residues (blue) are false positives that are in close proximity to the experimentally identified residues, and this has decreased the TP fraction to 0.65. For both proteins' TP fractions, their specificity is >50%, which is satisfactory. There exists overlap between CSP (active and passive) and TH identified residues, indicating that these experimentally identified residues and their neighbours are conserved. As described previously, the presence of overlapping passive residues that are conserved designates them to active residue status (section 3.9.1).

All runs except the TH run produce more NOCs, which are statistically significant results, when compared with the *ab initio* run under CAPRI assessment. All the TH run's models had orientational restraint errors exceeding CAPRI cut-offs for L-rmsd and I-rmsd, rendering them all as incorrect according to CAPRI criteria. These errors may stem from the fact that TP rate contribution for the receptor protein is low (0.11), preventing adequate docking sampling in the correct binding region of the receptor protein in order to generate a pool of CAPRI acceptable binding poses during docking. The NOCs produced for the 'ALL' (180) and CSP/RDCs (183) are similar in number, indicating the absence of statistical significance in the difference between them (Table 7-5). The possible reason for this is that the addition of TH data only contributed to increase in TP rate for the ligand protein, but not the receptor protein. This lack of increase of receptor TP rate does not improve upon existing information provided by the CSP data to improve docking sampling's generation of NOCs. Indeed, it was demonstrated that when TH data improved the TP rates for both ligand and receptor proteins this resulted in a boost in the NOCs generated in both 1OO9 and 1J6T docking runs that coupled consensus data with RDCs (sections 7.4.1 and 7.4.2).

The NOC results of CSP/TH (30) and CSP (69) runs indicate a statistically significant difference in favour of the CSP run. This may be due to the random removal of restraints in HADDOCK that possibly removed more true positives during docking for the CSP/TH run coupled with the effect of the false positive restraints when the CSP and TH restraints were combined. This reduced the number of correct solutions produced because of the incorrect binding poses for the models produced relative to the known 1GGR experimentally solved complex. An examination of the docking data revealed that more models fulfilled the L-rmsd and I-rmsd criteria for CSP docking than CSP/TH docking, indicating an increase of models with orientational errors in CSP/TH, suggesting that indeed this may be the case.

The I-rmsds of all 1GGR runs' models as a function of HADDOCK score are shown in Figure 7-7. The RDC-incorporating runs produced the majority of models that converge in their first cluster (*i.e.* with the lowest I-rmsd average) than the RDC lacking runs. The consensus data/RDC and standard CSP/RDC runs produced I-rmsd cluster means of 2.7 Å (STDV of 0.82 Å) and 2.1 Å (STDV of 0.75 Å), respectively. The CSP and CSP/TH

runs have similar I-rmsd means (CSP/TH 2.0 Å and STDV of 0.30 Å, CSP 2.3 Å and STDV of 0.34 Å) for their best clusters to the RDC-incorporating runs, but have a lower convergence of models that populate these clusters. The clusters for the TH run are all above the 4.0 Å threshold for the I-rmsd criterion and hence the best cluster has an I-rmsd mean of 5.9 Å (STDV of 0.46 Å), indicating the orientational errors of the models.

For all runs that produced a greater amount of correct models compared to the *ab initio* run, the HADDOCK score discriminates between correct and incorrect models such that the best clusters are clearly identified. This is due to a charged patch of the interface of the HPr protein coupled with the shape complementarity between the two interfaces of the ligand and receptor proteins, allowing HADDOCK to discriminate between correct and incorrect models (Dominguez *et al*., 2003; Wang *et al*., 2000).

Both RDC-incorporating runs enrich the acceptable (*) and medium (**) categories more than the other runs with the exception of the CSP run, which enriched the medium quality category with 49 models. More models of acceptable category are produced for the 'ALL' run, whereas the standard run produces more medium quality models. This difference may be attributed to increased false positives from the TH data for the receptor protein, which may have had an impact on sampling of docking binding poses thereby enriching the acceptable category for the 'ALL' run. Both RDC-incorporating runs produce a significant number of high (***) models, indicating the success of the docking for this complex. Examination of the top ten solutions ranked by the HADDOCK score indicates that the RDC-incorporating runs produce 10 correct models in the top-ten solutions, according to CAPRI criteria. The CSP/TH and CSP runs produced 2 and 9 correct solutions ranked in the top ten models, respectively.

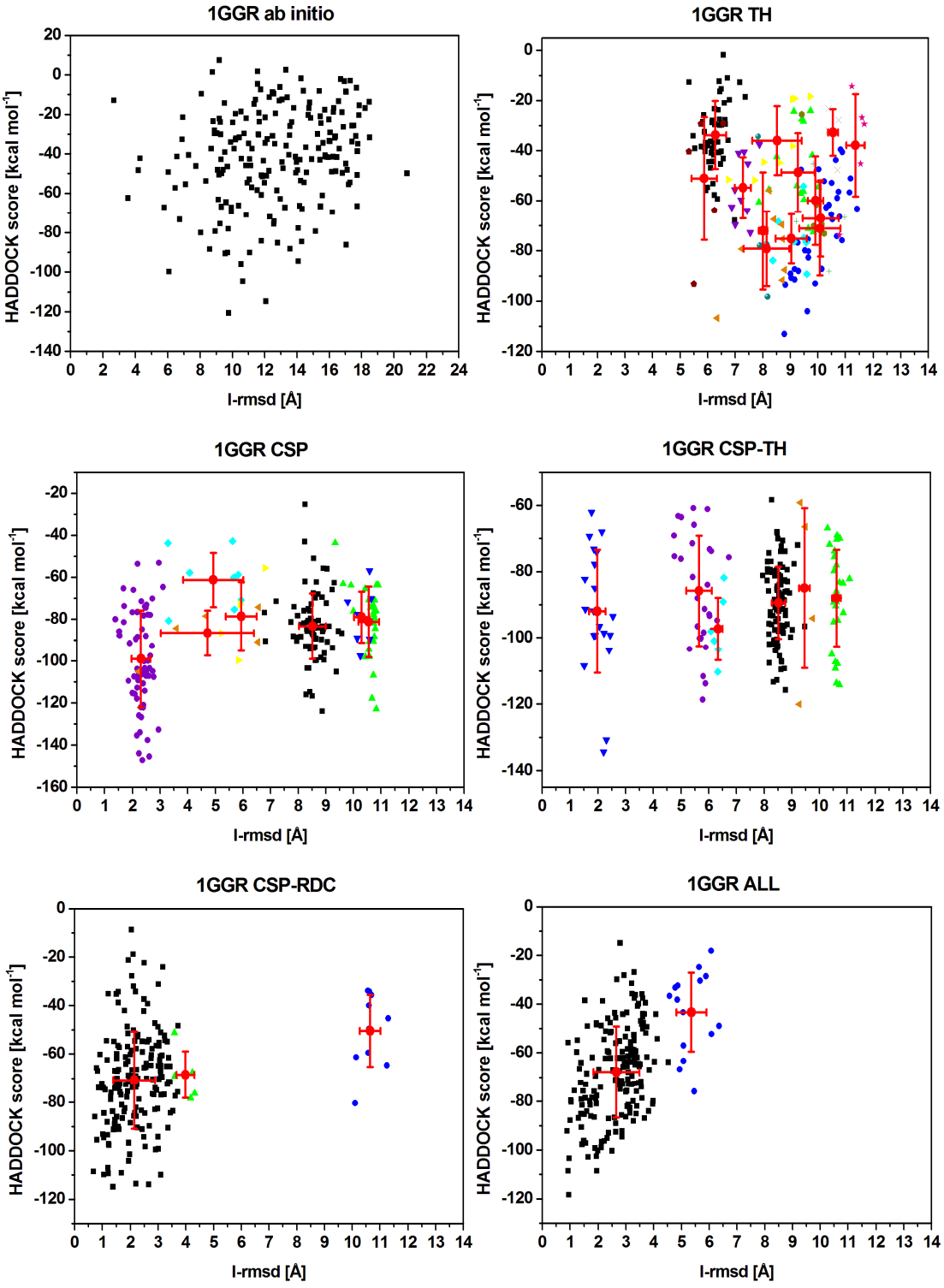

Figure 7-7: HADDOCK score versus I-rmsd for consensus-data (ALL = CSPs, RDCs, and theoretical restraints) and experimental data-driven docking (CSPs/RDCs restraints) for the 1GGR complex. See the legend of Figure 7-5 for further details.

This difference in the ranking of correct models may be attributed to the smaller amount of correct models produced in the CSP/TH run. *Ab initio* docking produces only one acceptable quality model and it is ranked 41st by HADDOCK. Although the 'ALL', CSP/RDC, and CSP runs produced a similar amount of correct models ranked in the top ten solutions, the quality of the models ranked in the top ten solutions differs somewhat. For the 'All' run the 1st ranked (*i.e.* best scored) model is of high quality, whereas for the other two runs their best-ranked model is of medium quality. Indeed, there are 3 high quality models ranked within the top-four models for the 'ALL' run, whereas the CSP/RDC run had only one high model in the top ten models and it was ranked 8th best overall with other models ranked acceptable or medium quality scoring better than it. The CSP run only produced 9 medium quality models in the top ten solutions. Therefore, even though the overall difference in NOCs among the RDC-incorporating runs is not significant, the enrichment of better quality models in the 'ALL' docking run is improved in the top-ten ranked solutions compared to the standard CSP/RDC run. This is also the case when the 'ALL' run is compared to the CSP run.

For the F_{nat} evaluation of the runs, the 'ALL' and the CSP/RDC runs produce 180 and 185 correct near-native models, respectively. Here the difference in 5 near-native models is not significant. The RDC-incorporating runs' results are statistically significant when compared to the runs that do not utilize RDC data. The CSP/TH (158 near-native models) and CSP (161 near-native models) runs do not have a significant difference compared to each other. However they both excel the TH and *ab initio* runs in NOCs produced. In model quality terms, the RDC-including runs have similar numbers for the acceptable F_{nat} category, but differ in NOCs produced for the medium and high quality F_{nat} categories. The 'ALL' run has higher NOCs (91) of medium quality than the CSP/RDC run, whereas the CSP/RDC has higher NOCs (77) for the high category of models. Like the CAPRI evaluation, this difference in the spread of models may be due to higher false positives incorporated for the receptor protein's docking restraints from the TH data, which may have influenced docking sampling in the 'ALL' run, resulting in less high quality models and greater medium quality models in contrast to the CSP/RDC run.

The ranking of the top-ten scored models for the RDC-incorporating runs under F_{nat} evaluation parallels the results of the CAPRI evaluation in that 10 correct models in the

top-ten ranked solutions are found. This was also the result for the CSP run. The CSP/TH run had 6 correct models ranked in the top ten models. Although the TH run produced 10 near-native models overall, 2 of them are ranked in the top-ten. The *ab initio* run does not produce any models ranked in the top-ten solutions. The quality of the models ranked in the top-ten solutions is similar for the runs. It can be seen that all runs (except the TH and *ab initio* runs) produced a high quality model ranked 1st and this is because the F_{nat} restraints are less stringent as they do not take into account orientational errors. Overall, the difference in the generation of NOCs for the 1GGR complex between the 'ALL' and CSP/RDC runs is not significant because the TH data when combined with the CSP data only increases the TP rate for the ligand protein and not the receptor. The TH data for the receptor increased the number of false positive residues, which influenced the docking sampling. Such residues may be conserved because of structural reasons. Additional interface residue discriminators are required to combine with the current TH data to predict the interface residues not predicted via TH and CSP data sets. This would be important for the receptor protein and would likely result in a non-overlapping contribution by both data sets, increasing the TP rate. The outcome would be significant NOCs being generated when compared to the RDC/CSP 1GGR run.

7.5 Conclusion

The application of theoretical restraints of the PROTIN_ID method to guide protein-protein docking to improve its performance was examined. This was compared to *ab initio* docking. The docking results were assessed using standard performance measures of the protein docking community (CAPRI), including the less stringent F_{nat} measure. Docking runs were performed using a docking dataset (26 complexes) comprising unbound proteins that fulfilled a minimum (≥10%) TP rate cut-off when theoretical restraints were generated for them using PROTIN_ID. Complexes under the conservation TP rate cut-off were removed since they were assumed to have few restraints to generate successful results in data-driven docking. Data analysis based on the docking dataset and extrapolated to account for the excluded complexes (<10% TP rate cut-off) demonstrated that docking is improved in the generation of correct models through the use of theoretical restraints compared to *ab initio* docking for both CAPRI

and F_{nat} measures. Specifically, a 16% success rate for all cases (*i.e.* conservative estimate) was determined for data-driven docking based on CAPRI criteria, which is four times more successful, compared to 4% in *ab initio* docking. For the less stringent F_{nat} evaluation, a success rate of 29% (*i.e.* conservative estimate) was observed relative to no success (0%) in *ab initio* docking. The difference in success rates for *ab initio* docking for both CAPRI or F_{nat} measures is because the analysis for significant results in the docking dataset cases is dependent on the number of correct models produced by both data-driven and *ab initio* runs per case that are compared to ascertain if significance exists. Only one example produced a significant number of CAPRI correct models in favour of *ab initio* docking in the docking dataset, resulting in a 4% success rate under CAPRI criteria when extrapolated to the entire dataset of (63) complexes. The same example did not produce a significant result when analysed by the F_{nat} criterion; however, its data-driven run counterpart produced significant results. When extrapolating to the entire dataset, this resulted in no cases for *ab initio* docking that produced significant results based on a comparison between the run types according to the F_{nat} measure.

The success of using theoretical restraints to improve docking compared to *ab initio* docking paved the way for further examination of using theoretical and CSP data (consensus data) combined with RDC orientational restraints data to assess its impact on docking performance. This was compared to standard CSP/RDC docking simulations. Three case studies were examined for which RDC and CSP data were obtainable. In general the performance of consensus data/RDC-driven docking improves the generation of correct docking solutions compared to standard CSP/RDC docking based on CAPRI and F_{nat} measures. The improvement in docking is applicable specifically when CSP and theoretical data both map the same area on a protein's surface, resulting in the further inclusion of true positive (interface) residues for both ligand and receptor proteins. This is because there are theoretical data that do not overlap with the CSP data, increasing the number of true positives. If different areas on a protein's surface are identified by both restraints, then it is more likely that the theoretical restraints have possibly predicted another binding site. In this scenario, they are not applicable for use with CSP data to drive docking. When both CSP and theoretical restraints localize on the same protein surface's region, passive residues, which have insignificant CSPs and/or are in close proximity to (active) residues with

significant CSPs, were converted to active residues since they were conserved based on the theoretical restraints that overlapped with them. To recapitulate, the improvement in docking using consensus-data was demonstrated when CSP and theoretical data are restricted to the same region of a protein's surface. Upon satisfying this condition, this allows the HADDOCK docking program to further restrict the docking sampling to the area of interest and boost docking performance by significantly increasing the number of correct models produced compared to standard CSP/RDC docking.

Chapter 8

Conclusions

Determining interfaces of proteins is a significant step for characterizing protein complexes and contextualizing their functions within the wider protein interactome. Prior to the onset of this research study, protein interface predictors were found to be limited with respect to interface prediction through use of sequence data and clustering of prediction data (see table A-3). Firstly, no attempt was made to systematically examine explicit transient protein multiple sequence alignment improvement and its effect on conservation signal retrieval by previous predictors, which used conservation as an interface predictive feature, to improve interface prediction. Secondly, the effect of three-dimensional clustering of interface prediction data accuracy was not explored systematically in previous predictors to ascertain its effect on interface prediction accuracy. In this work a new protein-protein interface predictor (PROTIN_ID) was introduced that addressed these deficiencies and sought to identify protein interface residues through implementing explicit MSA sequence data editing and interface prediction data clustering heuristics to improve interface prediction quality. These heuristics were tested on the latest Benchmark 4.0 dataset of transient protein complexes (Hwang *et al*., 2010). The results of this study corroborate the hypotheses that both heuristics significantly improve interface prediction accuracy compared to not applying them. Thus the PROTIN_ID predictor implements novel and useful features for predicting interface residues. Compared to current interface predictors (WHISCY and CCRXP) with similar interface prediction goals, PROTIN_ID was found to perform as well as WHISCY and outperform CCRXP (Ahmad *et al*., 2010; de Vries *et al*., 2006).

In the context of whether interface residues were more conserved compared to rest of surface residues (ROS), it was demonstrated that interface residues were indeed more conserved, using seven different conservation scores, which is in agreement with previous findings (Choi *et al*., 2009; see section 5.5). However, previous work indicated the absence of prediction significance to distinguish interface patches from ROS patches, using evolutionary conservation (Burgoyne and Jackson, 2006; Caffrey *et al*.,

2004; see section 5.5). In this work, an evolutionary conservation-based clustering heuristic, implemented in PROTIN_ID, was shown to significantly predict interfaces residue clusters. This approach exploits the conservation signal of interface residues more effectively than in previous work. This is in agreement with recent findings, which showed that interfaces are non-homogenously conserved, and those residues, which are conserved, are clustered together (Guharoy and Chakrabarti, 2010; Guharoy and Chakrabarti, 2005).

During this study and after its completion, other interface predictors have been published. Like PROTIN_ID, they also implemented novel concepts for predicting protein interfaces. For example, Li *et al*., (2008) trained an SVM-based predictor that utilized neighbour residue profiles at the sequence and structural levels during training. Both structural and sequence profiles concepts have been implemented separately in the previous predictors of Wang *et al*., (2006) and Ofran and Rost (2003a), respectively. This predictor combined both concepts to construct a holistic neighbour profile utilizing 8 predictive features (ex. hydrophobicity, sequence conservation, physicochemical properties, solvent accessibility, side-chain environment, secondary structure, and sequence and spatial distance) to predict core interface residues. This predictor uses a PSI-BLAST profile to compute conservation. It could be combined with PROTIN_ID's sequence editing heuristic and conservation score implementation to enhance the conservation signal retrieval during its training stage, as it was determined that the PSI-BLAST profile conservation predictive feature of the predictor contributed most to the accuracy measure compared to the rest of the predictive features during cross-validation of their predictor (Li *et al*., 2008). PROTIN_ID's clustering heuristic could be applied to cluster top-N ranked surface residues predicted by the method of Li *et al*., (2008) to improve core interface prediction accuracy through elimination of potential noise created by false positive residues. Another notable predictor used only electrostatic desolvation profiles for prediction of interface sites (Fiorucci and Zacharias, 2010a). This predictor scanned protein surfaces to identify sites with low desolvation penalties, which are predicted as putative interface sites, using the finite-difference Poisson-Boltzmann method. When implemented with other interface discriminative features, it has a potential to improve interface prediction. Applied with PROTIN_ID, this desolvation profile concept could be used to develop PROTIN_ID's cluster ranking and improve its interface prediction success rate.

The application of PROTIN_ID's theoretical restraints to docking performance was investigated using the HADDOCK method. It was demonstrated that theoretical restraints-driven docking was more successful than *ab initio* docking, when evaluated with stringent or relaxed metrics, generating significantly more correct models for a higher percentage of the test dataset than *ab initio* docking. The current findings agree with previous work by de Vries *et al*. (2006) who used WHICSY's prediction data to guide docking, and this was contrasted to *ab initio* docking. Additionally, PROMATE prediction data was also combined with WHISCY data to repeat the same runs, which showed greater guided-docking performance improvement. However, in their work de Vries *et al*. (2006) conduct docking using the initial rigid-body docking stage of HADDOCK, and used a relaxed evaluation metric for docking performance assessment. In contrast, in this work the full HADDOCK protocol was applied and stringent CAPRI criteria were used to evaluate docking results, providing a more realistic framework for the study. Although, combination of PROTIN_ID restraints with another predictor's restraints (ex. WHISCY) was not assessed on docking performance here, the same principle of consensus data was explored in this study through the novel combination of NMR data (CSP/RDC) with theoretical restraints, achieving an improvement over docking runs using only PROTIN_ID's restraints. It was demonstrated that this approach improved the performance of data-driven docking compared to standard experimental (CSP/RDC) docking overall. Recent work showed that using heterogeneous experimental data to filter docking solutions (*i.e.* back-end docking) in a novel meta-docking approach led to improved ranking of acceptable quality models or higher (61% ranked 1st) for Benchmark 4.0 cases (Schneidman-Duhovny *et al*., 2012; Hwang *et al*., 2010). Although both studies showed that docking performance can be improved, they differ in the sources of data integrated in the step of their docking procedures (front-end vs. back-end docking). It would be valuable to extend the work of Schneidman-Duhovny *et al*., (2012) to include more data sources such as consensus interface predictor data (including PROTIN_ID predictor data) with the experimental data (RDC and CSP) to improve their back-end docking approach further. Thus using more experimental and/or theoretical sources to expand consensus-data would make significant inroads in docking performance improvement. Recent steps have been taken to achieve this. For example, de Vries and Bonvin (2011a) used prediction data

generated from a consensus predictor (CPORT) composed of six individual interface predictors (including WHISCY) to guide HADDOCK docking, and showed that docking performance was better than HADDOCK *ab initio* docking and as competitive to ZDOCK. Schneider and Zacharias (2012) combined the ATTRACT docking method with the meta-PPISP (composed of three predictors) prediction data to drive docking of unbound proteins (Qin and Zhou, 2007a; Fiorucci and Zacharias, 2010b; Mintseris *et al*., 2005b; Zacharias, 2003). It was found that data-driven docking was more successful than *ab initio* docking (77% vs. 65% success rate). Compared to this current work and others, the main difference in this study is the effect ATTRACT has on docking performance. For example, Schneider and Zacharias (2012) compared their results to CPORT prediction data-driven docking and *ab initio* docking both using HADDOCK, which achieved 41% and 15% success rates, respectively. The prediction performance is more significant using both ATTRACT docking approaches compared to CPORT-driven or *ab initio* HADDOCK docking. This is because input prediction restraints are applied as force field weights to bias sampling to the predicted region on a protein's surface, while also permitting other surface regions to be sampled (Schneider and Zacharias, 2012). In contrast, HADDOCK data-driven docking sampling is exclusive to the region of interest, and can produce incorrect results if the prediction data are completely incorrect. In another study, Li and Kihara (2012) used the CPORT interface predictor with a different docking method, PI-LZerD, to analyze their approach on Benchmark 2.0 (de Vries and Bonvin, 2011a; Venkatraman *et al*., 2009). PI-LZerD allows more flexible sampling (unlike HADDOCK) by sampling around prediction data defined on two input proteins' surfaces. Compared to CPORT-HADDOCK, their approach's success rate was better (24.6% vs. 15.8%). Both ATTRACT and PI-LZerD are more sampling tolerant and would be useful to apply in future work to extend this study using all Benchmark 4.0 complexes (Hwang *et al*., 2010). It is anticipated that prediction performance would improve for most cases due to their sampling efficiency. This would also be true when combining prediction data with other sources of experimental data.

The field of interface prediction is growing rapidly. Upon the conclusion of this work, future research directions were identified to improve and extend this field of interface prediction. For example, current predictors are designed on the basis that protein

interactions have negligible conformation change, which is reflected in the training and testing datasets used to develop them. This is not always the case as some protein interfaces are difficult to predict because of high conformational change upon complex formation. Correctly predicted interface residues in their unbound forms would be far apart, where it would not be apparent that they are proximal interface residues. Further studies are required to address this. For instance, a starting point would be to generate conformational 'snap shots' of a protein (ex. from Molecular Dynamics) to allow ensemble interface prediction (e.g., via PROTIN_ID) and merge their results (or provide conformational prediction 'snap shots') to determine if proximity associations exist between predicted residues. Another problem is the lack of standardization in terms of training and testing datasets and interface residue definitions. Most predictors utilize different interface definitions and developmental datasets, when reporting their predictors' performance evaluation (ex. specificity, sensitivity, etc.). This makes it unsuitable to compare their reported performances. A way forward is to create standard non-redundant transient protein datasets based on datasets reported in previous work. The standard datasets can be categorized based on biological functionality of the transient complexes and the number of known interaction partners per protein to determine all possible interface residues from the ROS residue group.

Another way of improving interface predictors would be the incorporation of more useful interface residue predictive features. Swapna *et al*., (2012) have shown that interface residues of transient proteins with the lowest B-factors were mostly interface core residues, indicating their rigidity. Additionally, interface core residues were observed to be significantly more rigid than ROS residues. To compute the rigidness of surface residues, they applied a normalized backbone B-factor measure. This interface residue predictive feature can be combined with others implemented in current interface predictors to improve interface prediction quality. Using their approach is advantageous as the predictive data is derived from unbound proteins unlike the approach of Chung *et al.,* (2006) who in contrast utilized B-factor data derived from bound models, which can introduce bias in terms of 'inflated' predictor performance (see section 1.9.2). The novel features introduced in this work and recent studies suggest that practical application of such features has the potential to boost prediction performance. An important achievement has recently been made in a recent study published after the completion of this current work. Segura *et al*., (2011) used multiple sources of interface predictive

features (ex. structural, sequence, and energy based) to develop a predictor. This predictor produced ROC area under the curve value (AUC) of 0.85 on Benchmark 3.0 (excluding antibody-antigen complexes) (Segura *et al*., 2011). This indicates the benefit of applying many heterogeneous predictive features in interface prediction. While the AUC cannot be meaningfully compared to PROTIN_ID's AUC value due to differing protein complexes in their datasets and interface definitions used, it is not surprising that the predictor of Segura *et al*., (2011) produced a high AUC due to multiple heterogeneous interface residue predictive features being implemented. As such, these heterogeneous features would be beneficial in future development of PROTIN_ID and other predictors to improve their prediction performance. Another dimension to enhance and improve interface predictors would be to add a component based on known homologous protein complexes (*i.e.* protein complex-level predictive feature) and use the data as an interface predictive feature in combination with other heterogeneous interface residue predictive features. Indeed using structural data was shown to improve prediction performance in a recent study (Xue *et al*., 2011a). Such new and improved interface predictors would be useful especially in their combination with experimental data in the application of high-throughput docking for predicting protein complexes.

Appendix

Table A-1: The advantages (green) and disadvantages (red) of predictors for categories regarding training and testing datasets and predictor performances on these datasets predictor are shown. A dash (-) indicates no information could be obtained. A category not applicable for a predictor is indicated as 'N/A'.

Predictor	Developmental datasets (including testing): Transient only or mixed (obligate/permanent and transient) complexes	Developmental dataset: transient proteins bound or unbound?	Specificity and sensitivity/ or accuracy performance % (developmental testing dataset)	Independent testing dataset: transient, obligate, or mixed complexes	Independent testing dataset: transient proteins bound or unbound?	Specificity and sensitivity performance % (Independent testing dataset)
ET- Mihalek *et al.*, 2004; Litcharge *et al.*, 1996a	N/A	N/A	N/A	N/A	N/A	N/A
SHARPE^2- Murakami and Jones, 2006; Jones and Thornton, 1997a	Mixed	Bound models used	Specificity: >70 Sensitivity: -	N/A	N/A	N/A
Landgraf *et al.*, (2001)	Obligate complexes Transient complexes	Bound models used	N/A	N/A	N/A	N/A
Cons-PPISP Chen and Zhou (2005)	Mixed (Obligate homodimer complexes present both): i. Training set (1156 proteins) ii. Testing set (100	Bound models used	Specificity: 80 Sensitivity: 51	Transient (68 proteins) Transient (8 NMR complexes)	unbound unbound	Specificity: 61.4 Sensitivity: 38 (8 NMR proteins) Specificity: 69 Sensitivity: 47

	proteins)					
Fariselli *et al*., (2002)	Mixed (Obligate complexes present)	Bound models used	Specificity: 72 Sensitivity: 56	N/A	N/A	N/A
ISIS- Ofran and Rost (2006, 2003)	Predicted as transient (may contain obligates)	N/A	Specificity: ~61 Sensitivity: 20	N/A	N/A	N/A
ProMate- Neuvirth *et al*., (2004)	Transient only (Obligate complexes are removed)	Unbound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 67% of dataset	N/A	N/A	N/A
Crescendo- Chelliah *et al*., (2006)	Transient only	Mixed bound/ unbound dataset	Specificity: > 50 for 85% of dataset	N/A	N/A	N/A
Koike and Takagi (2004)	Mixed (Obligate complexes present)	Bound models used	Specificity: 56.1 Sensitivity: 44.6	N/A	N/A	N/A
Keil *et al*., (2004)	Mixed (Obligate complexes present)	Bound models used	Specificity: N/A Sensitivity: 44	N/A	N/A	N/A
PPI-Pred- Bradford and Westhead (2005)	Mixed (Obligate complexes present) – Leave-one-out (LOO) cross validation dataset (180 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 76% of dataset	Mixed (47 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 72% of dataset
	Subset of dataset: Obligate (114 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 82% of dataset	Transient (57 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 53% of dataset
	Subset of dataset: Transient (66 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 65% of			

			dataset			
Bordner and Abagyan (2005)	Mixed (Obligate complexes present): i. 5-fold cross validation dataset (632 protein complexes)	Bound models used	Specificity: 34 Sensitivity: 64	Transient (43 protein complexes)	-	Specificity: 22 Sensitivity: 67
Bradford and Westhead (2006)	Mixed (Obligate complexes present) – Leave-one-out (LOO) cross validation dataset (180 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 82% of dataset	N/A	N/A	N/A
	Subset of dataset: Obligate (114 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 84% of dataset			
	Subset of dataset: Transient (66 proteins)	Bound models used	Specificity: ≥ 50 Sensitivity: ≥ 20 both for 79% of dataset			
Hoskins *et al.*, (2006)	-	Mixed bound/ unbound dataset	Specificity: > 50 for 79% protein interfaces in the dataset	N/A	N/A	N/A
Chung *et al.*, (2006)	Mixed (Obligate complexes present) i. 3-fold cross validation dataset (274 protein complexes)	Bound models used	Specificity: 50 Sensitivity: 67.3	N/A	N/A	N/A
Wang *et al.*, (2006)	Mixed (Obligate complexes present) – (LOO) cross validation dataset (69 protein complexes)	Unbound models used	Specificity: 49.7 Sensitivity: 66.3	N/A	N/A	N/A

PINUP- Liang *et al*., (2006)	Transient only (Obligate complexes are removed)- (LOO) cross validation dataset (57 proteins)	Unbound models used	Specificity: 44.5 Sensitivity: 42.2	Transient (68 proteins)	Unbound models used	Specificity: 29.4 Sensitivity: 30.5
WHISCY- de Vries *et al.*, (2006)	Transient complexes only (57 proteins)	Bound models used	Specificity: 33 Sensitivity: 30	Transient complexes (38 proteins)	Unbound models used	Specificity: 40.8 Sensitivity: 26.7
SPPIDER- Porollo and Meller (2007)	Mixed i. k-fold cross validation dataset (435 proteins)	Bound models used	Specificity: 67 Sensitivity: 52.7	Transient (86 proteins)	-	Specificity: 47 Sensitivity: 43
				Mixed (149 proteins)	Bound models used	Specificity: 63.7 Sensitivity: 60.3
HotPatch- Pettit *et al*., (2007)	Obligate complexes are present (tested separately from transient complexes)	Mixed (Unbound and bound models)	Specificity: ≥ 33 Sensitivity: -	N/A	N/A	N/A
	Transient complexes		Specificity: ≥ 33 Sensitivity: -			
PIER- Kufareva *et al*., (2007)	Mixed (Permanent complexes present): i. 3-fold cross validation dataset (748 proteins)	Bound models used	Specificity: 60 Sensitivity: 50	Mixed (Obligate complexes present) – (180 proteins)	Bound models used	Specificity: 61.8 Sensitivity: 50
	Subset of dataset: Permanent (552 proteins)	Bound models used	Specificity: 65.5 (average) Sensitivity: 58 (average)			
	Subset of dataset: Transient (196 proteins)	Bound models used	Specificity: 49 (average) Sensitivity: 50 (average)	Transient complexes (91 complexes)	-	Specificity: ≥25 Sensitivity: 80 both for 82% of dataset

Konc and Janežič (2007)	Mixed (Obligate complexes present)	Bound models used	Specificity: ~ 42 Sensitivity: ~ 46.6	N/A	N/A	N/A
Negi and Braun (2007)	-	Unbound models used	Accuracy: ~ 70	N/A	N/A	N/A

Table A-2: The advantages (green) and disadvantages (red) of predictors for categories based on dataset essentials for predictor training, the use of structural and/or sequence predictive features for predictor development, miscellaneous details, and availability of a predictor webserver or download for each interface residue predictor are indicated. A dash (-) indicates no information could be obtained. A category not applicable for a predictor is indicated as 'N/A'.

Predictor	Biological or crystal packing interaction	Antibody-antigen (Ab/Ag) interaction exclusions	ROS residue removal during training (with cross-validation)	Benchmarking to other predictors	Sequence and/or Structural data usage	Miscellaneous	Webserver or download available
ET-Mihalek *et al.*, 2004; Litcharge *et al.*, 1996a	-	N/A	N/A	No	Sequence (latest) and structure data used	-	Webserver available
SHARPE^2-Murakami and Jones, 2006; Jones and Thornton, 1997a	Crystallization interactions avoided (manually curated dataset)	(Ab/Ag) not removed from dataset	N/A	No	Only structural data used	Manually curated dataset used	Webserver available
Landgraf *et al.*, (2001)	-	(Ab/Ag) not included in their analysis	N/A	No	Sequence (latest) and structure data used	-	No
Cons-PPISP Chen and Zhou (2005)	Crystallization interactions present	Removed (Ab/Ag) from training and testing dataset	No	No	Sequence (latest) and structure data used	Four FPs are converted to TPs in their prediction based on proximity to TPs	Webserver available
Fariselli *et al.*,	-	-	No	No	Sequence	i. Highly	Only source

(2002)					(HSSP- not latest) Structural data	redundant training set; ii. Permissive interface definition used (~40%)	code available (upon request)
ISIS- Ofran and Rost (2007, 2003)	N/A	-	No	No	Only sequence (latest) data used	-	Webserver available
ProMate- Neuvirth *et al*., (2004)	N/A	Removed (Ab/Ag) from training dataset	No	No	Sequence (latest) and structure data used	i. Able to accept potentially new interface predictive properties as input (User-friendly) ii. Manually curated dataset used	Webserver available
Crescendo- Chelliah *et al*., (2006; 2004)	N/A	(Ab/Ag) not included	N/A	No	Sequence (user-input) and structure data used	-	Webserver available
Koike and Takagi (2004)	Crystallization interactions present	(Ab/Ag) not removed from training dataset	Yes	Direct benchmarking performed	Sequence (latest) and structure data used	-	No
Keil *et al*., (2004)	Crystallization interactions not removed	(Ab/Ag) not removed from dataset	No	No	Only structural data used	Not a manually curated dataset (77% of the PDB-1999 is used)	No

PPI-Pred- Bradford and Westhead (2005)	Crystallization interactions removed	-	No	Direct benchmarking performed	Sequence (latest) and structure data used	Manually curated dataset used	Webserver available
Bordner and Abagyan (2005)	Crystallization interactions avoided(Swiss-Prot ver.)	(Ab/Ag) not included	No (avoided)	No	Sequence (latest) and structure data used	Manually curated dataset used (Swiss-Prot verification)	No
Bradford and Westhead (2006)	Crystallization interactions removed	-	No	Direct benchmarking performed	Sequence (latest) and structure data used	Manually curated dataset used	No
Hoskins *et al.*, (2006)	N/A	(Ab/Ag) not included	N/A	No	Only structural data used	-	No
Chung *et al.*, (2006)	-	-	Yes	No	Sequence (latest) and structure data used	An FP is converted to a TP in their prediction based on proximity to 3 TPs Structural conservation is used	No
Wang *et al.*, (2006)	-	-	Yes	Direct benchmarking performed	Sequence (latest) and structure data used	Permissive interface definition used	No
PINUP- Liang *et al.*, (2006)	N/A	Removed (Ab/Ag) from training dataset	No	Direct benchmarking performed	Sequence (latest) and structure data used	Manually curated datasets used for training and independent	Webserver available

						testing	
WHISCY- de Vries *et al.*, (2006)	N/A	Removed (Ab/Ag) from training dataset	No	Direct benchmarking performed	Sequence (HSSP- not latest) Structural data used	i. Manually curated dataset used	Webserver available
SPPIDER- Porollo and Meller (2007)	Crystallization interactions removed (via PQS server)	-	No	Direct benchmarking performed	Sequence (latest) and structure data used	i. Difference between predicted and actual solvent accessibility used ii. All available alternative interfaces are assigned to Positive or Negative classes prior to training	Webserver available
HotPatch- Pettit *et al.*, (2007)	Crystallization interactions avoided (Manually curated)	(Ab/Ag) not included	No	No	Only structural data used	Manually curated dataset used	Webserver available
PIER- Kufareva *et al.*, (2007)	Crystallization interactions avoided (Swiss-Prot verification)	Removed (Ab/Ag) from training dataset	No	Direct benchmarking performed	Only structural data used	Manually curated datasets used for training (Swiss-Prot verification) and independent testing	Webserver available

Konc and Janežič (2007)	N/A	N/A	N/A	No	Only structural data used	A small developmental dataset was used	No
InterProSurf- Negi and Braun (2007)	-	(Ab/Ag) not removed from dataset	N/A	No	Only structural data used	-	Webserver available

Table A-3: The limitations (red) of interface residue predictors are indicated. These are addressed in this study (see section 1.11 for details). A category not applicable for a predictor is indicated as 'N/A'.

Predictor	Conservation data source: Multiple sequence alignment (MSA) or PSI-BLAST PSSM	Explicit sequence data editing heuristic generated MSA vs. automatically generated MSA: A comparison by systematic analysis of transient proteins (hetero-complexes)	Interface residue prediction data clustering vs. non-residue clustering: A comparison by systematic analysis of unbound transient proteins (hetero-complexes)	Application of interface prediction and experimental (NMR) data-driven protein-protein docking?
ET- Mihalek *et al.*, 2004; Litcharge *et al*., 1996a	MSA	No	No	No
SHARPE^2- Murakami and Jones, 2006; Jones and Thornton, 1997a	N/A	N/A	N/A	N/A
Langraf *et al*., (2001)	MSA	No	No	No
Cons-PPISP Chen and Zhou (2005) (PPISP, Zhou and Shan, 2001)	PSI-BLAST PSSM	N/A	No	No
Fariselli *et al*., (2002)	MSA (HSSP sequence profile)	No	No	No

ISIS-Ofran and Rost (2006, 2003)	PSI-BLAST PSSM	N/A	N/A	No
ProMate-Neuvirth *et al*., (2004)	PSI-BLAST PSSM	N/A	No	No
Crescendo-Chelliah *et al*., (2006; 2004)	MSA	No	No	No
Koike and Takagi (2004)	PSI-BLAST PSSM	N/A	No	No
Keil *et al*., (2004)	N/A	N/A	No	No
PPI-Pred-Bradford and Westhead (2005)	MSA	No	No	No
Bordner and Abagyan (2005)	MSA	No	No	No

Bradford and Westhead (2006)	MSA	No	No	No
Hoskins *et al.*, (2006)	N/A	N/A	No	No
Chung *et al.*, (2006)	MSA^{struc} and PSI-BLAST PSSM	No	No	No
Wang *et al.*, (2006)	MSA and MSA (HSSP sequence profile)	No	No	No
PINUP- Liang *et al.*, (2006)	PSI-BLAST PSSM	N/A	No	No
WHISCY- de Vries *et al.*, (2006)	MSA (HSSP)	No	No	No
SPPIDER- Porollo and Meller (2007)	MSA and PSI-BLAST PSSM	No	No	No

HotPatch- Pettit *et al*., (2007)	N/A	N/A	No	No
PIER- Kufareva *et al*., (2007)	N/A	N/A	No	No
Konc and Janežič (2007)	N/A	N/A	No	No
Negi and Braun (2007)	N/A	N/A	No	No

Table A-4

Complex	Non-editing heuristic: unrefined MSA (Top-20 hit score)	Editing heuristic: refined MSA (Top-20 hit score)	Top-20 hit score fractional difference ($\Delta_{top\text{-}20}$)
1E6E_A:B_r	0	0	0
1E6E_A:B_l	0.3	0.6	0.3
1EWY_A:C_r	0.2	0.25	0.05
1EWY_A:C_l	0.3	0.3	0
2O8V_A:B_r	0.1	0	-0.1
2O8V_A:B_l	0.6	0.65	0.05
2PCC_A:B_r	0.15	0	-0.15
2PCC_A:B_l	0.25	0.25	0
7CEI_A:B_r	0.2	0.2	0
7CEI_A:B_l	0.1	0.05	-0.05
1B6C_A:B_r	0.2	0.3	0.1
1B6C_A:B_l	0	0.05	0.05
1BUH_A:B_r	0.05	0.05	0
1BUH_A:B_l	0.4	0.35	-0.05
1E96_A:B_r	0.1	0	-0.1
1E96_A:B_l	0.3	0.45	0.15
1FQJ_A:B_r	0.2	0.15	-0.05
1FQJ_A:B_l	0.3	0.75	0.45
1GLA_G:F_r	0	0	0
1GLA_G:F_l	0.1	0.25	0.15
1GPW_A:B_r	0.3	0.25	-0.05
1GPW_A:B_l	0.55	0.55	0
1K74_A:D_r	0.5	0.5	0
1K74_A:D_l	0.4	0.15	-0.25
1KTZ_A:B_r	0	0.1	0.1
1KTZ_A:B_l	0	0.5	0.5
1QA9_A:B_r	0.15	0.1	-0.05
1QA9_A:B_l	0.05	0.15	0.1
1S1Q_A:B_r	0.3	0.15	-0.15
1S1Q_A:B_l	0.35	0.4	0.05
1XD3_A:B_r	0.15	0.6	0.45
1XD3_A:B_l	0.7	0.55	-0.15
1Z0K_A:B_r	0.35	0.45	0.1
1Z0K_A:B_l	0.4	0.45	0.05
1Z5Y_D:E_r	0	0.45	0.45
1Z5Y_D:E_l	0.2	0.45	0.25
1ZHI_A:B_r	0.05	0.05	0

2HQS_A:H_r	0.1	0.2	0.1
2HQS_A:H_l	0.25	0.7	0.45
2OOB_A:B_r	0.05	0.25	0.2
2OOB_A:B_l	0.3	0.4	0.1
1M10_A:B_r	0	0.25	0.25
1M10_A:B_l	0.25	0.05	-0.2
1NW9_B:A_r	0.15	0.1	-0.05
1NW9_B:A_l	0.5	0.25	-0.25
1GRN_A:B_r	0.5	0.6	0.1
1GRN_A:B_l	0.3	0.55	0.25
1HE8_B:A_r	0.25	0.25	0
1HE8_B:A_l	0	0	0
1WQ1_R:G_r	0.6	0.6	0
1WQ1_R:G_l	0.9	0.9	0
1XQS_A:C_r	0.05	0.65	0.6
1XQS_A:C_l	0	0.05	0.05
2CFH_A:C_r	0.1	0.45	0.35
2CFH_A:C_l	0.1	0	-0.1
2HRK_A:B_r	0.3	0.3	0
2HRK_A:B_l	0.1	0.35	0.25
2NZ8_A:B_r	0.5	0.6	0.1
2NZ8_A:B_l	0.25	0.65	0.4
1FQ1_A:B_r	0.1	0.15	0.05
1FQ1_A:B_l	0.1	0.4	0.3
1BKD_R:S_r	0.6	0.55	-0.05
1BKD_R:S_l	0	0.5	0.5
1IRA_Y:X_r	0.15	0.05	-0.1
1IRA_Y:X_l	0.35	0.2	-0.15
1JMO_A:H_r	0.1	0.05	-0.05
1JMO_A:H_l	0.1	0.45	0.35
1R8S_A:E_r	0.55	0.55	0
1R8S_A:E_l	0.6	0.85	0.25
2OT3_B:A_r	0.15	0.5	0.35
2OT3_B:A_l	0.45	0.5	0.05
1GXG_A:C_r	0	0	0
1GXG_A:C_l	0.25	0.1	-0.15
1OC0_A:B_r	0	0.05	0.05
1OC0_A:B_l	0.4	0.5	0.1
2J0T_A:D_r	0.15	0.5	0.35
2J0T_A:D_l	0.2	0.35	0.15
1FFW_A:B_r	0.1	0.1	0
1FFW_A:B_l	0.35	0.55	0.2
1H9D_A:B_r	0.1	0	-0.1
1H9D_A:B_l	0.45	0.5	0.05
1PVH_A:B_r	0	0	0

1PVH_A:B_l	0.05	0.15	0.1
1ZHH_A:B_r	0.05	0.05	0
1ZHH_A:B_l	0.2	0.15	-0.05
2A5T_A:B_r	0.15	0.2	0.05
2A5T_A:B_l	0.05	0.2	0.15
2FJU_B:A_r	0	0	0
2FJU_B:A_l	0.35	0.35	0
1JIW_P:I_r	0.3	0.2	-0.1
1JIW_P:I_l	0.05	0.3	0.25
1MQ8_A:B_r	0	0.15	0.15
1MQ8_A:B_l	0	0.25	0.25
1R6Q_A:C_r	0.35	0.4	0.05
1R6Q_A:C_l	0.15	0.3	0.15
1SYX_A:B_r	0.25	0.4	0.15
1SYX_A:B_l	0.45	0.45	0
2AYO_A:B_r	0.1	0.55	0.45
2AYO_A:B_l	0.55	0.75	0.2
2J7P_A:D_r	0.45	0.45	0
2J7P_A:D_l	0.15	0.45	0.3
3CPH_G:A_r	0.25	0.45	0.2
3CPH_G:A_l	0.2	0.65	0.45
1F6M_A:C_r	0.1	0.25	0.15
1F6M_A:C_l	0.55	0.5	-0.05
2O3B_A:B_r	0.2	0.45	0.25
2O3B_A:B_l	0.45	0.45	0
1JK9_B:A_r	0.55	0.65	0.1
1JK9_B:A_l	0.4	0.45	0.05
2I9B_E:A_r	0.15	0.15	0
2I9B_E:A_l	0.15	0.3	0.15
1GGR_A:B_r	0.05	0.3	0.25
1GGR_A:B_l	0.35	0.45	0.1
1J6T_A:B_r	0.1	0.45	0.35
1J6T_A:B_l	0.35	0.45	0.1
1O2F_A:B_r	0	0.25	0.25
1O2F_A:B_l	0.3	0.55	0.25
1P9D_S:U_r	0.35	0.65	0.3
1P9D_S:U_l	0.65	0.65	0
1UR6_A:B_r	0.05	0.2	0.15
1UR6_A:B_l	0.6	0.6	0
3EZA_A:B_r	0	0	0
3EZA_A:B_l	0.55	0.65	0.1

Table A-5: Overview of the intra-species binary protein complexes of Benchmark4.0.

Complex function	Complex	Receptor	Ligand
Enzyme/electron transport protein	1E6E_A:B	Adrenoxin reductase	Adrenoxin
Enzyme/electron transport protein	1EWY_A:C	Ferredoxin reductase	Ferredoxin
Enzyme/electron transport protein	2O8V_A:B	PAPS reductase	Thioredoxin
Enzyme/electron transport protein	2PCC_A:B	Cyt C peroxidase	Cytochrome C
Enzyme/enzyme inhibitor protein	7CEI_A:B	Colicin E7 nuclease	Im7 immunity protein
Receptor inhibitor/Receptor	1B6C_A:B	FKBP binding protein	TGFbeta receptor
Enzyme/enzyme regulatory protein	1BUH_A:B	CDK2 kinase	Ckshs1
Components of the enzyme complex NADPH oxidase	1E96_A:B	Rac GTPase	p67 Phox
Signal transducer protein/Signal transducer inhibitor protein	1FQJ_A:B	Gt-alpha	RGS9
Enzyme/Non-competitive enzyme inhibitor	1GLA_G:F	Glycerol Kinase	Glucose specific phosphocarrier
Molecular bienzyme complex's subunits	1GPW_A:B	HISF protein	Amidotransferase HISH
Hetero-dimeric receptors of transcriptional factor complex ARF6	1K74_AB:DE	RXR-alpha	PPAR-gamma
Cytokine (signalling) protein/Receptor	1KTZ_A:B	TGF-beta	TGF-beta receptor
Receptor/ligand interaction	1QA9_A:B	CD2	CD58
Protein transport complex	1S1Q_A:B	UEV domain	Ubiquitin
Enzyme/Enzyme substrate	1XD3_A:B	UCH-L3	Ubiquitin
Protein transport enzyme (PTE)/PTE effector protein	1Z0K_A:B	Rab4A GTPase	RAB4 binding domain of Rabenosyn
Electron transport protein/enzyme	1Z5Y_D:E	N-term of DsbD	E.coli CCMG protein
Chromatin silencing proteins	1ZHI_A:B	BAH domain of Orc1	Sir Orc-interaction domain
Maintenance of bacterial outer membrane integrity	2HQS_A:H	TolB	Pal
Ubquitin-binding enzyme (UBE) /Ubiquitin (promoter of protein dimerization of UBE	2OOB_A:B	Ubiquitin ligase	Ubiquitin

and hence UBE biological activity)			
Cause reduction in platelet velocity at vascular damaged areas and are important in haemostasis and thrombosis	1M10_A:B	Von Willebrand Factor dom. A1	Glycoprotein IB-alpha
Enzyme/enzyme inhibitor protein	1NW9_B:A	Capase-9	BIR3-XIAP
Enzyme/enzyme activating protein	1GRN_A:B *	CDC42 GTPase	CDC42 GAP
Enzyme activating protein/activated enzyme	1HE8_B:A	Ras GTPase	PIP3 kinase
Enzyme/enzyme regulatory protein (inactivates enzyme)	1WQ1_R:G *	Ras GTPase	Ras GAP
Protein chaperone (PC)/Nucleotide exchange protein which inhibits PC nucleotide affinity	1XQS_A:C	HspBP1	Hsp70 ATPase domain
Core sub-complex component of the transport protein particle (TRAPP) complex	2CFH_A:C	BET3	TPC6
Components of the aminoacyl-tRNA synthetase (aaRS) protein complex	2HRK_A:B	Glutamyl-t-RNA synthetase	GU-4 nucleic binding protein
Enzyme/enyme activator protein (activates the enzyme by the exchange of bound GDP for GTP)	2NZ8_A:B	Rac GTPase	DH/PH domain of TRIO
Enzyme/enzyme inhibitor (inactivates enzyme)	1FQ1_A:B	CDK2 kinase	CDK inhibitor 3

Enzyme/enzyme inhibitor	1PXV_A:C	Cystein protease	Cystein protease inhibitor
Enzyme/ Nucleotide exchange protein (allows nucleotide exchange for the enzyme)	1BKD_R:S	Ras GTPase	Son of Sevenless
Receptor/receptor protein antagonist (inhibits Il-1 by binding to the receptor)	1IRA_Y:X	Interleukin-1 receptor	Interleukin-1 receptor antagonist protein
Thrombin protease inhibitor/Protease	1JMO_A:HL	Heparin cofactor	Thrombin
GTP-binding protein (enzyme)/guanine-nucleotide exchange factor	1R8S_A:E	Arf1 GTPase	Sec 7 domain
GTP-binding protein (enzyme)/nucleotide exchange factor	2OT3_B:A	Rab21 GTPase	Rabex-5 VPS9
Enzyme/enzyme inhibitor	1GXD_A:C	proMMP2 type IV collagenase	Metalloproteinase inhibitor 2
Protein (PAI-1)/ PAI-1 activator protein (causes fibrinolysis inhibition when PAI-1 is active)	1OC0_A:B	Plasminogen activator inhibitor-1	Vitronectin Somatomedin B domain
Enzyme/enzyme inhibitor	2J0T_A:D	MMP1 Intersitial collagenase	Metalloproteinase inhibitor 1
Sensory signal transmission proteins (chemoreceptors to the flagellar motors transmission)	1FFW_A:B	Chemotaxis protein CheY	Chemotaxis protein CheA
Components of heterodimeric transcription factor known as core binding factors	1H9D_A:B	Runx1 domain of CBFa1	Dimerisation domain of CBF-beta
Receptor/receptor protein (cytokine)	1PVH_A:B	IL6 receptor beta chain D2-D3 domains	Leukemia inhibitory factor
Signal transduction proteins in the quorum sensing communication process: periplasmic receptor/ inner membrane sensor protein	1ZHH_A:B	Autoinducer 2-binding periplasmic protein LuxP	Autoinducer 2 sensor kinase/phosphatase LuxQ
Components of NMDA (N-methyl-D-aspartate)	2A5T_A:B	NMDA receptor R1-4A subunit ligand-	NMDA receptor R2A subunit

receptor		binding core	ligand-binding core
Participate in signalling cascade: Enzyme/ Protein activator of enzyme	2FJU_A:B	Phospolipase beta 2	Rac GTPase
Enzyme/enzyme inhibitor	1JIW_P:I	Alkaline metalloproteinase	Proteinase inhibitor
Intercellular adhesion proteins	1MQ8_A:B	ICAM-1 domain 1-2	Integrin a-L I domain
Enzyme/activity modulator of the enzyme	1R6Q_A:C	Clp protease subunit ClpA	Clp protease adaptor protein ClpS
Components of the spliceosome	1SYX_A:B	Spliceosomal U5 15 kDa protein	CD2 receptor binding protein 2 C-ter fragment
Deubiquitinating enzyme/ubiquitin	2AYO_A:B	Ubiquitin carboxyl-terminal hydrolase 14	Ubiquitin
GTPases sub-units of the signal recognition particle co-translational targeting complex	2J7P_A:D	SRP GTPase Ffh	Cell division protein FtsY
Rab small GTPase/GDP/GTP exchange reaction regulator of Rab	3CPH_A:G	Ras-related protein Sec4	Rab GDP-dissociation inhibitor
Redox reaction proteins protein	1F6M_A:C	Thioredoxin reductase	Thioredoxin 1
Enzyme/enzyme inhibitor	2O3B_A:B	NucA nuclease	NuiA nuclease inhibitor
Metallochaperone (trafficking factors), which delivers copper co-factor to activate the enzyme (i.e. apoenzyme)	1JK9_A:B	CCS metallochaperone	SOD1 superoxide dismutase
Receptor/receptor protein	2I9B_A:E	Urokinase plasminogen activator surface receptor	Urokinase-type plasminogen activator
Signal transduction proteins involved in the phosphoenolpyruvate sugar phosphotransferase system (PTS) signal transduction pathway	1GGR	Glucose-specific phosphotransferase enzyme IIA component	Phosphocarrier protein HPr
Signal transduction proteins involved in the phosphoenolpyruvate sugar phosphotransferase	1J6T	PTS system mannitol-specific EIICBA component	Phosphocarrier protein HPr

system (PTS) signal transduction pathway			
Signal transduction proteins involved in the phosphoenolpyruvate sugar phosphotransferase system (PTS) signal transduction pathway	1O2F	Glucose-specific phosphotransferase enzyme IIA component	PTS system glucose-specific EIICB component
Proteosomal subunit/modulator of subunit	1P9D	26S proteasome non-ATPase regulatory subunit 4	UV excision repair protein RAD23 homolog A
Proteins involved in the ubiquitination pathway	1UR6	Ubiquitin-conjugating enzyme E2 D2	CCR4-NOT transcription complex subunit 4
Signal transduction proteins involved in the phosphoenolpyruvate sugar phosphotransferase system (PTS) signal transduction pathway	3EZA	Phosphoenolpyruvate-protein phosphotransferase	Phosphocarrier protein HPr

Table A-6: PROTIN_ID default

Complex	TP count	FP count	Cluster count	Total Interface_res	Surface only_res	Total Surface	Cluster TP_frac	Cluster FP_frac	TP rate	FP rate	Specificity	Accuracy	F-measure	MCC
1A0O_A:B_r	0	10	10	12	76	88	0.00	1.00	0.00	0.13	0.00	0.75	0.00	-0.14
1A0O_A:B_l	8	7	15	14	43	57	0.53	0.47	0.57	0.16	0.53	0.77	0.55	0.40
1BRS_A:D_r	3	7	10	20	66	86	0.30	0.70	0.15	0.11	0.30	0.72	0.20	0.06
1BRS_A:D_l	13	5	18	16	51	67	0.72	0.28	0.81	0.10	0.72	0.88	0.76	0.69
2PTC_E:I_r	8	3	11	20	124	144	0.73	0.27	0.40	0.02	0.73	0.90	0.52	0.49
2PTC_E:I_l	10	10	20	13	35	48	0.50	0.50	0.77	0.29	0.50	0.73	0.61	0.44
1FIN_A:B_r	0	5	5	41	160	201	0.00	1.00	0.00	0.03	0.00	0.77	0.00	-0.08
1FIN_A:B_l	5	0	5	27	131	158	1.00	0.00	0.19	0.00	1.00	0.86	0.31	0.40
1SPB_S:P_r	4	1	5	33	129	162	0.80	0.20	0.12	0.01	0.80	0.81	0.21	0.26
1SPB_S:P_l	5	13	18	17	42	59	0.28	0.72	0.29	0.31	0.28	0.58	0.29	-0.02
1E6E_A:B_r	0	7	7	26	279	305	0.00	1.00	0.00	0.03	0.00	0.89	0.00	-0.05
1E6E_A:B_l	12	4	16	24	55	79	0.75	0.25	0.50	0.07	0.75	0.80	0.60	0.49
1EWY_A:C_r	4	6	10	18	177	195	0.40	0.60	0.22	0.03	0.40	0.90	0.29	0.25
1EWY_A:C_l	3	7	10	16	56	72	0.30	0.70	0.19	0.13	0.30	0.72	0.23	0.08
7CEI_A:B_r	0	12	12	19	52	71	0.00	1.00	0.00	0.23	0.00	0.56	0.00	-0.27
7CEI_A:B_l	0	8	8	17	81	98	0.00	1.00	0.00	0.10	0.00	0.74	0.00	-0.14
2PCC_A:B_r	0	5	5	15	177	192	0.00	1.00	0.00	0.03	0.00	0.90	0.00	-0.05
2PCC_A:B_l	1	4	5	16	63	79	0.20	0.80	0.06	0.06	0.20	0.76	0.10	0.00
1GLA_G:F_r	0	4	4	15	250	265	0.00	1.00	0.00	0.02	0.00	0.93	0.00	-0.03
1GLA_G:F_l	2	5	7	16	82	98	0.29	0.71	0.13	0.06	0.29	0.81	0.17	0.09
1WQ1_R:G_r	11	5	16	27	86	113	0.69	0.31	0.41	0.06	0.69	0.81	0.51	0.43
1WQ1_R:G_l	10	0	10	32	182	214	1.00	0.00	0.31	0.00	1.00	0.90	0.48	0.53
1FQ1_A:B_r	2	5	7	16	179	195	0.29	0.71	0.13	0.03	0.29	0.90	0.17	0.14
1FQ1_A:B_l	7	4	11	19	93	112	0.64	0.36	0.37	0.04	0.64	0.86	0.47	0.41

1BXI_A:B_r	0	10	10	18	78	96	0.00	1.00	0.00	0.13	0.00	0.71	0.00	-0.16
1BXI_A:B_l	7	1	8	21	45	66	0.88	0.13	0.33	0.02	0.88	0.77	0.48	0.44

Table A-7: CCRXP

Complex	TP count	FP count	Cluster count	Total Interface_res	Surface only_res	Total Surface	Cluster TP_frac	Cluster FP_frac	TP rate	FP rate	Specificity	Accuracy	F-measure	MCC
1A0O_A:B_r	0	5	5	12	76	88	0.00	1.00	0.00	0.07	0.00	0.81	0.00	-0.10
1A0O_A:B_l	9	3	12	14	43	57	0.75	0.25	0.64	0.07	0.75	0.86	0.69	0.61
1BRS_A:D_r	0	7	7	20	66	86	0.00	1.00	0.00	0.11	0.00	0.69	0.00	-0.16
1BRS_A:D_l	2	1	3	16	51	67	0.67	0.33	0.13	0.02	0.67	0.78	0.21	0.22
2PTC_E:I_r	18	33	51	20	124	144	0.35	0.65	0.90	0.27	0.35	0.76	0.51	0.46
2PTC_E:I_l	7	2	9	13	35	48	0.78	0.22	0.54	0.06	0.78	0.83	0.64	0.55
1FIN_A:B_r	18	36	54	41	160	201	0.33	0.67	0.44	0.23	0.33	0.71	0.38	0.19
1FIN_A:B_l	0	9	9	27	131	158	0.00	1.00	0.00	0.07	0.00	0.77	0.00	-0.11
1SPB_S:P_r	11	9	20	33	129	162	0.55	0.45	0.33	0.07	0.55	0.81	0.42	0.32
1SPB_S:P_l	3	5	8	17	42	59	0.38	0.63	0.18	0.12	0.38	0.68	0.24	0.08
1E6E_A:B_r	1	10	11	26	279	305	0.09	0.91	0.04	0.04	0.09	0.89	0.05	0.00
1E6E_A:B_l	4	13	17	24	55	79	0.24	0.76	0.17	0.24	0.24	0.58	0.20	-0.08
1EWY_A:C_r	1	16	17	18	177	195	0.06	0.94	0.06	0.09	0.06	0.83	0.06	-0.04
1EWY_A:C_l	11	10	21	16	56	72	0.52	0.48	0.69	0.18	0.52	0.79	0.59	0.47
7CEI_A:B_r	5	1	6	19	52	71	0.83	0.17	0.26	0.02	0.83	0.79	0.40	0.39
7CEI_A:B_l	0	4	4	17	81	98	0.00	1.00	0.00	0.05	0.00	0.79	0.00	-0.09
2PCC_A:B_r	0	7	7	15	177	192	0.00	1.00	0.00	0.04	0.00	0.89	0.00	-0.06
2PCC_A:B_l	1	10	11	16	63	79	0.09	0.91	0.06	0.16	0.09	0.68	0.07	-0.11
1GLA_G:F_r	0	194	194	15	250	265	0.00	1.00	0.00	0.78	0.00	0.21	0.00	-0.40
1GLA_G:F_l	4	10	14	16	82	98	0.29	0.71	0.25	0.12	0.29	0.78	0.27	0.14
1WQ1_R:G_r	19	39	58	27	86	113	0.33	0.67	0.70	0.45	0.33	0.58	0.45	0.21
1WQ1_R:G_l	3	3	6	32	182	214	0.50	0.50	0.09	0.02	0.50	0.85	0.16	0.17
1FQ1_A:B_r	14	38	52	16	179	195	0.27	0.73	0.88	0.21	0.27	0.79	0.41	0.41
1FQ1_A:B_l	0	5	5	19	93	112	0.00	1.00	0.00	0.05	0.00	0.79	0.00	-0.10

1BXI_A:B_r	0	6	6	18	78	96	0.00	1.00	0.00	0.08	0.00	0.75	0.00	-0.12
1BXI_A:B_l	5	1	6	21	45	66	0.83	0.17	0.24	0.02	0.83	0.74	0.37	0.35

Table A-8: WHISCY default

Complex	TP count	FP count	Cluster count	Total Interface_res	Surface only_res	Total Surface	Cluster TP_frac	Cluster FP_frac	TP rate	FP rate	Specificity	Accuracy	F-measure	MCC
1A0O_A:B_r	0	13	13	12	76	88	0.00	1.00	0.00	0.17	0.00	0.72	0.00	-0.17
1A0O_A:B_l	1	1	2	14	43	57	0.50	0.50	0.07	0.02	0.50	0.75	0.13	0.11
1BRS_A:D_r	2	1	3	20	66	86	0.67	0.33	0.10	0.02	0.67	0.78	0.17	0.20
1BRS_A:D_l	7	0	7	16	51	67	1.00	0.00	0.44	0.00	1.00	0.87	0.61	0.61
2PTC_E:I_r	6	2	8	20	124	144	0.75	0.25	0.30	0.02	0.75	0.89	0.43	0.43
2PTC_E:I_l	1	1	2	13	35	48	0.50	0.50	0.08	0.03	0.50	0.73	0.13	0.11
1FIN_A:B_r	14	24	38	41	160	201	0.37	0.63	0.34	0.15	0.37	0.75	0.35	0.20
1FIN_A:B_l	12	13	25	27	131	158	0.48	0.52	0.44	0.10	0.48	0.82	0.46	0.36
1SPB_S:P_r	5	1	6	33	129	162	0.83	0.17	0.15	0.01	0.83	0.82	0.26	0.31
1SPB_S:P_l	5	0	5	17	42	59	1.00	0.00	0.29	0.00	1.00	0.80	0.45	0.48
1E6E_A:B_r	11	29	40	26	279	305	0.28	0.73	0.42	0.10	0.28	0.86	0.33	0.26
1E6E_A:B_l	16	5	21	24	55	79	0.76	0.24	0.67	0.09	0.76	0.84	0.71	0.60
1EWY_A:C_r	8	18	26	18	177	195	0.31	0.69	0.44	0.10	0.31	0.86	0.36	0.29
1EWY_A:C_l	6	5	11	16	56	72	0.55	0.45	0.38	0.09	0.55	0.79	0.44	0.33
7CEI_A:B_r	1	1	2	19	52	71	0.50	0.50	0.05	0.02	0.50	0.73	0.10	0.09
7CEI_A:B_l	0	2	2	17	81	98	0.00	1.00	0.00	0.02	0.00	0.81	0.00	-0.07
2PCC_A:B_r	0	10	10	15	177	192	0.00	1.00	0.00	0.06	0.00	0.87	0.00	-0.07
2PCC_A:B_l	3	11	14	16	63	79	0.21	0.79	0.19	0.17	0.21	0.70	0.20	0.01
1GLA_G:F_r	0	7	7	15	250	265	0.00	1.00	0.00	0.03	0.00	0.92	0.00	-0.04
1GLA_G:F_l	6	2	8	16	82	98	0.75	0.25	0.38	0.02	0.75	0.88	0.50	0.47
1WQ1_R:G_r	6	3	9	27	86	113	0.67	0.33	0.22	0.03	0.67	0.79	0.33	0.30
1WQ1_R:G_l	18	7	25	32	182	214	0.72	0.28	0.56	0.04	0.72	0.90	0.63	0.58
1FQ1_A:B_r	7	27	34	16	179	195	0.21	0.79	0.44	0.15	0.21	0.82	0.28	0.21
1FQ1_A:B_l	2	4	6	19	93	112	0.33	0.67	0.11	0.04	0.33	0.81	0.16	0.10

1BXI_A:B_r	0	5	5	18	78	96	0.00	1.00	0.00	0.06	0.00	0.76	0.00	-0.11
1BXI_A:B_l	0	3	3	21	45	66	0.00	1.00	0.00	0.07	0.00	0.64	0.00	-0.15

Table A-9: PROTIN_ID HSSP

Complex	TP count	FP count	Cluster count	Total Interface_res	Surface only_res	Total Surface	Cluster TP_frac	Cluster FP_frac	TP rate	FP rate	Specificity	Accuracy	F-measure	MCC
1A0O_A:B_r	0	10	10	12	76	88	0.00	1.00	0.00	0.13	0.00	0.75	0.00	-0.14
1A0O_A:B_l	8	4	12	14	43	57	0.67	0.33	0.57	0.09	0.67	0.82	0.62	0.51
1BRS_A:D_r	5	4	9	20	66	86	0.56	0.44	0.25	0.06	0.56	0.78	0.34	0.26
1BRS_A:D_l	11	5	16	16	51	67	0.69	0.31	0.69	0.10	0.69	0.85	0.69	0.59
2PTC_E:I_r	8	1	9	20	124	144	0.89	0.11	0.40	0.01	0.89	0.91	0.55	0.56
2PTC_E:I_l	5	14	19	13	34	47	0.26	0.74	0.38	0.41	0.26	0.53	0.31	-0.02
1FIN_A:B_r	0	10	10	41	160	201	0.00	1.00	0.00	0.06	0.00	0.75	0.00	-0.12
1FIN_A:B_l	0	5	5	27	131	158	0.00	1.00	0.00	0.04	0.00	0.80	0.00	-0.08
1SPB_S:P_r	4	2	6	33	129	162	0.67	0.33	0.12	0.02	0.67	0.81	0.21	0.23
1SPB_S:P_l	4	14	18	17	42	59	0.22	0.78	0.24	0.33	0.22	0.54	0.23	-0.10
1E6E_A:B_r	0	7	7	26	279	305	0.00	1.00	0.00	0.03	0.00	0.89	0.00	-0.05
1E6E_A:B_l	12	4	16	24	55	79	0.75	0.25	0.50	0.07	0.75	0.80	0.60	0.49
1EWY_A:C_r	0	10	10	18	177	195	0.00	1.00	0.00	0.06	0.00	0.86	0.00	-0.07
1EWY_A:C_l	2	7	9	16	57	73	0.22	0.78	0.13	0.12	0.22	0.71	0.16	0.00
7CEI_A:B_r	0	12	12	19	52	71	0.00	1.00	0.00	0.23	0.00	0.56	0.00	-0.27
7CEI_A:B_l	0	4	4	17	81	98	0.00	1.00	0.00	0.05	0.00	0.79	0.00	-0.09
2PCC_A:B_r	0	8	8	15	178	193	0.00	1.00	0.00	0.04	0.00	0.88	0.00	-0.06
2PCC_A:B_l	2	11	13	16	65	81	0.15	0.85	0.13	0.17	0.15	0.69	0.14	-0.05
1GLA_G:F_r	0	3	3	15	251	266	0.00	1.00	0.00	0.01	0.00	0.93	0.00	-0.03
1GLA_G:F_l	2	6	8	16	82	98	0.25	0.75	0.13	0.07	0.25	0.80	0.17	0.07
1WQ1_R:G_r	9	3	12	27	86	113	0.75	0.25	0.33	0.03	0.75	0.81	0.46	0.41
1WQ1_R:G_l	11	0	11	32	182	214	1.00	0.00	0.34	0.00	1.00	0.90	0.51	0.56
1FQ1_A:B_r	0	16	16	16	179	195	0.00	1.00	0.00	0.09	0.00	0.84	0.00	-0.09
1FQ1_A:B_l	6	5	11	19	93	112	0.55	0.45	0.32	0.05	0.55	0.84	0.40	0.33

1BXI_A:B_r	0	3	3	18	78	96	0.00	1.00	0.00	0.04	0.00	0.78	0.00	-0.09
1BXI_A:B_l	7	1	8	21	45	66	0.88	0.13	0.33	0.02	0.88	0.77	0.48	0.44

Table A-10: WHICSY UniRef90

Complex	TP count	FP count	Cluster count	Total Interface_res	Surface only_res	Total Surface	Cluster TP_frac	Cluster FP_frac	TP rate	FP rate	Specificity	Accuracy	F-measure	MCC
1A0O_A:B_r	1	15	16	12	76	88	0.06	0.94	0.08	0.20	0.06	0.70	0.07	-0.10
1A0O_A:B_l	0	0	0	14	43	57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1BRS_A:D_r	1	2	3	20	66	86	0.33	0.67	0.05	0.03	0.33	0.76	0.09	0.05
1BRS_A:D_l	8	0	8	16	51	67	1.00	0.00	0.50	0.00	1.00	0.88	0.67	0.66
2PTC_E:I_r	6	1	7	20	124	144	0.86	0.14	0.30	0.01	0.86	0.90	0.44	0.47
2PTC_E:I_l	0	0	0	13	35	48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1FIN_A:B_r	11	23	34	41	160	201	0.32	0.68	0.27	0.14	0.32	0.74	0.29	0.13
1FIN_A:B_l	12	19	31	27	131	158	0.39	0.61	0.44	0.15	0.39	0.78	0.41	0.28
1SPB_S:P_r	4	2	6	33	129	162	0.67	0.33	0.12	0.02	0.67	0.81	0.21	0.23
1SPB_S:P_l	3	0	3	17	42	59	1.00	0.00	0.18	0.00	1.00	0.76	0.30	0.36
1E6E_A:B_r	10	36	46	26	279	305	0.22	0.78	0.38	0.13	0.22	0.83	0.28	0.20
1E6E_A:B_l	17	6	23	24	55	79	0.74	0.26	0.71	0.11	0.74	0.84	0.72	0.61
1EWY_A:C_r	8	18	26	18	177	195	0.31	0.69	0.44	0.10	0.31	0.86	0.36	0.29
1EWY_A:C_l	4	4	8	16	56	72	0.50	0.50	0.25	0.07	0.50	0.78	0.33	0.24
7CEI_A:B_r	1	1	2	19	52	71	0.50	0.50	0.05	0.02	0.50	0.73	0.10	0.09
7CEI_A:B_l	0	1	1	17	81	98	0.00	1.00	0.00	0.01	0.00	0.82	0.00	-0.05
2PCC_A:B_r	1	15	16	15	177	192	0.06	0.94	0.07	0.08	0.06	0.85	0.06	-0.02
2PCC_A:B_l	3	10	13	16	63	79	0.23	0.77	0.19	0.16	0.23	0.71	0.21	0.03
1GLA_G:F_r	0	15	15	15	250	265	0.00	1.00	0.00	0.06	0.00	0.89	0.00	-0.06
1GLA_G:F_l	8	3	11	16	82	98	0.73	0.27	0.50	0.04	0.73	0.89	0.59	0.54
1WQ1_R:G_r	15	8	23	27	86	113	0.65	0.35	0.56	0.09	0.65	0.82	0.60	0.49
1WQ1_R:G_l	17	4	21	32	182	214	0.81	0.19	0.53	0.02	0.81	0.91	0.64	0.61
1FQ1_A:B_r	6	26	32	16	179	195	0.19	0.81	0.38	0.15	0.19	0.82	0.25	0.17
1FQ1_A:B_l	2	6	8	19	93	112	0.25	0.75	0.11	0.06	0.25	0.79	0.15	0.06

1BXI_A:B_r	0	1	1	18	78	96	0.00	1.00	0.00	0.01	0.00	0.80	0.00	-0.05
1BXI_A:B_l	0	0	0	21	45	66	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

References

AHMAD, S., KESKIN, O., MIZUGUCHI, K., SARAI, A. & NUSSINOV, R. 2010. CCRXP: exploring clusters of conserved residues in protein structures. *Nucleic Acids Research,* 38**,** W398-W401.

ALBER, F., FORSTER, F., KORKIN, D., TOPF, M. & SALI, A. 2008. Integrating diverse data for structure determination of macromolecular assemblies. *Annual Review of Biochemistry.*

ALEKSIC, M., DUSHEK, O., ZHANG, H., SHENDEROV, E., CHEN, J.-L., CERUNDOLO, V., COOMBS, D. & VAN DER MERWE, P. A. 2010. Dependence of T Cell Antigen Recognition on T Cell Receptor-Peptide MHC Confinement Time. *Immunity,* 32**,** 163-174.

ALOY, P. & RUSSELL, R. B. 2004. Ten thousand interactions for the molecular biologist. *Nature Biotechnology,* 22**,** 1317-1321.

ALOY, P., QUEROL, E., AVILES, F. X. & STERNBERG, M. J. E. 2001. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *Journal of Molecular Biology,* 311**,** 395-408.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology,* 215**,** 403-410.

ANIBA, M. R., POCH, O. & THOMPSON, J. D. 2010. Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Research,* 38**,** 7353-7363.

APWEILER, R., MARTIN, M. J., O'DONOVAN, C., MAGRANE, M., ALAM-FARUQUE, Y., ANTUNES, R., BARRELL, D., BELY, B., BINGLEY, M., BINNS, D., BOWER, L., BROWNE, P., CHAN, W. M., DIMMER, E., EBERHARDT, R., FAZZINI, F., FEDOTOV, A., FOULGER, R., GARAVELLI, J., CASTRO, L. G., HUNTLEY, R., JACOBSEN, J., KLEEN, M., LAIHO, K., LEGGE, D., LIN, Q. A., LIU, W. D., LUO, J., ORCHARD, S., PATIENT, S., PICHLER, K., POGGIOLI, D., PONTIKOS, N., PRUESS, M., ROSANOFF, S., SAWFORD, T., SEHRA, H., TURNER, E., CORBETT, M., DONNELLY, M., VAN RENSBURG, P., XENARIOS, I., BOUGUELERET, L., AUCHINCLOSS, A., ARGOUD-PUY, G.,

AXELSEN, K., BAIROCH, A., BARATIN, D., BLATTER, M. C., BOECKMANN, B., BOLLEMAN, J., BOLLONDI, L., BOUTET, E., QUINTAJE, S. B., BREUZA, L., BRIDGE, A., DECASTRO, E., COUDERT, E., CUSIN, I., DOCHE, M., DORNEVIL, D., DUVAUD, S., ESTREICHER, A., FAMIGLIETTI, L., FEUERMANN, M., GEHANT, S., FERRO, S., GASTEIGER, E., GATEAU, A., GERRITSEN, V., GOS, A., GRUAZ-GUMOWSKI, N., HINZ, U., HULO, C., HULO, N., JAMES, J., JIMENEZ, S., JUNGO, F., KAPPLER, T., KELLER, G., LARA, V., LEMEREIER, P., LIEBERHERR, D., MARTIN, X., MASSON, P., MOINAT, M., MORGAT, A., PAESANO, S., PEDRUZZI, I., PILBOUT, S., POUX, S., POZZATO, M., REDASCHI, N., RIVOIRE, C., ROECHERT, B., SCHNEIDER, M., SIGRIST, C., SONESSON, K., STAEHLI, S., STANLEY, E., et al. 2011. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research,* 39**,** D214-D219.

ARCHAKOV, A. I., GOVORUN, V. M., DUBANOV, A. V., IVANOV, Y. D., VESELOVSKY, A. V., LEWI, P. & JANSSEN, P. 2003. Protein-protein interactions as a target for drugs in proteomics. *Proteomics,* 3**,** 380-391.

ARUMUGAM, S. & VAN DOREN, S. R. 2003a. Global orientation of bound MMP-3 and N-TIMP-1 in solution via residual dipolar couplings. *Biochemistry,* 42**,** 7950-7958.

ARUMUGAM, S., GAO, G. H., PATTON, B. L., SEMENCHENKO, V., BREW, K. & VAN DOREN, S. R. 2003b. Increased backbone mobility in beta-barrel enhances entropy gain driving binding of N-TIMP-1 to MMP-3. *Journal of Molecular Biology,* 327**,** 719-734.

ARUMUGAM, S., HEMME, C. L., YOSHIDA, N., SUZUKI, K., NAGASE, H., BEJANSKII, M., WU, B. & VAN DOREN, S. R. 1998. TIMP-1 contact sites and perturbations of stromelysin 1 mapped by NMR and a paramagnetic surface probe. *Biochemistry,* 37**,** 9650-9657.

AUDIE, J. 2009. Development and validation of an empirical free energy function for calculating protein-protein binding free energy surfaces. *Biophysical Chemistry,* 139**,** 84-91.

AZIZ, M. M., MALEKI, M., RUEDA, L., RAZA, M. & BANERJEE, S. 2011. Prediction of biological protein-protein interactions using atom-type and amino acid properties. *Proteomics,* 11**,** 3802-3810.

BAHADUR, R. P., CHAKRABARTI, P., RODIER, F. & JANIN, J. 2003. Dissecting

subunit interfaces in homodimeric proteins. *Proteins-Structure Function and Genetics,* 53**,** 708-719.

BAHADUR, R. P., CHAKRABARTI, P., RODIER, F. & JANIN, J. 2004. A dissection of specific and non-specific protein - Protein interfaces. *Journal of Molecular Biology,* 336**,** 943-955.

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. F. & NIELSEN, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics,* 16**,** 412-424.

BARTON, G. J. 1993, 2002. OC - A cluster analysis program. Scotland, UK: University of Dundee.

BERA, I. & RAY, S. 2009. A study of interface roughness of heteromeric obligate and non-obligate protein-protein complexes. *Bioinformation,* 4**,** 210-5.

BERMAN, H. M., KLEYWEGT, G. J., NAKAMURA, H. & MARKLEY, J. L. 2013. How community has shaped the protein data bank. *Structure (London, England : 1993),* 21**,** 1485-91.

BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research,* 28**,** 235-242.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. 1977. Protein Data Bank - Computer-Based Archival File for Macromolecular Structures. *Journal of Molecular Biology,* 112**,** 535-542.

BISSANTZ, C., KUHN, B. & STAHL, M. 2010. A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry,* 53**,** 5061-5084.

BOGAN, A. A. & THORN, K. S. 1998. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology,* 280**,** 1-9.

BONETTA, L. 2010. Interactome under construction. *Nature,* 468**,** 851-854.

BONVIN, A. M. J. J. 2010. *Ambiguous Interaction Restraints (AIRs)* [online]. The Netherlands: Utrecht University. Available at: <http://www.nmr.chem.uu.nl/haddock/> [Accessed 1 June 2010].

BORDNER, A. J. & ABAGYAN, R. 2005. Statistical analysis and prediction of protein-protein interfaces. *Proteins-Structure Function and Bioinformatics,* 60**,** 353-366.

BOURQUARD, T., BERNAUER, J., AZE, J. & POUPON, A. 2011. A Collaborative Filtering Approach for Protein-Protein Docking Scoring Functions. *Plos One,* 6.

BRADFORD, J. R., NEEDHAM, C. J., BULPITT, A. J. & WESTHEAD, D. R. 2006. Insights into protein-protein interfaces using a Bayesian network prediction method. *Journal of Molecular Biology,* 362**,** 365-386.

BRADFORD, J. R. & WESTHEAD, D. R. 2005. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics,* 21**,** 1487-1494.

BRADY, G. P. & SHARP, K. A. 1997. Entropy in protein folding and in protein-protein interactions. *Current Opinion in Structural Biology,* 7**,** 215-221.

BROWN, N. P., LEROY, C. & SANDER, C. 1998. MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics,* 14**,** 380-381.

BURGOYNE, N. J. & JACKSON, R. M. 2006. Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics,* 22**,** 1335-1342.

CAFFREY, D. R., SOMAROO, S., HUGHES, J. D., MINTSERIS, J. & HUANG, E. S. 2004. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science,* 13**,** 190-202.

CAPRA, J. A. & SINGH, M. 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics,* 23**,** 1875-1882.

CARUGO, O. 2007. Detailed estimation of bioinformatics prediction reliability through the Fragmented Prediction Performance Plots. *Bmc Bioinformatics,* 8.

CAZALS, F., PROUST, F., BAHADUR, R. P. & JANIN, J. 2006. Revisiting the Voronoi description of protein-protein interfaces. *Protein Science,* 15**,** 2082-2092.

CHAKRABARTI, P. & JANIN, J. 2002. Dissecting protein-protein recognition sites. *Proteins-Structure Function and Genetics,* 47**,** 334-343.

CHELLIAH, V., BLUNDELL, T. L. & FERNANDEZ-RECIO, J. 2006. Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. *Journal of Molecular Biology,* 357**,** 1669-1682.

CHELLIAH, V., CHEN, L., BLUNDELL, T. L. & LOVELL, S. C. 2004. Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *Journal of Molecular Biology,* 342**,** 1487-1504.

CHEN, H. L. & 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**,** 21-35.

CHEN, R., MINTSERIS, J., JANIN, J. & WENG, Z. P. 2003a. A protein-protein docking benchmark. *Proteins-Structure Function and Genetics,* 52**,** 88-91.

CHEN, R., LI, L. & WENG, Z. P. 2003b. ZDOCK: An initial-stage protein-docking algorithm. *Proteins-Structure Function and Genetics,* 52**,** 80-87.

CHEN, R. & WENG, Z. P. 2002. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins-Structure Function and Genetics,* 47**,** 281-294.

CHOI, Y. S., YANG, J. S., CHOI, Y., RYU, S. H. & KIM, S. 2009. Evolutionary conservation in multiple faces of protein interaction. *Proteins-Structure Function and Bioinformatics,* 77**,** 14-25.

CHOTHIA, C. & JANIN, J. 1975. Principles of Protein-Protein Recognition. *Nature,* 256**,** 705-708.

CHUNG, J. L., WANG, W. & BOURNE, P. E. 2006. Exploiting sequence and structure homologs to identify protein-protein binding sites. *Proteins-Structure Function and Bioinformatics,* 62**,** 630-640.

CLACKSON, T. & WELLS, J. A. 1995. A Hot-Spot of Binding-Energy in a Hormone-Receptor Interface. *Science,* 267**,** 383-386.

CLORE, G. M. & SCHWIETERS, C. D. 2003. Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from H-1(N)/N-15 chemical shift mapping and backbone N-15-H-1 residual dipolar couplings using conjoined rigid body/torsion angle dynamics. *Journal of the American Chemical Society,* 125**,** 2902-2912.

COLE, C. & WARWICKER, J. 2002. Side-chain conformational entropy at protein-protein interfaces. *Protein Science,* 11**,** 2860-2870.

COLLINS, F. S., LANDER, E. S., ROGERS, J. & WATERSTON, R. H. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-945.

COMEAU, S. R., GATCHELL, D. W., VAJDA, S. & CAMACHO, C. J. 2004. ClusPro: An automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics,* 20**,** 45-50.

CORNILESCU, G., LEE, B. R., CORNILESCU, C. C., WANG, G. S., PETERKOFSKY, A. & CLORE, G. M. 2002. Solution structure of the phosphoryl transfer complex between the cytoplasmic A domain of the mannitol transporter IIMannitol and HPr of the Escherichia coli phoshotransferase system. *Journal of Biological Chemistry,* 277**,** 42289-42298.

DAYHOFF, M. O., SCHWARTZ, R. M. & ORCUTT, B. C. 1978. A model of evolutionary change in proteins. *Atlas of protein sequence and structure,* 5**,** 345-351.

DE, S., KRISHNADEV, O., SRINIVASAN, N. & REKHA, N. 2005. Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *Bmc Structural Biology,* 5.

DE VRIES, S. J. & BONVIN, A. 2011a. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. *Plos One,* 6.

DE VRIES, S. J. & BONVIN, A. M. J. J. 2011b. *HADDOCK web server tutorial* [online]. The Netherlands: Utrecht University. Available at: <http://www.wenmr.eu/wenmr/haddock-web-server-tutorial> and <http://haddock.chem.uu.nl/services/HADDOCK/haddockserver-demo.html> [Accessed 9 January 2012].

DE VRIES, S. J., VAN DIJK, M. & BONVIN, A. 2010. The HADDOCK web server for data-driven biomolecular docking. *Nature Protocols,* 5**,** 883-897.

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 & Peptide Science,* 9**,** 394-406.

DE VRIES, S. J., VAN DIJK, A. D. J. & BONVIN, A. 2006. WHISCY: What information does surface conservation yield? Application to data-driven docking. *Proteins-Structure Function and Bioinformatics,* 63**,** 479-489.

DE VRIES, S. J., VAN DIJK, A. D. J., KRZEMINSKI, M., VAN DIJK, M., THUREAU, A., HSU, V., WASSENAAR, T. & BONVIN, A. 2007. HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins-Structure Function and Bioinformatics,* 69**,** 726-733.

DELANO, W. L. 2002. The PyMOL Molecular Graphics System. USA: DeLano Scientific.

DEY, S., PAL, A., CHAKRABARTI, P. & JANIN, J. 2010. The Subunit Interfaces of Weakly Associated Homodimeric Proteins. *Journal of Molecular Biology,* 398**,** 146-160.

DOBBINS, S. E., LESK, V. I. & STERNBERG, M. J. E. 2008. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 10390-10395.

DOBRODUMOV, A. & GRONENBORN, A. M. 2003. Filtering and selection of structural

models: Combining docking and NMR. *Proteins-Structure Function and Genetics,* 53**,** 18-32.

DODGE, C., SCHNEIDER, R. & SANDER, C. 1998. The HSSP database of protein structure sequence alignments and family profiles. *Nucleic Acids Research,* 26**,** 313-315.

DOMINGUEZ, C., BOELENS, R. & BONVIN, A. 2003. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society,* 125**,** 1731-1737.

DOMINGUEZ, C., BONVIN, A., WINKLER, G. S., VAN SCHAIK, F. M. A., TIMMERS, H. T. M. & BOELENS, R. 2004. Structural model of the UbcH5B/CNOT4 complex revealed by combining NMR, mutagenesis, and docking approaches. *Structure,* 12**,** 633-644.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* 32**,** 1792-1797.

EZKURDIA, L., BARTOLI, L., FARISELLI, P., CASADIO, R., VALENCIA, A. & TRESS, M. L. 2009. Progress and challenges in predicting protein-protein interaction sites. *Briefings in Bioinformatics,* 10**,** 233-246.

FARISELLI, P., PAZOS, F., VALENCIA, A. & CASADIO, R. 2002. Prediction of protein-protein interaction sites in heterocomplexes with neural networks. *European Journal of Biochemistry,* 269**,** 1356-1361.

FAWCETT, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* 27**,** 861-874.

FERNANDEZ-RECIO, J. 2011. Prediction of protein binding sites and hot spots. *Wiley Interdisciplinary Reviews-Computational Molecular Science,* 1**,** 680-698.

FERNANDEZ-RECIO, J., TOTROV, M. & ABAGYAN, R. 2002. Soft protein-protein docking in internal coordinates. *Protein Science,* 11**,** 280-291.

FIORUCCI, S. & ZACHARIAS, M. 2010a. Prediction of Protein-Protein Interaction Sites Using Electrostatic Desolvation Profiles. *Biophysical Journal,* 98**,** 1921-1930.

FIORUCCI, S. & ZACHARIAS, M. 2010b. Binding site prediction and improved scoring during flexible protein-protein docking with ATTRACT. *Proteins-Structure Function and Bioinformatics,* 78**,** 3131-3139.

FISCHER, D., LIN, S. L., WOLFSON, H. L. & NUSSINOV, R. 1995. A Geometry-Based Suite of Molecular Docking Processes. *Journal of Molecular Biology,* 248**,** 459-

477.

FLEISHMAN, S. J., WHITEHEAD, T. A., STRAUCH, E. M., CORN, J. E., QIN, S. B., ZHOU, H. X., MITCHELL, J. C., DEMERDASH, O. N. A., TAKEDA-SHITAKA, M., TERASHI, G., MOAL, I. H., LI, X. F., BATES, P. A., ZACHARIAS, M., PARK, H., KO, J. S., LEE, H., SEOK, C., BOURQUARD, T., BERNAUER, J., POUPON, A., AZE, J., SONER, S., OVALI, S. K., OZBEK, P., BEN TAL, N., HALILOGLU, T., HWANG, H., VREVEN, T., PIERCE, B. G., WENG, Z. P., PEREZ-CANO, L., PONS, C., FERNANDEZ-RECIO, J., JIANG, F., YANG, F., GONG, X. Q., CAO, L. B., XU, X. J., LIU, B., WANG, P. W., LI, C. H., WANG, C. X., ROBERT, C. H., GUHAROY, M., LIU, S. Y., HUANG, Y. Y., LI, L., GUO, D. C., CHEN, Y., XIAO, Y., LONDON, N., ITZHAKI, Z., SCHUELER-FURMAN, O., INBAR, Y., POTAPOV, V., COHEN, M., SCHREIBER, G., TSUCHIYA, Y., KANAMORI, E., STANDLEY, D. M., NAKAMURA, H., KINOSHITA, K., DRIGGERS, C. M., HALL, R. G., MORGAN, J. L., HSU, V. L., ZHAN, J., YANG, Y. D., ZHOU, Y. Q., KASTRITIS, P. L., BONVIN, A., ZHANG, W. Y., CAMACHO, C. J., KILAMBI, K. P., SIRCAR, A., GRAY, J. J., OHUE, M., UCHIKOGA, N., MATSUZAKI, Y., ISHIDA, T., AKIYAMA, Y., KHASHAN, R., BUSH, S., FOUCHES, D., TROPSHA, A., ESQUIVEL-RODRIGUEZ, J., KIHARA, D., STRANGES, P. B., JACAK, R., KUHLMAN, B., HUANG, S. Y., ZOU, X. Q., WODAK, S. J., JANIN, J. & BAKER, D. 2011. Community-Wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology. *Journal of Molecular Biology,* 414**,** 289-302.

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**,** 106-120.

GAO, G. H., PRUTZMAN, K. C., KING, M. L., SCHESWOHL, D. M., DEROSE, E. F., LONDON, R. E., SCHALLER, M. D. & CAMPBELL, S. L. 2004. NMR solution structure of the focal adhesion targeting domain of focal adhesion kinase in complex with a paxillin LD peptide - Evidence for a two-site binding model. *Journal of Biological Chemistry,* 279**,** 8441-8451.

GOMISRUTH, F. X., MASKOS, K., BETZ, M., BERGNER, A., HUBER, R., SUZUKI, K., YOSHIDA, N., NAGASE, H., BREW, K., BOURENKOV, G. P., BARTUNIK, H. & BODE, W. 1997. Mechanism of inhibition of the human matrix metalloproteinase stromelysin-1 by TIMP-1. *Nature,* 389**,** 77-81.

GONG, S., PARK, C., CHOI, H., KO, J., JANG, I., LEE, J., BOLSER, D. M., OH, D., KIM, D. S. & BHAK, J. 2005. A protein domain interaction interface database: InterPare. *Bmc Bioinformatics,* 6.

GOODSELL, D. S. & OLSON, A. J. 2000. Structural symmetry and protein function. *Annual Review of Biophysics and Biomolecular Structure,* 29**,** 105-153.

GOTTSCHALK, K. E., NEUVIRTH, H. & SCHREIBER, G. 2004. A novel method for scoring of docked protein complexes using predicted binding sites protein-protein binding sites. *Protein Engineering Design & Selection,* 17**,** 183-189.

GRAPHPAD-SOFTWARE GraphPad Prism. USA: GraphPad Software.

GRAY, J. J. 2006. High-resolution protein-protein docking. *Current Opinion in Structural Biology,* 16**,** 183-193.

GRAY, J. J., MOUGHON, S., WANG, C., SCHUELER-FURMAN, O., KUHLMAN, B., ROHL, C. A. & BAKER, D. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology,* 331**,** 281-299.

GRISHIN, N. V. & PHILLIPS, M. A. 1994. The Subunit Interfaces of Oligomeric Enzymes Are Conserved to a Similar Extent to the Overall Protein Sequences. *Protein Science,* 3**,** 2455-2458.

GRUNBERG, R., LECKNER, J. & NILGES, M. 2004. Complementarity of structure ensembles in protein-protein binding. *Structure,* 12**,** 2125-2136.

GUHAROY, M. & CHAKRABARTI, P. 2005. Conservation and relative importance of residues across protein-protein interfaces. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 15447-15452.

GUHAROY, M. & CHAKRABARTI, P. 2007. Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein-protein interactions. *Bioinformatics,* 23**,** 1909-1918.

GUHAROY, M. & CHAKRABARTI, P. 2010. Conserved residue clusters at protein-protein interfaces and their use in binding site identification. *Bmc Bioinformatics,* 11.

HALPERIN, I., MA, B. Y., WOLFSON, H. & NUSSINOV, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins-Structure Function and Genetics,* 47**,** 409-443.

HARPAZ, Y., GERSTEIN, M. & CHOTHIA, C. 1994. Volume Changes on Protein-Folding. *Structure,* 2**,** 641-649.

HEIFETZ, A., KATCHALSKI-KATZIR, E. & EISENSTEIN, M. 2002. Electrostatics in protein-protein docking. *Protein Science,* 11**,** 571-587.

HENIKOFF, S. & HENIKOFF, J. G. 1994. Position-Based Sequence Weights. *Journal of Molecular Biology,* 243**,** 574-578.

HERSHKO, A., HELLER, H., ELIAS, S. & CIECHANOVER, A. 1983. Components of Ubiquitin-Protein Ligase System - Resolution, Affinity Purification, and Role in Protein Breakdown. *Journal of Biological Chemistry,* 258**,** 8206-8214.

HERSHKO, A., LESHINSKY, E., GANOTH, D. & HELLER, H. 1984. Atp-Dependent Degradation of Ubiquitin-Protein Conjugates. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences,* 81**,** 1619-1623.

HOSKINS, J., LOVELL, S. & BLUNDELL, T. L. 2006. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Science,* 15**,** 1017-1029.

HOVSEPIAN, D. M., ZIPORIN, S. J., SAKURAI, M. K., LEE, J. K., CURCI, J. A. & THOMPSON, R. W. 2000. Elevated plasma levels of matrix metalloproteinase-9 in patients with abdominal aortic aneurysms: A circulating marker of degenerative aneurysm disease. *Journal of Vascular and Interventional Radiology,* 11**,** 1345-1352.

HUANG, B. & SCHROEDER, M. 2008. Using protein binding site prediction to improve protein docking. *Gene,* 422**,** 14-21.

HUBBARD, S. J. & THORNTON, J. M. 1993. 'NACCESS', Computer Program. London: Department of Biochemistry and Molecular Biology, University College London.

HWANG, H., PIERCE, B., MINTSERIS, J., JANIN, J. & WENG, Z. 2008. Protein-protein docking benchmark version 3.0. *Proteins-Structure Function and Bioinformatics,* 73**,** 705-709.

HWANG, H., VREVEN, T., JANIN, J. & WENG, Z. 2010. Protein-protein docking benchmark version 4.0. *Proteins-Structure Function and Bioinformatics,* 78**,** 3111-3114.

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**,** 265-285.

JANIN, J. 1995. Principles of Protein-Protein Recognition from Structure to

Thermodynamics. *Biochimie,* 77**,** 497-505.

JANIN, J. 2009. Basic Principles of Protein–Protein Interaction. *In:* NUSSINOV, R. & SCHREIBER, G. (eds.) *Computational Protein-Protein Interactions.* Boca Raton: CRC Press.

JANIN, J. 2010. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Molecular Biosystems,* 6**,** 2351-2362.

JANIN, J., HENRICK, K., MOULT, J., TEN EYCK, L., STERNBERG, M. J. E., VAJDA, S., VASKER, I. & WODAK, S. J. 2003. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins-Structure Function and Bioinformatics,* 52**,** 2-9.

JANIN, J., RODIER, F., CHAKRABARTI, P. & BAHADUR, R. P. 2007. Macromolecular recognition in the Protein Data Bank. *Acta Crystallographica Section D-Biological Crystallography,* 63**,** 1-8.

JOHNSON, R. J., LAVIS, L. D. & RAINES, R. T. 2007. Intraspecies regulation of ribonucleolytic activity. *Biochemistry,* 46**,** 13131-13140.

JOHNSON, S. 1967. Hierarchical clustering schemes. *Psychometrika,* 32**,** 241-254.

JONES, S. & THORNTON, J. M. 1997a. Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology,* 272**,** 133-143.

JONES, S. & THORNTON, J. M. 1997b. Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology,* 272**,** 121-132.

JOOSTEN, R. P., BEEK, T., KRIEGER, E., HEKKELMAN, M. L., HOOFT, R. W. W., SCHNEIDER, R., SANDER, C. & VRIEND, G. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Research,* 39**,** D411-D419.

KABSCH, W. & SANDER, C. 1983. Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers,* 22**,** 2577-2637.

KARACA, E. & BONVIN, A. 2011. A Multidomain Flexible Docking Approach to Deal with Large Conformational Changes in the Modeling of Biomolecular Complexes. *Structure,* 19**,** 555-565.

KASTRITIS, P. L. & BONVIN, A. M. J. J. 2013. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of the Royal Society, Interface / the Royal Society,* 10**,** 20120835-20120835.

KASTRITIS, P. L., MOAL, I. H., HWANG, H., WENG, Z. P., BATES, P. A., BONVIN, A. & JANIN, J. 2011. A structure-based benchmark for protein-protein binding affinity. *Protein Science,* 20**,** 482-491.

KASTRITIS, P. L. & BONVIN, A. 2010. Are Scoring Functions in Protein-Protein Docking Ready To Predict Interactomes? Clues from a Novel Binding Affinity Benchmark. *Journal of Proteome Research,* 9**,** 2216-2225.

KATCHALSKIKATZIR, E., SHARIV, I., EISENSTEIN, M., FRIESEM, A. A., AFLALO, C. & VAKSER, I. A. 1992. Molecular-Surface Recognition - Determination Of Geometric Fit Between Proteins And Their Ligands By Correlation Techniques. *Proceedings Of The National Academy Of Sciences Of The United States Of America,* 89**,** 2195-2199.

KEIL, M., EXNER, T. E. & BRICKMANN, J. 2004. Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. *Journal of Computational Chemistry,* 25**,** 779-789.

KESKIN, O., MA, B. Y. & NUSSINOV, R. 2005. Hot regions in protein-protein interactions: The organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology,* 345**,** 1281-1294.

KESKIN, O., TSAI, C. J., WOLFSON, H. & NUSSINOV, R. 2004. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Science,* 13**,** 1043-1055.

KOIKE, A. & TAKAGI, T. 2004. Prediction of protein-protein interaction sites using support vector machines. *Protein Engineering Design & Selection,* 17**,** 165-173.

KONC, J. & JANEZIC, D. 2007. Protein-protein binding-sites prediction by protein surface structure conservation. *Journal of Chemical Information and Modeling,* 47**,** 940-944.

KUFAREVA, I., BUDAGYAN, L., RAUSH, E., TOTROV, M. & ABAGYAN, R. 2007. PIER: Protein interface recognition for structural proteomics. *Proteins-Structure Function and Bioinformatics,* 67**,** 400-417.

LANDGRAF, R., XENARIOS, I. & EISENBERG, D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *Journal of Molecular Biology,* 307**,** 1487-1502.

LARKIN, M. A., BLACKSHIELDS, G., BROWN, N. P., CHENNA, R., MCGETTIGAN, P. A., MCWILLIAM, H., VALENTIN, F., WALLACE, I. M., WILM, A., LOPEZ, R., THOMPSON, J. D., GIBSON, T. J. & HIGGINS, D. G. 2007. Clustal W and clustal X version 2.0. *Bioinformatics,* 23**,** 2947-2948.

LARRANAGA, P., CALVO, B., SANTANA, R., BIELZA, C., GALDIANO, J., INZA, I., LOZANO, J. A., ARMANANZAS, R., SANTAFE, G., PEREZ, A. & ROBLES, V.

2006. Machine learning in bioinformatics. *Briefings in Bioinformatics,* 7**,** 86-112.

LAWRENCE, M. C. & COLMAN, P. M. 1993. Shape Complementarity at Protein-Protein Interfaces. *Journal of Molecular Biology,* 234**,** 946-950.

LEE, B. & RICHARDS, F. M. 1971. Interpretation of Protein Structures - Estimation of Static Accessibility. *Journal of Molecular Biology,* 55**,** 379-&.

LENSINK, M. F., MENDEZ, R. & WODAK, S. J. 2007. Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins-Structure Function and Bioinformatics,* 69**,** 704-718.

LENSINK, M. F. & WODAK, S. J. 2010. Docking and scoring protein interactions: CAPRI 2009. *Proteins-Structure Function and Bioinformatics,* 78**,** 3073-3084.

LESSENE, G., CZABOTAR, P. E., SLEEBS, B. E., ZOBEL, K., LOWES, K. N., ADAMS, J. M., BAELL, J. B., COLMAN, P. M., DESHAYES, K., FAIRBROTHER, W. J., FLYGARE, J. A., GIBBONS, P., KERSTEN, W. J. A., KULASEGARAM, S., MOSS, R. M., PARISOT, J. P., SMITH, B. J., STREET, I. P., YANG, H., HUANG, D. C. S. & WATSON, K. G. 2013. Structure-guided design of a selective BCL-X-L inhibitor. *Nature Chemical Biology,* 9**,** 390-+.

LEVY, E. D. 2010. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. *Journal of Molecular Biology,* 403**,** 660-670.

LI, B. & KIHARA, D. 2012. Protein docking prediction using predicted protein-protein interface. *Bmc Bioinformatics,* 13.

LI, J.-J., HUANG, D.-S., WANG, B. & CHEN, P. 2006. Identifying protein-protein interfacial residues in heterocomplexes using residue conservation scores. *International Journal of Biological Macromolecules,* 38**,** 241-247.

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.

LI, Y. L., HUANG, Y. P., SWAMINATHAN, C. P., SMITH-GILL, S. J. & MARIUZZA, R. A. 2005. Magnitude of the hydrophobic effect at central versus peripheral sites in protein-protein interfaces. *Structure,* 13**,** 297-307.

LIANG, S., ZHANG, C., LIU, S. & ZHOU, Y. 2006. Protein binding site prediction using an empirical scoring function. *Nucleic Acids Research,* 34**,** 3698-3707.

LICHTARGE, O., BOURNE, H. R. & COHEN, F. E. 1996a. An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology,* 257**,** 342-358.

LICHTARGE, O., BOURNE, H. R. & COHEN, F. E. 1996b. Evolutionarily conserved

G(alpha beta gamma) binding surfaces support a model of the G protein-receptor complex. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 7507-7511.

LO CONTE, L., CHOTHIA, C. & JANIN, J. 1999. The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology,* 285**,** 2177-2198.

MADABUSHI, S., YAO, H., MARSH, M., KRISTENSEN, D. M., PHILIPPI, A., SOWA, M. E. & LICHTARGE, O. 2002. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology,* 316**,** 139-154.

MAGRANE, M., & UniProt Consortium. 2011. UniProt Knowledgebase: a hub of integrated protein data. *DATABASE The Journal of Biological Databases and Curation,* 2011**,** 1-13.

MANNING, J. R., JEFFERSON, E. R. & BARTON, G. J. 2008. The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. *Bmc Bioinformatics,* 9.

MARIANS, K. J. 2008. Understanding how the replisome works. *Nature Structural & Molecular Biology,* 15**,** 125-127.

MARTIN, A. C. 1998. ProFit. London: SciTech Software.

MARTIN, O. & SCHOMBURG, D. 2008. Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines. *Proteins-Structure Function and Bioinformatics,* 70**,** 1367-1378.

MATTHEWS, B. W. 1975. Comparison of Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica Et Biophysica Acta,* 405**,** 442-451.

MAYNE, S. L. N. & PATTERTON, H. G. 2011. Bioinformatics tools for the structural elucidation of multi-subunit protein complexes by mass spectrometric analysis of protein-protein cross-links. *Briefings in Bioinformatics,* 12**,** 660-671.

MCCOY, M. A. & WYSS, D. F. 2002. Structures of protein-protein complexes are docked using only NMR restraints from residual dipolar coupling and chemical shift perturbations. *Journal of the American Chemical Society,* 124**,** 2104-2105.

MEILER, J., BLOMBERG, N., NILGES, M. & GRIESINGER, C. 2000. A new approach for applying residual dipolar couplings as restraints in structure elucidation (vol 16, pg 245, 2000). *Journal of Biomolecular Nmr,* 17**,** 185-185.

MELQUIOND, A., KARACA, E., KASTRITIS, P. & BONVIN, A. 2011. Next challenges in protein–protein docking: from proteome to interactome and beyond.

Computational Molecular Science, 00**,** 1-10.

MENDEZ, R., LEPLAE, R., LENSINK, M. F. & WODAK, S. J. 2005. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins-Structure Function and Bioinformatics,* 60**,** 150-169.

MENDEZ, R., LEPLAE, R., DE MARIA, L. & WODAK, S. J. 2003. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins-Structure Function and Bioinformatics,* 52**,** 51-67.

MIHALEK, I., RES, I. & LICHTARGE, O. 2004. A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology,* 336**,** 1265-1282.

MINTSERIS, J. & WENG, Z. P. 2005a. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 10930-10935.

MINTSERIS, J., WIEHE, K., PIERCE, B., ANDERSON, R., CHEN, R., JANIN, J. & WENG, Z. P. 2005b. Protein-protein docking benchmark 2.0: An update. *Proteins-Structure Function and Bioinformatics,* 60**,** 214-216.

MIRNY, L. A. & SHAKHNOVICH, E. I. 1999. Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function. *Journal of Molecular Biology,* 291**,** 177-196.

MONOD, J., WYMAN, J. & CHANGEUX, J. P. 1965. On the Nature of Allosteric Transitions: A Plausible Model. *Journal of molecular biology,* 12**,** 88-118.

MONTALVAO, R. W., CAVALLI, A., SALVATELLA, X., BLUNDELL, T. L. & VENDRUSCOLO, M. 2008. Structure Determination of Protein-Protein Complexes Using NMR Chemical Shifts: Case of an Endonuclease Colicin-Immunity Protein Complex. *Journal of the American Chemical Society,* 130**,** 15990-15996.

MOONT, G., GABB, H. A. & STERNBERG, M. J. E. 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins-Structure Function and Genetics,* 35**,** 364-373.

MORELLI, X. J., PALMA, P. N., GUERLESQUIN, F. & RIGBY, A. C. 2001. A novel approach for assessing macromolecular complexes combining soft-docking calculations with NMR data. *Protein Science,* 10**,** 2131-2137.

MOTULSKY, H. 2007. *Graphpad prism Version 5.0 Statistics Guide* [online]. USA:

GraphPad Software, inc. Available at: <http://www.graphpad.com/downloads/docs/Prism5Stats.pdf> [Accessed 5 May 2011].

MOTULSKY, H. 2007. *Graphpad prism Version 5.0 Regression Guide* [online]. USA: GraphPad Software, inc. Available at: <http://www.graphpad.com/downloads/docs/Prism5Regression.pdf> [Accessed 25 January 2012].

MURAKAMI, Y. & JONES, S. 2006. SHARP(2): protein-protein interaction predictions using patch analysis. *Bioinformatics,* 22**,** 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**,** 1841-1848.

NEEDLEMAN, S. B. & WUNSCH, C. D. 1970. A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins. *Journal of Molecular Biology,* 48**,** 443-&.

NEGI, S. S. & BRAUN, W. 2007. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. *Journal of Molecular Modeling,* 13**,** 1157-1167.

NEUVIRTH, H., HEINEMANN, U., BIRNBAUM, D., TISHBY, N. & SCHREIBER, G. 2007. ProMateus - an open research approach to protein-binding sites analysis. *Nucleic Acids Research,* 35**,** W543-W548.

NEUVIRTH, H., RAZ, R. & SCHREIBER, G. 2004. ProMate: A structure based prediction program to identify the location of protein-protein binding sites. *Journal of Molecular Biology,* 338**,** 181-199.

NOOREN, I. M. A. & THORNTON, J. M. 2003a. Diversity of protein-protein interactions. *Embo Journal,* 22**,** 3486-3492.

NOOREN, I. M. A. & THORNTON, J. M. 2003b. Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology,* 325**,** 991-1018.

OFRAN, Y. & ROST, B. 2007a. Protein-protein interaction hotspots carved into sequences. *Plos Computational Biology,* 3**,** 1169-1176.

OFRAN, Y. & ROST, B. 2007b. ISIS: interaction sites identified from sequence. *Bioinformatics,* 23**,** E13-E16.

OFRAN, Y. & ROST, B. 2003a. Predicted protein-protein interaction sites from local

sequence information. *Febs Letters,* 544**,** 236-239.

OFRAN, Y. & ROST, B. 2003b. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology,* 325**,** 377-387.

OZBABACAN, S. E. A., ENGIN, H. B., GURSOY, A. & KESKIN, O. 2011. Transient proteinprotein interactions. *Protein Engineering Design & Selection,* 24**,** 635-648.

PALMA, P. N., KRIPPAHL, L., WAMPLER, J. E. & MOURA, J. J. G. 2000. BiGGER: A new (soft) docking algorithm for predicting protein interactions. *Proteins-Structure Function and Genetics,* 39**,** 372-384.

PARK, H. M. 2008. *Univariate Analysis and Normality Test Using SAS, Stata, and SPSS* [online]. Indiana Indiana University. Available at: <http://www.indiana.edu/~statmath/stat/all/normality/normality.pdf> [Accessed 10 October 2011].

PARRISH, J. R., GULYAS, K. D. & FINLEY, R. L. 2006. Yeast two-hybrid contributions to interactome mapping. *Current Opinion in Biotechnology,* 17**,** 387-393.

PERKINS, J. R., DIBOUN, I., DESSAILLY, B. H., LEES, J. G. & ORENGO, C. 2010. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure,* 18**,** 1233-1243.

PERUMAL, S. K., YUE, H., HU, Z., SPIERING, M. M. & BENKOVIC, S. J. 2010. Single-molecule studies of DNA replisome function. *Biochimica Et Biophysica Acta-Proteins and Proteomics,* 1804**,** 1094-1112.

PETTIT, F. K., BARE, E., TSAI, A. & BOWIE, J. U. 2007. HotPatch: A statistical approach to finding biologically relevant features on protein surfaces. *Journal of Molecular Biology,* 369**,** 863-879.

PONSTINGL, H., THOMAS, K. B., GORSE, D. & THORNTON, J. M. 2005. Morphological aspects of oligorneric protein structures. *Progress in Biophysics & Molecular Biology,* 89**,** 9-35.

PONTIUS, J., RICHELLE, J. & WODAK, S. J. 1996. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *Journal of Molecular Biology,* 264**,** 121-136.

POROLLO, A. & MELLER, J. 2007. Prediction-based fingerprints of protein-protein interactions. *Proteins-Structure Function and Bioinformatics,* 66**,** 630-645.

QIN, S. & ZHOU, H.-X. 2007a. meta-PPISP: a meta web server for protein-protein interaction site prediction. *Bioinformatics,* 23**,** 3386-3387.

QIN, S. & ZHOU, H. X. 2007b. A holistic approach to protein docking. *Proteins-Structure Function and Bioinformatics,* 69**,** 743-749.

QIU, Z. & WANG, X. 2012. Prediction of protein-protein interaction sites using patch-based residue characterization. *Journal of Theoretical Biology,* 293**,** 143-150.

RAFFERTY, J. B., SOMERS, W. S., STGIRONS, I. & PHILLIPS, S. E. V. 1989. 3-Dimensional Crystal-Structures of Escherichia-Coli Met Repressor with and without Corepressor. *Nature,* 341**,** 705-710.

RAMANI, A. K., BUNESCU, R. C., MOONEY, R. J. & MARCOTTE, E. M. 2005. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology,* 6.

REDDY, B. V. B. & KAZNESSIS, Y. N. 2005. A quantitative analysis of interfacial amino acid conservation in protein-protein hetero complexes. *Journal of Bioinformatics and Computational Biology,* 3**,** 1137-1150.

RENNIE, J. D. M. 2004. *Derivation of the F-Measure* [online]. Boston: Massachusetts Institute of Technology. Available at: <http://people.csail.mit.edu/jrennie/writing/fmeasure.pdf> [Accessed 9 October 2011].

RICE, P., LONGDEN, I. & BLEASBY, A. 2000. EMBOSS: The European molecular biology open software suite. *Trends in Genetics,* 16**,** 276-277.

RICHARDS, F. M. 1974. Interpretation of Protein Structures - Total Volume, Group Volume Distributions and Packing Density. *Journal of Molecular Biology,* 82**,** 1-14.

RITCHIE, D. W. 2008. Recent progress and future directions in protein-protein docking. *Current Protein & Peptide Science,* 9**,** 1-15.

ROSS, G. J. S. 1969. Single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* 18**,** 106-110.

RYU, T. 2009. Benchmarking of BioPerl, Perl, BioJava, Java, BioPython, and Python for Primitive Bioinformatics Tasks and Choosing a Suitable Language. *International Journal of Contents,* 5**,** 1-65.

SANDER, C. & SCHNEIDER, R. 1991. Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment. *Proteins-Structure Function and Genetics,* 9**,** 56-68.

SCHNEIDER, S. & ZACHARIAS, M. 2012. Scoring optimisation of unbound protein-

protein docking including protein binding site predictions. *Journal of Molecular Recognition,* 25**,** 15-23.

SCHNEIDMAN-DUHOVNY, D., ROSSI, A., AVILA-SAKAR, A., KIM, S. J., VELAZQUEZ-MURIEL, J., STROP, P., LIANG, H., KRUKENBERG, K. A., LIAO, M., KIM, H. M., SOBHANIFAR, S., DOETSCH, V., RAJPAL, A., PONS, J., AGARD, D. A., CHENG, Y. & SALI, A. 2012. A method for integrative structure determination of protein-protein complexes. *Bioinformatics,* 28**,** 3282-3289.

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.

SHAFIQ, M. I., STEINBRECHER, T. & SCHMID, R. 2012. FASCAPLYSIN as a Specific Inhibitor for CDK4: Insights from Molecular Modelling. *Plos One,* 7.

SHANNON, C. E. A Mathematical Theory of Communication. *The Bell System Technical Journal,* 27**,** 379-423 and 623-656.

SHENKIN, P. S., ERMAN, B. & MASTRANDREA, L. D. 1991. Information-Theoretical Entropy as a Measure of Sequence Variability. *Proteins-Structure Function and Genetics,* 11**,** 297-313.

SONAVANE, S. & CHAKRABARTI, P. 2008. Cavities and Atomic Packing in Protein Structures and Interfaces. *Plos Computational Biology,* 4.

SONEGO, P., KOCSOR, A. & PONGOR, S. 2008. ROC analysis: applications to the classification of biological sequences and 3D structures. *Briefings in Bioinformatics,* 9**,** 198-209.

SPIRIN, V. & MIRNY, L. A. 2003. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences of the United States of America,* 100**,** 12123-12128.

STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G. R., KORF, I., LAPP, H., LEHVASLAIHO, H., MATSALLA, C., MUNGALL, C. J., OSBORNE, B. I., POCOCK, M. R., SCHATTNER, P., SENGER, M., STEIN, L. D., STUPKA, E., WILKINSON, M. D. & BIRNEY, E. 2002. The bioperl toolkit: Perl modules for the life sciences. *Genome Research,* 12**,** 1611-1618.

STARK, C., BREITKREUTZ, B.-J., CHATR-ARYAMONTRI, A., BOUCHER, L., OUGHTRED, R., LIVSTONE, M. S., NIXON, J., VAN AUKEN, K., WANG, X., SHI, X., REGULY, T., RUST, J. M., WINTER, A., DOLINSKI, K. & TYERS, M.

2011. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research,* 39**,** D698-D704.

STATACORP. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP.

STITES, W. E. 1997. Protein-protein interactions: Interface structure, binding thermodynamics, and mutational analysis. *Chemical Reviews,* 97**,** 1233-1250.

STRATMANN, D., BOELENS, R. & BONVIN, A. M. J. J. 2011. Quantitative use of chemical shifts for the modeling of protein complexes. *Proteins-Structure Function and Bioinformatics,* 79**,** 2662-2670.

SUZEK, B. E., HUANG, H. Z., MCGARVEY, P., MAZUMDER, R. & WU, C. H. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics,* 23**,** 1282-1288.

SWAPNA, L. S., BHASKARA, R. M., SHARMA, J. & SRINIVASAN, N. 2012. Roles of residues in the interface of transient protein-protein complexes before complexation. *Scientific Reports,* 2.

THIRUPATHI, R., SRAVANTHI, S., KUMAR, A. & PRABHAKARAN, E. 2011. Protein-Protein Complexes. *Journal of the Indian Institute of Science,* 91**,** 497-520.

THORN, K. S. & BOGAN, A. A. 2001. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics,* 17**,** 284-285.

TORCHALA, M., MOAL, I. H., CHALEIL, R. A. G., FERNANDEZ-RECIO, J. & BATES, P. A. 2013. SwarmDock: a server for flexible protein-protein docking. *Bioinformatics,* 29**,** 807-809.

TRESS, M., DE JUAN, D., GRANA, O., GOMEZ, M. J., GOMEZ-PUERTAS, P., GONZALEZ, J. M., LOPEZ, G. & VALENCIA, A. 2005. Scoring docking models with evolutionary information. *Proteins-Structure Function and Bioinformatics,* 60**,** 275-280.

TSAI, C. J., LIN, S. L., WOLFSON, H. J. & NUSSINOV, R. 1997. Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Science,* 6**,** 53-64.

TUNCBAG, N., GURSOY, A. & KESKIN, O. 2009. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics,* 25**,** 1513-1520.

ULRICH, E. L., AKUTSU, H., DORELEIJERS, J. F., HARANO, Y., IOANNIDIS, Y. E.,

LIN, J., LIVNY, M., MADING, S., MAZIUK, D., MILLER, Z., NAKATANI, E., SCHULTE, C. F., TOLMIE, D. E., WENGER, R. K., YAO, H. Y. & MARKLEY, J. L. 2008. BioMagResBank. *Nucleic Acids Research,* 36**,** D402-D408.

VALDAR, W. S. J. 2002. Scoring residue conservation. *Proteins-Structure Function and Bioinformatics,* 48**,** 227-241.

VALDAR, W. S. J. & THORNTON, J. M. 2001. Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins-Structure Function and Bioinformatics,* 42**,** 108-124.

VAN DIJK, A. D. J., FUSHMAN, D. & BONVIN, A. 2005a. Various strategies of using residual dipolar couplings in NMR-driven protein docking: Application to Lys48-linked Di-ubiquitin and validation against N-15-relaxation data. *Proteins-Structure Function and Bioinformatics,* 60**,** 367-381.

VAN DIJK, A. D. J., DE VRIES, S. J., DOMINGUEZ, C., CHEN, H., ZHOU, H. X. & BONVIN, A. 2005b. Data-driven docking: HADDOCK's adventures in CAPRI. *Proteins-Structure Function and Bioinformatics,* 60**,** 232-238.

VAN RIJSBERGEN, V. 1979. *Information Retireval,* London, Butterworths.

VAKSER, I. A. 2004. Protein-protein interfaces are special. *Structure,* 12**,** 910-912.

VENKATESAN, K., RUAL, J.-F., VAZQUEZ, A., STELZL, U., LEMMENS, I., HIROZANE-KISHIKAWA, T., HAO, T., ZENKNER, M., XIN, X., GOH, K.-I., YILDIRIM, M. A., SIMONIS, N., HEINZMANN, K., GEBREAB, F., SAHALIE, J. M., CEVIK, S., SIMON, C., DE SMET, A.-S., DANN, E., SMOLYAR, A., VINAYAGAM, A., YU, H., SZETO, D., BORICK, H., DRICOT, A., KLITGORD, N., MURRAY, R. R., LIN, C., LALOWSKI, M., TIMM, J., RAU, K., BOONE, C., BRAUN, P., CUSICK, M. E., ROTH, F. P., HILL, D. E., TAVERNIER, J., WANKER, E. E., BARABASI, A.-L. & VIDAL, M. 2009. An empirical framework for binary interactome mapping. *Nature Methods,* 6**,** 83-90.

VENKATRAMAN, V., YANG, Y. D., SAEL, L. & KIHARA, D. 2009. Protein-protein docking using region-based 3D Zernike descriptors. *Bmc Bioinformatics,* 10.

WALL, L., CHRISTIANSEN, T. & ORWANT, J. 2000. *Programming Perl,* USA, O'Reilly Media, Inc.

WANG, B., CHEN, P., HUANG, D. S., LI, J. J., LOK, T. M. & LYU, M. R. 2006. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *Febs Letters,* 580**,** 380-384.

WANG, C., BRADLEY, P. & BAKER, D. 2007. Protein-protein docking with backbone

flexibility. *Journal of Molecular Biology,* 373**,** 503-519.

WANG, G. S., LOUIS, J. M., SONDEJ, M., SEOK, Y. J., PETERKOFSKY, A. & CLORE, G. M. 2000. Solution structure of the phosphoryl transfer complex between the signal transducing proteins HPr and IIA(Glucose) of the Escherichia coli phosphoenolpyruvate : sugar phosphotransferase system. *Embo Journal,* 19**,** 5635-5649.

WANG, K. & SAMUDRALA, R. 2006. Incorporating background frequency improves entropy-based residue conservation measures. *Bmc Bioinformatics,* 7.

WASS, M. N., FUENTES, G., PONS, C., PAZOS, F. & VALENCIA, A. 2011a. Towards the prediction of protein interaction partners using physical docking. *Molecular Systems Biology,* 7.

WASS, M. N., DAVID, A. & STERNBERG, M. J. E. 2011b. Challenges for the prediction of macromolecular interactions. *Current Opinion in Structural Biology,* 21**,** 382-390.

WILLIAMS, L. 2011. *Molecular Interaction* [online]. Available at: <http://ww2.chemistry.gatech.edu/~lw26/structure/molecular_interactions/mol_int.html#G> [Accessed 1 February 2012].

WILLIAMSON, R. A., CARR, M. D., FRENKIEL, T. A., FEENEY, J. & FREEDMAN, R. B. 1997. Mapping the binding site for matrix metalloproteinase on the N-terminal domain of the tissue inhibitor of metalloproteinases-2 by NMR chemical shift perturbation. *Biochemistry,* 36**,** 13882-13889.

WILLIAMSON, R. M. 1995. Information-Theory Analysis of the Relationship between Primary Sequence Structure and Ligand Recognition among a Class of Facilitated Transporters. *Journal of Theoretical Biology,* 174**,** 179-188.

WODAK, S. J. & JANIN, J. 1978. Computer-Analysis of Protein-Protein Interaction. *Journal of Molecular Biology,* 124**,** 323-342.

WOOD, E. & MYERS, A. 1991. *Essential Chemistry for Biochemistry,* London, The Biochemical Society.

WU, F., TOWFIC, F., DOBBS, D. & HONAVAR, V. 2007. Analysis of protein protein dimeric interfaces. *2007 Ieee International Conference on Bioinformatics and Biomedicine, Proceedings.*

XU, D., TSAI, C. J. & NUSSINOV, R. 1997a. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering,* 10**,** 999-1012.

XU, D., LIN, S. L. & NUSSINOV, R. 1997b. Protein binding versus protein folding: The

role of hydrophilic bridges in protein associations. *Journal of Molecular Biology,* 265**,** 68-84.

XUE, L. C., DOBBS, D. & HONAVAR, V. 2011a. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *Bmc Bioinformatics,* 12.

XUE, L., JORDAN, R., EL-MANZALAWY, Y., DOBBS, D. & HONAVAR, V. 2011b. Ranking Docked Models of Protein-Protein Complexes Using Predicted Partner-Specific Protein-Protein Interfaces: A Preliminary Study. *Proceedings of the Second ACM International Conference on Bioinformatics and Computational Biology.* ACM.

YU, Y. H., LU, B. Z., HAN, J. G. & ZHANG, P. F. 2004. Scoring protein-protein docked structures based on the balance and tightness of binding. *Journal of Computer-Aided Molecular Design,* 18**,** 251-260.

ZACHARIAS, M. 2003. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Science,* 12**,** 1271-1282.

ZHANG, C., LIU, S. & ZHOU, Y. Q. 2005. Docking prediction using biological information, ZDOCK sampling technique, and clustering guided by the DFIRE statistical energy function. *Proteins-Structure Function and Bioinformatics,* 60**,** 314-318.

ZHOU, H.-X. & QIN, S. 2007. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics,* 23**,** 2203-2209.

ZHOU, H. X. & SHAN, Y. B. 2001. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins-Structure Function and Genetics,* 44**,** 336-343.

ZHU, H. B., DOMINGUES, F. S., SOMMER, I. & LENGAUER, T. 2006. NOXclass: prediction of protein-protein interaction types. *Bmc Bioinformatics,* 7.

ZIMMERMANN, O. 2002. Untersuchungen zur Vorhersage der nativen Orientierung von Protein-Komplexen mit Fourier-Korrelationsmethoden: Universität zu Köln

ZWECKSTETTER, M. 2008. NMR: prediction of molecular alignment from structure using the PALES software. *Nature Protocols,* 3**,** 679-690.

ZWECKSTETTER, M. & BAX, A. 2000. Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. *Journal of the American Chemical Society,* 122**,** 3791-3792.