# Computational modelling of multidomain proteins with covarying residue pairs

Stuart Tetchner

Bioinformatics Group

Department of Computer Science

University College London

A thesis submitted in partial fulfilment of

the requirements for the degree of

Doctor of Philosophy

December 2015

I, Stuart Tetchner, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

## Abstract

The vast majority of known protein sequences have no solved three-dimensional structure at all, and the remaining ones usually have not been completely characterised, due to the limitations of experimental structural biology techniques. Structural genomics projects have helped increase the coverage of the protein structure universe, but most available structures still consist of either individual domains or sets of relatively small ones. This has prompted the development of computational methods for protein structure prediction, as well as for multidomain architecture modelling.

One appealing idea to achieve this goal consists of detecting residue-residue contacts from multiple sequence alignments, under the assumption that they covary in order to maintain the local microenvironment and the overall stability of protein structures. After early limited success, this type of analysis has lately witnessed substantial progress, thanks to theoretical advances in disentangling genuine from spurious instances of correlation. Unsurprisingly, structural bioinformatics has promptly and successfully applied these improved tools to model globular and transmembrane proteins, along with guiding the assembly of protein complexes. However, the efficacy of these methods in the context of multidomain protein modelling has not yet been investigated.

In this thesis state-of-the-art methods for predicting contacts from sequence data have been evaluated and used to build models of two-domain protein structures. Firstly, the ability of alternative methods to identify interdomain contacts was examined in a reference set of experimentally solved structures. Secondly, predicted contacts were employed to score docking models and select near-native solutions accordingly. Finally, predicted contacts were used to guide the assembly of individual domains in a multidomain modelling protocol.

# Acknowledgements

Firstly I'd like to thank my supervisor Professor David Jones, for all of his help and guidance over the past 3 years. I would also like to thank my secondary supervisor, Dr Renos Savva and thesis chair, Dr Maya Topf, for many useful suggestions throughout the course of my studies.

I would also like to thank all the members of the Bioinformatics group, past and present for many informal chats about my work and the wider field as well as making it such an enjoyable place to be. A special mention must go to Dr Domenico Cozzetto for tirelessly providing assistance with the R package and teaching me a great deal about Italian cuisine. Jim Rohn once said "you are the average of the 5 people you spend the most time with" and while the Bioinformatics group may not be exactly 5 people, I believe the sentiment holds true. I've been fortunate to undertake my PhD in such a fantastic lab.

I am also grateful for The Institute of Structural and Molecular Biology (ISMB) for providing such a stimulating work environment and the Wellcome Trust for funding, making this all possible.

Lastly, I'd also like to thank my fiancée Jess, my friend and seemingly long-term housemate Lucian, along with the rest of my friends and family for their continued support over the last 4 years.

# Contents

## List of Figures

## List of Tables

# 1. Introduction

## 1.1 Proteins and their structure

Proteins are the “machinery” of the cell, involved in the expansive range of biological processes necessary for life, including molecular transport, transcription, translation, metabolism and cell signalling, and countless others (Alberts, 1998). This myriad of functions is achieved through structural differences of the proteins performing each role. The three-dimensional (3D) structure of each protein is formed by the folding of a linear chain of amino acids.

The structure of a protein can be described by four levels of increasing complexity. The sequence of amino acids comprising the protein chain is termed the primary structure. During protein synthesis and the subsequent folding process, amino acids form local secondary structures - alpha helices and beta strands - interspersed with regions of coil. These secondary structural elements fold into a specific 3D arrangement, forming the protein's tertiary structure. Some proteins further assemble with other chains, which can be identical or different, in order to form quaternary structures. As amino acids in distant regions of the chain are often colocated after folding, knowledge of the folded structure can provide insight into local functional features, such as active sites and binding regions. Therefore, gaining an understanding of the protein structure is necessary to shed light on the precise mechanisms enabling a protein to perform its function. Furthermore, protein structures are a key starting point for rational drug design (Mandal et al., 2009).

### 1.1.1 Experimental structure determination

In order to unveil the molecular mechanisms underpinning protein function, great emphasis has been placed on the elucidation of novel structures since the first one was solved (Kendrew et al., 1958). The structure of proteins can be determined by a number of different experimental approaches, with the most commonly used being X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) and cryo-electron microscopy (cryo-EM), each with different advantages and disadvantages. For example, X-ray crystallography is capable of generating high-resolution structures, provided that high-quality crystals can be produced of pure samples of the protein under investigation. NMR is able to capture information relating to molecular motions in solution, but requires relatively large samples of protein which are stable at room temperature. Cryo-EM methods are able to generate structures of larger macromolecules from a frozen sample, bypassing the requirement for high-quality crystals necessary for X-ray crystallography. Due to this, cryo-EM has proven to be especially useful in the study of larger multiprotein assemblies. Previously, cryo-EM has been regarded as a lower-resolution technique, but recent advancements in the processing of data and hardware improvements have enabled high-resolution structures to be generated, rivalling that achievable by X-ray crystallography (Scheres, 2014; Bai et al., 2015; Callaway, 2015).

Protein structures determined by these experimental techniques are deposited in the Protein Data Bank (PDB) (Bernstein et al., 1978; Berman et al., 2000), which acts as a central repository for the storage of protein structures. In recent years there has been considerable growth in the number of deposited structures within the PDB, largely thanks to technical advances made by structural genomics initiatives (SGIs), which aimed to increase the number and diversity of experimental structures (Brenner, 2001) (Figure 1.1).

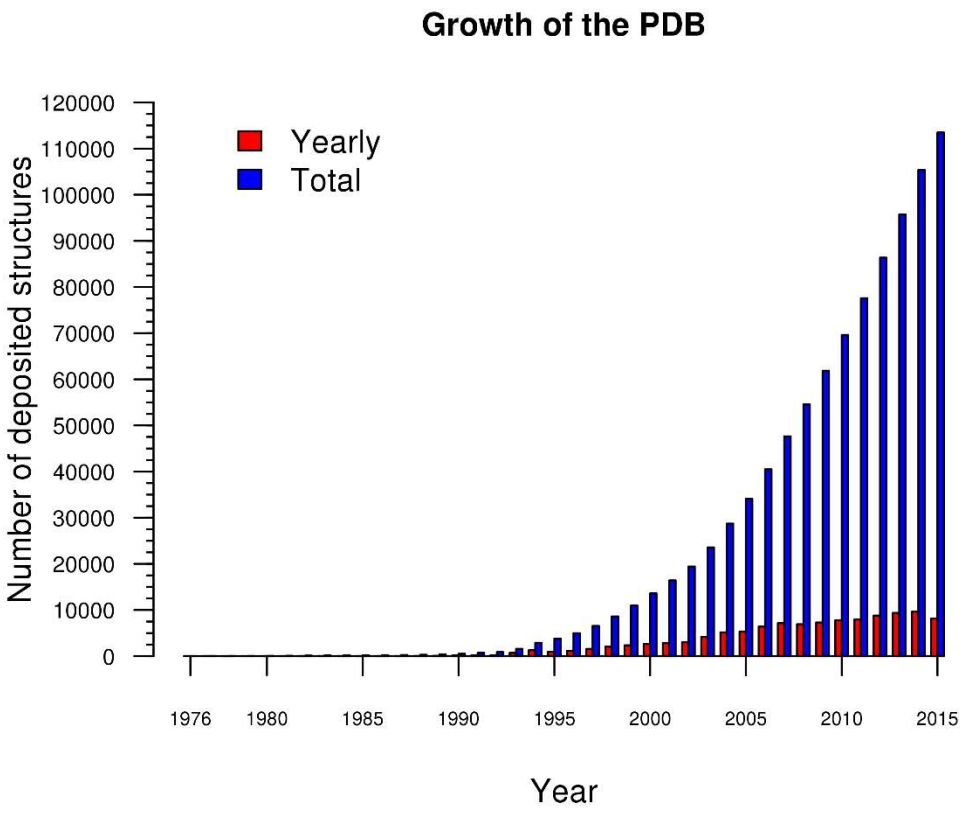



**Figure 1.1: Yearly growth of the Protein Data Bank.** Data retrieved from www.rcsb.org/pdb/results/contentGrowthCsv.do?format=total

SGIs sought to expand the coverage of the protein “structural universe” by identifying novel protein structures. The aim of these efforts was to identify the structures of proteins which were thought to be unlike the structure of those which were already known. Driven by this goal, this led the SGIs to reveal a large number of novel protein folds during their course (Figure 1.2). However, since the heyday of the SGIs after the turn of the millennium, the number of novel folds discovered has essentially halted, with the last novel fold identified in 2012. These protein folds are the underlying “architecture” which form the core structure of a protein. Considering the reduction in the number of novel folds which are now being identified, one may conclude that we have now determined the most

common folds which create the most prevalent structures. However, it could also be argued that these structures are just the most feasible to characterise experimentally. Either way, current evidence would suggest that the diverse world of protein structures is formed of a small set of folds.

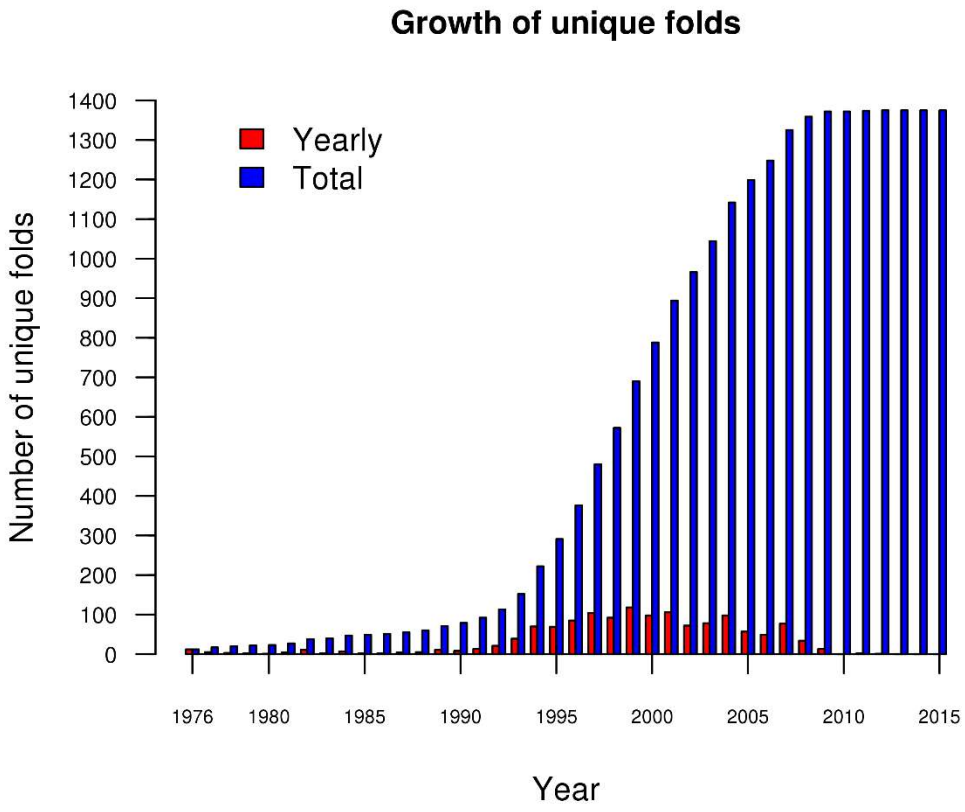



**Figure 1.2: Yearly growth of unique folds in the Protein Data Bank.** Data retrieved from http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=fold-cath

## 1.2 Protein domains

### 1.2.1 Overview

By observing the structures of experimentally-determined proteins, it can be seen that these complex biological molecules are often formed of smaller modules, termed domains (Figure 1.3). The observation of such substructures was made early on, with Donald Wetlaufer (1973) identifying “distinct globular units” (i.e. domains) as a routine structural feature in a survey of 18 protein structures, expanding upon the earlier observation raised in a study of immunoglobulins (Cunningham et al., 1971).

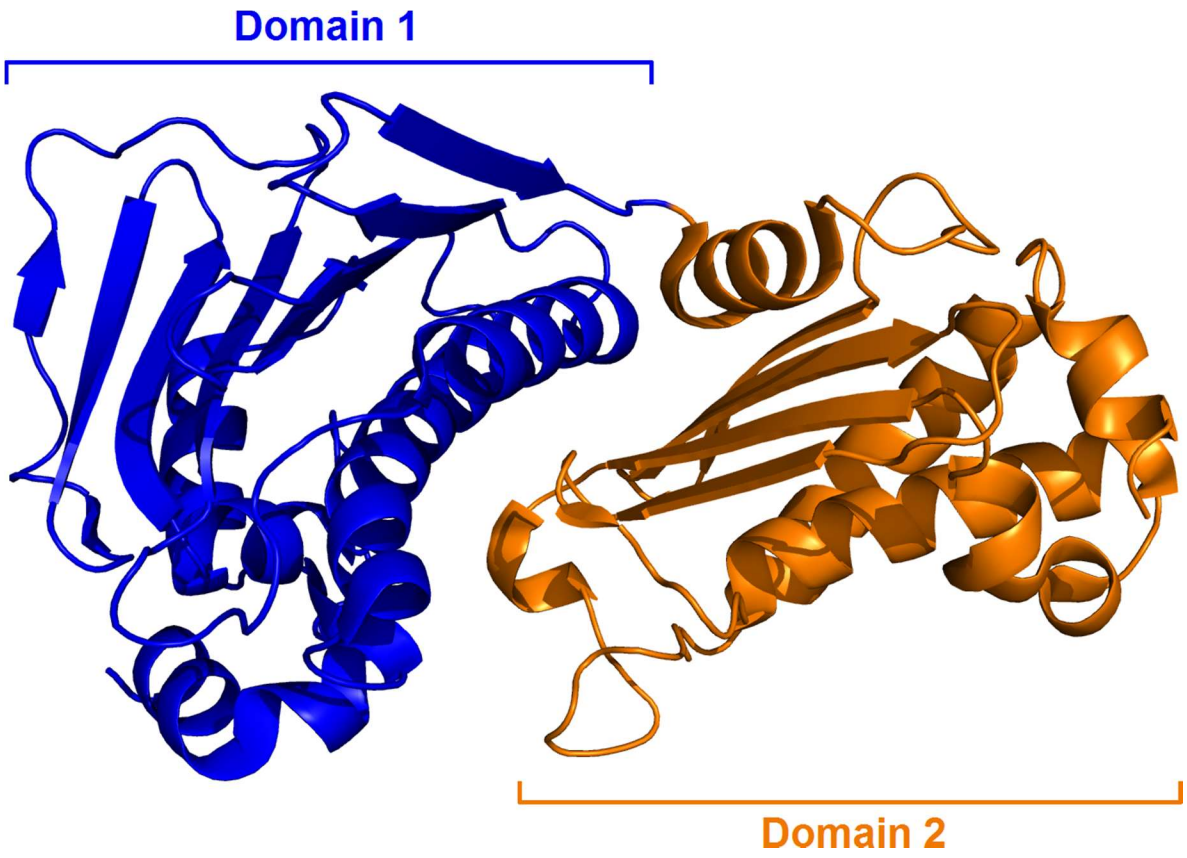



**Figure 1.3: The X-ray crystallographic structure of the DNA gyrase B-subunit from *Myxococcus xanthus* comprising two structural domains.** Domains are numbered in order of appearance from the N-terminus. Image generated using PyMOL (Schrödinger LLC).

Domains are the unit of protein evolution (Vogel et al., 2004). Whilst a general definition of what constitutes a domain is not universally agreed, broadly they are thought to be spatially distinct substructures which are able to fold and function in isolation (Ponting and Russell, 2002). Throughout the course of evolution, organisms have grown in complexity, and in order to do so, proteins have had to acquire novel functionality. Evolution has proceeded through the duplication, divergence, fusion and recombination of genes and their encoded proteins (Chothia and Gough, 2009; Vogel et al., 2005). Individual domains are often associated with a specific function, such as interacting with DNA or binding a particular substrate. By combining different domains, novel proteins can be formed with increased functional complexity, either through incorporating multiple separate functions or forming novel functional sites between domains (Apic and Russell, 2010; Bashton and Chothia, 2007; Han et al., 2007). By combining domains in this manner, it is possible to create substantial functional diversity from a much smaller set of units (Moore et al., 2008). Almost all novelty in newly discovered proteins comes from alternative combinations of already known domains (Levitt, 2009). Considering domains as functional units, it should come as little surprise that the majority of proteins across all known life are believed to comprise multiple domains. Based on the analysis of sequenced genomes, it has been estimated that approximately four-fifths of metazoan and two-thirds of prokaryotic proteins are multidomain (Apic et al., 2001).

Over the course of domain evolution, the underlying sequence is inherited by subsequent organismal generations. Domains which share a common ancestor are termed "homologous", and related domains form “homologous superfamilies”. The number of members in each of these domain superfamilies varies considerably; some superfamilies are highly populated, whereas the majority have far fewer representatives. For example, the 9 most abundant human domain superfamilies have been estimated to account for 20% of all human domains. This is in contrast to another 220 superfamilies which were

observed to appear only once; accounting for less than 1% of identified domains (Chothia and Gough, 2009). This frequency distribution is more widely observed across multiple genomes, which approximately follows a power-law distribution (Qian et al., 2001).

Domain sizes are highly variable, though the majority contain an average of approximately 100 residues (Figure 1.4). Considering the extremes of the distribution, the smallest domain superfamily is that of the “single helix bin” (CATH code 1.20.5.460) with an average length of 15 amino acids and the largest observed domain superfamily is the “Photosystem I subunits PsaA/PsaB” (CATH code 1.20.1130.10), with an average domain length of 740 amino acids. An overview of the CATH domain classification system is given in the next section.

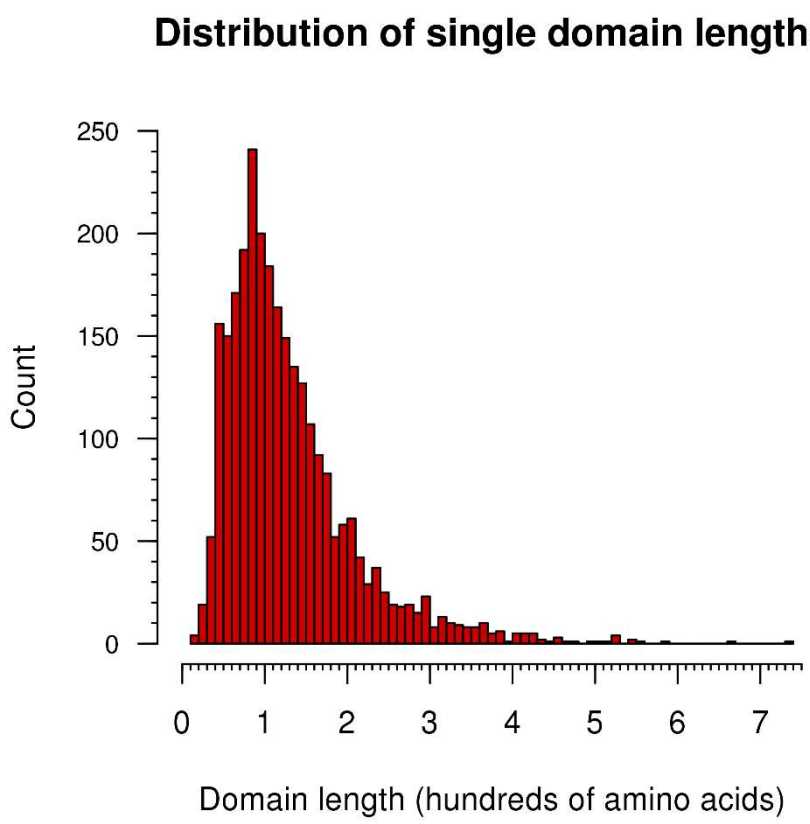



**Figure 1.4: Mean domain length distribution of the 2738 unique CATH homologous superfamilies.** All 235,858 CATH domains (release 4.0.0) were analysed and mean domain lengths were calculated after grouping at the homologous superfamily level.

Before the advent of genomic sequencing, early studies focused on investigating structural features of multidomain PDB entries. These studies found a number of different features relating to the interdomain interface. The interdomain interface is generally hydrophobic, but the level of hydrophobicity is intermediate between that of the solvent-accessible surface, and that observed within domain cores (Jones et al., 2000; Argos, 1988). Interdomain interfaces also display remarkable differences in surface area, from small interfaces permitting interdomain motion restrained by an interdomain linker, to much larger interfaces where little interdomain motion is observed. Unsurprisingly, due to the

obligate nature of the interdomain interaction, domain-domain interfaces have been shown to have similar hydrophobic residue propensities as the interchain interface of permanent dimers (Jones et al., 2000). Residues at the domain-domain interface have also been shown to be more conserved than other solvent-exposed residues (Littler and Hubbard, 2005).

With the advent of genomic sequencing, it became possible to more broadly observe how multidomain proteins have evolved, and investigate differences between species. The order in which domains appear along the protein chain is termed the “domain architecture”. By comparing between different organisms it became clear that domain architectures tend to be strongly conserved once established (Han et al., 2007; Bashton and Chothia, 2002; Vogel et al., 2004; Apic et al., 2001). Whilst in theory domain architectures could be repeatedly formed through convergent evolution, this scenario is rare (Gough, 2005).

Different domain families have different propensities to interact with other families. Some families, such as those involved in key cellular processes (for example, the “P-loop nucleotide triphosphate hydrolase” family involved in nucleotide binding (Iyer et al., 2004)), frequently interact with numerous other types of domains, whereas most others strictly interact with one or two others (Apic et al., 2001). Domains which are observed to interact with many other domains are termed “promiscuous”, and represent a major source of functional novelty. Domain promiscuity has been demonstrated to be widely observed within protein interaction networks, where novel protein-protein interactions are often created at the domain level (Basu et al., 2008). Domains interacting with a single other partner typically interact using the same interface, whereas domains which interact with multiple different partner domains are commonly observed to do so by making use of different interfaces (Littler and Hubbard, 2005).

Within a single multidomain chain, the orientation of interacting domains tends to be evolutionarily conserved, in order to maintain the global structure. In a study of multidomain chains, Aloy and colleagues (2003) calculated the conservation of orientation among domains, finding that homologous proteins sharing sequence similarity above a threshold of 30-40% typically maintain a similar interdomain orientations. These findings were later corroborated in a study explicitly investigating two-domain proteins (Han et al., 2006). However, such interdomain interactions can have substantial differences in orientation even amongst highly similar sequences (Aloy et al., 2003; Han et al., 2006). Investigations into such examples observed long interdomain linkers in conjunction with small interfaces, permitting greater ranges of motion (Han et al., 2006). On the other hand, interdomain orientation can be conserved between distant homologues if the interface is of particular functional importance, for example, if it contains the active site (Han et al., 2006). Considering domain-domain interactions between different protein chains, the picture is less clear-cut, with the orientation of interchain interactions considerably less conserved than their intrachain counterparts (Aloy et al., 2003).

### 1.2.2 The classification of domain structures

Whilst there are approximately 113,000 protein structures available in the PDB at the time of writing, each of these proteins can be described using a more limited repertoire of roughly 2,700 different domain superfamilies (Lees et al., 2014). By grouping domains based on structural and sequential similarity as a means of inferring relatedness, the evolution of domains and the proteins they comprise can be investigated.

Over the years different databases have been developed to categorise domains based on structural similarity, but the two most widely used and comprehensive resources are SCOP (Murzin et al., 1995) and CATH (Orengo et al., 1997).

The SCOP (Structural Classification Of Proteins) hierarchical classification system is based on a manual inspection protocol with the aim of categorising domains based on structural and evolutionary relationships. SCOP organises protein domains into four chief categories: Class, Fold, Superfamily and Family. The intensive manual assignments used in the SCOP database meant updates to the database proved problematic with the rapid increase in deposited structural data, and the last update to the SCOP database was released in 2009.

The CATH classification shares a similar hierarchical structure and derives its name from the 4 highest levels of its classification system: Class, Architecture, Topology and Homologous superfamily. The CATH classification is mainly based on structural features, but employs greater automated analysis than the manual approach taken by SCOP. However, CATH retains the use of expert manual curation in cases where the automated protocol cannot make an automated decision. At each level of the hierarchy, domains are clustered according to structural and sequence similarities (Figure 1.5).

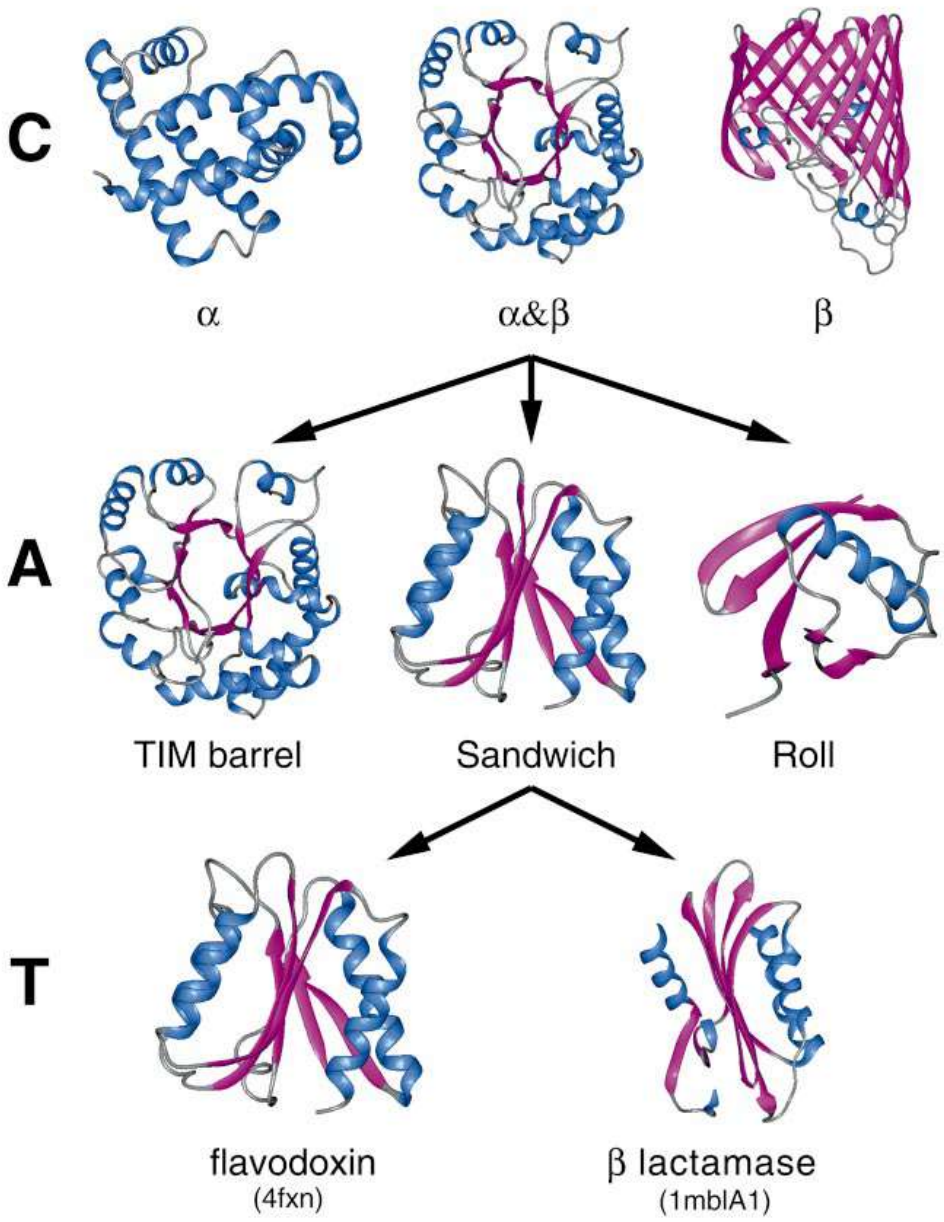



**Figure 1.5: The three highest levels of the CATH hierarchy.** Alpha helices are drawn as blue helices and beta sheets as magenta arrows, with the arrow indicating the direction of the chain from start to end. Image reproduced from (Orengo et al., 1997).

At the highest level of the hierarchy, relating to the protein “class” or “C” level, structures are grouped according to the predominant secondary structure composition. Four such classes are assigned: mainly alpha helices (class 1), mainly beta sheets (class 2), mixed

alpha helices and beta sheets (class 3) and few secondary structures (class 4). Following the class level is the “architecture” (A) level. The architecture describes the approximate arrangement of the secondary structural elements in 3D space. Beneath this is the “topology” (T) level, describing the specific connections between the secondary structures, describing the protein fold. The fourth level describes “homologous superfamilies” (H), and contains proteins with evidence of a common ancestor determined by the analysis of similarities between sequence, structural and function (Sillitoe et al., 2013). At each level of the hierarchy, each group is assigned a numeric identifier. These numerical identifiers specify a particular “CATH code” which is usually written in the form C.A.T.H.

As mentioned previously, domains can be defined in a number of different ways. This carries over to how computational approaches apply different definitions for categorisation. For example, the manual approach used by SCOP defines domains based on evidence that a domain occurs in the structure of more than one different protein. The approach used by CATH employs automated procedures to identify sequence and structural similarities. CATH, and the methods employed in the automated protocol, exploit general knowledge about the distinct, globular structure of domains. The underlying principle is that domains should contain more residue-residue interactions within the domain than observed between domains. To apply this, firstly, novel sequences are compared to existing, categorised CATH domains. Using sequence comparison and domain boundary prediction, the number of domains present within an uncategorised structure are estimated. In cases where the underlying automated methods do not come to a consensus agreement, manual review and assignation are performed.

However, such domain definitions have their limitations. The manual approach used by SCOP is inherently slow, and manual approaches are at risk of subjective decisions, particularly in edge cases. CATH attempts to minimise the reliance on human intervention, but that also comes with its own complications. For example, in the development of such

automated protocols, large numbers of smaller, experimentally cleaved, domains are used and comparatively few large and complex structures are available, which can result in automated approaches handling large and complex domains poorly (Veretnik et al., 2004). While these complications do arise in such methods, the use of automation has the considerable benefit of speed, necessary when thousands of novel structures are structurally characterised each year.

Despite categorising the same structures (those deposited in the PDB) and aiming to achieve similar goals (structural classification of protein domains), there are minor differences in classifications, due to the different approaches to identify domains and the domain definitions used. However, in general, CATH and SCOP have broad agreement (Hadley and Jones, 1999; Jefferson et al., 2008; Csaba et al., 2009).

Later developments looked to use CATH and SCOP data to classify all available sequences from completed genome projects. The Gene3D (Buchan et al., 2002) and SUPERFAMILY (Gough et al., 2001) resources use the domain assignments from CATH and SCOP respectively to assign domains to sequences based upon sophisticated sequence comparison tools (discussed in Section 1.5.3.1.1).

## 1.3 Protein structure prediction

Thanks to large-scale structural genomics initiatives, structural coverage of the protein universe has increased (Khafizov et al., 2014), but we are still a long way from having an experimental structure for every known protein sequence (Lees et al., 2014). Currently, sequence databases store approximately 55 million chains (http://www.uniprot.org/), of which only about 37,000 have been structurally characterised, either completely or partially (http://www.rcsb.org/). Major developments in genomic sequencing technologies have led to novel protein sequences being identified at an incredible rate, with which experimental structural characterisation cannot keep pace (Schwede, 2013; The Uniprot Consortium, 2015). As such, the gap between the number of known sequences and the number of available structures is ever increasing.

In an attempt to bridge the “sequence-structure gap”, computational approaches have long been devised with the aim of predicting the tertiary structure of proteins from their sequence. A number of different approaches have been proposed over the years.

A major area of protein structure prediction uses knowledge of protein structures already within the PDB in order to guide the modelling of a sequence without a known structure. The most popular approach is termed “comparative” or “homology” modelling which is based on the observation that the structure of a protein is more conserved than the encoding sequence during evolution (Chothia and Lesk, 1986). Therefore, the unknown structure of a protein can be approximated in the first instance by known homologous structures. Homologous templates are typically identified on the basis of sequence similarity using approaches such as those discussed in Section 1.5.3.1.1 (Söding and Remmert, 2011). If a template structure can be identified, the target and the template sequence are aligned and the atomic coordinates of the template are assigned to

corresponding target residues. When applicable, comparative modelling typically offers a reliable means of generating accurate structural models (Huang et al., 2014). Many different implementations of this methodology are available which have slight variations on the detection, alignment and use of template structures for modelling. Of these, Modeller (Šali and Blundell, 1993) is the most cited, though many other web servers and downloadable tools are available (some of the most widely used are outlined by Schwede (2013)). A more detailed account of the Modeller approach is provided in Section 4.2.1.

An alternative to template-based modelling is template-free modelling (also known as "*ab initio*" or "*de novo"* modelling) which is typically performed if a suitable template cannot be identified. Template-free modelling aims to generate the model of a protein without the use of a homologous template, generally attempting to mimic the folding of a protein chain based upon physicochemical principles. The concept underlying this approach is that the typical folded state of a protein should correspond to the lowest kinetically accessible energy state (Dobson, 2003). Therefore, by generating models and approximating their energy, a crude simulation of the traversal down the "folding funnel" may be achieved. The most successful of these methods generally use protein "fragments"; small peptides extracted from experimental structures in order to reduce the otherwise vast possible search-space, which are combined to form the model (Jones, 1997; Simons et al., 1997; Jones, 2001). However, exploring the conformational landscape of protein chains by fragment assembly generally requires considerable computational resources, and progress has remained modest, with sequences longer than 150 amino acids still posing considerable challenges (Kryshtafovych et al., 2014).

## 1.4 Modelling multidomain proteins

Despite the advances of experimental structural biology, the structures which have been deposited are not representative of those inferred to exist in nature. As mentioned previously, the majority of all proteins are estimated to be formed of at least two domains (Apic et al., 2001). However, within the PDB, single domain structures vastly outnumber multidomain ones, with multidomain structures comprising just 32.7% of the PDB after accounting for redundancy (Xu et al., 2015). This can largely be explained by the experimental approaches used for structure determination. Experimentally solving the structure of multidomain proteins is often problematic due to their large size and interdomain motions. In order to circumvent such issues, multidomain proteins are regularly cleaved at domain boundaries, and the smaller, more stable, individual domains are solved in isolation (Savitsky et al., 2010).

Protein domain assignment is therefore the first essential step towards the structural characterisation of proteins, either experimentally or computationally. Building on resources such as Gene3D (Buchan et al., 2002), SUPERFAMILY (Gough et al., 2001) and Pfam (Bateman et al., 2000), often it is possible to easily identify the domain architecture of the proteins of interest, using pre-calculated results from advanced sequence comparison tools (described in Section 1.5.3.1.1). Given the increasing coverage of completely sequenced genomes and the possibility to assign novel sequences to known superfamilies, these databases have outdated most of previous homology-based predictors for this task (e.g. Bryson et al., 2007). Homology-free methods are based on statistical analyses of sequence features that distinguish globular domains from linker regions or machine learning approaches to distinguish these two classes in a supervised way (Wheelan et al., 2000; Dumontier et al., 2005).

### 1.4.1 Multidomain comparative modelling

If a homologous structure can be found for a query sequence, with significant sequence similarity over the entire length of the protein of interest, then this can be used as a single template to model the target sequence, akin to the modelling of single domains. However, for multidomain targets, the variability of interdomain orientations adds further complications. If a template can be identified with a sequence identity greater than 30-40%, the orientation between domains should resemble the query sequence (Aloy et al., 2003). However, the orientation between domains can be substantially different, even at high sequence identity (Han et al., 2006). Whilst this means of comparative modelling may seem like an attractive strategy, suitable templates will simply not be available for the majority of query sequences.

### 1.4.2 Domain docking

As multidomain structures are often cleaved into individual domain components, in order to generate the full multidomain structure, separate domains must often be recombined. If a structure of a domain is not available within the PDB, it may be possible to generate a model using either template-based or template-free methods. Once a structure for each domain has been obtained, the next step consists of combining the individual domains into the full multidomain structure.

One approach to achieve this makes use of *in silico* “docking” which is intended to generate models of multicomponent structures from the unbound, constituent parts. Docking can be performed on a number of different biological systems, for example, between protein domains in order to model multidomain structures, between entire proteins in order to model protein complexes or between proteins and other biological

macromolecules, such as nucleotides (Inbar et al., 2005; van Zundert et al., 2015; Huang, 2015). Protein docking approaches were initially developed for the purpose of modelling protein complexes from separate protein chains. However, due to the similarities between interchain and interdomain interfaces (Jones et al., 2000), the problem of docking chains and domains can be considered to be equivalent. This is evidenced by the successful application of protein docking programs for domain docking (Lise et al., 2006; Inbar et al., 2005).

Docking takes place in two broad steps. Firstly, a large number of putative bound conformations are generated by sampling the conformational space. In order to reduce the large possible search space, backbone and sidechain motions are usually ignored during sampling, an approach known as “rigid body” docking (Huang, 2015). By keeping each component static, the search-space is reduced to 6 dimensions – 3 translational and 3 rotational. In order to combine constituent components, docking methods often further improve computational tractability by combining structures based on simplifications of shape complementarity. The most common of these approaches are based upon the fast Fourier transform (FFT), which was first introduced by Katchalski-Katzir and colleagues (1992), though other simplifications of molecular surfaces have also been employed (Duhovny et al., 2002; Huang, 2015). Whilst such approaches are computationally attractive, the rigid body simplification impacts the ability of these methods to successfully model cases where large conformational changes occur upon binding (Janin, 2010).

In the second step, generated models are ranked according to a scoring function. The sampling step generates a large number of putative models, some of which will hopefully resemble the native structure, with the vast majority of them not. Scoring functions attempt to rank models according to how well they are believed to resemble the native structure, often employing knowledge about expected levels of shape or chemical complementarity from structures observed in the PDB. However, whilst current functions are generally able

to identify near-native models from those unlike the native structure, they are not accurate enough to consistently identify optimal models from a set of near-native decoys, and model scoring remains an active field of study (Lensink and Wodak, 2013).

Once a small set of candidate structures have been identified, model refinement can be performed using more computationally intensive post-processing procedures permitting sidechain and backbone motions (Huang, 2015).

### 1.4.3 Domain assembly

In a similar vein to domain docking, domain assembly attempts to recapitulate the structure of a multidomain protein from separate domains, except these approaches make use of the knowledge that domains are connected via the chain (Cheng et al., 2008; Wollacott et al., 2007; Xu et al., 2014). By considering the linker between domains as a tether, the range of possible binding modes is considerably reduced in comparison to docking. Within this framework, a number of approaches have been proposed, which typically keep the individual domain structures unchanged, and alter the conformation of the interdomain linker in order to sample available tethered motions. After generating a variety of models, those with the lowest pseudo-energy scores are taken as the final solution, similar to model selection in docking.

The simplest approach to tackle this problem is the MultiDomain Assembler (MDA) (Hertig et al., 2015), which first finds close non-overlapping templates through a BLAST search, and then maps the local alignments onto the target sequence. An initial model is built by placing the individual templates at relative distances, which depend on the length of the inter-domain gaps observed in the alignment from the previous step, so that clashes and knots can be avoided. Finally, Modeller is used to build the missing linker regions and resolve interdomain packing and interactions.

A more sophisticated approach is the *Ab Initio* Domain Assembly (AIDA) (Xu et al., 2014; Xu et al., 2015) method, which generates an initial full-length model, where the linkers are modelled based on the secondary structure predicted by PSIPRED (Jones, 1999). Linker backbone torsion angles are subsequently perturbed in order to sample the range of possible motions. The final model is generated by minimising the model energy function which includes terms both to score the linker conformation and the resulting interdomain interactions.

A related approach has been implemented using the ROSETTA method and demonstrated on a set of two-domain proteins (Wollacott et al., 2007). Here, starting structures consisted of the two domains with the interdomain linker in a fully extended conformation. The conformational space of the linker was initially sampled using a low-resolution search, with the chain represented as the backbone and side-chain centroids. After this low-resolution pass, more intensive refinement was conducted after residue side-chains were restored via further small random backbone changes within the linker.

Work has also been conducted to use the location of the domain linker to aid in the ranking of docking models. Cheng and co-workers (2008) proposed to rank rigid body docking results with additional restraints derived from the conformations of domain linkers found in the PDB. To this end, they first collated a set of 542 regions from highly resolved X-ray structures spanning between 2 and 18 residues. For each linker, they calculated the end-to end distance as the distance between the Cα atoms of the N- and C-terminal residues, and they summarised the data through the mean and standard deviation in a length-dependent manner. Simple pseudo-energy terms were then defined to reward linker conformations with end-to-end distances within 1 standard deviation of the mean previously observed.

### 1.4.4 Guiding and scoring modelling procedures through predictions of interface features

Modelling can be assisted by including additional information indicating where an interaction is likely to occur (Wodak and Méndez, 2004). In theory, this can include experimental data, though this is seldom available. Alternatively, interface features can be predicted and used as restraints during modelling, or as a means of identifying native-like models from a set of alternatives. Such restraints can come from a number of different sources and some examples are briefly outlined below.

The most obvious starting point is sequence conservation, based on the observation that residues at the domain interface are more evolutionarily conserved than those exposed to the solvent (Littler and Hubbard, 2005). Therefore, solvent-exposed conserved residues may indicate putative binding sites (Glaser et al., 2003) and this information has been exploited by different groups in protein-protein docking (Oliva et al., 2013; Duan and Reddy, 2005).

Score based methods make use of various sequence and interface features to generate a scoring function relating to the likelihood of a surface region being an interface. These features can then be used to develop machine learning approaches to predict binding surfaces. Work in this area has predominantly been focused around the prediction of protein-protein interfaces (Hamer et al., 2010; Liang et al., 2006; Fariselli et al., 2002; Bradford et al., 2005), though similar scoring functions for interdomain prediction have been almost entirely neglected in the literature except just two studies by Lise and colleagues (2006) and Bhaskara and colleagues (2013).

Lise and colleagues (2006) used a range of different interface features (shape complementarity, residue-pair potentials, interface propensity, residue conservation and

correlated mutations) in order to select native multidomain structures from a set of docking models. Bhaskara and colleagues (2013) conducted a study of domain-domain interfaces and incorporated sequence conservation along with limited structural features (residue solvent-accessibility, protrusion and depression terms) in order to train a classifier for the prediction of intramolecular domain interfaces.

The next section outlines recent developments in the analysis of inter-residue covariation which has been demonstrated to help guide the docking (Hopf et al., 2014; Ovchinnikov et al., 2014) and scoring (Tress et al., 2005) of protein-protein models. To date, these approaches have not been evaluated in the context of interdomain modelling, though they offer a promising alternative source of restraints to those mentioned above.

## 1.5 Using covarying residues for protein structure prediction

### 1.5.1 Residue-residue contacts

During protein folding, distant parts of the chain come into close proximity and interacting pairs of amino acids form “residue-residue contacts”, or simply “contacts”. The relevance of contacting residues is due to their role in the stabilisation of native states as well as favouring or disfavouring non-native like states along the folding pathway (Gromiha and Selvaraj, 2004).

Contacts are typically defined using a distance threshold between specific atom types (e.g. Cα, Cβ or side chain heavy atoms). The most widely used cut-off value is the one employed in the Residue-Residue (RR) prediction category of the Critical Assessment of techniques for protein Structure Prediction (CASP) experiment, where residues are considered in contact if their Cβ atoms (or Cα in the case of Glycine) are within 8Å in the

experimental structure (Monastyrskyy et al., 2014). However, it is difficult to ascertain when two residues are no longer interacting given the dynamic nature of proteins, and the different chemical structures of the amino acids. A major limitation of the above definition is that it fails to account for size differences in the amino acid side chains. As such, other works have used alternative thresholds, which are more or less permissive, for instance accepting residues as contacting if any pair of heavy atoms (HA) is within 5Å (Marks et al., 2011; Martin et al., 2005; Skwark et al., 2013; Janin, 2010), HA-HA < 6Å (Weigt et al., 2009; Jones et al., 2012), HA-HA < 7Å (Jardin et al., 2013), HA-HA < 8Å (Morcos et al., 2011), HA-HA < 8.5Å (Feinauer et al., 2014) or with an inter-residue Cα-Cα distance < 12Å (Ovchinnikov et al., 2014).

The above binary distance cutoffs broadly relate to approximations of interatomic distances involved in biochemical interactions, such as Hydrogen bonds, VdW interactions and salt bridges. These cutoffs are broadly more generous than typically accepted maximum bond lengths as they also incorporate an element of tolerance which accounts for molecular motions exhibited in the aqueous environment.

Historically, contacts have been categorised into different groups based upon the sequence separation ($x$) between the two considered residues. Whilst exact boundaries differ slightly, contacts are generally categorised into short ($4 < x \leq 8$), medium ($8 < x \leq 23$) and long range ($x > 23$) (Tetchner et al., 2014). Short range contacts guide local interactions within secondary structure elements, whilst long range contacts provide information regarding how distant parts of the chain interact to form the global fold (Tanaka and Scheraga, 1975).

### 1.5.2 Correlated mutations and protein contacts

The structure of a protein is more conserved than the underlying sequence (Chothia and Lesk, 1986), with many diverse homologous sequences adopting the same fold. As residue contacts are responsible for maintaining the structure, the mutation of a residue will disrupt some local interactions and affect structural stability, or the efficacy of the protein to perform its function. If a mutation occurs, there are 3 potential outcomes: catastrophic mutations may cause the organism to die, and the mutation is not inherited; otherwise, the mutation may be reverted in subsequent generations, or the residues in immediate proximity may adapt in order to tolerate it (Maisnier-Patin et al., 2002).

The last scenario is called “correlated, “concerted” or “compensatory” mutation, and allows for the maintenance of the native structure and function of proteins during evolution (Poon and Chao, 2005; Altschuh et al., 1987; Yanofsky et al., 1964; Ohta, 1973; Vernet et al., 1992). Amino acids at the interface of interacting proteins can also covary in order to maintain favourable binding (Goh et al., 2000; Sandler et al., 2013; Urano et al., 2015; Mintseris and Weng, 2005). Compensatory mutations are surprisingly common, and occur more frequently than subsequent generations of organisms simply reverting the initial change (Poon and Chao, 2005; Maisnier-Patin et al., 2002; DePristo et al., 2005). This can be explained by the fact that by chance it is more likely to create viable neighbouring substitutions than to reverse at the originally mutated position (Maisnier-Patin et al., 2002).

With this in mind, homologous sequences can be considered as a record of the natural sampling of the sequence space available to folded functional proteins. By inverting the observation that positions covary in order to maintain the structure, structural or functional interdependencies between amino acids can be inferred from patterns of correlated mutations within homologues. Importantly, this concept provides a direct link between sequence and 3D structure, and can be turned into a predictive method.

### 1.5.3 Prediction of contacts using sequence covariation

#### 1.5.3.1 Multiple sequence alignments

In order to detect covarying positions, a comparison of the sequences between a protein of interest and its homologues must be performed. Firstly, homologous sequences are collected from public databases such as UniProt (Apweiler et al., 2004) by looking for similarities in their amino acid composition. Once a set of homologous sequences have been identified, a multiple sequence alignment (MSA) is built so that each row corresponds to a different homologous protein, and each column reports a set of amino acids that are evolutionarily related (i.e. they have evolved from the same residue in the last common ancestor and tend to preserve the relative structural position or functional role) (Edgar and Batzoglou, 2006). Reading down each column, conserved residues would be seen as the same amino acid and would suggest intolerance to mutations for critical structural or functional reasons. Mutations would be seen as alternative amino acids, and insertions or deletions seen as a “gap”, denoted by the dash symbol. An example MSA subsection is shown in Figure 1.6, where a conserved proline is highlighted in blue and covarying residues are coloured green and red.


**Figure 1.6: Subsection of a multiple sequence alignment.** A subsection of 6 aligned protein sequences. A conserved proline present in all sequences is shown in blue. Covarying residues (arginine and glutamic acid) are shown in green and red respectively.

#### 1.5.3.1.1 Homology detection and multiple sequence alignments

**BLAST and PSI-BLAST**

The Basic Local Alignment Search Tool (BLAST) method is a heuristic approach to identify homologous proteins (Altschul et al., 1990) based on the levels of sequence similarity to database matches (e.g. Henikoff and Henikoff, 1992). The BLAST algorithm starts by reducing a query sequence into a series of short “words”, which are compared against equivalently-formed words from all sequences within a database. Initial matches are extended in both directions in order to maximise the alignment. This approach identifies a set of putative homologues, each of which is aligned pairwise to the query sequence.

As the initial list of homologous sequences contains specific information about individual positions, such as whether it is conserved or has a particular charge preference; this information is useful to repeat the search and find further, more remote, homologues. The Position-Specific Iterated BLAST program (PSI-BLAST) (Altschul et al., 1997) uses information from an initial BLAST search in order to inform subsequent searches. For this purpose, BLAST results are converted into a Position-Specific Scoring Matrix (PSSM), containing the propensity for each amino acid to appear at each position. The PSSM then guides the subsequent iteration of the database search by replacing the standard BLOSUM62 substitution matrix in the calculation of the alignment score. This procedure can be iterated multiple times (usually three) and permits PSI-BLAST to be more sensitive to remote homology than the standard BLAST approach.

In order to quantify the reliability of the retrieved sequences to represent genuine homologues of the input sequence, E-values for the alignment scores are calculated. The E-value represents the number of hits with a similarity score greater than or equal to that

under consideration, which one would obtain by chance given the size of the database, the length of the input sequence, as well as the scoring system used.

**HMM-based approaches**

An advancement in sequence comparison was made with the use of hidden Markov models (HMMs), which build a more sophisticated statistical model of an observed series of data (in this case the mutating amino acids of proteins) (Eddy, 1996). HMMs capture information for a protein family relating to site-specific propensities for accepted amino acid types and tendencies for insertion or deletions. By using site-specific gap penalties, HMM-based approaches are able to penalise non-homologous sequences more appropriately than homologous sequences which tend to have gaps in the same location as the model built for the family (Söding, 2005). HMM-based methods have been demonstrated to provide greater capability for detecting remote homologues than using PSSM-based approaches (Park et al., 1998).

In a similar fashion to PSI-BLAST, this process can be iterated in order to identify more distant homologous sequences by incorporating identified homologues into a search profile (Remmert et al., 2011; Johnson et al., 2010). The two most prevalent programs which perform iterative profile-HMM comparison are HHblits (Remmert et al., 2011) and jackHMMer (Johnson et al., 2010).

HHblits performs HMM-HMM comparison for a query protein HMM (generated by HHsearch (Söding, 2005)) against a database of template profile-HMMs of sequences clustered at low sequence identity. After each iteration, the search profile HMM incorporates sequences from matched database HMMs in order to guide further steps.

JackHMMer executes profile-HMM comparisons against standard sequence databases using a series of database filtering steps to reduce the otherwise considerable search

space. Once again, homologous sequences identified in initial searches are incorporated into the profile-HMM to inform subsequent iterations.

#### 1.5.3.2 Initial approaches to identify covarying positions

In early work, residue covariation was calculated from MSAs using measures taken from information theory, such as Mutual Information (MI), or other similar concepts (Altschuh et al., 1987; Neher, 1994; Taylor and Hatrick, 1994; Göbel et al., 1994; Tress et al., 2005). Whilst these initial studies were able to identify some structural contacts, they also commonly identified residue pairs which displayed covariation, yet were observed at long distance within the experimental structure.

Later, protein-specific measures tailored specifically for the analysis of amino acid covariation were introduced, such as the McLachlan Based Substitution Correlation (McBASC) (Göbel et al., 1994; Olmea et al., 1999), Statistical Coupling Analysis (SCA) (Süel et al., 2003) and the Observed Minus Expected Squared (OMES) (Kass and Horovitz, 2002). These approaches improved the ability to identify structural contacts, though many incorrect predictions were still present for the reasons listed below.

### 1.5.4 Problems of bias within the analysis of MSAs

At the turn of the millennium, Lapedes et al. (1999) realised that MI calculations were affected by two major sources of bias within MSAs: phylogenetic bias and the chaining effect. The following year, Wollenberg and Atchley (2000) reported that predictions were additionally affected by chance occurrences of covariation (entropic bias). From this, the authors concluded that the observed level of covariation between a pair of MSA positions

was the result of covariation due to chance, the underlying phylogenetic relationship between source species as well as relevant structural and functional constraints.

#### 1.5.4.1 Phylogenetic bias

Due to the evolutionary relationships among species, aligned sequences do not represent independent samples (Felsenstein, 1985), and this tends to skew MI estimates of covariation, especially when closely related sequences are considered. As a demonstration of this effect, the background phylogeny of an artificially “evolved” set of sequences was demonstrated to generate substantial MI signal, even in the absence of structural or functional dependencies (Lapedes et al., 1999).

Figure 1.7A shows a toy MSA which is intended for covariation analysis and the inference of residue contacts (represented as coloured boxes). Naïve MI estimates assume the simplest underlying phylogeny, where all observed sequences evolved independently after a single instantaneous creation event (represented as a “star” phylogeny), though this is obviously incorrect (Figure 1.7B). In fact, the observed covarying positions may have been caused by either one very ancestral covariation event which has then stabilised in subsequent generations (Figure 1.7C), or it may be that it occurred independently twice, in separate branches of the phylogeny (Figure 1.7D). Therefore, to avoid incorrect interpretations, analyses of recurring features such as conservation or correlated mutations need to account for the underlying phylogeny of the aligned sequences.

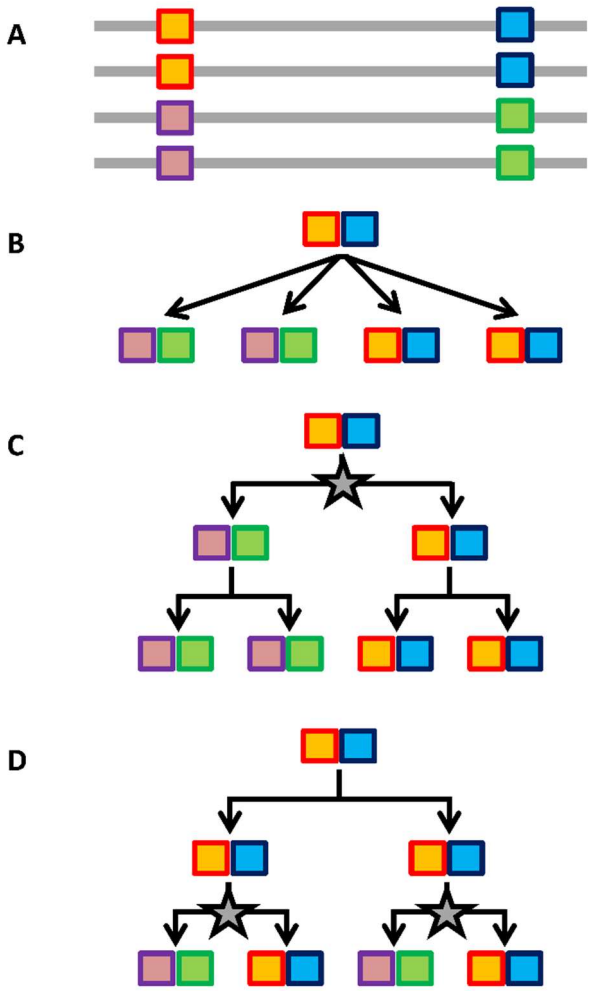



**Figure 1.7: Overview of phylogenetic bias.** A) Covariation analysis seeks to identify pairs of residues which are covarying in a sequence (e.g. orange and blue, purple and green), indicative of a structural interaction. B) The simplest model (a “star” phylogeny) assumes that all observed sequences have descended from a single ancestral sequence, evolving independently. Depending on the true underlying phylogeny, the same observed sequences may be caused by either a single correlated mutation event (as in scenario C) or from two separate events (as in D). Correlated mutation events are displayed as grey stars. Image reproduced from (Tetchner et al., 2014).

#### 1.5.4.2 Entropic bias

Entropy is a measure of uncertainty in an observed sample (Cover and Thomas, 1991); within a MSA column it indicates the level of variation at a specific site. Fully conserved positions would have an entropy score of 0, whilst variable positions would have higher scores. The MI between two positions can be calculated in terms of the entropy observed at both sites (Cover and Thomas, 1991) and due to this relationship, MI strongly correlates with the entropy of the two considered positions (Martin et al., 2005; Fodor and Aldrich, 2004). In fact, the MI between a pair of positions within an MSA will only have a value of 0 if the observed pair frequencies reflect all possible pairings for the observed single position amino acid frequencies (Martin et al., 2005). As such, there will almost always be a residual level of MI between any pair of positions. In order to quantify the level of background MI arising from these finite sampling effects, Martin et al. (2005) simulated the evolution of artificial sequences and measured the level of background MI which arose. The authors report that the level of background MI reduced with increasing numbers of sequences and these effects were small if the number was greater than approximately 150.

#### 1.5.4.3 Reducing phylogenetic and entropic biases

Once the above sources of spurious covariation between MSA columns had been identified, attempts were made to reduce their effects. Dunn et al. (2008) proposed that observed covariation can be considered as the result of structural and functional effects (sf) as well as the background phylogeny and positional entropy (b) in a study using MI, such that:

$$MI = MI_{sf} + MI_{b}$$

In order to approximate $MI_b$, the authors introduced the Average Product Correction (APC), which can be calculated from the MSA. The APC for two positions, $i$ and $j$, is calculated by:

$$APC(i,j) = \frac{MI(i,\bar{x})MI(j,\bar{x})}{\overline{MI}}$$

Where $MI(i,\bar{x})$ is the mean MI for column $i$ and the average MI for all other columns (except $i$), calculated by:

$$MI(i,\bar{x}) = \frac{1}{n-1}\sum MI(i,x)$$

Where $n$ is the number of columns in the alignment and $x = 1$ to $n$, where $x \neq i$. The equivalent calculation is performed for $j$.

$\overline{MI}$ represents the overall mean MI calculated by:

$$\overline{MI} = \frac{2}{n(n-1)}\sum MI(i,j)$$

Finally, in order to apply the APC, the calculated value is subtracted from the standard MI score for the two positions:

$$MIp(i,j) = MI(i,j) - APC(i,j)$$

The APC was applied to the standard MI measure under the name MIp, and was demonstrated to identify contacting residues more precisely than unaltered MI. Since its inception, the APC has become the most widely used approach to handle entropic and phylogenetic biases for contact prediction, though other approaches do exist (Little and Chen, 2009; Gloor et al., 2010; Martin et al., 2005).

Whilst simple, approaches such as the APC have been demonstrated to perform as well as more sophisticated methods, which directly attempt to infer the underlying phylogenetic

tree (Caporaso et al., 2008). APC-like approaches are often many orders of magnitude faster than tree-based approaches, making them suitable for analysing many thousands of protein sequences.

### 1.5.5 The chaining problem

Identifying genuine instances of covariation is further complicated due to the effect of multiple instances of covariation occurring simultaneously. For a set of covarying elements within a system (e.g. covarying positions within an MSA), the effect of multiple covarying pairs causes “chained” covariance (Lapedes et al., 1999). Chained covariance is the observation that covariance can propagate along a chain of interacting elements (Figure 1.8).

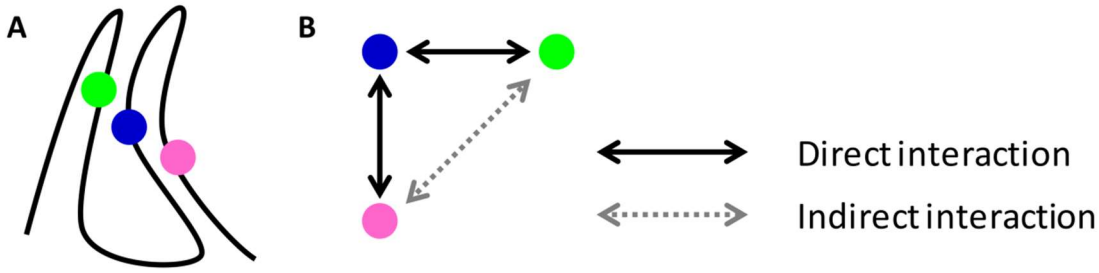



**Figure 1.8: Overview of the chaining problem using a toy example.** A) Three residues are spaced along the chain. The residues green and blue are in contact, and blue and pink are in contact (“direct” interactions). B) Due to the shared interaction with the blue residue, the green and pink residues display covariation (an “indirect” interaction), even though there is no direct interaction between them.

Consider the example in Figure 1.8A, where there are three residues interacting linearly, where green interacts with blue, and blue also interacts with pink. In this example, covariance displayed between green and blue, and blue and pink (termed “direct”

interactions) will also display covariance between the green and pink residues. Without further knowledge, it would appear that all three residues interact, whereas in reality this is not the case (Figure 1.8B). The observed covariance between the non-contacting green and pink residues is termed “indirect” correlation. In larger systems, such as in protein MSAs where there may be hundreds of interacting positions, these chains of covariance are obviously much more complicated. One might assume that differentiating between direct and indirect correlations could be dealt with simply by removing weak instances of correlation. However, multiple weak direct correlations can lead to strong indirect interactions – an effect which has been termed “superadditive correlation”, so this would prove futile (Giraud et al., 1999).

### 1.5.6 Recent approaches to tackle the chaining problem

While biases arising due to the underlying phylogeny and entropy can be reduced by utilising approaches such as sequence weighting and the APC, practical methods to address the chaining problem remained elusive until more recently. A breakthrough was made with the use of “global” statistical models of covariation, which attempt to isolate the effect of single pairs of covarying amino acids after accounting for the effect of all other instances of covariation. Global statistical modelling techniques treat pairs of correlating residues as dependent on one another, which minimises the effect of chaining and noise within the data (Marks et al., 2012). The basis of these approaches is that a global model of covariation would be able to explain the observed covariation within an alignment. However, calculating the exact form of such a model would require vast computational time (Balakrishnan et al., 2011), so instead simplified models are generated from the aligned sequences. In recent years a number of different approaches to approximate such models of covariation have been proposed, which are summarised below.

#### 1.5.6.1 Maximum entropy approaches

A major turning point towards tackling chaining effects was brought about by the theoretical work of Lapedes and co-workers (1999), who proposed a maximum-entropy approach for generating a model to distinguish between direct and indirect coupling effects. The concept of maximum entropy states that the statistical model which best describes a set of given data (here, the observed amino acid states in an alignment) should be the one with maximal entropy. By maximising the model entropy, the selected model is the least biased to unobserved data (Jaynes, 1957). However, the high computational demands and the lack of available sequence data meant that the suggested approach was not feasible for real test cases, and so the promise and potential of the approach went broadly unappreciated.

Since then, a number of different groups have devised similar approaches built upon the same maximum entropy principle (Stein et al., 2015). The prospect of using the maximum entropy approach was resurrected by Weigt and colleagues (2009), who used a more computationally efficient message-passing approach to approximate the global model, employing the maximum entropy principle in order to prevent overfitting (Weigt et al., 2009). Due to still considerable computational demands, this work analysed 60 positions (pre-selected using standard MI) of a paired alignment from the highly abundant bacterial two-component regulatory proteins. This approach was named “Direct Coupling Analysis” (DCA) after the ability of the method to differentiate between the direct and indirect covariation.

Maximum entropy-based approaches aim to represent an observed MSA as a 21-state Potts model, with each sequence position described as the frequency of observing one of the standard 20 amino acids or a gap (Ekeberg et al., 2013). In this model, each protein

sequence within the MSA can be thought to represent an independent sampling event taken from an underlying Potts-model probability distribution (Ekeberg et al., 2014). Therefore, if a general model can be approximated which could give rise to the observed sequences, this model could be interrogated to identify positions which covary. As such, approaches of this ilk attempt to generate such a model of the following form:

$$P(A_1 \dots A_L) = \frac{1}{Z} \exp\left\{ \sum_{i<j} e_{ij}(A_i A_j) + \sum_i h_i(A_i) \right\}$$

Where $A$ represents an entire protein sequence, $i$ and $j$ are positions within each sequence, $e_{ij}$ relates to pairwise couplings between positions which in this application mean covarying residue pairs, and finally $h_i$ which corresponds to local biases within the models relating to amino acid conservation.

However, the $Z$ term, known as the "partition function", is not directly computable due to the large numbers of parameters, requiring unreasonable time and data requirements necessary to calculate the exact solution. Therefore, approximations of the term must be computed instead. Approaches to approximate the partition function have received a lot of attention in recent years. Some of the key methods which have introduced alternative approaches to approximate this function are outlined below.

**Mean field approaches**

The first of the newest generation of methods to reduce the chaining effect used a mean field (MF) approximation approach to estimate a global model of covariation. The approach devised by Morcos and colleagues (2011) enabled the costly parameter learning process of the original message passing approach to be dealt with in a single efficient step. The exponential of the $e_{ij}(A_iA_j)$ term can be expanded into a Taylor series, from which the mean-field equations can be obtained:

$$\frac{f_i(A)}{f_i(q)} = exp\left\{h_i(A) + \sum_A \sum_{j \neq i} e_{ij}(A,B) f_j(B)\right\}$$

where

$$e_{ij}(A,B) = -(C^{-1})_{ij}(A,B)$$

and

$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)f_j(B)$$

This enables the mfDCA calculation to be orders of magnitude faster than the approach proposed by Weigt and colleagues, permitting the analysis of many hundreds of positions, enabling full length protein sequences to be analysed. Around the same time, the EVfold (Marks et al., 2011; Marks et al., 2012) method employed the same mfDCA calculation and demonstrated that the predicted contacts were sufficiently accurate to generate reliable protein structure models for single domains.

**Pseudolikelihood maximisation approaches**

Global models were further refined with the use of pseudolikelihood maximisation (PLM) methods (Balakrishnan et al., 2011; Ekeberg et al., 2013). The maximum likelihood estimate for model parameters is guaranteed to recover the true parameters as the

quantity of data increases. However, again, such a calculation is computationally intractable, as mentioned previously. In order to make this calculation feasible and avoid parameter overfitting, the pseudo log-likelihood is calculated instead, which replaces the approximation of the global partition function with local partition functions (Balakrishnan et al., 2011).

Here the pseudo log-likelihood (pll) of the model parameters $\theta$ is calculated by:

$$pll(\theta) = \frac{1}{n}(\sum_{X^i \in \mathrm{X}} \sum_{j=1}^{p} \log(P(X_j^i | X_{\_j}^i))$$

which is equal to

$$pll(\theta) = \frac{1}{n} \sum_{X^i \in \mathrm{X}} \sum_{j=1}^{p} \times \left[ log\phi_j(X_j^i) + \sum_{\mathrm{k} \in v_j'} log\psi_{jk}(X_j^i, X_k^i) - logZ_j \right]$$

Where $X_j^i$ is the residue at the j[th] position of the i[th] MSA sequence. $X_{\_j}^i$ is the "Markov blanket" of $X_j^i$, $Z_j$ is a local normalising contact for each node of the calculated Markov Random Field and $v_j'$ is the set of all vertices which connect to vertex $j$ in the model (Balakrishnan et al., 2011).

Not long after the initial implementation, the CCMpred method was introduced which reimplemented the standard plmDCA approach, but focused on improving the speed of calculations (Seemayer et al., 2014). By improving calculation runtime, it was hoped that analyses using plmDCA methods could be applied to larger systems, such as multidomain proteins and protein complexes. The authors report significant improvements over the original implementation for both processor-based (CPU) and graphics card-based (GPU) implementations.

An attempt to extend the PLM approach was later suggested incorporating structural priors to aid in cases where little sequence information is available, but the effect of these priors on performance appeared minimal (Kamisetty et al., 2013). Later, the gplmDCA method introduced specific terms to account for stretches of gaps in MSAs which were identified as a common cause of false positive predictions in the earlier plmDCA approaches, thus improving the accuracy of the results (Feinauer et al., 2014).

#### 1.5.6.2 Sparse inverse covariance estimation approaches

Partial correlation coefficients are a means of distinguishing between direct and indirect coupling effects (Jones et al., 2012). Partial correlation coefficients assess the amount of dependency between two variables, after removing the effect of all other variables (Jones et al., 2012; Friedman et al., 2008; Meinshausen and Bühlmann, 2006). Typically, partial correlation coefficients would be extracted by inverting the covariance matrix calculated for a dataset, to obtain the "precision" or "concentration" matrix. However, a covariance matrix calculated from protein sequence data is guaranteed to be singular as not every amino acid will be observed at every position (Banerjee et al., 2008; Jones et al., 2012). As the covariance matrix cannot be directly inverted, in order to gain access to these partial correlation coefficients, an approximation of the precision matrix can be made using statistical methods such as the graphical Least Absolute Shrinkage and Selection Operator (LASSO (Friedman *et al.*, 2008; Tibshirani, 1996)).

This approach was first implemented by the Protein Sparse Inverse COVariance (PSICOV) method (Jones et al., 2012). PSICOV employs the LASSO to minimise the objective function:

$$\sum_{ij=1}^{d} S_{ij}\theta_{ij} - \ logdet\theta + \rho \sum_{ij=1}^{d} |\theta_{ij}|$$

Where $i$ and $j$ represent two columns of the considered MSA, $S$ represents the empirical covariance matrix calculated from a sequence of $d$-dimensional vectors and $\theta$, the concentration matrix.

The third term of the above expression is the so-called “shrinkage parameter”, a type of regularisation or penalty term. $\rho$ is a positive parameter which controls how many of the components in the matrix $\hat{\theta}$ will be set to zero. This exploits the knowledge that most of the theoretically possible contacts are not formed within folded protein chains. Consequently, the regularisation term can be included to explicitly introduce sparsity into the calculated solution, increasing computational efficiency. Based on the same underlying idea of approximating the inverse of the covariance matrix, an alternative approach using regularized least squares regression was devised, which showed similar performance to PSICOV, but with improved calculation time on a small set of proteins (Andreatta et al., 2013).

A further development related to the PSICOV approach incorporated related sequence families using the group graphical LASSO into covariation models (Ma et al., 2015). Under the assumption that related sequence families are likely to share the same fold, the number of sequences available for analysis can be increased. For each alignment a distinct graphical model is generated, and models are combined using a random forest classifier.

#### 1.5.6.3 Effect of global statistical methods for the prediction of contacts

Recently developed global statistical approaches have proven to be effective at identifying structural contacts, substantially improving upon previous “local” methods. Thus far, comparisons between these different approaches have been tested exclusively on the ability to identify contacts within domains. In these comparisons, all of the “chaining-aware”

approaches have been demonstrated to improve upon methods such as MI and bias-corrected approaches such as MIp (Jones et al., 2012; Kamisetty et al., 2013). These studies are broadly consistent in finding that PLM-based methods (such as plmDCA and CCMpred) generally outperform PSICOV's sparse inverse covariance approach, both of which in turn improve upon the mean field approach employed by EVfold (Ma et al., 2015; Jones et al., 2015; Tetchner et al., 2014).

#### 1.5.6.4 Scoring of predicted contacts

The aforementioned methods generate large numbers of contact predictions for a given MSA, with each method generating approximately equal numbers of predictions in total. All methods rank generated predictions according to a score, which differs depending on the approach. The output score of MIp is simply the APC-adjusted MI value, CCMpred uses an APC-adjusted Frobenius norm, EVfold uses a "Direct Information" metric and PSICOV makes use of the APC-regularised L1-norm.

#### 1.5.6.5 Combining covariation-based approaches with machine learning

As the above methods are able to increase the ability to identify structural contacts, groups began to investigate the similarities - and perhaps more importantly - the differences in the contacts predicted by each approach. Figure 1.9 shows the overlap of correctly predicted contacts from 3 methods employing different models of covariation for the same set of MSAs. From the figure it is clear that the different approximations used to generate models result in a number of correct predictions which are unique to each method, even if the majority of correct contacts (57.3%) are identified by all three.

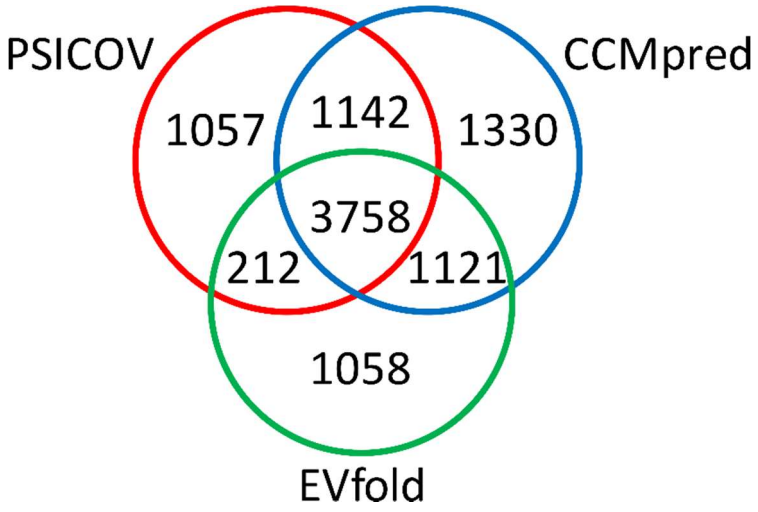



**Figure 1.9: Venn diagram of the overlap for 19,669 correct contact predictions, predicted by PSICOV, EVfold and CCMpred.** The data represent the top L/2 correct intradomain predictions (sequence separation > 4) for a set of 150 proteins (Jones et al., 2012), where L represents the number of amino acids within the protein chain (Jones et al., 2015).

In an attempt to make use of the correct predictions which are unique to each method, meta-predictors (methods combining predictions from different approaches) have been developed with the aim of increasing contact precision (Jones et al., 2015; Skwark et al., 2013; Skwark et al., 2014). The first developed meta-predictors were the PconsC1 (Skwark et al., 2013) and PconsC2 (Skwark et al., 2014) methods. These methods combined predictions from plmDCA and PSICOV along with 8 different MSAs for a query protein using a random forest classifier. Whilst this improved performance compared to the individual methods, the requirement of generating 8 separate alignments is time-consuming. This prompted the development of the MetaPSICOV approach (Jones et al., 2015), which combines predictions from EVfold, CCMpred and PSICOV for a single MSA along with other structural predictions such as secondary structure and solvent accessibility using a neural network. All of these methods were demonstrated to improve the quality of intradomain contact predictions. MetaPSICOV (submitting predictions under

the name “consip2”) was the best performing method in the most recent CASP RR assessment (Kosciolek and Jones, 2015; Monastyrskyy et al., 2015).

Both the PconsC2 and MetaPSICOV approaches employ machine learning in order to learn typical patterns of contact formation from experimental protein structures. By learning these patterns, the methods are able to remove predicted contacts that are likely to be incorrect, as well as adding contacts which are likely to be present within the native structure, though not identified as covarying. Such contacts may include adjacent residues within predicted beta strands or residues with a sequence separation of 4 within predicted alpha helices. Whilst revising the set of predicted contacts in this manner was demonstrated to improve precision scores, the gains when applied to protein modelling were small due to the introduced redundancy (Jones et al., 2015).

### 1.5.7 Applications for predicted contacts

Alongside the development of methods to reduce the effects of phylogeny and chaining, there has been considerable work employing contacts identified by these methods for the prediction of protein structure. Because these analyses provides a direct link from sequence to structural restraints, there has been a lot of excitement in recent years about the applications for these contacts.

The most obvious application is in tertiary structure prediction. Intrachain contacts derived from covariation analyses have been successfully used to restrain the *ab initio* modelling of globular protein structures, with a focus on single protein domains (Marks et al., 2011; Kosciolek and Jones, 2014; Michel et al., 2014). These predictions have also proven to be particularly useful to accurately model transmembrane protein structures; both alpha helical (Hopf et al., 2012; Nugent and Jones, 2012; Hopf et al., 2015) and beta barrels (Hayat et al., 2015).

There has also been work demonstrating that residue covariation is detectable not only between residues within the same protein, but also between interacting proteins. By aligning interacting pairs of protein sequences in a “joined” alignment, contacts can be detected in a similar manner to those within the chain, and can be used to guide protein-protein docking procedures (Ovchinnikov et al., 2014; Hopf et al., 2014).

Groups have also made use of predicted contacts to study proteins which undergo large conformational changes to perform their typical function (Morcos et al., 2013). In these cases correlated mutations can identify residues which come into close proximity within each of these functional states, providing insight into complex molecular motions, along with transient, intermediate states.

## 1.6 Thesis overview

The aim of this thesis is to investigate the capability of recent advances in the analysis of covarying residues for the task of identifying interdomain contacts from sequence. In order to make the problem more tractable, work will be focused on the simplest multidomain case – proteins comprising two domains. Previous work has demonstrated that covarying residues are present at the domain-domain interface and can be identified using MI-based approaches (Gomes et al., 2012). Since then, other studies have shown that the best performing approaches employed by Gomes and colleagues are outperformed by global statistical approaches for intradomain contact prediction (Jones et al., 2012; Kamisetty et al., 2013). However, whether chaining-aware approaches improve interdomain contact prediction has not previously been investigated, and provides the motivation for the work conducted in this thesis.

The next chapter describes the development of an approach to identify interdomain contacts. Four different approaches to detect covarying residues (CCMpred, EVfold,

PSICOV and MIp) are tested and the ability of each method to identify correct interdomain contacts is assessed. Procedure development is based upon a generated dataset of two-domain protein structures gathered from the PDB.

Chapter 3 focuses on using predicted contacts as a means of modelling proteins using a domain docking approach. After generating a set of alternative models using the PatchDock docking program, near-native structures are identified from the set of alternatives using predicted contacts to rank each generated model based upon the number of observed predictions.

The fourth chapter describes the use of predicted contacts as *a priori* constraints in a modelling procedure. The widely-used comparative modelling method Modeller is used to assemble separate domain structures, making use of predicted interdomain contacts in the form of additional distance restraints.

Finally, the fifth chapter reviews the work contained in this thesis and speculates about potential avenues for future studies to expand upon the findings outlined.

# 2. Analysis of covarying residue pairs spanning protein domains

## 2.1 Introduction

With recent developments in approaches to distinguish between direct and indirect contacts, numerous groups have applied these methods for the analysis of covarying residues in two main areas: within single domains, and across proteins chains. These studies have found that covarying residues are generally observed to be in close structural proximity. However, thus far, no group has explicitly assessed the ability of these methods to detect contacts between intramolecular domains.

This chapter describes an investigation into whether interdomain contacts can be identified from sequence data using current covariation-based techniques. In order to make this investigation more tractable, this study was conducted on the simplest multidomain case – two-domain proteins. Here we describe the development and evaluation of an approach to identify interface contacts using large multiple sequence alignments (MSAs), and outline the testing procedure. In order to evaluate the developed approach, a dataset of experimentally solved two-domain protein structures was assembled from the PDB. Using this dataset for reference, the performance of four different methods to identify covarying residues was assessed.

## 2.2 Method

### 2.2.1 Dataset

In order to verify if observed covariation relates to structural proximity, a dataset of experimentally-determined protein structures was required. The CATH database (Orengo et al., 1997) was used for domain assignments, and acted as a natural entry-point to identify structures relevant for analysis.

Inclusion criteria for proteins used in the dataset for benchmarking and evaluation are outlined in Table 2.1.

| Criterion | Number of examples |
|---|---|
| All CATH chains | 187,125 |
| Two-domain chains | 54,417 |
| Continuous two-domain chains | 32,062 |
| Structure solved by crystallography | 26,537 |
| Resolution ≤ 2.3Å | 12,526 |
| 50 ≤ x ≤ 500 residues | 11,736 |
| Chains with < 25 residues missing compared to UniProt sequence | 6,259 |
| Largest chain from each PDB file selected | 3,234 |
| Standard and unmodified amino acids | 2,917 |
| Unique UniProt entry | 1,063 |
| Unique homologous superfamily pairing | 332 |
| Exclude homomultimers | 116 |
| Minimum of 10 interface contacts | 90 |
| Large initial sequence alignments | 37 |

**Table 2.1: Overview of selection criteria for the dataset used to benchmark alignment parameters.**

Each criterion is explained in further detail below.

**All CATH chains**

Information relating to structural domains was obtained from the latest release of the CATH database (version 4.0.0, (Orengo et al., 1997)). The CATH database identifies structural domains for proteins with experimental structures available within the PDB (Bernstein et al., 1978; Berman et al., 2000).

**Two-domain chains**

The protein chains in the CATH database consisting of only 2 domains, and each domain has a full CATH classification.

**Continuous two-domain chains**

The 2 domains were continuous along the chain. Discontinuous domains were excluded from consideration in order to avoid complications of identifying such domains using sequence-similarity approaches (Bateman et al., 2004). The simultaneous development of two different alignment protocols was not attempted, and research was focused on the more common continuous domains (Jones et al., 1998).

**Structure solved by crystallography**

The experimental structure was solved by X-ray crystallography.

**Structure resolution ≤ 2.3Å**

High resolution structures provide detailed information about a protein, necessary to reliably place each amino acid. As contacts are determined by inter-residue distances, low-resolution structures would introduce ambiguity about whether two residues form a contact.

**Sequence length between 50 and 500 residues**

Protein sequences were selected using the above criteria to gain a diverse range of structures. An upper-bound on the length of sequence was used to exclude structures which are unrepresentative of most two-domain proteins (Figure 2.1). Additionally, as the memory usage of the covariance methods scales primarily with sequence length, limiting the number of analysed positions also has practical benefits. 45 unique pairings were excluded in this manner.

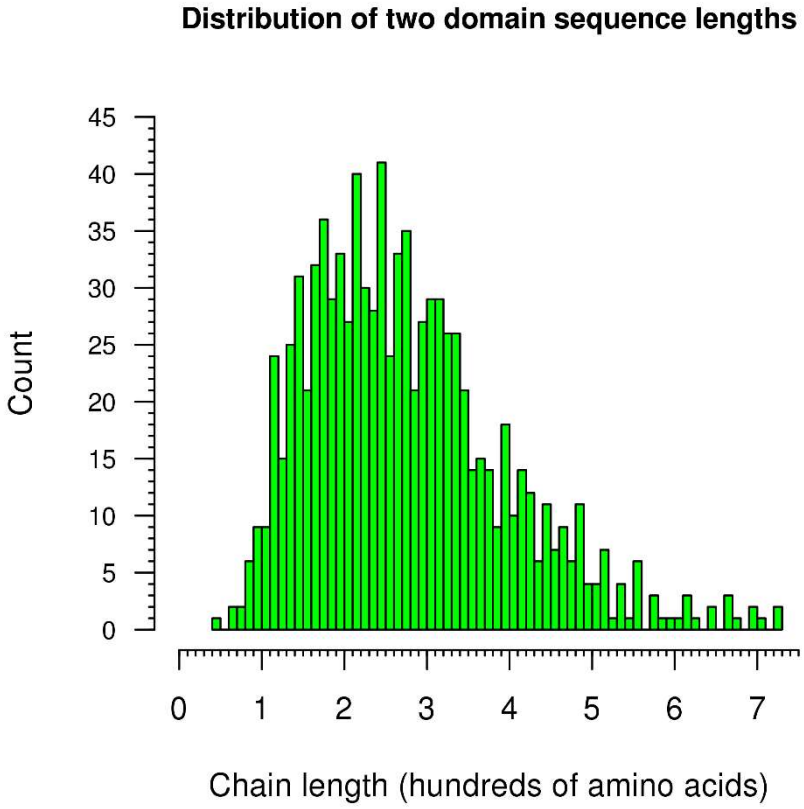



**Figure 2.1: Length distribution of all 916 unique CATH two-domain pairs.** All 2-domain proteins from the CATH database (release 4.0.0) were analysed and average protein lengths were calculated for unique pairings at the homologous superfamily level.

**Chains with fewer than 25 residues missing compared to UniProt sequence**

Proteins were excluded if they were missing a total of 25 or more amino acids after comparing the SEQRES record of the PDB file to the sequence deposited in UniProt (Apweiler et al., 2004). By doing this, unobserved lengths of sequence which may fold into additional domains were accounted for. Whilst domains do exist with fewer than 25 residues (Figure 1.4), these are rare, representing just 0.04% of the 235,858 domains within the CATH database.

**Largest chain from each PDB file selected**

Each PDB file may contain multiple copies of the crystallised protein. If this was the case, the structure with the longest chain was selected. If all structures contained the same number of residues, the chain occurring first alphabetically was chosen.

**Standard amino acids**

Proteins were required to contain standard amino acids for analysis with the covariance methods, so chains with unnatural or modified natural amino acids were excluded. Two exceptions were made in the cases of selenomethionine and selenocysteine, which are routinely used within crystallography to aid structure determination (Hendrickson et al., 1990; Strub et al., 2003). Where applicable, these residues were converted to methionine and cysteine in the query sequence, respectively.

**Unique UniProt entry**

The PDB contains many duplicates of the same proteins. As a first step to reduce redundancy within the dataset, the UniProt primary accession number for each PDB entry was obtained using the SIFTS resource (Velankar et al., 2013) and one structure for each UniProt identifier was selected, prioritising longer chains.

Many of the proteins which passed the criteria to this point had numerous alternative deposited structures. For example, Human cyclin-dependent kinase 2 (UniProt Accession Number P24941) had 177 suitable high resolution structures.

**Unique homologous superfamily pairing**

To further reduce redundancy, proteins were selected so that each CATH homologous superfamily level pairing was unique. A domain from a particular homologous superfamily may appear more than once in the dataset if the partner domain in each case was different. This was performed rather than filtering sequences based on percentage sequence identity in order to take advantage of the manually curated domain information from CATH. Diverse sequences within the same superfamily may be missed based upon sequence similarity. A similar approach to reduce dataset redundancy has been employed previously (Jones et al., 2000).

**Exclude homomultimers**

The covariation signal of homomultimeric interactions will contaminate the signal arising from intrachain covariation. Proteins which perform their normal function as homomultimers are likely to have evolved interchain interfaces which also exhibit covariation. If homomultimers were not excluded from this dataset, then covariation signal will arise both from intrachain interactions, as well as those occurring between chains.

The biological unit was determined using the “REMARK 350” field of the PDB file header. Within the “REMARK 350” field, there are typically two suggestions of the likely biological unit: one provided by the depositing authors, and the other predicted by the Proteins, Interfaces, Structures and Assemblies (PISA) program (Krissinel and Henrick, 2007). Where possible, the biological unit of a protein was taken from the depositing author’s remark field. In cases where the author proposes multiple biological units, the biological unit in agreement with the current version of the PISA program via the web server

(http://www.ebi.ac.uk/pdbe/pisa/) was accepted, as the PISA prediction in the “REMARK 350” field is not automatically updated with changes in the PISA software (RCSB PDB, pers. comm., 19/2/2014). If a biological unit was not specified by the depositing author, the biological unit predicted by PISA was taken.

Protein chains forming part of heteromeric structures were permitted, provided that the heteromeric structure contained only a single copy of the chain, in order to remove potential homooligomeric effects within larger assemblies.

**Minimum of 10 interface contacts**

Structures were required to contain a minimum of 10 interdomain van der Waals (vdW) contacts (with a vdW contact threshold defined as the sum of the two vdW radii of a pair of atoms + 0.5Å (Bondi, 1964)), in line with other work examining interdomain interactions (Aloy, et al., 2003; Jefferson, et al., 2008; Park, et al., 2001). This criterion was applied in order to ensure there was a significant interdomain interface for prediction.

One example of a protein removed by this criterion is shown in Figure 2.2.

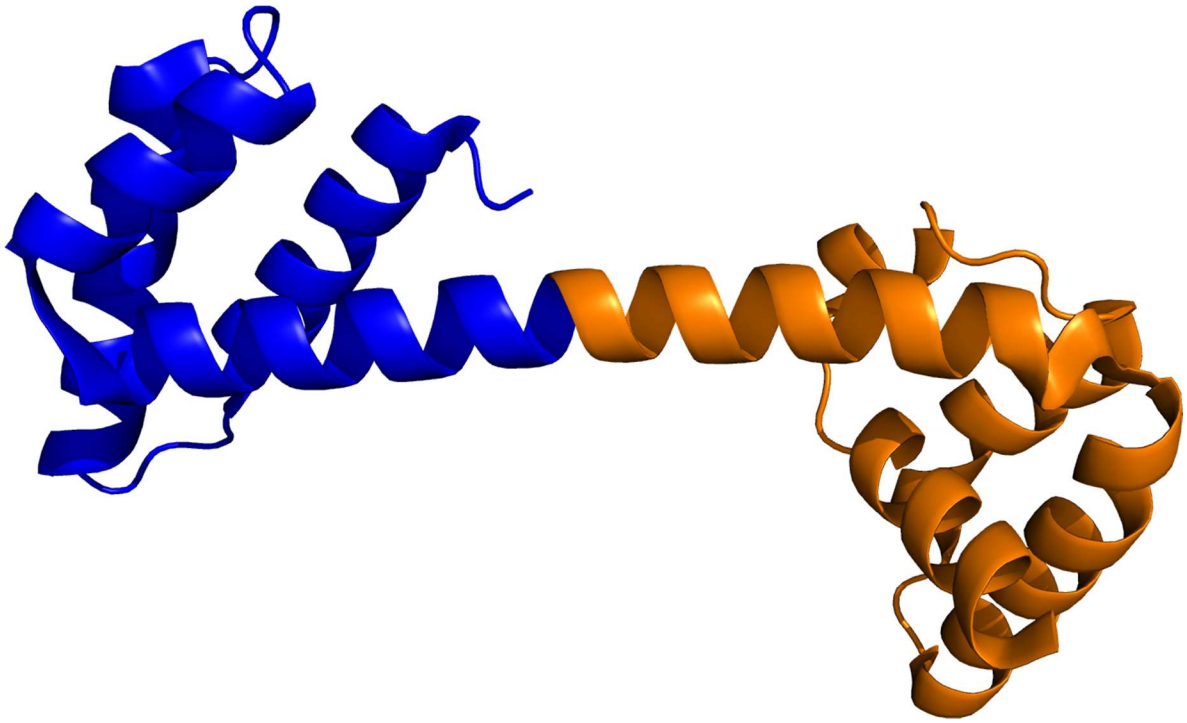


**Figure 2.2: An example of a protein removed under the minimum number of interdomain contacts criterion.** The protein (UPF0307 protein PSPTO_4464, PDB ID: 2P0TA, UniProt primary accession number: Q87WS9) is coloured according to the CATH domain assignments, with domain 1 coloured blue and domain 2 coloured orange. The CATH domain break occurs along a long helix. Given the location of the domain boundary, no interdomain contacts (with a minimum sequence separation of 5) exist between the two domains.

### Large initial sequence alignments

Large MSAs are required for reliable results from covariation-based approaches. A target was deemed to have a sufficiently large MSA if at least 150 sequences were aligned (at which point the effect of positional entropy is reduced in the calculation of correlated mutations (Martin et al., 2005)), and at least 1 sequence was present for each amino acid. Of the 90 monomeric or heteromeric proteins, 37 proteins were deemed to have sufficient numbers of homologous sequences in an initial MSA procedure. These 37 proteins were then retained for benchmarking and analysis.

#### 2.2.1.1 Table of proteins used for analysis

The 37 proteins used in this thesis are summarised in Table 2.2. A gallery of each experimental structure is provided in Section 6.2.

| PDB ID | UniProt identifier | Protein name | Domain 1 CATH code | Domain 1 length | Domain 2 CATH code | Domain 2 length |
|---|---|---|---|---|---|---|
| 1AF7A | P07801 | Chemotaxis protein methyltransferase | 1.10.155.10 | 80 | 3.40.50.150 | 194 |
| 1AQTA | P0A6E6 | ATP synthase epsilon chain | 2.60.15.10 | 88 | 1.20.5.440 | 47 |
| 1BL0A | P0ACH5 | Multiple antibiotic resistance protein MarA | 1.10.10.60 | 56 | 1.10.10.60 | 60 |
| 1EE8A | O50606 | Formamidopyrimidine-DNA glycosylase | 3.20.190.10 | 120 | 1.10.8.50 | 138 |
| 1EH6A | P16455 | Methylated-DNA--protein-cysteine methyltransferase | 3.30.160.70 | 71 | 1.10.10.10 | 90 |
| 1GRJA | P0A6W5 | Transcription elongation factor GreA | 1.10.287.180 | 74 | 3.10.50.30 | 77 |
| 1H8PA | P02784 | Seminal plasma protein PDC-109 | 2.10.10.10 | 40 | 2.10.10.10 | 44 |
| 1JDBF | P0A6F1 | Carbamoyl-phosphate synthase small chain | 3.50.30.20 | 150 | 3.40.50.880 | 229 |
| 1KSLA | P0AA43 | Ribosomal small subunit pseudouridine synthase A | 3.10.290.10 | 66 | 3.30.2350.10 | 167 |
| 1LI5A | P21888 | Cysteine--tRNA ligase | 3.40.50.620 | 273 | 1.20.120.640 | 87 |
| 1MGPA | Q9X1H9 | Fatty acid-binding protein TM_1468 | 3.40.50.10170 | 155 | 3.30.1180.10 | 121 |
| 1OI7A | P09143 | Succinyl-CoA ligase [ADP-forming] subunit alpha | 3.40.50.720 | 122 | 3.40.50.261 | 147 |
| 1PUJA | O31743 | Ribosome biogenesis GTPase A | 3.40.50.300 | 156 | 1.10.1580.10 | 93 |
| 1T6CA | O67040 | Exopolyphosphatase | 3.30.420.40 | 125 | 3.30.420.150 | 181 |
| 1U98A | P0A7G6 | Protein RecA | 3.40.50.300 | 228 | 3.30.250.10 | 59 |
| 1V0BA | Q07785 | Cell division control protein 2 homolog | 3.30.200.20 | 83 | 1.10.510.10 | 203 |
| 1VHNA | Q9WXV1 | tRNA-dihydrouridine synthase | 3.20.20.70 | 234 | 1.10.1200.80 | 71 |
| 1VMAA | Q9WZ40 | Signal recognition particle receptor FtsY | 1.20.120.140 | 83 | 3.40.50.300 | 211 |
| 1WF3A | Q5SM23 | GTPase Era | 3.40.50.300 | 182 | 3.30.300.20 | 114 |
| 1WJ9A | Q53WG9 | CRISPR-associated endoribonuclease Cse3 | 3.30.70.1200 | 87 | 3.30.70.1210 | 101 |
| 2B6CB | Q82ZI8 | Uncharacterized protein | 1.20.1660.10 | 116 | 1.25.40.290 | 99 |
| 2CGJA | P32171 | L-Rhamnulokinase | 3.30.420.40 | 235 | 3.30.420.40 | 244 |
| 2DYIA | Q5SJH5 | Ribosome maturation factor RimM | 2.40.30.60 | 84 | 2.30.30.240 | 71 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2HIYC | Q97RI5 | Uncharacterized protein | 3.30.70.1280 | 89 | 3.30.70.1260 | 92 |
| 2QFLA | P0ADG4 | Inositol-1-monophosphatase | 3.30.540.10 | 140 | 3.40.190.80 | 122 |
| 2RA9A | A3D5G6 | Uncharacterized protein | 3.10.540.10 | 54 | 2.30.270.10 | 73 |
| 2W6PB | P24182 | Biotin carboxylase | 3.40.50.20 | 131 | 3.30.470.20 | 248 |
| 2WHYA | P40409 | Iron-uptake system-binding protein precursor | 3.40.50.1980 | 124 | 3.40.50.1980 | 159 |
| 3A4TA | Q60343 | tRNA (cytosine(48)-C(5))-methyltransferase | 3.30.70.1170 | 60 | 3.40.50.150 | 198 |
| 3CI0J | Q8VPC2 | Hypothetical type II secretion protein | 3.10.610.10 | 104 | 2.10.70.20 | 50 |
| 3CWVA | Q1CZR7 | DNA gyrase, B subunit, truncated | 3.30.565.10 | 198 | 3.30.230.10 | 151 |
| 3FUXC | Q5SM60 | Ribosomal RNA small subunit methyltransferase A | 3.40.50.150 | 200 | 1.10.8.100 | 65 |
| 3HP7A | Q5M3Z4 | Hemolysin, putative | 3.10.290.10 | 65 | 3.40.50.150 | 210 |
| 3NZKB | A1JJJ9 | UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase | 3.30.230.20 | 128 | 3.30.1700.10 | 172 |
| 3QCZA | Q0WDC2 | Bifunctional protein FolC | 3.40.1190.10 | 278 | 3.90.190.20 | 128 |
| 3VO8B | P0A029 | Cell division protein FtsZ | 3.40.50.1440 | 209 | 3.30.1330.20 | 95 |
| 3VRDA | D0G7Q3 | Flavocytochrome c heme subunit precursor | 1.10.760.10 | 78 | 1.10.760.10 | 96 |

**Table 2.2: Summary table of the 37 proteins analysed.** PDB ID: The Protein Data Bank identifier for the protein (first 4 characters) and chain identifier (5$^{th}$ character). UniProt identifier: The UniProt Primary Accession Number. Protein name: The name of the protein provided by UniProt. CATH code: The 4 highest CATH levels in the domain hierarchy. Domain length: The number of amino acids observed in the experimental crystal structure.

### 2.2.2 Alignment procedure development

As covariation-based methods rely solely on MSAs to predict interacting residues, obtaining good quality MSAs is fundamental. Parameter space was searched to identify the set of parameters which produces the best MSAs for identifying covarying residues which relate to interface contacts.

Two leading HMM-based MSA programs were evaluated: jackHMMer (Johnson et al., 2010) and HHblits (Remmert et al., 2011). Both programs are HMM-based, iterative procedures, which incorporate identified homologous sequences into the search profile. The updated profile is then used to inform the search in the subsequent iteration, allowing the methods to identify more remote homologues. For both programs the following parameters were trialled:

- Number of iterations: 1-8
- E-value for accepted homologous sequences: $1x10^{-6}$, $1x10^{-3}$, 1
- Minimum sequence coverage: 50, 60, 70, 80, 90% and variable coverages: $n$ = 15, 20, 25

Minimum variable coverages were calculated based on the length of the query sequence ($L$), and expressed as a percentage by:

$$variable\ coverage\ (\%) = \left(\frac{L}{L-n}\right) \cdot 100$$

E-values for results to be included in subsequent iterations and the final reported list were kept the same in each trial.

The input protein sequence for both MSA methods was extracted from the ATOM record of the corresponding PDB file, and stored in FASTA format.

**HHblits-specific usage information**

HHblits requires a program-specific database of pre-calculated HMM profiles, with each HMM relating to a sequence cluster from the UniProt sequence database (Apweiler et al., 2004). The most recent version of this database available (20th March 2013 release) was used.

The full set of HHblits parameters used were:

-n <iteration> -e <E-value> -E <E-value> -maxfilt 1,000,000 -diff inf

The "diff" parameter of HHblits filters generated MSAs, retaining the $n$ most divergent sequences. However, this filtering step can lead to underrepresented regions in multidomain alignments (HHsuite user guide version 2.0.15, page 42) so the parameter was set to infinite, in order to prevent sequences from being removed in this way.

The "maxfilt" parameter specifies the maximum number of sequences which are permitted to pass a pre-filtering step. By default, this value is 20,000, so this value was set arbitrarily high in order to prevent sequences from being excluded in this manner and maximise the size of generated alignments.

MSAs produced by HHblits were initially recorded in 'a3m' format, then subsequently transformed into 'aln' format, suitable as input for all covariance methods.

**JackHMMer-specific usage information**

JackHMMer can use the standard UniRef100 database, so the latest available database was used (downloaded 26th September 2014).

Full jackHMMer parameters for homology searches were:

-N <iteration> -E <E-value> --incE <E-value>

JackHMMer results were initially recorded in Stockholm ('sto') format, then subsequently transformed into 'a3m' format (using the "reformat.pl" script provided with the HHsuite package) and finally into 'aln' format, as above.

### 2.2.3 MSA sequence counts

The number of sequences identified within each MSA generated by HHblits and jackHMMer were assessed using two metrics. The first metric is a measure of the total number of aligned sequences, which is simply a count of the number of rows present within each MSA.

The second metric accounts for sequence redundancy. The "effective sequence count" measures the number of sequence groups after all aligned sequences have been clustered at 62% sequence identity; a threshold used in other contact prediction studies (Jones et al., 2012; Jones et al., 2015). The 62% sequence identity thresholds was selected for clustering sequences based on previous work demonstrating the importance of sequence weighting, which employed the same sequence identity threshold (Buslje et al., 2009), and has been shown to provide maximal performance (Shackelford and Karplus, 2007).

### 2.2.4 Trialled covariation methods

Four different approaches to identify covarying residues were evaluated. Three approaches, EVfold, PSICOV and CCMpred, differentiate between direct and indirect contacts using different models of covariation. As a comparison, MIp, a widely-used

Mutual Information-based approach which does not attempt to distinguish between these two types of covariation, was also evaluated.

#### 2.2.4.1 MIp

Mutual Information with the APC (MIp), as described by Dunn et al. (2008), was calculated using an in-house program. This MIp implementation also includes the same sequence-weighting method used by PSICOV, which has been shown to improve results (Jones, et al., 2012). MIp was selected instead of alternative corrections to MI as the APC is also employed by EVfold, PSICOV and CCMpred, enabling a fairer comparison between each method.

#### 2.2.4.2 PSICOV

PSICOV (version 2.1beta3) calculations were generated using recommended parameters: -o (forced override of alignment diversity check), -d 0.03 (3% target density of covariance matrix).

#### 2.2.4.3 EVfold

EVfold (Marks et al., 2011) calculations were generated using the implementation provided by the FreeContact package (Kaján et al., 2014) with default parameters.

#### 2.2.4.4 CCMpred

CCMpred (Seemayer et al., 2014) was run using the standard implementation, also with default parameters.

### 2.2.5 Extracting interdomain contacts

The aforementioned approaches identify all instances of covariation between columns of an MSA, both within and between domains. Interdomain contacts were extracted from the list of contact predictions using CATH structural domain boundaries, as reported in the 'CathDomall' file, available from the CATH FTP site (ftp://ftp.biochem.ucl.ac.uk/pub/cath/v4_0_0/CathDomall). A list of interdomain predictions was generated by excluding predictions which do not span the CATH domain boundaries.

### 2.2.6 Assessment of predicted contacts

#### 2.2.6.1 Definition of contacting residues

There have been many thresholds used to determine whether two residues are in contact. For this work, two different contact thresholds were used. The first contact threshold is the one employed in the CASP Residue-Residue (RR) contact assessment, which considers a contact to be formed if the inter-residue Cβ distance (Cα for Glycine) is within 8Å in the reference crystal structure (Monastyrskyy et al., 2014), hereafter referred to as a "CB8A" contact. The second contact threshold considers a contact to be formed if two residues possess any inter-residue heavy-atom pair (that is, any non-Hydrogen atom) closer than 6Å ("HA6A" contacts).

#### 2.2.6.2 Experimental structure-derived interdomain contacts

The PDB file of each protein in the dataset was downloaded from the PDB and was renumbered starting from 1, with numbering continuous across any gaps formed by missing residues. Structurally observed contacts were determined and numbered from the renumbered files.

#### 2.2.6.3 Number of contacts to assess

Previous studies assessing intradomain contact prediction typically assess a number of predictions based on differing fractions of the length of the chain length, $L$ (e.g. $L/10$, $L/5$, $L/2$, $L$, etc.). This approach is suitable due to the large number of intra-chain contacts which are formed (in the region of 3% of all possible contact pairs (Jones *et al.*, 2012), though this number varies by fold type). In contrast to the large number of contacts formed within a fold, the number of contacts present at the domain interface is comparatively small (an average of 0.5% of all possible interdomain contacts are observed under the CB8A threshold for this dataset).

An alternative approach to determine the number of contacts to analyse is to consider the number of contacts which are required in order to benefit the modelling problem at hand. In order to reliably fold a protein chain, many contacts, evenly distributed along the length of the chain are required, with a lower bound of roughly 1 contact for every 12 residues required to reliably model the protein topology (Kim et al., 2013). This figure is closer to 30% of the total number of experimental contacts if the contacts are randomly distributed (Konopka et al., 2014; Marks et al., 2011). In contrast to this, in order to constrain the way two objects interact in 3-dimensional space, a single correct contact is sufficient to identify

the binding interface, drastically reducing the number of possible interactions. Additional contacts then act to further constrain the orientation of the interaction.

Previous work using chemical cross-link data has demonstrated that accurate docking can be achieved with relatively few restraints, and improvements rapidly diminish in excess of 5 (Kahraman et al., 2013). The cross-links employed in this study provide upper-bound distances in the region of 30Å. With covarying positions widely observed to occur at much shorter distances, one could expect that the same number of contacts would impose much heavier restrictions on the orientations possible between two interacting structures.

This work concentrated on predicting a small number of interdomain contacts, prioritising precision over recall, in the knowledge that even small numbers of incorrect contacts can dramatically reduce the accuracy of modelling (Konopka et al., 2014). In theory, the perfect prediction of just 3 non-collinear contacts would be sufficient to properly dock two structures. However, a more realistic goal is to concentrate on predicting a slightly larger set of contacts, which can be used to guide modelling procedures, understanding that false positive predictions are inevitable. However, once a set of predicted contacts have been generated, it may be possible to exclude likely false positives. To this end, our efforts were focused on the prediction of 10 interdomain contacts.

#### 2.2.6.4 Assessment of covarying residue pairs

A simple measure for the assessment of predicted contacts is to calculate the precision score. Precision is calculated using the formula:

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

where True Positives (TPs) are predictions observed to be in contact within the reference structure, and False Positives (FPs) are predicted contacts separated by a distance above the same threshold. Precision scores have a range between 0 and 1, with 0 indicating all predictions were incorrect, and a score of 1 denoting all predictions were made correctly.

For this work, predicted contacts were required to have a minimum sequence separation of 5 residues, in order to remove simple contact prediction along structures spanning the domain boundary (e.g. within alpha helices).

Although binary distance cutoffs have the benefit of simplicity, they paint a rather simplistic picture of whether two residues are in contact. Residue pairs may be classified as non-contacting if they are fractions of an Angstrom over the binary distance threshold. This will therefore affect calculated measures of accuracy for the prediction methods. Alternative, more permissive approaches, would avoid these scenarios at the expense of complicating accuracy metrics. Although the use of binary contact definitions is flawed, their use is standard procedure for the evaluation of contact predictions (Monastyrskyy et al., 2014),.

#### 2.2.6.5 Selection of a single alignment parameter set

The set of alignment parameters enabling the most precise contact prediction was selected by calculating the mean top-10 contact precision value for CCMpred, EVfold and PSICOV contacts, for both CB8A and HA6A contact definitions.

### 2.2.7 Overlap of predictions by PSICOV, CCMpred and EVfold

The overlap of predictions generated by PSICOV, CCMpred and EVfold was assessed, considering the top 10 contacts identified by each method. The 35 targets for which interdomain contacts were successfully identified were considered (excluding targets 1H8PA and 2W6PB from the set of 37 proteins listed in Table 2.2), resulting in the evaluation of 1050 contacts.

### 2.2.8 Simple consensus of different models of covariation

Previous studies have shown that the different models of covariation underlying PSICOV, CCMpred and EVfold identify largely overlapping sets of contacts, though some contacts are unique to each approach (Tetchner et al., 2014; Jones et al., 2015). An attempt was made to combine predictions from the 3 methods using a simple consensus with the goal of increasing precision scores. The set of consensus predictions was generated by considering the ranking of each contact within the list of predictions from the individual methods. The approach assigned the highest rank to the contact appearing first within each of the 3 sets of predictions. This procedure was repeated in order to generate the set of consensus predictions and the performance of this approach was evaluated in the same manner as each separate method.

## 2.3 Results and discussion

### 2.3.1 Benchmark results

After generating MSAs for the 37 proteins in the data set (Table 2.2), instances of interdomain covariation were calculated using the four considered methods and evaluated using the CB8A and HA6A contact definitions. A summary of the best performing alignment parameters for each method are presented in Tables 2.3 and 2.4.

| Covariance method | Alignment method | Number of iterations | E-value | Coverage | Mean precision (CB8A) |
|---|---|---|---|---|---|
| CCMpred | jackHMMer | 2 | $1x10^{-6}$ | 70% | 0.595 |
| PSICOV | jackHMMer | 2 | 1 | L-20 | 0.519 |
| EVfold | jackHMMer | 4 | 1 | 80% | 0.514 |
| MIp | jackHMMer | 3 | $1x10^{-6}$ | 80% | 0.281 |
| MIp | jackHMMer | 2 | 1 | 70% | 0.281 |
| MIp | jackHMMer | 1 | $1x10^{-3}$ | 80% | 0.281 |
| CCMpred | HHblits | 1 | 1 | 80% | 0.541 |
| PSICOV | HHblits | 2 | $1x10^{-6}$ | 70% | 0.454 |
| EVfold | HHblits | 1 | $1x10^{-3}$ | 80% | 0.438 |
| MIp | HHblits | 1 | 1 | 80% | 0.273 |

**Table 2.3: The best performing parameter set for each of the covariance methods, evaluated using the CB8A contact definition.** The mean precision score was calculated across the top 10 interdomain contacts.

| Covariance method | Alignment method | Number of iterations | E-value | Coverage | Mean precision (HA6A) |
|---|---|---|---|---|---|
| CCMpred | jackHMMer | 2 | $1x10^{-3}$ | 80% | 0.681 |
| PSICOV | jackHMMer | 2 | 1 | 80% | 0.603 |
| EVfold | jackHMMer | 2 | 1 | 80% | 0.589 |
| MIp | jackHMMer | 2 | 1 | 70% | 0.335 |
| CCMpred | HHblits | 1 | 1 | 80% | 0.616 |
| PSICOV | HHblits | 1 | 1 | 80% | 0.511 |
| EVfold | HHblits | 1 | $1x10^{-3}$ | 80% | 0.503 |
| MIp | HHblits | 1 | 1 | 80% | 0.322 |

**Table 2.4: The best performing parameter set for each of the covariance methods, evaluated using the HA6A contact definition.** The mean precision score was calculated across the top 10 interdomain contacts.

The first thing to note from Tables 2.3 and 2.4 is that the contact definition used affects both the precision score of the covarying pairs, as well as the parameters that produce maximal contact precision. However, over both contact definitions, the covariance programs maintain their relative rankings, with CCMpred performing with the highest precision, followed by PSICOV, EVfold and MIp, respectively.

It is also evident that jackHMMer outperforms HHblits at generating alignments to predict interdomain contacts. Whilst HHblits is reported to have better performance than jackHMMer (Remmert et al., 2011), the disparity between the database releases used here is the most obvious reason why HHblits achieves worse performance in this study. As jackHMMer is capable of using standard sequence databases, it was possible to use the most current version of this database, whereas the most recent release of the HHblits database was from a year and a half earlier. Due to the rapid expansion of sequence repositories, the existing protocol for clustering sequences at 20%, required to generate the HHblits HMM database, was insufficient to keep pace with the monthly release schedule (Söding, J, pers. comm., 14/10/14). With the release of an up-to-date HMM

database, HHblits may outperform jackHMMer for interdomain contact prediction, though that would have to be tested. In order to aid future work using future releases of the HHblits database, the best performing parameter set for HHblits observed here may act as a reasonable starting point.

As the parameters used to generate the highest scoring interdomain contacts varies between methods, an approach was devised to select the single parameter set which confers the best average performance across PSICOV, CCMpred and EVfold. MIp was excluded from consideration due to the considerably lower performance achieved throughout all trials (Tables 2.3 and 2.4). In order to balance between the two different contact definitions, the average precision score for PSICOV, CCMpred and EVfold was calculated over both CB8A and HA6A contact definitions to reduce bias towards a particular definition. The parameters which produce the mean highest precision contacts are shown in Table 2.5.

| Method | Number of iterations | E-value | Coverage | Mean precision CB8A | Mean precision HA6A | Mean precision CB8A + HA6A |
|---|---|---|---|---|---|---|
| JackHMMer | 2 | $1x10^{-6}$ | 70% | 0.535 | 0.618 | **0.577** |
| JackHMMer | 2 | 1 | 80% | 0.532 | 0.621 | **0.576** |
| JackHMMer | 2 | $1x10^{-3}$ | 70% | 0.530 | 0.617 | **0.573** |
| JackHMMer | 2 | 1 | 70% | 0.526 | 0.619 | **0.573** |
| JackHMMer | 2 | $1x10^{-3}$ | 80% | 0.526 | 0.618 | **0.572** |
| HHblits | 1 | 1 | 80% | 0.472 | 0.541 | **0.506** |
| HHblits | 1 | $1x10^{-3}$ | 70% | 0.471 | 0.535 | **0.503** |
| HHblits | 1 | $1x10^{-3}$ | 80% | 0.469 | 0.536 | **0.503** |
| HHblits | 1 | 1 | 70% | 0.467 | 0.536 | **0.501** |
| HHblits | 1 | $1x10^{-6}$ | 70% | 0.465 | 0.532 | **0.498** |

**Table 2.5: The 5 highest precision parameter sets for jackHMMer and HHblits.** Mean precision values were calculated by averaging over the top-10 interdomain predictions from PSICOV, CCMpred and EVfold for both CB8A and HA6A contact definitions. The final best performing parameter set was selected using the mean performance of both HA6A and CB8A criteria, shown in bold.

Using this performance measure, the single best performing parameter set was identified as that generated by jackHMMer, using 2 iterations, an E-value threshold of $1x10^{-6}$ and a minimum sequence coverage of 70%. This parameter set is the highest scoring under the CB8A contact definition, and the third highest scoring under the HA6A contact definition. This is also the same parameter set which achieves the best performance for CCMpred under the CB8A contact definition (Table 2.3).

To establish if the parameter search benefitted contact prediction, a comparison was made with default jackHMMer parameters, the results of which are presented in Tables 2.6 and 2.7.

| Method | Test statistic (W) | *p*-value |
|---|---|---|
| CCMpred | 403 | **$2.96x10^{-5}$** |
| EVfold | 208 | **$4.18x10^{-3}$** |
| PSICOV | 325.5 | **$2.69x10^{-3}$** |
| MIp | 340 | **$4.07x10^{-3}$** |

**Table 2.6: Comparison of the best performing jackHMMer parameters with default jackHMMer parameters.** Default jackHMMer parameters are: 5 iterations, E-value 10 and no minimum coverage. Contacts were considered correct under the CB8A criterion. The statistical test was performed using a paired one-tailed Wilcoxon signed-rank test with a 95% confidence interval. $H_1$ = selected parameter set (2 iterations, E-value $1x10^{-6}$, minimum 70% coverage) is on average more precise than the default parameters (5 iterations, E-value 10). Statistically significant results are shown in bold.

Table 2.6 shows that the single best performing parameter set performs significantly better at identifying interdomain contacts than the default jackHMMer parameters for all methods. However, by default, jackHMMer does not impose a minimum sequence coverage to the query sequence, instead selecting sequences on the basis of E-values. In order to perform a fairer comparison, the same minimum 70% coverage parameter was applied, and the results are presented in Table 2.7.

| Method | Test statistic (W) | *p*-value |
|---|---|---|
| CCMpred | 301.5 | **$6.6 \times 10^{-4}$** |
| EVfold | 166 | **0.0127** |
| PSICOV | 220 | 0.0615 |
| MIp | 211.5 | **0.0129** |

**Table 2.7: Comparison of the best performing jackHMMer parameters with default parameters, also including a minimum 70% coverage.** The parameters compared against were: 5 iterations, E-value 10 and 70% minimum coverage. Contacts were considered correct under the CB8A criterion. The comparison was conducted using a paired one-tailed Wilcoxon signed-rank test with a 95% confidence interval. $H_1$ = selected parameter set (2 iterations, E-value $1 \times 10^{-6}$, minimum 70% coverage) is on average more precise than the default parameters with the minimum 70% coverage requirement (5 iterations, E-value 10, minimum 70% coverage). Statistically significant results are shown in bold.

Table 2.7 shows that even with the same minimum sequence coverage, the selected parameter set significantly improves upon the default parameters for contact prediction by CCMpred, EVfold and MIp. However, under this condition, PSICOV was only approaching significance.

### 2.3.2 Effect of MSA parameters on MSA size and contact prediction

Now that the best performing parameter set has been identified for interdomain contact prediction, we can investigate the effect of varying each separate MSA parameter. By fixing two of the three parameters, the effect of the third parameter on the aligned sequences can be observed, along the resultant predicted contacts. Firstly, the effect of varying the E-value for accepting sequences into the search profile, and final reported list of homologous sequences was investigated. The results are shown in Figure 2.3.

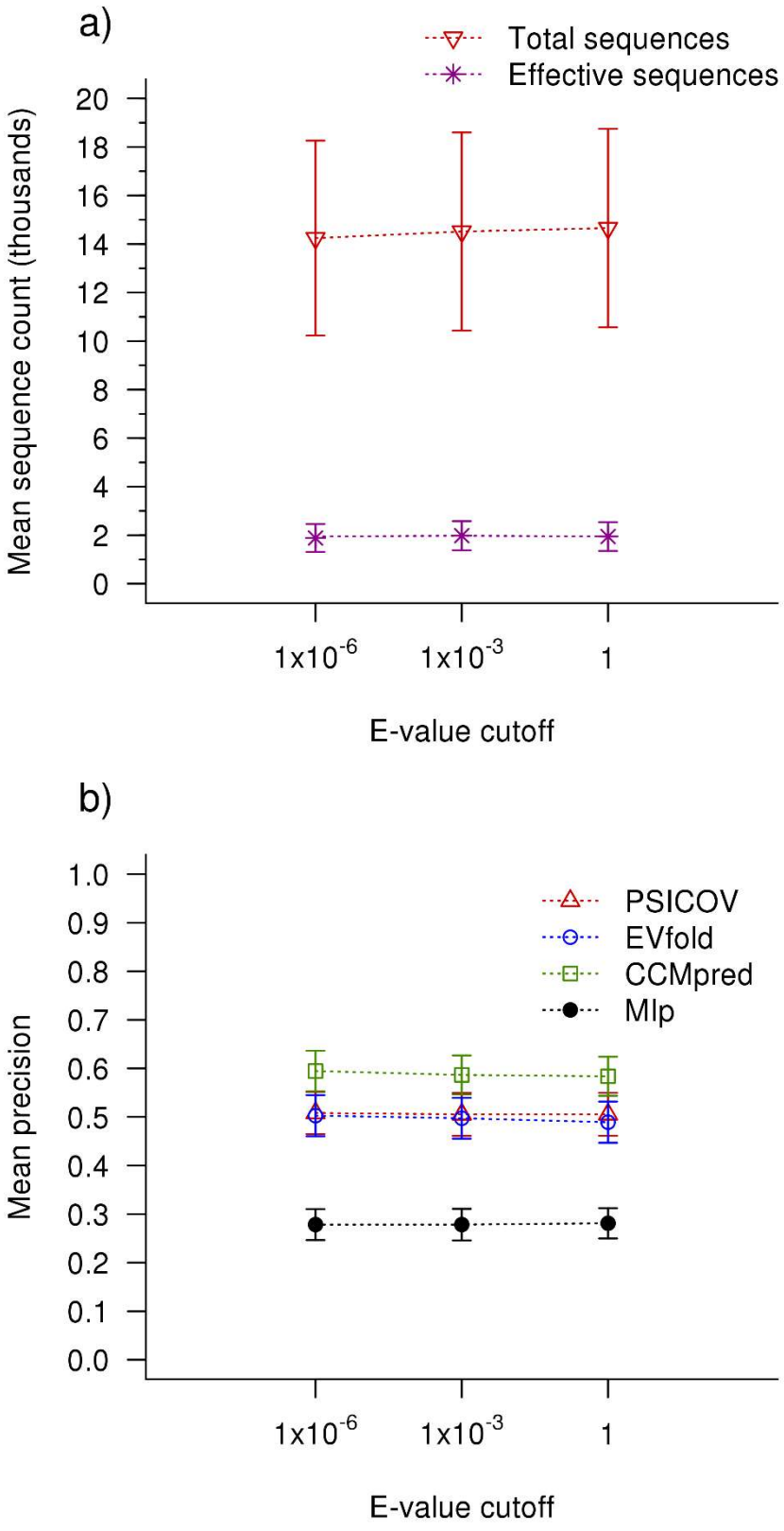



**Figure 2.3: Effect of different jackHMMer E-value cutoffs on contact precision.** The number of iterations is fixed at 2, and the coverage fixed at a minimum of 70% to the query sequence.

Figure 2.3a shows that more permissive E-value thresholds incorporate marginally more sequences into generated alignments (with the addition of 414 total and 97 effective sequences on average between the E-value = $1x10^{-6}$ and E-value = 1 conditions). However, the addition of these sequences has a negligible effect on performance, slightly reducing performance for PSICOV, EVfold and CCMpred (mean reduction in precision score = 0.01). The precision score of MIp increases by a value of 0.01 (Figure 2.3b).

Secondly, the effect of different numbers of search iterations was assessed. Figure 2.4a shows that with additional iterations of the jackHMMer search procedure, more sequences are identified on average, for both total and effective sequence counts. The number of available sequences has previously been shown to correlate with the precision of covariance approaches (Jones et al., 2012). However, in this study, additional sequences reduce precision scores (Figure 2.4b), showing that sheer sequence number does not give rise to better predictive performance.

However, additional sequences do appear to give rise to higher precision scores initially. Between the first and second iterations, a mean value of 7626 additional sequences are incorporated into the alignments (and a mean value of 1112 effective sequences). This results in an increase in precision score across all methods (PSICOV = 0.03; EVfold = 0.07; CCMpred = 0.06; MIp = 0.07). However, whilst the third iteration incorporates an additional 1682 sequences into the generated alignments on average (228 effective), precision scores drop for all methods (PSICOV = -0.03; EVfold = -0.01; CCMpred = -0.04; MIp = -0.01). From these observations, it would appear that there is a tradeoff between generating an alignment with as many diverse sequences as possible, and excluding homologues which are too divergent from the sequence under analysis. The inclusion of low percentage identity sequences may introduce additional noise into the generated MSAs. It is well known that the interface of low sequence identity multidomain proteins becomes less conserved as the level of sequence identity reduces (Aloy et al.,

2003; Han et al., 2006). Therefore, the incorporation of these sequences into the MSA may reduce the signal of the higher percentage identity sequences which do have a singly-orientated interface. The addition of a minimum sequence identity cutoff may reduce this effect in future investigations.

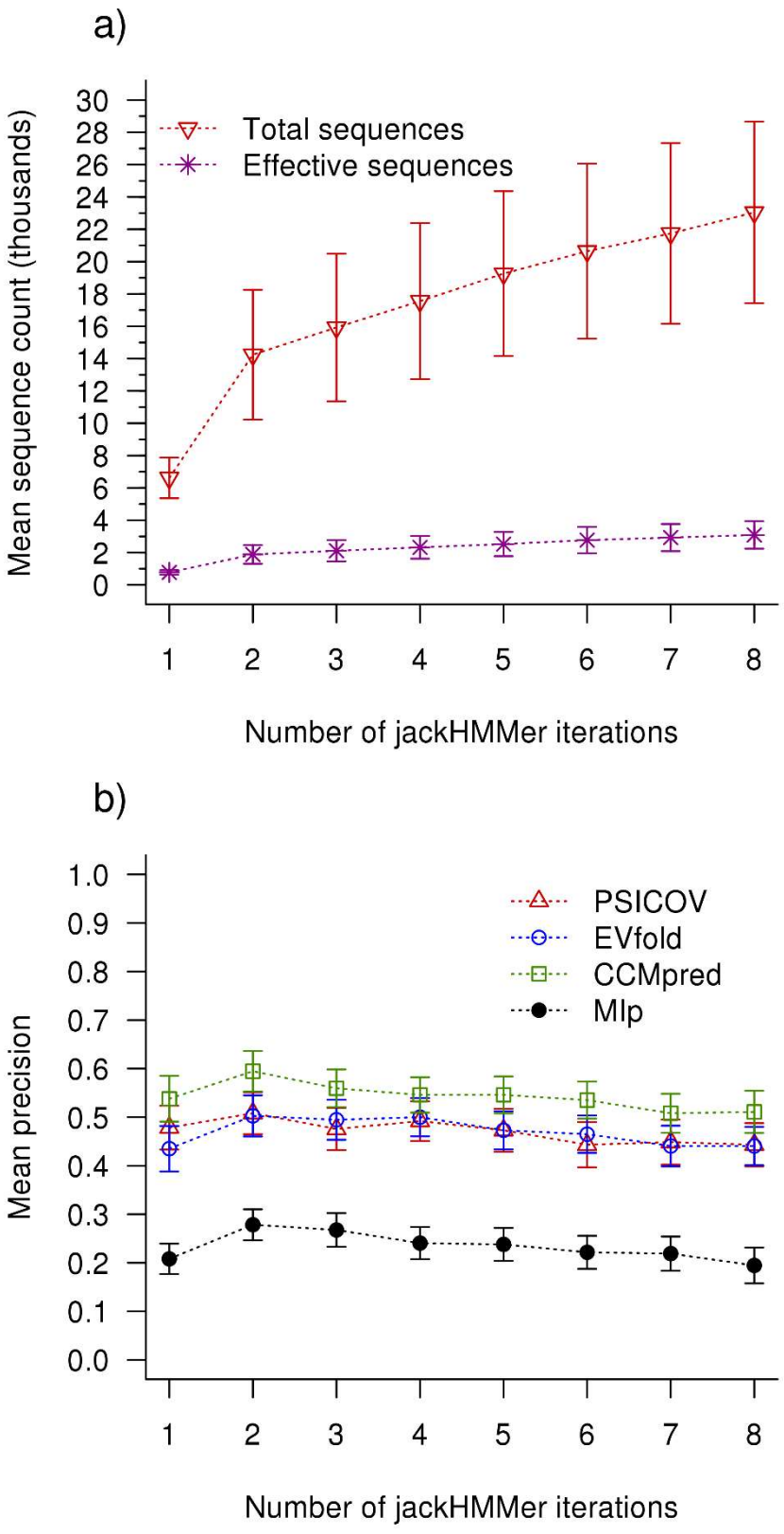



**Figure 2.4: Effect of different numbers of jackHMMer iterations on contact precision.** The E-value is fixed at $1x10^{-6}$, and the minimum coverage fixed at a minimum of 70%.

Downstream iterations incorporate homologous sequences identified in earlier iterations into the search HMM-profile in order to identify more diverse homologues. If the sequences identified in later iterations do not share the same overall structure and interdomain orientation as the query sequence and as those identified in earlier iterations, then these high-iteration sequences are likely to hinder predictive capabilities. As mentioned previously, differences in interdomain orientation can occur even at high sequence identity (Han et al., 2006) and may explain the observed effects.

In addition to improving performance, knowledge that fewer iterations are required for this type of analysis has beneficial practical implications, too. Additional iterations quickly become very computationally expensive, requiring large memory footprints and considerable additional time in order to incorporate large numbers of sequences into high iteration profiles.

Finally, the effect of different minimum coverages was assessed. Figure 2.5a shows that with higher minimum coverage requirements, the numbers of both total sequences and effective sequences reduce. The variable coverage parameters based on sequence length ensure that high coverage is met for aligned sequences. The variable coverage parameters align approximately as many sequences as the minimum 90% coverage parameter, in line with expectations. Despite the reduction in sequence number, performance over the trialled parameters remains fairly consistent (Figure 2.5b). Interestingly, as observed in Table 2.3, PSICOV achieves maximal performance using the variable $L - 20$ coverage parameter and is the only method where the variable coverage parameter confers performance higher than the static thresholds.

By setting the coverage parameter too low, many more sequences are incorporated, but these may only align over one of the two domains. Again, as this work assumes that the proteins in each row of the alignment relate to alternative sequences encoding the same

overall structure, it is imperative that both domains are present in the considered alignment. It is important to realise that the sizes of the domains in the dataset are not equal, that is, the total chain length is not evenly split with half of the residues appearing in domain 1, and half occurring in domain 2 (Table 2.2). By calculating the average length of the biggest domain as a percentage of the entire chain length, the biggest domain contributes 62.6% of the chain length on average (range 50.8-79.4%). By setting the minimum coverage at 60% or less (with the coarse granularity of the coverage parameters trialled), sequences possessing only one of the two domains may be included in the generated alignments for some of the proteins. This may explain why the 50 and 60% coverage parameters produce lower performance than the 70% coverage parameter for all methods.

In the opposite scenario, setting a high minimum coverage would be very exclusive, permitting few sequences in alignments. The ideal parameter is likely to be a compromise between these two opposing requirements.

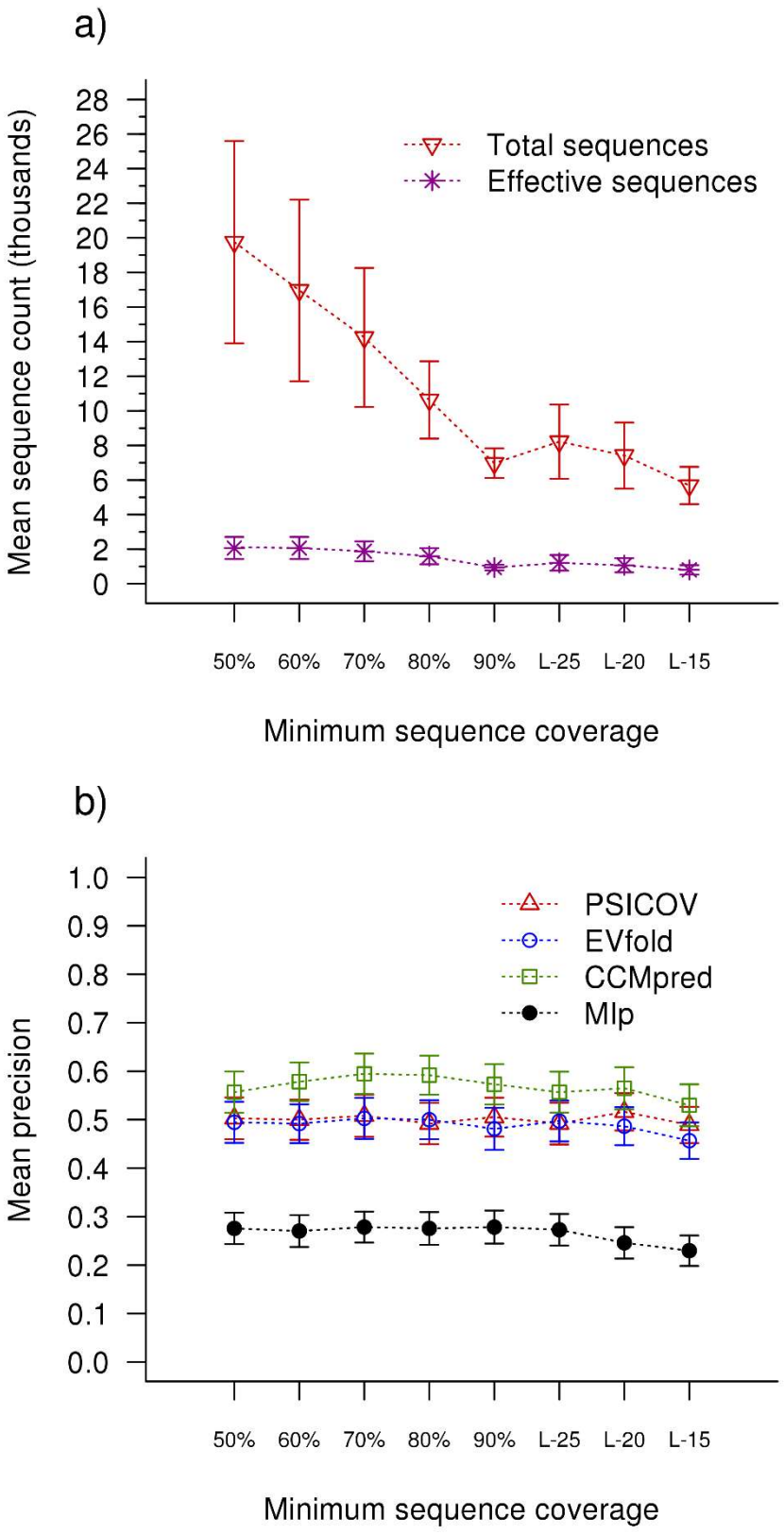



**Figure 2.5: Effect of different jackHMMer coverages on contact precision.** The E-value is fixed at $1x10^{-6}$, and the number of iterations at 2.

### 2.3.3 Analysis of predicted interdomain contacts from the best performing parameter set

This section looks in greater detail at the interdomain contacts generated using the best identified parameter set from Section 2.3.1. A summary of the results of interdomain contact prediction for the dataset are presented in Table 2.8.

| | | | HA6A contact threshold | | | | | CB8A contact threshold | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Top 10 contact precision | | | | | Top 10 contact precision | | | |
| PDB ID | Number of aligned sequences | Number of effective sequences | Number of inter-domain contacts | CCMpred | EVfold | PSICOV | MIp | Number of inter-domain contacts | CCMpred | EVfold | PSICOV | MIp |
| 1EE8A | 9286 | 1256 | 116 | 1 | 0.5 | 0.6 | 0 | 71 | 0.8 | 0.3 | 0.4 | 0 |
| 1VMAA | 16616 | 1059 | 179 | 1 | 0.8 | 1 | 0.2 | 126 | 1 | 0.8 | 1 | 0.3 |
| 2QFLA | 16918 | 3082 | 123 | 1 | 0.9 | 0.9 | 0.5 | 107 | 0.9 | 0.9 | 0.7 | 0.5 |
| 1JDBF | 9665 | 404 | 127 | 0.9 | 0.7 | 0.9 | 0.1 | 93 | 0.6 | 0.4 | 0.5 | 0 |
| 1LI5A | 9962 | 605 | 101 | 0.9 | 0.9 | 0.9 | 0.9 | 60 | 0.4 | 0.3 | 0.4 | 0.4 |
| 1MGPA | 7794 | 1866 | 183 | 0.9 | 1 | 0.8 | 0.9 | 120 | 0.8 | 0.9 | 0.8 | 0.8 |
| 1WF3A | 6962 | 635 | 107 | 0.9 | 0.6 | 0.6 | 0.4 | 57 | 0.5 | 0.3 | 0.4 | 0.3 |
| 3CI0J | 1036 | 240 | 120 | 0.9 | 0.8 | 0.7 | 0.5 | 79 | 0.6 | 0.6 | 0.4 | 0.4 |
| 1EH6A | 10166 | 1994 | 76 | 0.8 | 0.6 | 0.7 | 0.6 | 39 | 0.7 | 0.4 | 0.3 | 0.4 |
| 1GRJA | 7961 | 532 | 46 | 0.8 | 0.6 | 0.5 | 0.4 | 30 | 0.7 | 0.7 | 0.6 | 0.5 |
| 1KSLA | 24799 | 2611 | 35 | 0.8 | 0.6 | 0.7 | 0.4 | 30 | 0.9 | 0.6 | 0.7 | 0.2 |
| 1OI7A | 10885 | 871 | 153 | 0.8 | 0.5 | 0.3 | 0.1 | 106 | 0.6 | 0.4 | 0.3 | 0 |
| 1V0BA | 127055 | 12340 | 114 | 0.8 | 0.7 | 0.8 | 0.4 | 69 | 0.9 | 0.8 | 0.9 | 0.4 |
| 2B6CB | 1641 | 427 | 68 | 0.8 | 0.7 | 0.6 | 0.1 | 55 | 0.7 | 0.6 | 0.5 | 0.1 |
| 2CGJA | 23715 | 3319 | 237 | 0.8 | 0.7 | 0.6 | 0.1 | 157 | 0.8 | 0.6 | 0.5 | 0.1 |
| 2HIYC | 1940 | 538 | 114 | 0.8 | 0.7 | 0.5 | 0.4 | 74 | 0.7 | 0.5 | 0.5 | 0.2 |
| 2WHYA | 22297 | 5214 | 113 | 0.8 | 0.9 | 0.8 | 0.5 | 77 | 0.6 | 0.7 | 0.6 | 0.3 |
| 3A4TA | 12191 | 1857 | 97 | 0.8 | 0.8 | 0.8 | 0.4 | 50 | 0.8 | 0.8 | 0.9 | 0.4 |
| 3HP7A | 4546 | 416 | 65 | 0.8 | 0.8 | 0.8 | 0.5 | 46 | 0.9 | 0.9 | 0.7 | 0.5 |
| 3NZKB | 3800 | 327 | 228 | 0.8 | 0.8 | 0.9 | 0.7 | 157 | 0.7 | 0.7 | 0.8 | 0.4 |
| 1AF7A | 7467 | 1496 | 50 | 0.7 | 0.2 | 0.5 | 0 | 24 | 0.6 | 0.2 | 0.4 | 0 |
| 1BL0A | 93675 | 18161 | 63 | 0.7 | 0.6 | 0.8 | 0.5 | 23 | 0.7 | 0.6 | 0.7 | 0.5 |
| 3FUXC | 8997 | 998 | 51 | 0.7 | 0.9 | 0.9 | 0.3 | 27 | 0.8 | 0.7 | 1 | 0.4 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3QCZA | 10051 | 1989 | 69 | 0.7 | 0.9 | 0.9 | 0.3 | 45 | 0.8 | 0.9 | 0.9 | 0.3 |
| 3VO8B | 6516 | 354 | 124 | 0.7 | 0.5 | 0.6 | 0.3 | 89 | 0.7 | 0.5 | 0.6 | 0.3 |
| 1AQTA | 6210 | 864 | 37 | 0.6 | 0.5 | 0.4 | 0.2 | 33 | 0.7 | 0.6 | 0.5 | 0.3 |
| 2DYIA | 6493 | 1421 | 28 | 0.6 | 0.6 | 0.4 | 0.4 | 17 | 0.5 | 0.4 | 0.3 | 0.2 |
| 2RA9A | 1134 | 136 | 61 | 0.6 | 0.5 | 0.6 | 0.3 | 42 | 0.4 | 0.3 | 0.4 | 0.3 |
| 1U98A | 6791 | 121 | 45 | 0.5 | 0.3 | 0.5 | 0.5 | 28 | 0.3 | 0.3 | 0.5 | 0.5 |
| 1VHNA | 14773 | 1133 | 64 | 0.5 | 0.7 | 0.4 | 0.2 | 46 | 0.5 | 0.7 | 0.5 | 0.3 |
| 1PUJA | 5278 | 796 | 39 | 0.4 | 0.4 | 0.2 | 0.1 | 26 | 0.3 | 0.2 | 0.2 | 0.1 |
| 1T6CA | 8091 | 1335 | 151 | 0.4 | 0.4 | 0.4 | 0.5 | 125 | 0.4 | 0.4 | 0.4 | 0.5 |
| 3CWVA | 15764 | 406 | 68 | 0.4 | 0.3 | 0.4 | 0.2 | 40 | 0.3 | 0.3 | 0.3 | 0.3 |
| 3VRDA | 2222 | 432 | 77 | 0.3 | 0 | 0.2 | 0 | 53 | 0.3 | 0 | 0.1 | 0 |
| 1WJ9A | 905 | 365 | 78 | 0.2 | 0.2 | 0.1 | 0.2 | 45 | 0.1 | 0.2 | 0.1 | 0.1 |
| 1H8PA | 361 | 68 | 25 | 0 | 0 | 0 | 0 | 21 | 0 | 0 | 0 | 0 |
| 2W6PB | 3025 | 28 | 223 | 0 | 0.2 | 0 | 0 | 148 | 0 | 0.1 | 0 | 0 |
| **Mean value** | 14242.919 | 1883.676 | 98.784 | 0.678 | 0.589 | 0.586 | 0.327 | 65.811 | 0.595 | 0.503 | 0.508 | 0.278 |

**Table 2.8: Summary of interdomain contact prediction results using the best performing jackHMMer parameters.** PDB ID – Protein Data Bank identifier. Number of aligned sequences – Total number of aligned sequences by jackHMMer. Number of effective sequences – Number of sequence groups after clustering sequences at 62% sequence identity. Top 10 contact precision – Precision value of the top-10 interdomain predictions.

In general, the selected parameters are able to identify large numbers of homologous sequences, and the covariance methods perform well on the generated alignments. Considering each of the covariation-based predictors individually, CCMpred achieves the highest mean precision values, followed by EVfold and PSICOV which achieve approximately equal precision scores, and finally MIp.

However, not all targets are predicted so well under these parameters. One such example is the target 2W6PB. Whilst this case has 3025 total sequences in the generated alignment, these sequences are highly similar. Considering the diversity of the aligned sequences, the number of “effective” sequences is a mere 28, which is likely to explain the poor performance of all the methods on this case. EVfold identifies 2/10 correct contacts for this example, whilst the other 3 methods do not identify any contacts correctly. One could speculate that EVfold is able to make better use of the closely related sequences in this case, but this would require further investigation.

The situation is similar for the other case with low precision scores across the considered methods: 1H8PA. In this example, there is less redundancy observed in the alignment, with 68 effective sequences from the 361 total sequences. However, the lack of available sequence numbers again appears to hamper predictive efforts.

For these two cases, there are alignment parameters that produce improved results in the total set of parameters tested. In the case of 2W6PB, HHblits is able to identify 10/10 contacts correctly for 14 different parameter conditions, and all of these contact predictions were generated by CCMpred (though PSICOV and EVfold were able to generate set of contacts with a precision score of 0.9 in some cases). The number of effective sequences aligned in these cases ranges from 671 to 899 (up from 28 in the best overall parameter set). By increasing the number of identified homologous sequences, higher precision scores are achieved.

For the target 1H8PA, the highest scoring parameter set achieves just 3/10 correct contacts. The 7 sets of parameters achieving this results were generated by jackHMMer, and consist of a single iteration. The total number of effective sequences is similar to those generated with the general alignment parameter set (range: 45-74, in comparison to 68 in the best overall parameter set). From these results, it would appear that during the second iteration, sequences incorporated into the alignment reduce the ability of the covariance methods to identify interdomain contacts. Again, this may be due to different structures occurring in close homologues. Whilst we have identified a set of parameters that perform well across the tested set, better contact predictions can be achieved on a case-by-case basis. As the number of aligned sequences for 1H8PA is small (regardless of the parameters used), this may indicate that there are few homologous sequences available for this target in the sequence databases used, and greater predictive success may be obtained for this target in the future, with the availability of more sequence data.

There are also cases where targets with few effective sequences are predicted considerably better than others with approximately the same number of diverse sequences. For example, 1JDBF has an average precision score of 0.83 across PSICOV, CCMpred and EVfold, with 9665 total sequences and 404 effective sequences in the generated alignment. However, 3CWVA, with 15764 total sequences and 406 effective sequences, only achieves a mean precision score of 0.37, despite the similarities in the number of aligned sequences. From this, one could speculate that some proteins are more amenable to this type of analysis than others, such as those where the family of sequences possess only a single interface which interacts with partner domains, or where a single interface site is strongly maintained over relatives.

However, the average results achieved over the top 10 contacts only represents one way to interpret the data. Considering different numbers of contacts for assessment, it is possible to observe the overall trend when attempting to predict larger sets of interface

contacts. The effect of assessing different numbers of contacts on the precision score is shown in Figure 2.6.

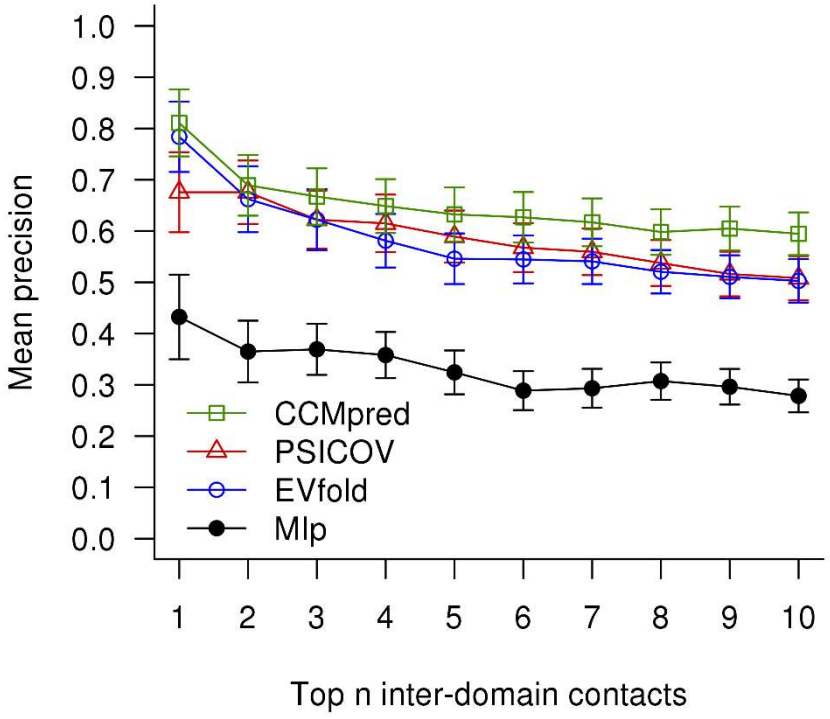



**Figure 2.6: Mean precision values for the set of 37 proteins in the dataset, varying the number of interdomain contacts evaluated.** The CB8A contact definition was used.

From Figure 2.6 we can observe a general trend when evaluating additional interdomain contacts, the overall precision decreases, in line with expectations. Additionally, all of the methods which account for covariance chaining (CCMpred, EVfold and PSICOV) are substantially more precise than the more simplistic Mlp method, regardless of how many contacts are evaluated. After the 10$^{th}$ evaluated contact, EVfold has a mean precision score higher than PSICOV.

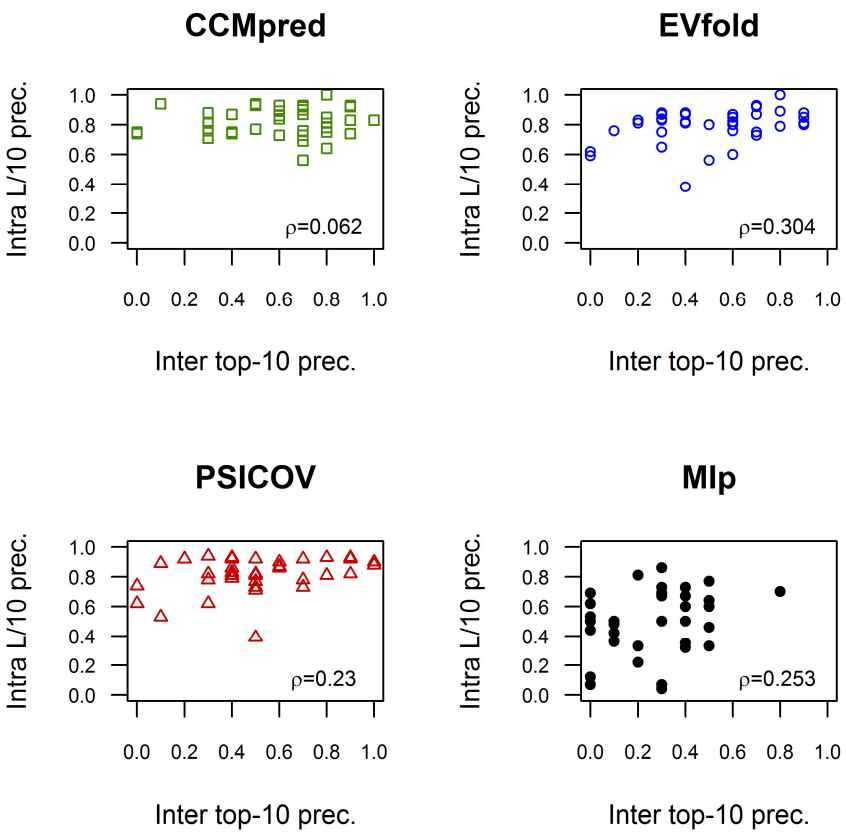



**Figure 2.7: Interdomain and intradomain precision values for each of the four assessed prediction methods.** The CB8A contact definition was used.

CCMpred, EVfold and PSICOV generate “global” models of covariation for a given MSA. As such, one may assume that low interdomain precision score may be the result of poor models being generated by the covariation methods. To investigate this possibility, the precision scores of intradomain predictions from the same models were calculated. One would assume that if the poor interdomain precision scores stemmed from poor covariation models, similarly poor intradomain scores would also be observed. However, as demonstrated in Figure 2.7, this is not the case. For all methods, low interdomain prediction precision scores may occur even when high precision intradomain predictions are observed, displaying weak correlation. This is particularly prominent in the global,

chaining-aware methods. If these global approaches failed to model covariation at all, one would expect intra- and interdomain precision scores to be strongly correlated. These findings may be explained by the level of maintenance of the interface throughout the MSA. In models where the interdomain interface is not well conserved throughout related proteins, these interfaces are unlikely to exhibit the same evolutionary pressure to coevolve. However, the structure of the separate domains will still need to be maintained, which would explain the much higher observed intradomain prediction capability. Unfortunately, this finding also suggests that intradomain prediction precision cannot be used as a means to infer the precision of interdomain predictions. MIp, which we know suffers from the effects of covariance chaining, generated considerably less precise intradomain predictions than the global approaches. As with the global approaches, the intra- and interdomain predictions generated by MIp are also only weakly correlated.

The interdomain predictions from the four methods are the results of explicitly removing intradomain predictions. As a consequence, the rank where the interdomain prediction lie within the full list is lost. As predictions are ranked according to their scores, one would naturally assume that predictions towards the top of the list, with higher scores, are more likely to be correct. To investigate the effect of full-list rank on resultant precision, the rank of the interdomain predictions within the full list was investigated; the results of which are predicted in Figure 2.8. From the figure it can be seen that the average rank of the interdomain predictions within the full list negligibly correlates with precision score. One would expect to see a strong negative correlation if mean rank did relate with scores.

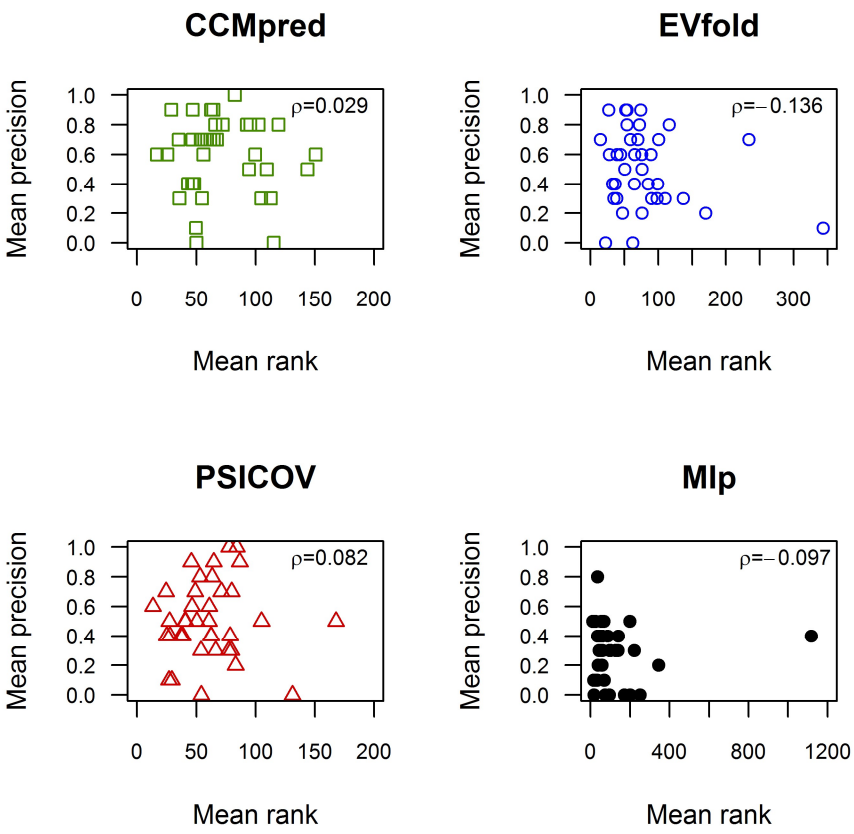



**Figure 2.8: Correlation of mean precision values for the top 10 interdomain contacts and mean interdomain rank.** The CB8A contact definition was used.

Considering the scores generated by each method, it is clear that the raw score values have clear differences (Figure 2.9). Looking at how these scores relate to prediction precision, one can see that, as hoped, high mean scores typically correspond to high mean precision values (Figure 2.10). The scores generated by CCMpred, EVfold and PSICOV are moderately correlated with precision values. MIp, which does not account for covariance chaining, produced a much larger number of erroneous predictions with high scores, highlighting the issue of superadditive correlation.

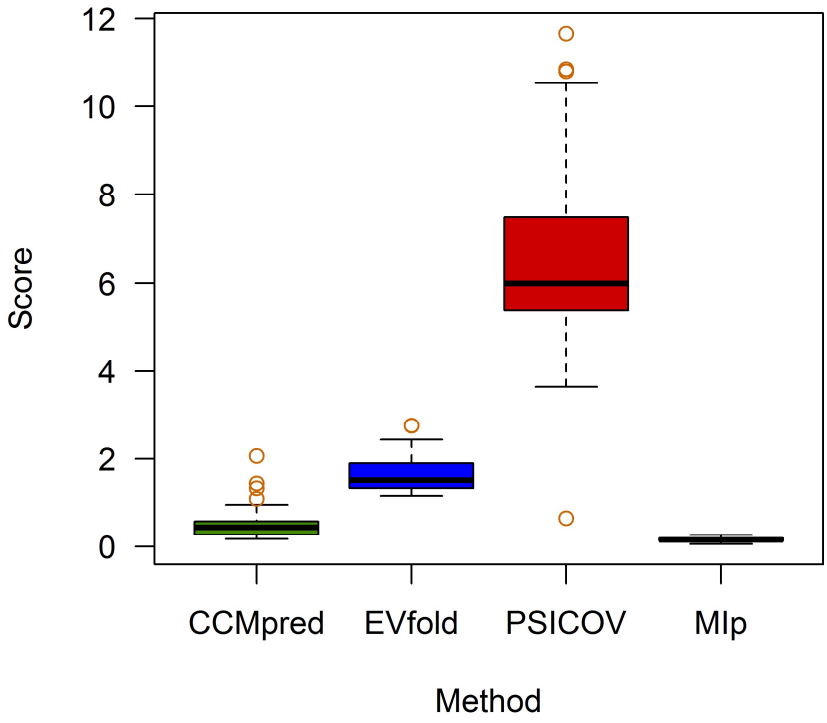



**Figure 2.9: Raw scores for the top 10 predictions from each of the assessed methods.**

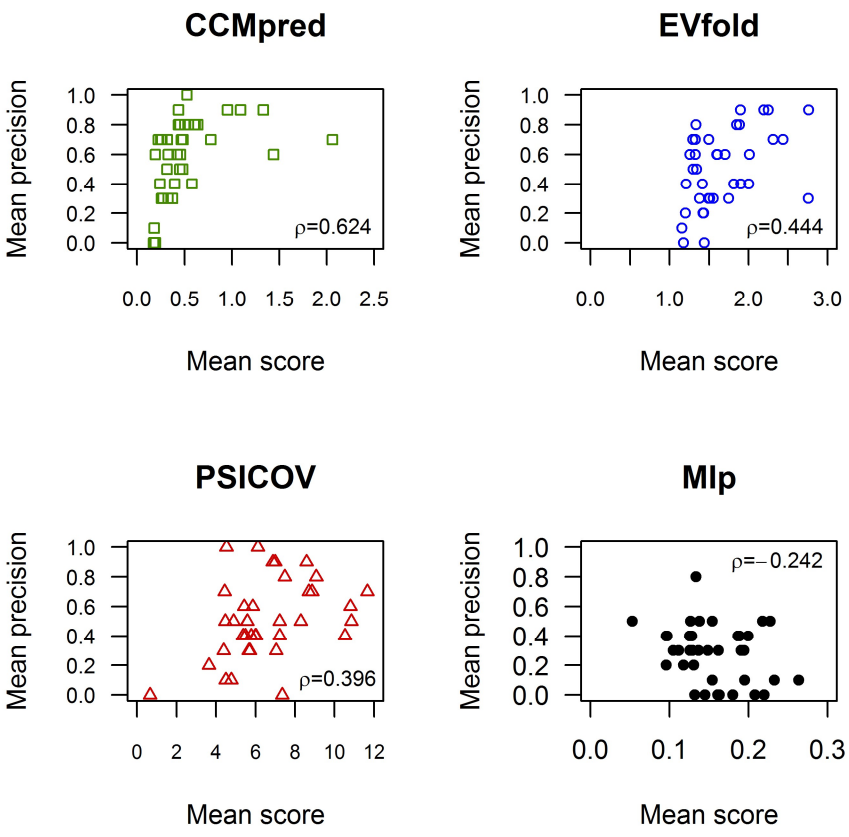



**Figure 2.10: Correlation between mean method score and mean precision values.**
The CB8A contact definition was used.

### 2.3.4 Statistical comparison between all methods

It is of course important to determine whether the differences in predictive capability observed between the assessed methods are statistically significant. A comparison between the four methods is presented in Table 2.9.

| | | MIp | EVfold | PSICOV |
|---|---|---|---|---|
| CCMpred | Test statistic (W) | **487** | **324.5** | **256** |
| | *p*-value | **$1.45 \times 10^{-6}$** | **$5.53 \times 10^{-4}$** | **$1.23 \times 10^{-3}$** |
| PSICOV | Test statistic (W) | **426** | 263 | |
| | *p*-value | **$3.34 \times 10^{-6}$** | 0.387 | |
| EVfold | Test statistic (W) | **475.5** | | |
| | *p*-value | **$4.18 \times 10^{-6}$** | | |

**Table 2.9: Table of *p*-values after an all-against-all comparison of methods using a paired one-tailed Wilcoxon signed-rank test with a 95% confidence interval.** The top 10 interdomain predictions for each method were evaluated, using the CB8A contact definition. Methods in each row are compared under the alternative hypothesis that the population mean rank is greater than that of the method for the column. Statistically significant results are shown in bold.

From Table 2.9 it can be seen that all of the approaches that account for the chaining effect significantly outperform the MIp method, which does not. Amongst the methods which differentiate between direct and indirect couplings, CCMpred performs with significantly higher precision than both EVfold and PSICOV, whereas the difference between PSICOV and EVfold is non-significant at the 95% confidence interval. The same pattern of significance between methods was observed when tested using the HA6A contact definition (data not shown).

### 2.3.5 Effect of alignment depth and diversity on predicted contact precision

Previous work on intradomain contact prediction has shown that predictive performance is moderately correlated with the number of sequences present in the analysed alignments (Jones et al., 2012). In our analysis, we also observe moderate correlation (Spearman's ρ = 0.464) between the total number of aligned sequences and precision (Figure 2.11). Considering individual methods, EVfold displays the strongest correlation between precision and total sequence count (ρ = 0.500), followed by PSICOV (ρ = 0.478) and CCMpred (ρ = 0.454).

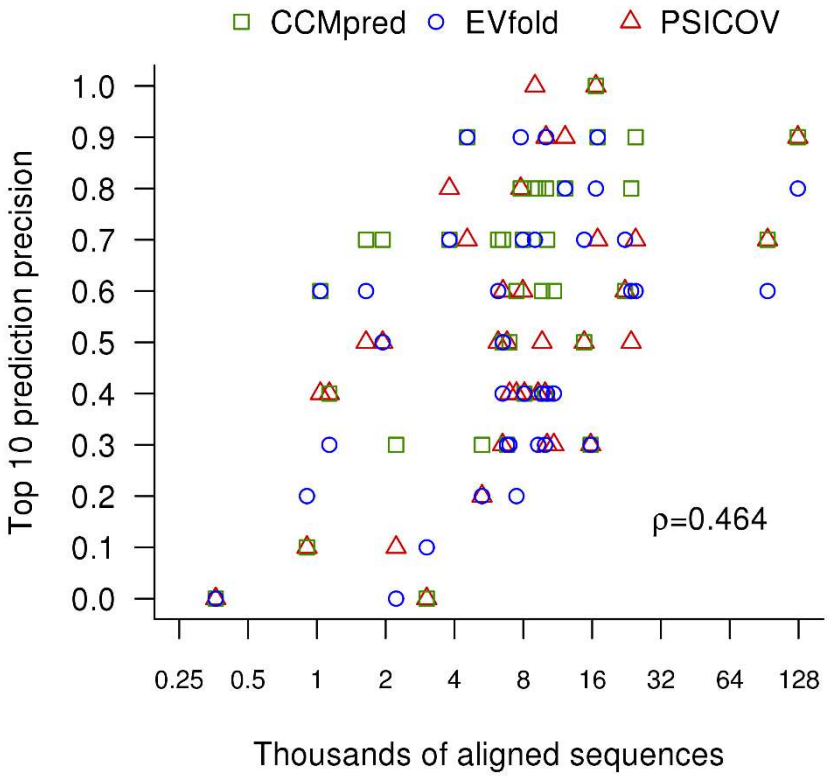



**Figure 2.11: Correlation between mean contact precision and total sequence count.** The CB8A contact definition was used.

Considering the number of effective sequences, again a moderate correlation is observed when considering all methods (Spearman's ρ = 0.491), marginally higher than that of total sequence count (Figure 2.12). Individual methods show slightly more variety in the levels of correlation observed, than that observed for the total sequence count (CCMpred ρ = 0.482; PSICOV ρ = 0.427 and EVfold ρ = 0.573). Both of these levels of correlation are smaller than their intradomain counterparts (intradomain total sequence correlation = 0.596, effective sequence correlation = 0.588; (Jones et al., 2012)), possibly caused by differences in interdomain orientation affecting precision scores for the work conducted here.

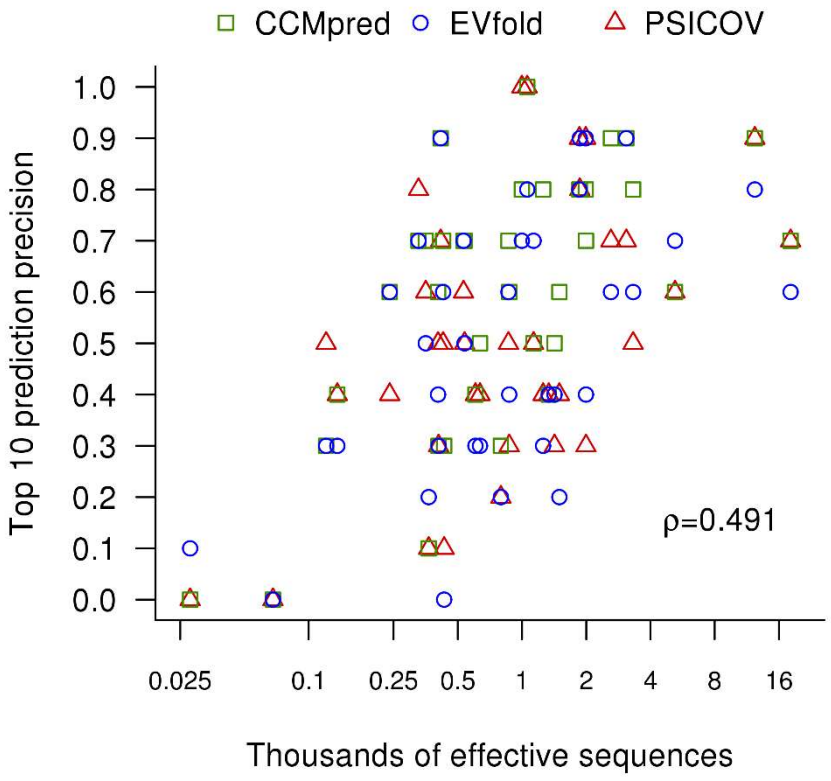



**Figure 2.12: Correlation between mean contact precision and effective sequence count.** The CB8A contact definition was used.

### 2.3.6 Overlap of predictions from alternative prediction approaches

PSICOV, EVfold and CCMpred employ different statistical models to distinguish between direct and indirect contacts. As the methods are clearly producing different results from their resultant precision scores provided with the same alignment (i.e. Figure 2.6), it is of interest to investigate the overlap between the generated predictions. In order to do so, we observed the similarities in the top 10 contacts predicted by each approach. The overlap of results is shown in Figure 2.13.

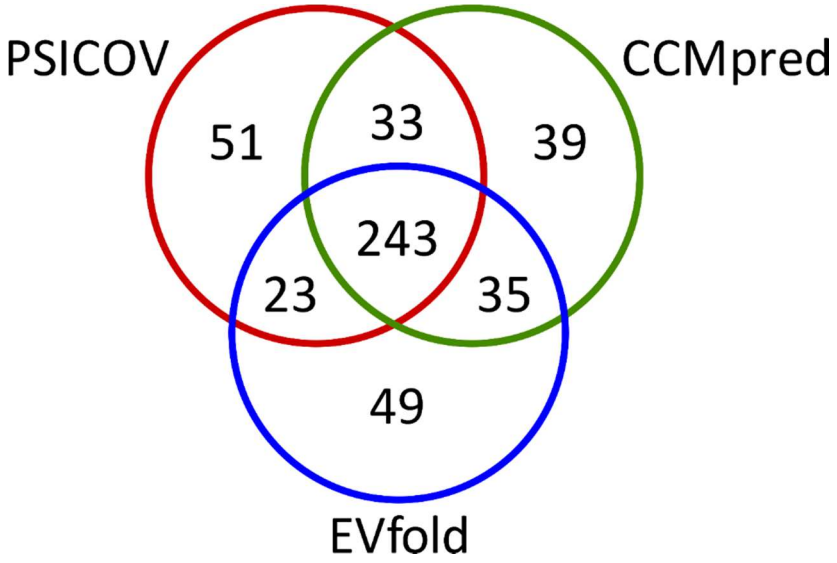



**Figure 2.13: Overlap of the top 10 correct contact predictions by PSICOV, CCMpred and EVfold.** 1050 contacts were assessed, and contacts were deemed correct under the CB8A criterion.

Figure 2.13 shows that contacts identified by different approaches share substantial overlap, with 69.4% of the top 10 contacts predicted by all three approaches. There is a minor level of overlap between pairs of methods, as well as correct contacts that are uniquely predicted by individual methods. These findings are in line with previous work

investigating the overlap of predicted contacts in intradomain cases (Tetchner et al., 2014; Jones et al., 2015).

However, to date, the overlap between incorrect contact predictions has not been investigated. The overlap between incorrect predictions identified by the methods is presented in Figure 2.14.

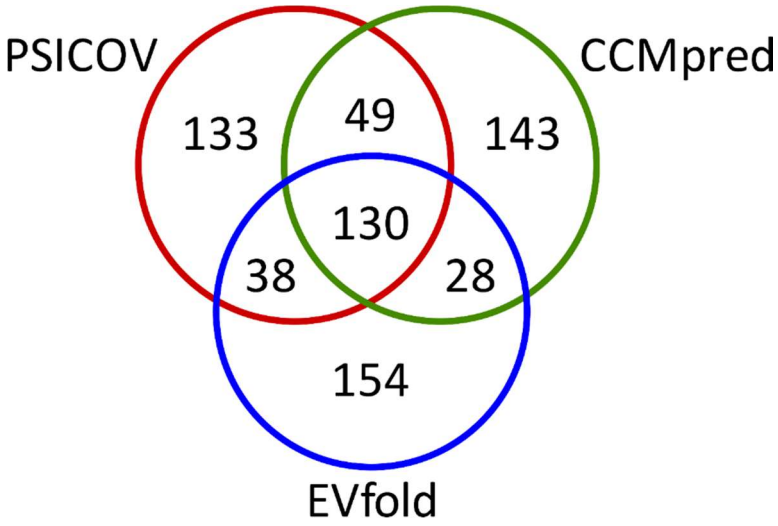



**Figure 2.14: Overlap of the top 10 incorrect contact predictions by PSICOV, CCMpred and EVfold.**1050 contacts were assessed and contacts were deemed incorrect under the CB8A criterion.

Here we can see that the overlap between incorrect contacts is less pronounced than the overlap between correct predictions. A large proportion (41.0%) of the incorrect predictions are isolated to individual methods. However, there remains a significant overlap (37.1%) of incorrect predictions that are identified by all three methods. The fact that these contacts are predicted by all three approaches despite the differences in the underlying models is interesting. It may be that some of these contacts fall slightly above the distance threshold used to determine whether a contact is correct or not. Alternatively, the methods may be identifying covarying residues which are not in close proximity within the experimental

structure, but may have some other functional role, such as involvement in folding, allosteric changes or interactions with other proteins.

A number of the contacts that are identified as incorrect using the CB8A contact threshold are actually in close proximity. This is shown in Figure 2.15 and highlights a key flaw with the CB8A contact threshold. The CB8A contact threshold does not account for the length of the side chains involved in an interaction, with the threshold merely being an approximation of the “average” Cβ-Cβ distance between a pair of residues. However, the Cβ-Cβ distance can be considerably longer than 8Å, whilst still maintaining an inter-residue heavy atom pairing within 5Å. The most extreme example of this would be a tryptophan-tryptophan contact, where the inter-Cβ distance can be 12.8Å, while still maintaining an inter heavy-atom contact between terminal oxygen atoms within 5Å. Investigating these contacts further, a number are in genuine contact, forming hydrogen bonds and salt bridges in the experimental structures. If instead the HA6A contact threshold is considered, 42 of the 130 incorrect contacts (under the CB8A criterion) would be reclassified as correct (Figure 2.15).

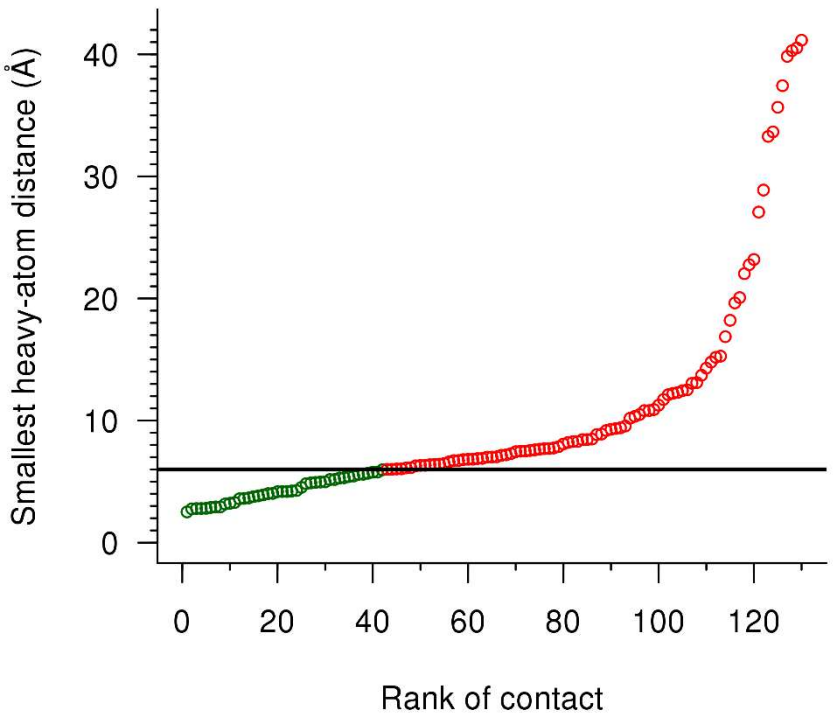



**Figure 2.15: Closest inter-heavy atom distance for the 130 contacts identified by EVfold, CCMpred and PSICOV, deemed incorrect by the CB8A contact threshold.** Inter-heavy atom distances less than 6Å are shown in green, and those above this threshold in red. The threshold at 6Å is shown in black for clarity. Distances are plotted in ascending order.

Repeating the analysis into the overlap of correct contacts with the HA6A contact definition, 66.3% of contacts overlap between all 3 methods, and other results are broadly similar to those observed in Figure 2.13 (data not shown). However, if we now look at the overlap of incorrect contacts (Figure 2.16), we can see that the overlap between incorrect contacts identified by all three methods reduces slightly, and more of the incorrect contacts are now isolated to individual methods. The 117 contacts identified by all 3 methods which

cannot be explained by close structural proximity are intriguing. The smallest inter-heavy atom distance of these 117 contacts is shown in Figure 2.17.

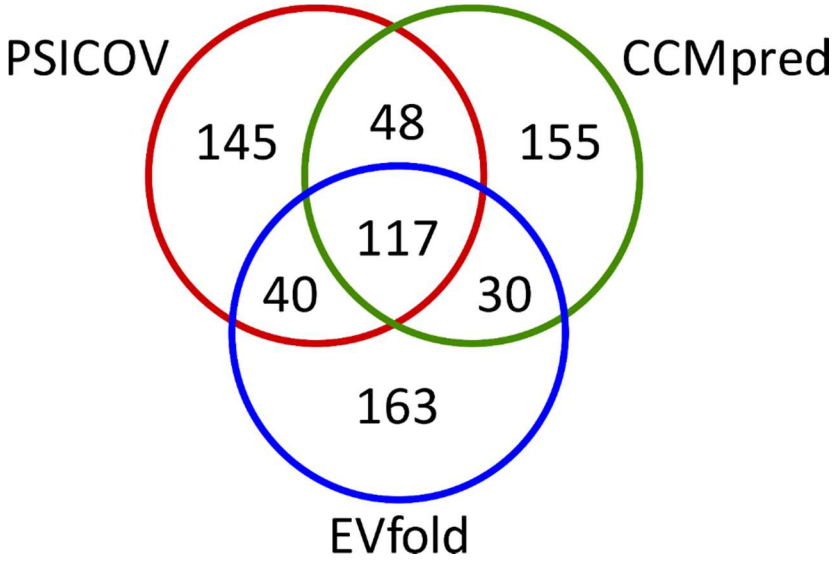



**Figure 2.16: Overlap of the top 10 incorrect contact predictions by PSICOV, CCMpred and EVfold.**1050 contacts were assessed and contacts were deemed incorrect under the HA6A criterion.

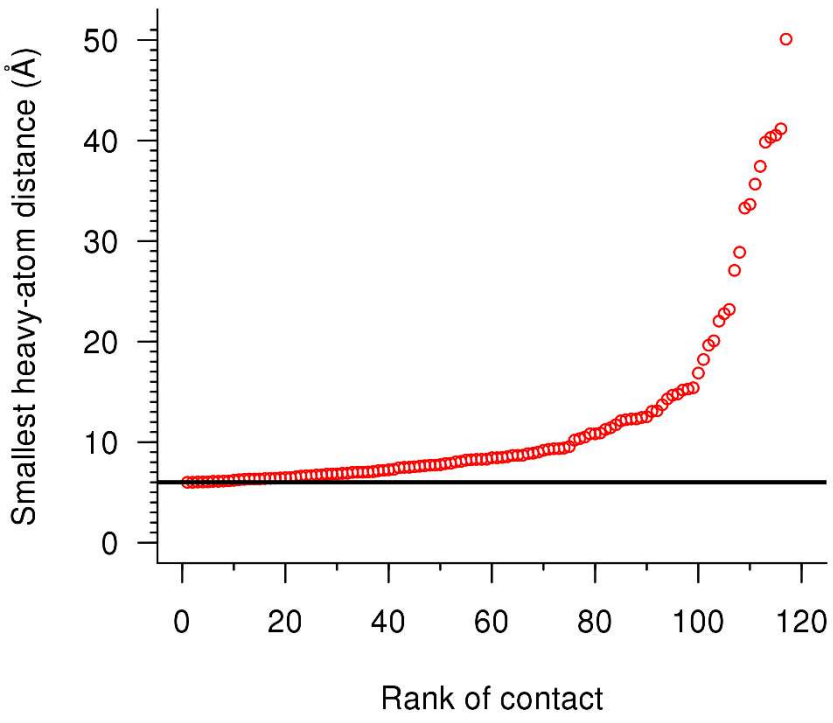



**Figure 2.17: Closest inter-heavy atom distance for the 117 contacts identified by the 3 covariance methods, deemed incorrect by the HA6A contact threshold (black line).** Distances are plotted in ascending order.

A number of the incorrect contacts fall slightly above the 6Å cutoff. These would be deemed correct using a slightly more permissive threshold, as used in other studies. However, there remains a large population of contacts observed at much larger distances which would still fall outside other, non-standard contact definitions.

If these larger distance contacts were limited to a single method, these could be explained as errors arising from inaccuracies in the approximations underlying the statistical models. However, as PSICOV, EVfold and CCMpred are based on different underlying models and approximations, this seems unlikely. There have been attempts to explain these longer-

range covariation events as structural effects. One explanation proposed is that these apparent long-distance covarying pairs may related to so-called “elastic” interactions involved in protein folding (Morcos et al., 2014). Further investigation into the reason why these long-distance contacts are identified by all three different methods would be interesting in future work.

### 2.3.7 Combining predictions using a simple consensus

As a large number of correct contacts are predicted by all 3 approaches, an attempt was made to combine the different predictors using a simple consensus approach, the results of which are shown in Figure 2.18.

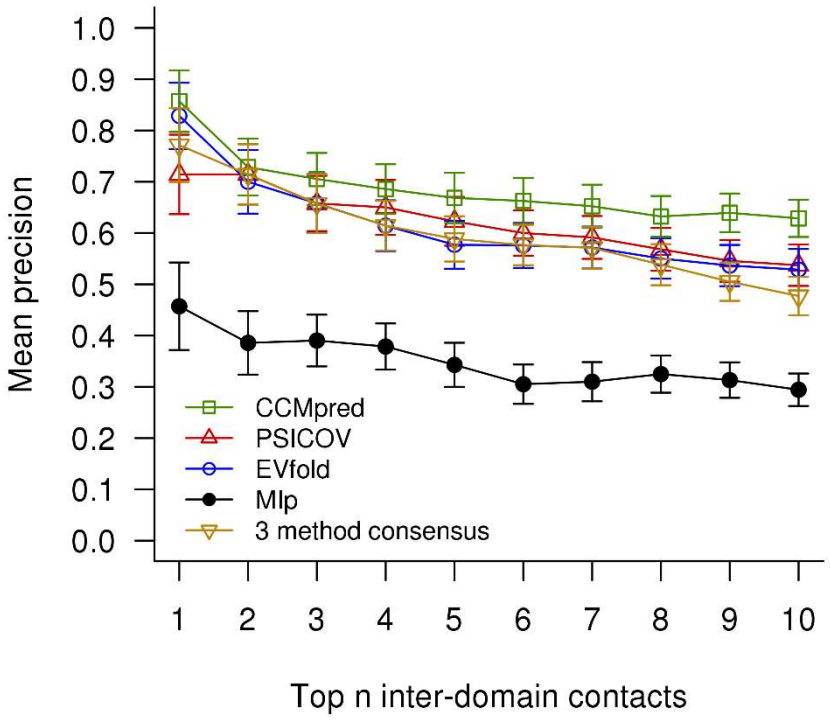



**Figure 2.18: Simple consensus approach compared to each single method.** Performance was assessed using the CB8A contact definition.

Figure 2.18 shows that the 3-method consensus is only able to improve upon the MIp method, and does not improve upon any of the individual methods on which it is based when assessed at 10 contacts, due to the large overlap of incorrect contacts observed in Figure 2.16. Whilst this simple method was unable to take advantage of the overlap between different methods, it is likely that more sophisticated machine learning approaches would be able to improve upon individual methods, as demonstrated in the development of methods for intradomain contact prediction (Jones et al., 2015; Skwark et al., 2013; Skwark et al., 2014; Ma et al., 2015). However, the small size of the current dataset is unlikely to be sufficient for reliable training and testing of a generalisable machine learning approach, due to overfitting. Training may be feasible when more of the 53/90 examples identified in Table 2.1 (which currently have small available alignments) gain additional sequences in the future.

## 2.4 Conclusions

This chapter has shown that covarying residue pairs between protein domains generally relate to residues located at the interdomain interface. Current covariation-based approaches are capable of identifying interdomain contacts with high precision across a diverse range of two-domain proteins, provided that large numbers of diverse homologous sequences are available.

Benchmark results have shown that maximal contact precision is achieved by generating MSAs using jackHMMer, rather than HHblits, likely due to the greater availability of sequences in searches. For jackHMMer, the best performing parameter set trialled used 2 search iterations, an E-value cutoff of $1x10^{-6}$ and a minimum sequence coverage of 70%.

After comparing four approaches based on different principles, it is apparent that all methods are capable of identifying interdomain contacts, though there are considerable

differences between achieved precision scores. The three methods which account for "covariance chaining" achieve the highest precision scores. Of these methods, CCMpred is the most precise (achieving a mean top-10 precision score (HA6A contact threshold) of 0.678), significantly improving upon the performance of PSICOV and EVfold (equivalent precision scores of 0.589 and 0.586, respectively), in line with observations of performance from previous studies of intradomain contact prediction (Jones et al., 2015; Skwark et al., 2013; Skwark et al., 2014). All of these approaches achieve significantly higher precision scores than the more basic MIp approach (which achieved an equivalent precision score of 0.327).

Overlap between the different prediction methods was investigated, and there is considerable overlap between correctly predicted contacts. Interestingly, a number of covarying residues are identified by PSICOV, EVfold and CCMpred which fall outside of typical thresholds for residues in direct contact. Whether these longer-distance covarying residues relate to unobserved functional or structural roles is unclear, and would be an interesting avenue of research in the future.

# 3. Using predicted contacts to select near-native docking models from a set of alternatives

## 3.1 Introduction

Docking algorithms attempt to model the bound form of multicomponent assemblies given the structures of each constituent. Commonly, this is performed at the chain level, where entire proteins are “docked” in order to generate models of the bound protein complex. In the same vein, docking can also be performed at the domain level, attempting to model the structure of multidomain proteins if the structures of individual domains can be obtained. Docking proceeds in two stages. Firstly, docking approaches generate a large set of alternative models. This assortment of models (or “decoys”) is then ranked in a second step according to a scoring function, attempting to place models thought to represent the native structure at, or at least near, the top of the ranked list.

Progress in the protein docking community is monitored through a community-wide experiment named CAPRI (Critical Assessment of PRediction of Interactions; http://www.ebi.ac.uk/msd-srv/capri/). In order to assess the current capability of approaches for identifying good quality docking models, a “scoring” category was introduced during the 3rd CAPRI assessment (Wodak and Lensink, 2007). In the scoring experiment, after initial docking models have been submitted by all groups, the models generated during the prediction stage are combined and redistributed for groups developing scoring functions to re-rank the decoys according to their own approaches. The performance of each scoring method is then assessed by evaluating the number of near-native decoys which are identified. However, despite 40 years of developing docking approaches since the first such work in 1975 (Levinthal et al., 1975) and 8 years of

specifically evaluating scoring approaches in CAPRI, correctly ranking decoys remains a difficult challenge (Lensink and Wodak, 2013).

As covarying residues between domains are often located at the interface (as demonstrated in the previous chapter), predicted contacts should be suitable for identifying models containing native-like interfaces. The idea to identify native-like decoys using covarying residues is not new, and local covariation approaches have previously been used for this purpose (Tress et al., 2005; Andreani et al., 2013; Madaoui and Guerois, 2008; Pazos et al., 1997). However, these studies did not take advantage of the recent developments in chaining-aware approaches, which forms the rationale for the work conducted here.

This chapter describes an evaluation of whether predicted contacts are sufficient to identify native-like docking models from a set of alternatives. In order to evaluate the devised approach, a set of decoys was generated using the PatchDock program (Schneidman-Duhovny et al., 2005; Duhovny et al., 2002), which has been used in other recent studies to assess decoy selection approaches (such as Bhaskara *et al.*, 2013; Ovchinnikov *et al.*, 2014).

# 3.2 Method

## 3.2.1 The PatchDock program

PatchDock is a rigid body docking program, which generates a set of alternative bound models based on the structures of two unbound proteins (Schneidman-Duhovny et al., 2005; Duhovny et al., 2002). PatchDock uses surface contours to guide the assembly of unbound molecules using a surface complementarity approach. Firstly, the surface of each unbound component is assessed as a series of concave, convex and flat sections (Figure 3.1). Concave and convex patches are then matched together, and flat patches matched

with any type of patch, producing an initial set of candidate structures. In the next stage, candidate structures are tested for steric clashes and models without substantial intrusions are ranked according to their level of surface complementarity (Schneidman-Duhovny et al., 2005).

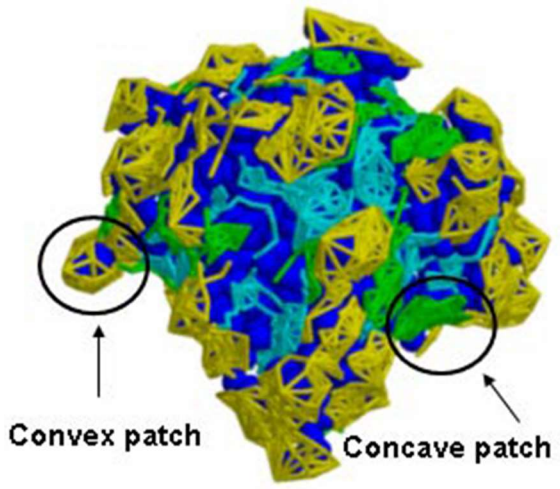



**Figure 3.1: Example of trypsin inhibitor (PDB code 1BA7), assigned convex, concave and flat patches by PatchDock.** Convex patches are shown in yellow, concave patches shown in green and flat patches shown in cyan. Image reproduced from (Duhovny et al., 2002).

The local matching of surface features enables the algorithm to be fast, as it does not require extensive searching of six-dimensional space (three rotational and three translational). However, the comparison of surface features can lead to multiple poses achieving similar surface complementarity scores. In order to reduce the similarity of reported models, PatchDock incorporates two clustering steps to enforce diversity. Models are first ranked according to their transformation parameters, followed by a more intensive root-mean-square deviation (RMSD) clustering step. An explanation of the RMSD measure is provided in Section 3.2.3. The RMSD clustering step reduces the similarity

between models, so that the minimum RMSD score between any two models is specified by a threshold, which by default is set to a value of 4Å.

Throughout this chapter, models will be referred to by the rank assigned from the PatchDock surface complementarity score, such that the model ranked with the highest score will be referred to as “decoy 1”.

### 3.2.2 Generating decoy structures with PatchDock

The two domains used for this study were cleaved from the experimental crystal structure obtained from the PDB. In order to include the full length of the protein chain for analysis, the two domains were split at the beginning of the second domain, to include a linker in the structure of the first domain, if present. The two cleaved domains were used as the starting structures for the docking procedure without further modification.

PatchDock decoy generation was performed using default parameters, including the default model clustering threshold of 4Å RMSD to ensure that a diverse range of alternative models was generated. 200 docking models were generated for each of the 37 proteins listed in Table 2.2.

### 3.2.3 Assessment of decoy models

To evaluate the performance of the devised scoring approach, the two scoring metrics employed for the assessment of docking models in the CAPRI experiment were used. Additionally, the CAPRI contact definition is used, which defines a contact to be formed between two residues when the distance between any two heavy atoms is less than 5Å (Lensink and Wodak, 2013). Hereafter, this contact definition will be referred to as “HA5A”.

**Interface RMSD (iRMSD) and RMSD**

The first of the assessment criteria used in CAPRI is the interface root-mean-square deviation (iRMSD) metric. iRMSD is related to the more common RMSD metric, but emphasises differences over interface residues, rather than the whole protein chain.

Both iRMSD and RMSD values are calculated using the equation:

$$(i)RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

Where $\delta$ is the distance between $N$ pairs of equivalent atoms between the reference experimental structure and a model to be evaluated, after structural superposition.

The difference between iRMSD and RMSD stems from the atoms which are measured between. The iRMSD measures the deviation between the backbone atoms of the residues lying at the interface of the reference structure (defined using the HA5A threshold) and the equivalent residues of a model after the two structures have been superposed (Lensink and Wodak, 2013). On the other hand, RMSD scores are calculated using all equivalent atom pairs, across the whole of both chains.

Calculating the iRMSD quantifies the distance between the interface of a decoy and that of the experimental structure. Using the iRMSD measure rather than the whole chain RMSD reduces the impact of the bound molecule size and shape on the calculated iRMSD value, increasing comparability between results (Janin *et al.*, 2003; Figure 3.2). Identical structures would have an (i)RMSD value of 0, with values increasing as structures increasingly diverge from the reference.

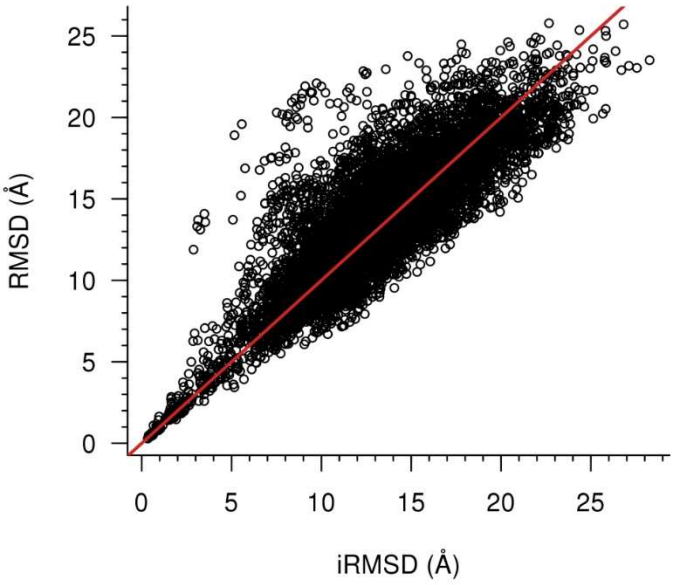



**Figure 3.2: Decoy RMSD plotted against decoy iRMSD for the 7400 models generated by PatchDock.** Spearman's ρ = 0.811. The red line shows $x = y$ for reference.

Figure 3.2 demonstrates that it is possible to achieve small iRMSD scores when considering differences at the interface that would be deemed as large RMSD differences over the protein as a whole. Whilst some differences can be observed, in general the iRMSD and RMSD values are strongly correlated.

Both RMSD and iRMSD values were calculated using ProFit (version 3.1, Martin, A.C.R., http://www.bioinf.org.uk/software/profit/).

**Fraction of native contacts (fNat)**

The second assessment criterion used in CAPRI is the fraction of native contacts (fNat) which are observed in a docked model. The fNat is simply the percentage of native HA5A contacts recalled within a decoy structure (Lensink and Wodak, 2013). The fNat score is generally reported as a decimal, but for this work fNat scores were converted into a

percentage to aid interpretation. Identical structures would score a value of 100, with structures not sharing any interface residues scoring 0.

**CAPRI assessment of model quality**

The assessors of the CAPRI experiment classify predictions based on model iRMSD and fNat scores. A summary of these classifications is shown in Figure 3.3.

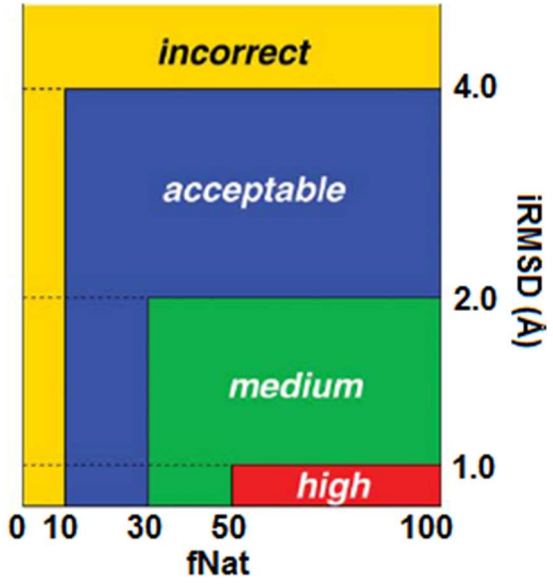



**Figure 3.3: CAPRI assessment criteria for the quality of docking models, based on iRMSD and fNat scores** Image adapted from (Lensink and Wodak, 2013).

## 3.2.4 Dataset

The 37 proteins listed in Table 2.2 acted as the basis for this study. To ensure that a near-native model was present within the set of decoy structures for each model after docking, targets were filtered by iRMSD score. For a target to be used in this study, at least 1 decoy within the set of 200 alternatives was required to have an iRMSD score ≤ 2Å, relating to CAPRI “medium” and “high” quality structures. PatchDock was able to generate a medium

or high quality model within the 200 decoys for 25 of the 37 proteins from Table 2.2. One protein, target 2W6PB, was excluded from analysis as the 3 of the 4 covariance methods were unable to identify any interdomain contacts correctly for this target (Table 2.8), leaving a final set of 24 proteins which were assessed in this study. The 24 remaining proteins and their best scoring PatchDock decoy are presented in Table 3.1.

| PDB ID | Lowest iRMSD decoy (Å) | fNat score |
|---|---|---|
| 1EE8A | 0.658 | 85 |
| 1EH6A | 0.785 | 94 |
| 1JDBF | 0.918 | 93 |
| 1LI5A | 0.609 | 95 |
| 1MGPA | 0.498 | 94 |
| 1OI7A | 0.383 | 97 |
| 1PUJA | 0.939 | 93 |
| 1T6CA | 0.455 | 95 |
| 1V0BA | 0.928 | 86 |
| 1VHNA | 0.355 | 96 |
| 1VMAA | 0.535 | 94 |
| 1WF3A | 0.890 | 88 |
| 1WJ9A | 0.841 | 85 |
| 2B6CB | 1.034 | 92 |
| 2CGJA | 0.540 | 95 |
| 2HIYC | 0.804 | 88 |
| 2QFLA | 0.388 | 94 |
| 2RA9A | 0.822 | 89 |
| 2WHYA | 0.331 | 100 |
| 3A4TA | 0.533 | 95 |
| 3CI0J | 0.479 | 100 |
| 3HP7A | 0.965 | 89 |
| 3NZKB | 0.639 | 87 |
| 3VO8B | 0.652 | 97 |

**Table 3.1: Summary table of the 24 proteins used for the decoy selection study.** PDB ID = Protein Data Bank identifier. Lowest iRMSD = Lowest observed iRMSD value within the 200 decoys generated by PatchDock. fNat score = Corresponding fNat score of the lowest scoring iRMSD decoy.

Table 3.1 shows that PatchDock was able to generate at least one CAPRI “high” quality model for 23/24 cases, with the remaining case generating a decoy of “medium” quality.

### 3.2.5 Proposed re-ranking procedure

We proposed to use a simple procedure to re-rank models using contacts. Provided with a list of contacts, the approach ranked each decoy structure based on the number of contacts observed. The decoy or decoys which contained the most predicted contacts were taken as our prediction of being native-like. As the set of predicted contacts mainly relate to residue pairs located at the domain interface (as demonstrated in Chapter 2), the decoys containing the highest number of predicted contacts should have a domain interface resembling the native structure.

As previously discussed, contacts can be defined using a number of different thresholds, calculating distances between different parts of each amino acid (see Sections 1.5.1 and 2.2.6.1). Three contact definitions were considered in this chapter; the two contact definitions used in the previous chapter (HA6A and CB8A) alongside the definition of a contact used in the CAPRI assessment: HA5A.

#### 3.2.5.1 Use of experimental contacts

In order to establish which of the three contact thresholds provides optimal selective performance, a simulation was performed emulating the decoy selection procedure using contacts extracted from the experimental structures. By applying experimental-structure derived contacts, we are able to assess the best-case scenario for this decoy selection approach as if the contact list was free of false positive predictions.

To perform these analyses, the decoy set described in Section 3.2.4 was used. For each protein, $n$ contacts were randomly sampled from the set of observed experimental contacts and used to rank decoys. This procedure was repeated 1,000 times for each of the 24 proteins in Table 3.1. For each repeat, a different random set of contacts was selected to avoid biasing the simulation for a particular arrangement of interface residues.

#### 3.2.5.2 Use of predicted contacts

For each of the proteins in the dataset, the 200 decoys were re-ranked according to the number of predicted contacts observed within the decoy structure. The predicted contacts used here are the same as those generated using the best-performing parameter set identified in Chapter 2.

### 3.2.6 Significance testing with bootstrapping

In order to evaluate the statistical significance of the differences in prediction accuracy between two methods, a bootstrap calculation was performed to test the null hypothesis that the proposed contact-based approach does not provide more accurate solutions (i.e. with lower iRMSD values) than PatchDock does. For the 24 test cases at hand, the paired differences in iRMSD values obtained by the models identified by PatchDock and by the proposed contact-based approach were calculated. When the contact-based approach identified multiple models as top ranked, the prediction with highest iRMSD was considered. Then $10^5$ samples were drawn with replacement, and for each sample the average value was stored to approximate the expected distribution of differences in iRMSD values. The *p*-value for the null hypothesis was finally estimated as the proportion of expected paired iRMSD differences less than or equal to zero.

## 3.3 Results and discussion

### 3.3.1 Decoy set properties

This section describes the range of observed iRMSD and fNat scores for the set of 4800 decoys used to evaluate the selection procedure. The aim of generating a large set of decoys was to increase the level of diversity among models, necessary in order to evaluate whether the contact-based selection method is able to identify native-like structures. The diversity of the decoy set, as measured by the iRMSD value, is shown in Figure 3.4.

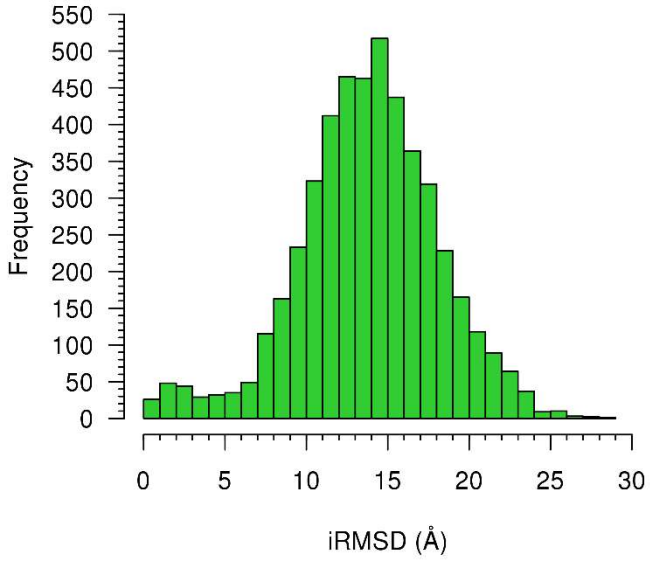



**Figure 3.4: iRMSD distribution of the 4800 decoy structures generated by PatchDock.**

From Figure 3.4 we can see that the generated decoys are highly diverse, with a wide range of iRMSD scores generated. The majority of decoys have high iRMSD scores,

indicating that the modelled interface is distant from that seen in the reference experimental structure. Referring back to the CAPRI evaluation criteria shown in Figure 3.3, 147/4800 models (3.1%) have an iRMSD score ≤ 4Å (“acceptable quality models”), 74 ≤ 2Å (1.5%) (“medium quality models”) and 26 ≤ 1Å (0.5%) (“high quality models”). The vast majority (96.9%) of models relate to “incorrect” models under the CAPRI criterion (iRMSD values > 4Å).

The distribution of decoy fNat scores is shown in Figure 3.5.

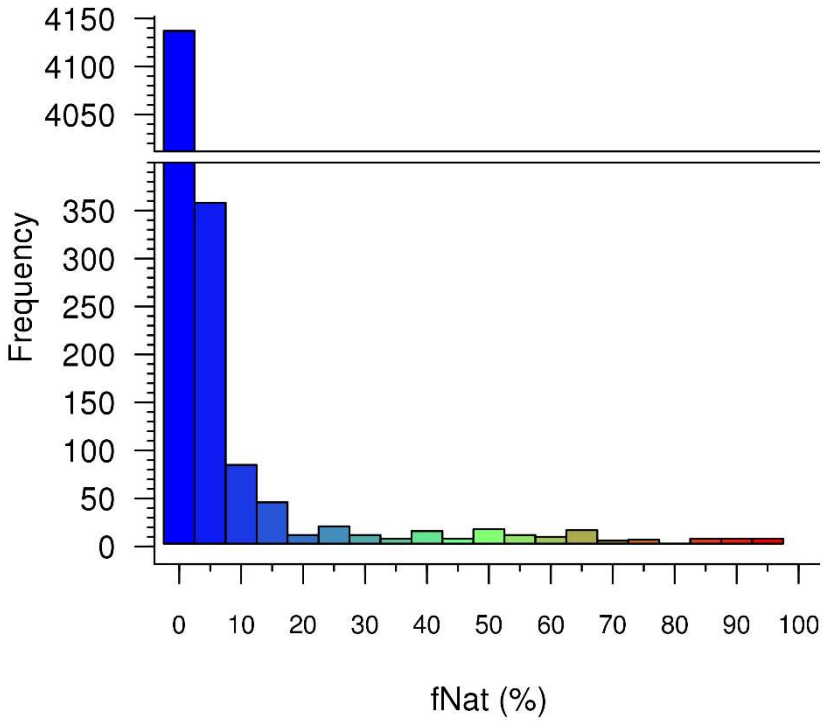



**Figure 3.5: fNat distribution of the 4800 decoy structures generated by PatchDock.**

Figure 3.5 shows that the majority of generated decoys recall 0-5% of the interface residues present in the experimental structure. In fact, most decoys (3170/4800 or 66.0%)

recall 0% of the native contacts. Using the CAPRI criterion shown in Figure 3.2, 305 decoys (6.4%) obtain an fNat score ≥ 10, 141 decoys (2.9%) obtain an fNat score ≥ 30, 97 decoys (2.0%) obtain an fNat score ≥ 50, 16 decoys (0.3%) obtain an fNat score ≥ 90 and just 2 decoys (0.04%) obtain an fNat score of 100, indicating all interface contacts observed in the experimental structure are recalled in the model.

As the set of decoys is diverse, with the majority of decoys not resembling the experimental structure (as measured by both iRMSD and fNat scores), the set was deemed suitable for testing our contact-based approach for near-native decoy selection.

### 3.3.2 Selection of contact definition for decoy selection

Figure 3.6 shows the effect of using true contacts, extracted from the reference structure, for selecting decoys assumed to resemble the native structure. Figure 3.6a shows that the contact definition affects the number of decoys which are selected by the procedure. The HA6A definition is the most permissive, selecting the highest number of decoys across all 10 trials. The CB8A threshold selected slightly fewer decoys, with the HA5A threshold being the most exclusive of the trialled thresholds. Also from Figure 3.6a, it can be seen that the approach is able to select a small subset of the decoy structures using only a single contact. Using a single contact, the HA5A threshold selects an average of just 3.6% of the 200 starting decoys. Provided with additional contacts, the selection procedure selects the decoys which match the most provided contacts, identifying the models that have interfaces closest to that seen in the reference structure. However, the added benefit from additional contacts can be seen to rapidly diminish after approximately 6 contacts. Using 10 contacts, all 3 contact thresholds identify just over 1 model on average (HA6A = 1.25 models, CB8A = 1.22 models and HA5A = 1.15 models).

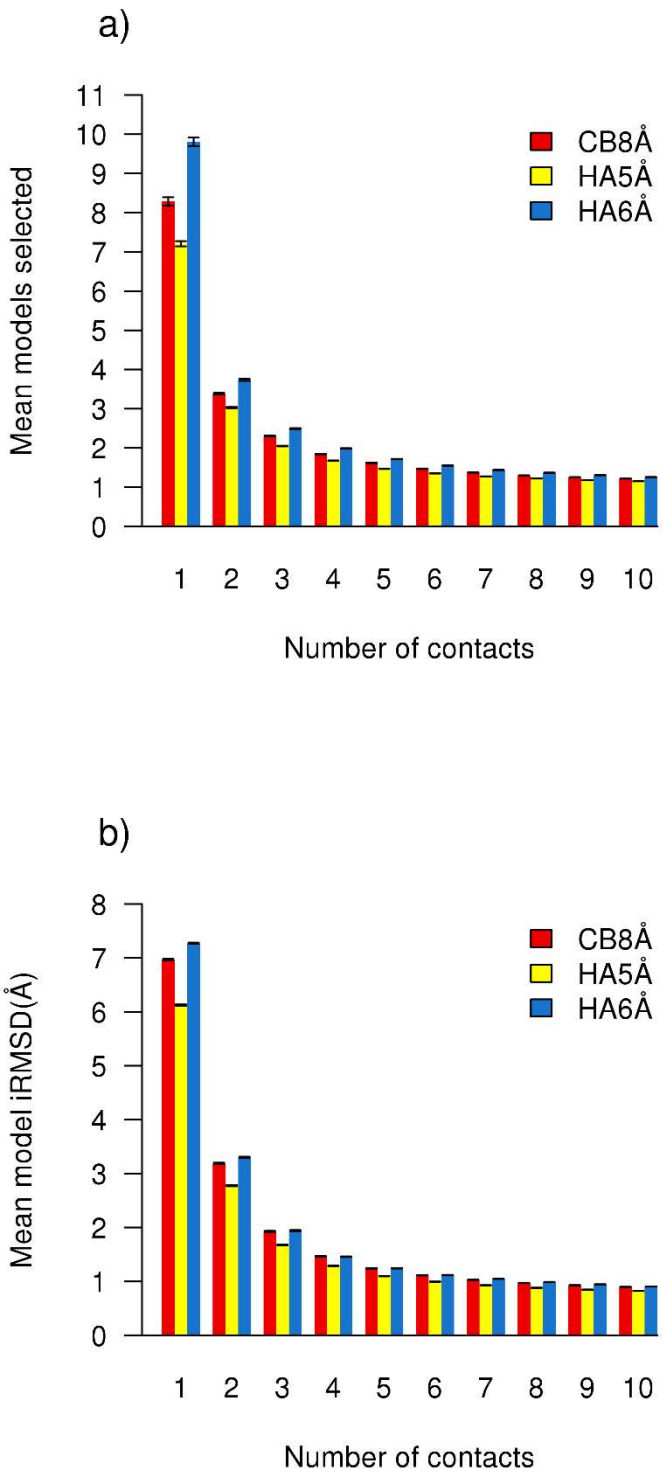



**Figure 3.6: Effect of the decoy selection procedure using increasing numbers of experimentally-observed contacts.** Effect shown for a) the number of models selected from the 200 decoys and b) iRMSD values for each selected model.

Figure 3.6b shows the corresponding iRMSD value for the selected models. Using a single contact, a number of very diverse models may contain this contact, resulting in a high average iRMSD value, regardless of the contact definition considered. With the addition of further contacts, more diverse decoys are removed, in turn reducing the mean iRMSD values of the selected set. Using 10 contacts, the three contact thresholds identify decoys with average iRMSD scores just below a value of 1Å (HA6A = 0.905Å, CB8A = 0.898Å and HA5A = 0.823Å).

As the HA5A distance identifies the fewest decoy structures (Figure 3.6a), which also have the lowest iRMSD values (Figure 3.6b), the HA5A threshold was selected as the threshold used for applying predicted contacts. Whilst the absolute differences between the iRMSD values generated by each contact threshold are small when 10 contacts are applied, the HA5A threshold significantly outperforms both CB8A and HA6A thresholds using the procedure (Wilcoxon rank sum tests, $p$-values < $2.2 \times 10^{-16}$).

### 3.3.3 Identification of near-native decoys using predicted contacts

In this section the decoy selection procedure was repeated using predicted contacts from CCMpred, EVfold, MIp and PSICOV. A predicted contact was considered to be observed if it formed a HA5A contact, the threshold identified as having the best selective performance in the previous section. Figure 3.7 shows the effect of applying predicted contacts on the number of decoys selecting using the decoy selection approach.

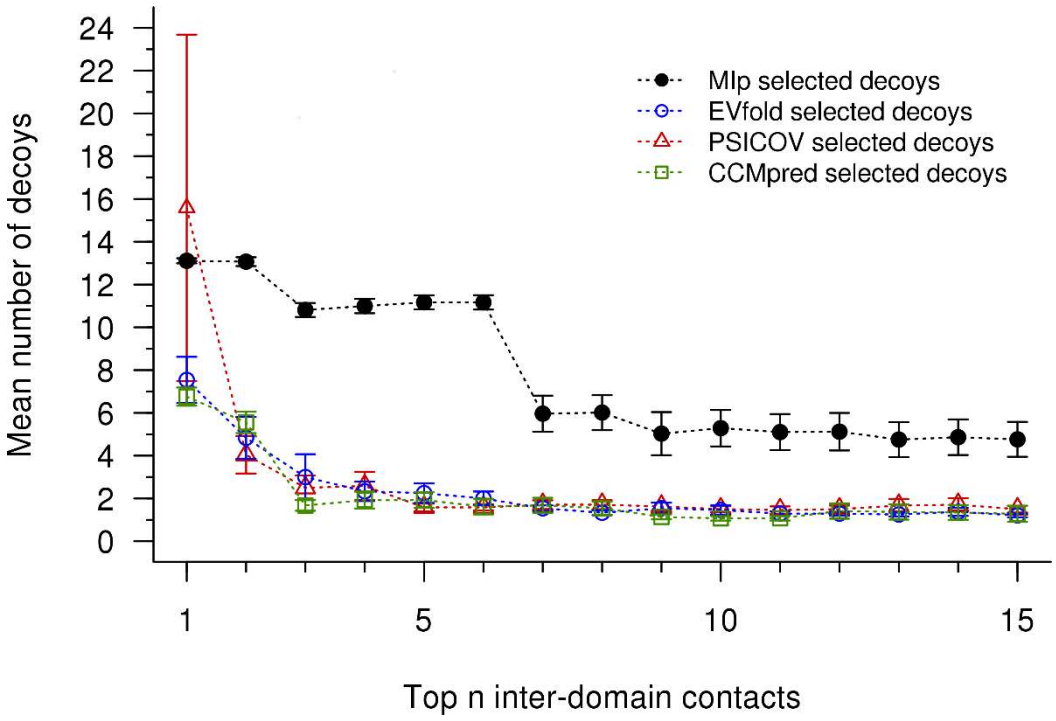



**Figure 3.7: The effect on the number of decoys selected by applying increasing numbers of predicted contacts.**

Figure 3.7 shows that using the generated predictions, we observed the same general trend as using the experimental contacts, as seen in Figure 3.6a. A single contact was sufficient to remove the majority of the 200 decoys, then additional contacts acted to further reduce the number of top-ranking models. Beyond 6 contacts, the difference in the number of decoys identified by CCMpred, PSICOV and EVfold was small, and additional contacts typically reinforce the selected models after this point. This trend occurred until roughly 11 contacts (though the exact value is method dependent), where the number of models then began to increase. This effect was the result of increasing proportions of false-positive predictions being incorporated into the ranking procedure. False-positive predictions which are not located at the true interface permit models which do not resemble the experimental structure to be included in the highest-scoring group. MIp, which produces less precise predictions, typically generated larger groups of models in the

final set of highest-scoring decoys. These results highlight the importance of high precision in the prediction of contacts in order to effectively distinguish between decoys.

Particularly notable is the atypical point using a single PSICOV prediction where the method selects the highest mean number of decoys. This is due to the target 2HIYC, where the single predicted contact is incorrect and not observed in any of the 200 structures. Due to this, all 200 models were observed to contain 0 contacts and ranked equally as the highest/lowest scoring group. Provided with a second (correct) contact, the procedure then identified 8 models which satisfied 1 of the 2 provided predictions.

Looking at the quality of the models identified by the approach, it can be seen that the predicted contacts are successful at identifying low iRMSD models (Figure 3.8). Again, the general trend is for model quality to increase with the consideration of additional contacts. As one may expect, MIp was unable to discriminate between decoys as effectively as the chaining-aware methods, selecting higher mean iRMSD scoring models.

As all 200 decoys were considered equal for the case of 2HIYC using a single PSICOV contact, the average iRMSD score for all 200 decoys (11.84Å) raised the overall mean value in that trial. However, as with the number of models, performance was quickly recovered with the use of additional contacts. The increase in mean iRMSD score seen for PSICOV using 13 and 14 contacts was caused by the target 1OI7A. Using 11 and 12 contacts, the top-ranked group contains models satisfying just 3 predictions. With the consideration of the 13$^{th}$ contact (a false positive prediction), 5 additional models which previously only satisfied 2 predictions were incorporated into the highest-scoring group. The addition of these 5 models, which do not resemble the experimental structure, acted to increase the mean iRMSD score. These 8 models remained in the selected group until the 15$^{th}$ contact was considered (a true positive prediction), which acted to remove the

higher iRMSD models, reinstating the same 2 models identified using 11 contacts as the highest ranking pair.

All three of the chaining-aware methods significantly outperform MIp at identifying low iRMSD models (Table 3.2). Amongst the chaining-aware methods, differences between CCMpred and EVfold are non-significant, although CCMpred significantly outperforms PSICOV (just passing the 95% significance threshold). Differences between PSICOV and EVfold are also shown to be statistically non-significant.

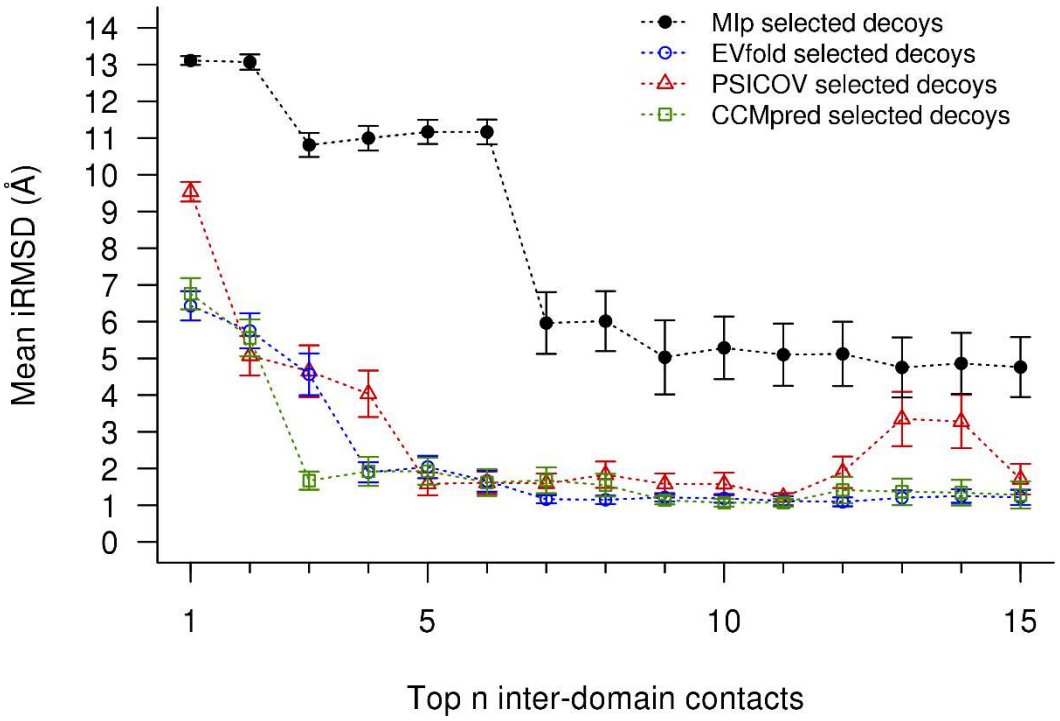



**Figure 3.8: Average iRMSD values for the contact-selected decoys from Figure 3.7.**

| | | MIp | EVfold | PSICOV |
|---|---|---|---|---|
| CCMpred | Test statistic (W) | **6** | 30 | **8** |
| | *p*-value | **$7.36 \times 10^{-4}$** | 0.412 | **0.049** |
| PSICOV | Test statistic (W) | **9** | 57 | |
| | *p*-value | **$2.06 \times 10^{-3}$** | 0.799 | |
| EVfold | Test statistic (W) | **13** | | |
| | *p*-value | **$1.43 \times 10^{-3}$** | | |

**Table 3.2: Table of *p*-values after an all-against-all comparison of methods using a one-tailed Wilcoxon signed-rank test with a 95% confidence interval.** As no single number of contacts provides general optimal performance across the methods, the iRMSD values for the best-performing number of contacts for each method was evaluated (10 contacts for CCMpred, 11 contacts for PSICOV, 12 contacts for EVfold and 13 contacts for MIp). For targets where multiple decoys were selected, the average iRMSD value was calculated. Methods in each row are compared under the alternative hypothesis that they identify lower iRMSD models than method in the column. Statistically significant results are shown in bold.

A similar result is observed when considering the fNat score (Figure 3.9). Due to the additional decoys which are selected using MIp, the approach achieves the lowest mean fNat score across the decoy set. Once again, the mean fNat score of PSICOV-selected decoys using a single contact was reduced as a result of the incorrect contact used for the 2HIYC target, as well as the 2 points for the case of 1OI7A, mentioned previously. Statistical comparisons between selected model iRMSD scores display the same patterns of significance shown in Table 3.2 (one-tailed Wilcoxon signed-rank tests with a 95% confidence interval; data not shown).

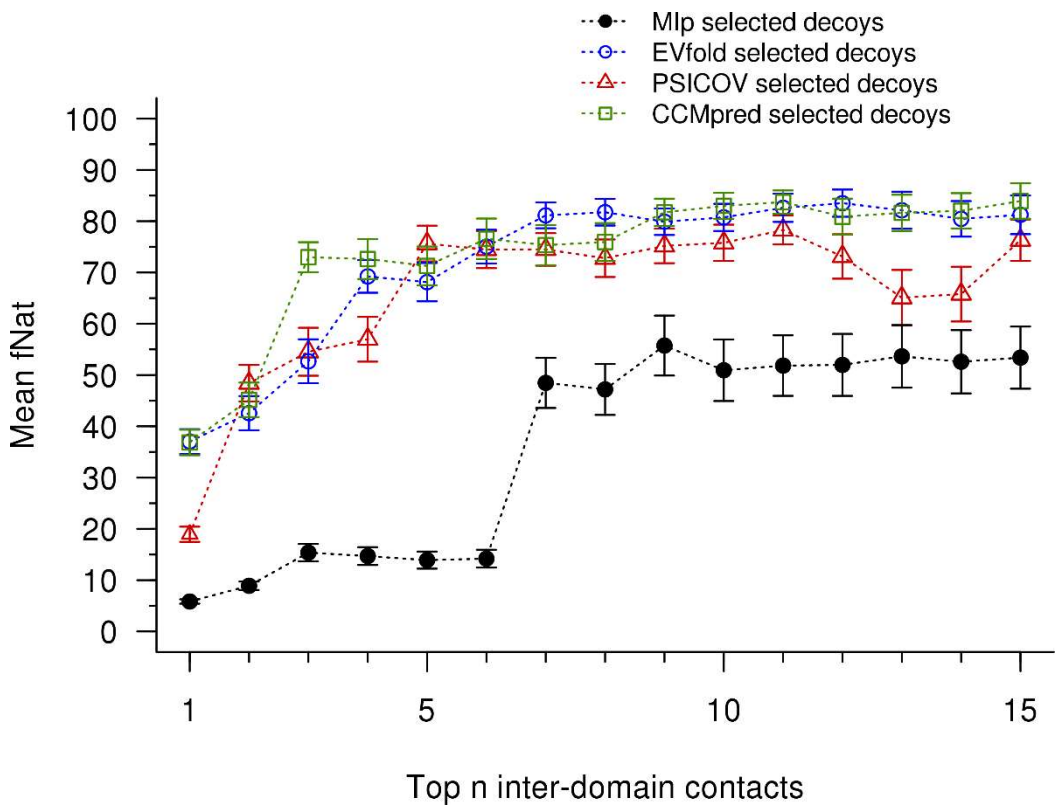



**Figure 3.9: Average fNat scores for the contact-selected decoys from Figure 3.7.**

Encouraged by the result that predicted contacts can be used to effectively identify native-like models from a set of alternatives when a native-like model exists, it is natural to wonder if the method would also be applicable when no near-native models are present. In order to evaluate this, the excluded targets where a sub-2A model was not generated by PatchDock were processed using the same procedure.

Applying the procedure to the sets of decoys which do not contain a CAPRI “medium” quality model or better, it is clear that in these cases, considerably lower numbers of contacts are satisfied (Figure 3.10). Testing for significance, these differences in the number of satisfied contacts are significant at the 95% confidence interval, irrespective of method (Mann-Whitney U Test; CCMpred: $p = 1.72 \times 10^{-5}$, PSICOV: $p = 1.29 \times 10^{-4}$, EVfold: $p = 2.37 \times 10^{-5}$, MIp: $p = 0.021$). As such, these results suggest that this observation

can be used to establish the credibility of a model identified by the contact-satisfaction approach. If a large percentage of the provided contacts are selected by the model, at the very least, this indicates that the contacts are approximately collocated, permitting the observed high percentage satisfaction. Ideally, one would hope that low numbers of satisfied contacts would indicate that a model is unlikely to be of good quality. However, there is considerable overlap between the two sets, making a clear distinction between the two groups difficult to establish.

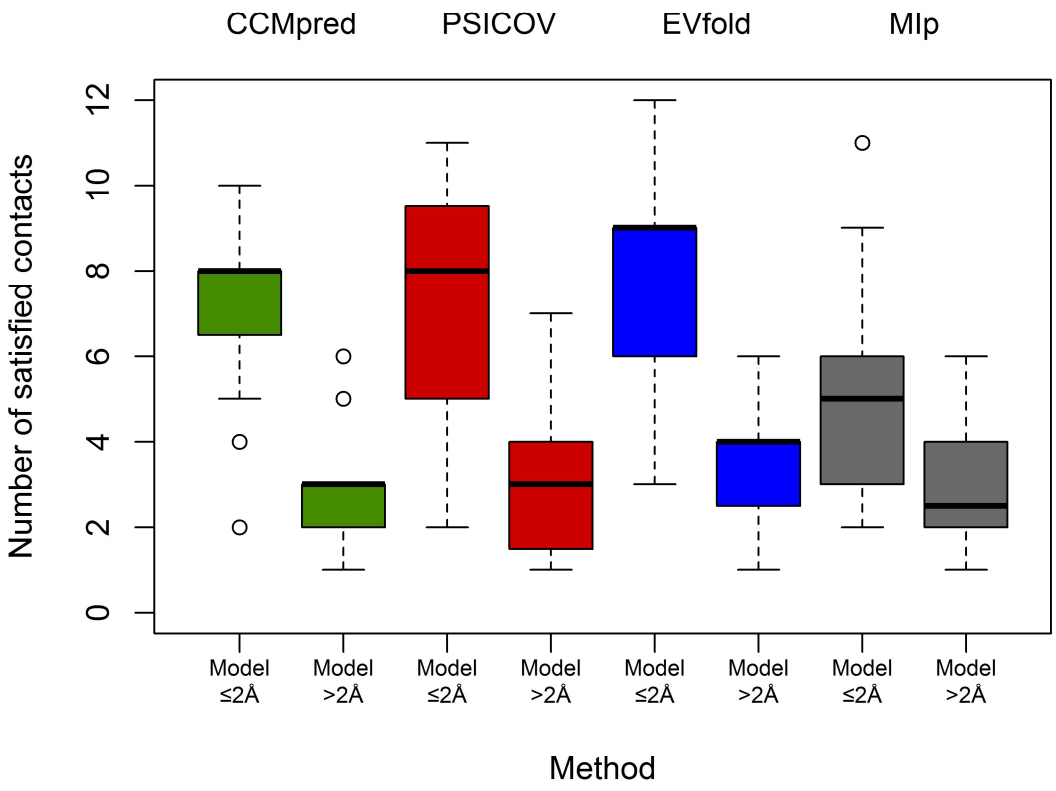



**Figure 3.10: Number of satisfied contacts in the set of decoy structures (best possible iRMSD model ≤2Å) and in the decoy sets where no high quality model is present (best possible iRMSD model >2Å).**The best performing number of contacts was used for each method: CCMpred with 10 contacts, PSICOV with 11 contacts, EVfold with 12 contacts and MIp with 13 contacts.

This reranking procedure was developed as an initial step to demonstrate how sequence-based predictions can be applied for filtering models. The results presented here demonstrate that even this simple approach is effective in identifying good quality models. The added advantage this brings is the simplicity for implementing the current approach by others. Of course, other, more sophisticated approaches could certainly be developed to take better advantage of the predictions for decoy selection. For example, it may be possible to give additional emphasis to high ranking contacts, and reducing the weight placed on lower ranked contacts, rather than treating all contacts as equal.

### 3.3.4 Comparison with the PatchDock scoring function

As mentioned in Section 3.2.1, during the generation of the decoy set, PatchDock ranked each generated model according to its shape complementarity score. In this section we directly compare the performance of our approach with the PatchDock scoring function. Table 3.3 provides a more detailed overview of the decoys selected using CCMpred contacts. CCMpred was chosen for further analysis as it selected the highest quality decoys, as measured by both iRMSD (Figure 3.8) and fNat (Figure 3.9) scores, though admittedly the increase was non-significant when compared to EVfold (Table 3.2). Work was conducted using 10 predicted contacts; the number which generated the set of lowest-iRMSD structures.

**Table 3.3: Summary table of the 24 proteins used for the docking study, selected using the contacts generated by CCMpred.** PDB ID - the Protein Data Bank identifier for the protein analysed. Lowest iRMSD decoy - the lowest iRMSD score observed for any model within the set of 200 alternative decoy structures. Rank 1 model iRMSD (fNat) – The iRMSD (and fNat) scores of the model with the highest PatchDock surface complementarity score. Top 10 precision score (HA5A) – Precision score of the 10 contacts used to rank contacts, using the HA5A contact definition. Number of contacts observed in selected model(s) – The number of contacts observed in the selected decoy group. Rank of PatchDock decoy selected using contacts – The PatchDock rank (based on shape complementarity) of the decoys selected using the contact selection approach. Selected model iRMSD (fNat) - The iRMSD (and fNat) scores of the models selected using predicted contacts.

Results are ordered according to the iRMSD score of the model selected by PatchDock. Targets where multiple models satisfy the same number of contacts are reported in ascending PatchDock rank, and iRMSD and fNat scores are presented in this respective order.

| | | PatchDock selection | Models selected using CCMpred contacts | | | |
|---|---|---|---|---|---|---|
| PDB ID | Lowest iRMSD decoy | Rank 1 model iRMSD (fNat) | Top 10 precision score (HA5A) | Number of contacts satisfied in selected model(s) | Rank of PatchDock decoy selected using contacts | Selected model iRMSD (fNat) |
| 2B6CB | 1.034 | 15.036 (0) | 0.7 | 8 | 18, 39 | 2.341 (55), 1.041 (92) |
| 1WJ9A | 0.841 | 12.346 (0) | 0.1 | 2 | 4 | 0.841 (85) |
| 1VHNA | 0.355 | 12.289 (0) | 0.5 | 5 | 175 | 0.355 (96) |
| 3HP7A | 0.965 | 12.259 (4) | 0.8 | 8 | 79 | 1.522 (89) |
| 1PUJA | 0.939 | 11.522 (0) | 0.4 | 4 | 13 | 0.939 (93) |
| 2RA9A | 0.822 | 10.806 (0) | 0.5 | 5 | 62 | 0.822 (89) |
| 1JDBF | 0.918 | 10.547 (5) | 0.7 | 8 | 2, 113 | 0.918 (93), 2.351 (65) |
| 1EE8A | 0.658 | 10.307 (3) | 0.8 | 9 | 3 | 0.658 (85) |
| 3A4TA | 0.533 | 5.162 (12) | 0.8 | 8 | 2 | 0.533 (95) |
| 1V0BA | 0.928 | 0.928 (86) | 0.7 | 8 | 1 | 0.928 (86) |
| 1WF3A | 0.890 | 0.890 (88) | 0.8 | 7 | 1 | 0.890 (88) |
| 2HIYC | 0.804 | 0.823 (83) | 0.7 | 7 | 1, 17 | 0.823 (83), 1.536 (54) |
| 1EH6A | 0.785 | 0.785 (94) | 0.8 | 8 | 45 | 1.352 (76) |
| 3VO8B | 0.652 | 0.652 (97) | 0.7 | 7 | 1 | 0.652 (97) |
| 3NZKB | 0.639 | 0.639 (87) | 0.5 | 6 | 2 | 1.539 (67) |
| 1LI5A | 0.609 | 0.609 (95) | 0.9 | 9 | 1 | 0.609 (95) |
| 2CGJA | 0.540 | 0.540 (95) | 0.8 | 8 | 1 2 | 0.540 (95), 1.301 (77) |
| 1VMAA | 0.535 | 0.535 (94) | 0.8 | 8 | 1, 36 | 0.535 (94), 1.703 (62) |
| 1MGPA | 0.498 | 0.498 (94) | 0.9 | 10 | 3 | 1.426 (78) |
| 3CI0J | 0.479 | 0.479 (100) | 0.9 | 9 | 1 | 0.479 (100) |
| 1T6CA | 0.455 | 0.455 (95) | 0.4 | 4 | 1, 3 | 0.455 (95), 1.647 (72) |
| 2QFLA | 0.388 | 0.388 (94) | 1 | 9 | 1 | 0.388 (94) |
| 1OI7A | 0.383 | 0.383 (97) | 0.7 | 7 | 65 | 2.108 (58) |
| 2WHYA | 0.331 | 0.331 (100) | 0.8 | 8 | 1, 57 | 0.331 (100), 1.576 (65) |

From Table 3.3, it can be seen that PatchDock was able to identify CAPRI “high quality” models for 15 of the 24 cases (63%). However, it must be noted that this was after the removal of 13 cases where PatchDock was unable to identify a sub-2Å iRMSD model amongst the 200 generated decoys. Of the 15 high quality cases, PatchDock was able to identify the optimal model (that is, the model with the lowest iRMSD within the set of 200 decoys) for 14 targets. In the one case where the optimal was not found (2HIYC), the optimal model had an iRMSD score slightly lower than the model identified by the PatchDock scoring function, though both would be categorised as high quality. For the remaining 9 targets, the decoy with the highest surface complementarity score achieves a high iRMSD score, indicating that the decoy interface is not located near to that seen in the experimental structure (iRMSD range = 5.162-15.036Å, all CAPRI “incorrect” results).

Looking at our contact-based approach, we were able to identify the optimal model in 17 of the 24 cases (71%), though this does include targets where 2 decoys could not be distinguished between. Unlike the PatchDock scoring function, even in cases where the optimal model was not identified in the top result, near-optimal results were achieved. All selected models fall within the CAPRI “correct” criteria, with the selected model with the highest iRMSD value (2.351Å) still falling within the CAPRI “acceptable” category.

Using the bootstrap procedure described in Section 3.2.6, the contact-based approach was observed to select significantly lower iRMSD decoys than the PatchDock scoring function ($p = 3.2 \times 10^{-4}$).

The following sections focus on some interesting examples which arose during the decoy selection procedure.

### 3.3.4.1 PatchDock's scoring function and our method are in agreement

In 10 of the 24 examples, PatchDock and the contact-based approach both identified the optimal model from the decoy set. Figure 3.11 shows one such example.

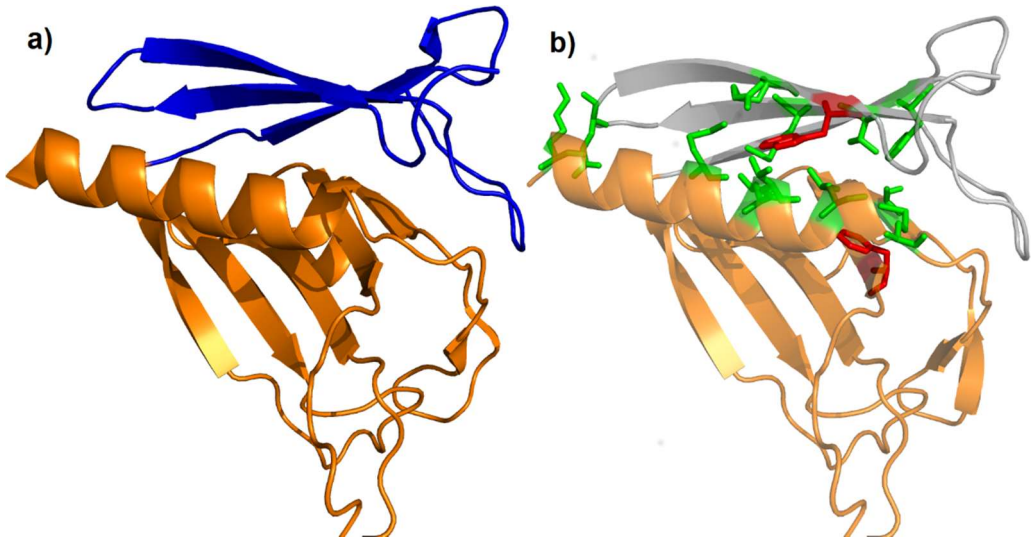



**Figure 3.11: Example where PatchDock's shape complementarity function and the contact-based approach are in agreement for target 3CI0J.** a) The experimental structure and b) the PatchDock selected structure based on shape complementarity and CCMpred contact-selected structure where 9/10 predicted contacts are observed. Contacts observed within the HA5A threshold are shown in green, and the residues forming the single contact outside this distance, shown in red. Model iRMSD = 0.479Å, fNat = 100. Structures are aligned over domain 2 (shown in orange) and images generated using the same orientation.

Figure 3.11 shows that the selected decoy closely resembles the experimental structure. From Figure 3.11b we can see why the contact based approach was successful in this instance. The model ranked first by the PatchDock scoring function is the only model which is observed to form 9 of the 10 predicted contacts.

### 3.3.4.2 Our method outperforms PatchDock's scoring function

More interestingly, the devised approach is capable of identifying near-native models that were not identified by the PatchDock scoring function. One such example is shown in Figure 3.12.

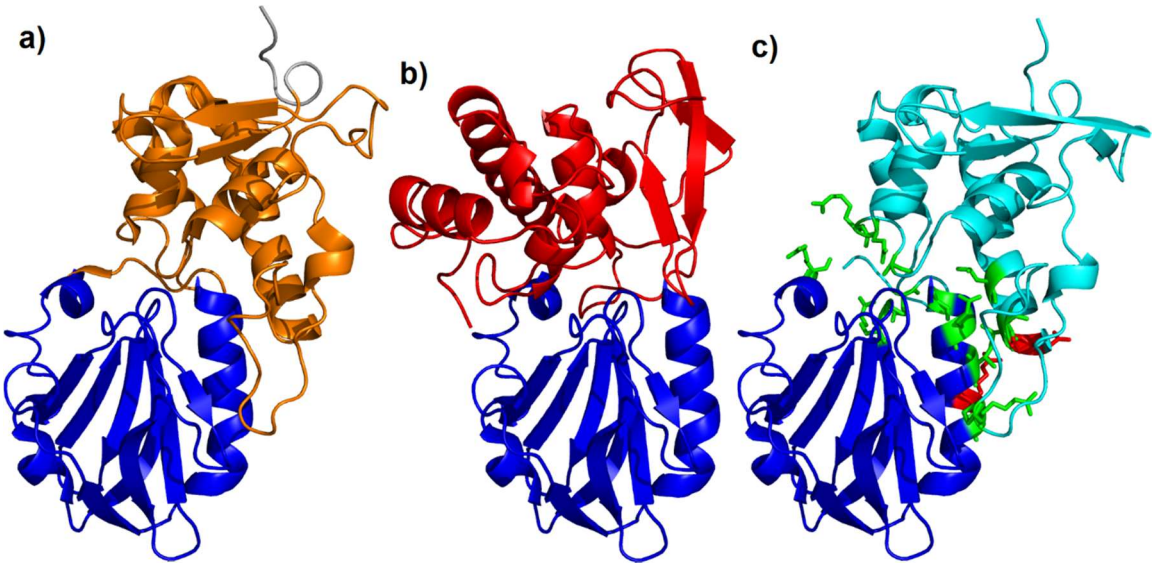



**Figure 3.12: Example where the contact-based approach outperforms PatchDock's shape complementarity function for target 1EE8A.** a) The experimental structure, b) the PatchDock selected structure based on shape complementarity and c) the CCMpred contact-selected structure where 9/10 predicted contacts are observed. Contacts observed within the HA5A threshold are shown in green, and the residues forming the single contact above this threshold, shown in red. PatchDock selected model: iRMSD = 10.307Å, fNat = 3. Contact-selected model: iRMSD = 0.658 Å, fNat = 85. Structures are aligned over domain 1 (shown in blue) and images generated using the same orientation.

The solution selected by PatchDock's shape complementarity function positions the domain in the wrong location and in addition to this, the orientation of the domain is incorrect. This can be seen by observing the location of the beta-hairpin located at the top of the experimental structure (Figure 3.12a), and contrasting with the location in the

PatchDock solution (Figure 3.12b). By considering the covarying residues between domains, we can identify the true domain interfaces, and subsequently select the model which best satisfies our predictions.

### 3.3.4.3 PatchDock's scoring approach outperforms our method

Not all targets were identified perfectly using the contact-based approach. For some examples, predicted contacts can guide the selection procedure away from the optimal solution.

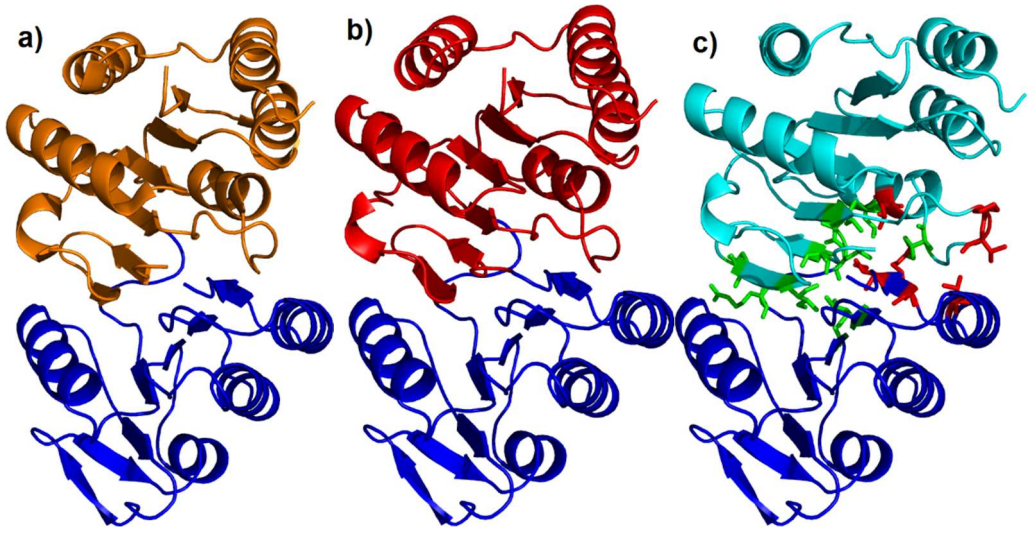



**Figure 3.13: Example where the PatchDock shape complementarity function outperforms the contact-based approach for target 1OI7A.** a) The experimental structure, b) the PatchDock selected structure based on shape complementarity which selected the optimal solution and c) the CCMpred contact-selected structure where 7/10 predicted contacts are observed. Contacts observed within the HA5A threshold are shown in green, and the residues forming the contacts above this threshold, shown in red. PatchDock selected model iRMSD = 0.383Å, fNat = 97, contact-selected model iRMSD = 2.108Å, fNat = 58. Structures are aligned over domain 1 (shown in blue) and images generated using the same orientation.

However, in all such cases (targets 1OI7A, 1MGPA and 3NZKB), the model selected using contacts is still of good quality, even if sub-optimal. For the example of 1OI7A (Figure 3.13), the sub-optimal model was selected as it formed 7 of the 10 predicted contacts. This decoy has a slight anticlockwise rotation when compared to the experimental structure, resulting in a model iRMSD of 2.108Å. This can be seen most easily by looking at the alpha helices located at the top of Figure 3.13a and contrasting them to Figure 3.13c. In this example, the optimal model was observed to satisfy 6 of the 10 CCMpred predictions. While a sub-optimal model was selected using 10 contacts, the optimal model can be retrieved by considering additional contacts. Using 12-20 contacts, the devised approach is able to identify the optimal model along with the model shown in Figure 3.13c. Using 21-40 contacts, the optimal model is identified exclusively. Similar results are achieved for the target 1MGPA with 15 or more predictions and target 3NZKB using 48 or more contacts.

In these cases where sub-optimal models are selected by the procedure, the selected models remain similar to the optimal due to the large overlap of contacts shared between the optimal and selected structures.

### 3.3.5 Comparison with a naïve approach based on domain termini

The devised approach only uses predicted contacts to identify near-native decoys and is able to identify lower iRMSD models than PatchDock's scoring function, as shown in the previous section. This section compares the performance of the devised approach with a naïve approach, which identifies the model with the smallest distance between the terminal carboxyl carbon of domain 1, and the terminal amino nitrogen of domain 2. No interdomain linkers were present within the set of proteins analysed.

Firstly, all carbon-nitrogen distances between the two terminal domain residues were measured for the 4800 decoys. However, only 9 models had a distance within the typical

peptide bond length of 1.32Å (Martin, 2001). Instead, the smallest distance between the two atoms was calculated, and the decoy with the smallest C-N distance was taken as the result of the naïve approach. These results are shown in Table 3.4.

Table 3.4 shows that simply using information regarding the location of the domain boundary, near-native models can be identified. Using this approach, the optimal model was selected for 14 of the 24 targets (58%), highlighting the selective power the domain boundary provides. In a realistic scenario of how model selection would be performed, it would be foolish to ignore available information regarding the location of the domain termini.

Using the same bootstrap procedure described in Section 3.2.6, the available evidence suggests that the contact-based approach does not identify significantly lower iRMSD models than the naïve approach ($p$ = 0.377). Therefore, using the distance between domain termini would be an obvious starting point for future development of the procedure.

**Table 3.4: Results of the decoy selection procedure, ranking models according to their observed terminal carbon-nitrogen distance.** Results are sorted by C-N distance.

| | | C-N distance approach | | | Contact-based approach |
|---|---|---|---|---|---|
| PDB | Lowest iRMSD decoy | Smallest C-N distance | Rank of PatchDock decoy selected | Model iRMSD (fNat) | Selected model iRMSD (fNat) |
| 1T6CA | 0.455 | 0.429 | 12 | 3.457 (42) | 0.455 (95), 1.647 (72) |
| 1JDBF | 0.918 | 0.877 | 2 | 0.918 (93) | 0.918 (93), 2.351 (65) |
| 3A4TA | 0.533 | 0.906 | 106 | 1.357 (67) | 0.533 (95) |
| 1VMAA | 0.535 | 0.908 | 1 | 0.535 (94) | 0.535 (94), 1.703 (62) |
| 3HP7A | 0.965 | 0.980 | 79 | 1.522 (89) | 1.522 (89) |
| 1MGPA | 0.498 | 0.986 | 1 | 0.498 (94) | 1.426 (78) |
| 1VHNA | 0.355 | 1.048 | 175 | 0.355 (96) | 0.355 (96) |
| 2CGJA | 0.540 | 1.080 | 2 | 1.301 (77) | 0.540 (95), 1.301 (77) |
| 1EE8A | 0.658 | 1.259 | 29 | 3.127 (33) | 0.658 (85) |
| 3NZKB | 0.639 | 1.378 | 3 | 1.009 (78) | 1.539 (67) |
| 1EH6A | 0.785 | 1.428 | 45 | 1.352 (76) | 1.352 (76) |
| 2B6CB | 1.034 | 1.469 | 39 | 1.041 (92) | 2.341 (55), 1.041 (92) |
| 3CI0J | 0.479 | 1.524 | 1 | 0.479 (100) | 0.479 (100) |
| 2WHYA | 0.331 | 1.562 | 1 | 0.331 (100) | 0.331 (100), 1.576 (65) |
| 2HIYC | 0.804 | 1.603 | 43 | 1.507 (55) | 0.823 (83), 1.536 (54) |
| 1WJ9A | 0.841 | 1.796 | 4 | 0.841 (85) | 0.841 (85) |
| 2QFLA | 0.388 | 1.821 | 1 | 0.388 (94) | 0.388 (94) |
| 1OI7A | 0.383 | 1.877 | 1 | 0.383 (97) | 2.108 (58) |
| 2RA9A | 0.822 | 2.208 | 117 | 7.180 (9) | 0.822 (89) |
| 3VO8B | 0.652 | 2.292 | 1 | 0.652 (97) | 0.652 (97) |
| 1LI5A | 0.609 | 2.416 | 1 | 0.609 (95) | 0.609 (95) |
| 1V0BA | 0.928 | 2.686 | 1 | 0.928 (86) | 0.928 (86) |
| 1PUJA | 0.939 | 2.699 | 13 | 0.939 (93) | 0.939 (93) |
| 1WF3A | 0.890 | 3.541 | 1 | 0.890 (88) | 0.890 (88) |

## 3.4 Conclusions

This chapter has described an investigation into whether interdomain contacts are sufficient to identify near-native docking models from a diverse set of alternatives. Using the predicted contacts from Chapter 2, here a simple method is described which is capable of identifying models with native-like interfaces, and by extension, resemble the experimental structure as a whole.

This approach was capable of identifying optimal models in 71% of the tested cases, with near-optimal models being selected for the remaining targets. Even in the presence of false-positive predictions, the simple approach is capable of identifying low iRMSD decoys. However, in some instances, the current method identifies multiple models which tie for the most predicted contacts observed. In these cases, it may be possible to differentiate between ties by considering the distances formed by the observed contacts, including rankings using other contact definitions or using inter-termini distances.

This chapter has also shown that the contact-based approach outperforms the PatchDock scoring function, despite not using any structural information. This is all the more remarkable as PatchDock has the capacity to use perfect shape complementarity information for the two cleaved domains, including correct bound-state side-chain orientations.

Whilst it is interesting that purely sequence-derived information is sufficient to identify near-native decoys, in a realistic scenario, available structural information should not be ignored. Explicitly filtering predicted contacts in the knowledge of the domain structure may be of benefit. Incorporating other sources of structural information in addition to predicted contacts will almost certainly benefit future developments of related decoy selection approaches. Other features which are likely to improve selective performance include (but

are not limited to): surface hydrophobicity, domain termini positions, conserved surface residues, electrostatics and predictions of “hot spot” residues.

Whilst this work has focused on the use of covarying residues to identify near-native domain docking models, the general approach should be applicable for discriminating between protein-protein docking decoys. Approaches of this ilk may be particularly useful in cases where proteins undergo large structural rearrangements upon binding, which remains a particularly challenging docking problem (Huang, 2015). Previous studies have shown that covarying residue pairs can be identified which relate to multiple structural conformations (Morcos et al., 2013; Jana et al., 2014). If a protein undergoes large conformational changes upon binding, it would not be unreasonable to assume that this alternate state will also have covarying residues relating to it. If covarying residues can be identified which appear to be incompatible in the unbound structures, these unsatisfied pairings may be able to shed light upon conformational changes in these difficult binding cases.

# 4. Applying predicted contacts as restraints for domain assembly

## 4.1 Introduction

The previous chapter explored the use of predicted contacts for identifying near-native structures from a set of diverse docked models. A large set of docking models was generated, and then subsequently filtered *post hoc* in order to identify models assumed to represent the native structure. However, it should be possible to bypass the generation of models unlikely to represent the native structure using the predicted contacts from the outset. Knowledge of the domain interface, as indicated by the predicted contacts, considerably reduces the search space, allowing modelling efforts to be focused on this region. In addition, the previous chapter also demonstrated the importance of considering the domain termini in modelling, and how knowledge of the domain termini restricts how domains can interact. With these points in mind, this chapter investigates how to incorporate predicted contacts prior to modelling, whilst also taking advantage of the conformational restraint imposed by the domain termini.

As an alternative to docking, the structures of isolated domains can be used as templates for comparative modelling in order to generate models of multidomain structures. Modelling in this manner constrains the process around the location of the linker and domain termini. This has been termed “domain assembly”, in contrast to docking studies where each molecule is unconstrained and free to bind anywhere on the surface (Xu et al., 2014; Wollacott et al., 2007).

A program routinely used to model proteins starting from template structures is Modeller (Šali and Blundell, 1993). During the modelling process, additional restraints can be

included from other sources, such as NMR spectroscopy and chemical cross-linking experiments, amongst others (Webb and Sali, 2014). This chapter describes a proof-of-principle investigation into the efficacy of applying predicted contacts as interdomain restraints to guide the assembly of two domain proteins. In this work, multidomain models were built from isolated structures, simulating how domains may be combined if a template covering the entire target protein is unavailable. As CCMpred has been shown to identify interdomain contacts with the highest precision (as demonstrated in Chapter 2), work will be focused on contacts generated from this method for this study.

## 4.2 Method

### 4.2.1 The Modeller program

The Modeller approach is based on the satisfaction of spatial restraints. Given sufficient structural restraints, models recapitulating a protein's native structure can be generated. In Modeller, these restraints are described in the form of probability density functions (PDFs), with each PDF representing the likelihood of a distance being observed. Using these PDFs, the most likely structure of the target sequence given one or more templates can be calculated and subsequently modelled.

Modelling proceeds in 3 steps. Firstly, the query sequence is aligned to a protein structure which acts as the template for modelling. Secondly, from this alignment, structural information including the topology, backbone dihedral angles and inter-residue distances are extracted from aligned regions of the template, and each of these structure-derived restraints is then expressed as a PDF. Supplementing this, stereochemical restraints are obtained using the query sequence, such as bond lengths, bond angles and van der Waals contact distances, taken from the CHARMM 22 force field (Eswar et al., 2006; MacKerell et al., 1998). In the third step, a solution is sought which attempts to satisfy all

restraints in a final “molecular PDF”, by combining restraints into an objective function which is optimised using conjugate gradients. The resultant molecular PDF represents the most probable structure of the query sequence given the alignment with the template and from this, a 3D model of the query sequence can be generated. Starting from an extended chain structure, increasing numbers of restraints are built into the model, starting with those local in structure, followed by those increasingly more distant. If specified, model refinement can be performed using molecular dynamics with simulated annealing (MD/SA), which simulates the movements of coarse-grained approximations of the modelled residues, with the goal of improving the quality of models (Durrant and McCammon, 2011).

As with docking, after a series of models have been generated, the estimated quality of each model must be established to identify models thought to be most like the native. The Modeller program provides two principal scores for determining the quality of the model. The first is the objective function which is maximised during the creation of the molecular PDF, which measures how well the model satisfies the generated PDFs (Webb and Sali, 2014). An alternative is the Discrete Optimized Protein Energy (DOPE) score, which was specifically designed for model quality assessment and selection, and demonstrated to outperform the objective function for these tasks (Shen and Sali, 2006; Webb and Sali, 2014). The DOPE score is based on a reference state corresponding to non-interacting atoms of a sphere, whose radius is equivalent to the size of the protein being evaluated. DOPE was shown to provide good performance in a variety of selective tasks, including identifying native structures amongst decoys, identifying the most accurate model within a set and importantly, was also shown to correlate with model error. Considering these benefits, DOPE was selected as the score used to select final models.

For this study, version 9.14 of the Modeller program was used.

### 4.2.2 Domain modelling

#### 4.2.2.1 Target sequences

The same set of 37 proteins from Table 2.2 for which predicted contacts were generated in Chapter 2 was used as the starting point for this study. Two targets were excluded where CCMpred failed to predict any interdomain contacts correctly (targets 1H8PA and 2W6PB), leaving a final set of 35 proteins used for analysis.

In order to eliminate complications arising from introducing chain breaks into the modelling procedure, the SEQRES sequence of the proteins from the first residue of domain 1 in the crystal structure was taken in full until the last residue of domain 2. This is inclusive of any residues missing from the crystal structure and those relating to an interdomain linker outside of CATH domain assignments.

#### 4.2.2.2 Domain model generation

The two domains of the experimental crystal structure were cleaved before the first residue of the second domain, again using CATH domain boundaries. The SEQRES sequence of the first domain in addition to a linker, if present, was aligned to the first domain (and linker if present), using the “align2d” method provided by Modeller (Madhusudhan et al., 2006). The SEQRES sequence of the second domain was simply aligned over the second domain structure. The align2d method implements a variable gap penalty, based on the template structure (Madhusudhan et al., 2006). This penalty favours inserting gaps in 3 areas: solvent-exposed regions, regions outside of secondary structures and nearby regions in the template structure. As the sequence and structure being aligned are identical, except for regions of the SEQRES sequence that are missing in the crystallographic experimental structure, any standard sequence-based alignment

approach with reasonable parameters would almost certainly produce an identical alignment to those used here.

After alignment to the crystal structure, models inclusive of any additional residues were produced using the standard “automodel” protocol. This produced an idealised model of each domain from a perfect template, except for residues missing in the experimental structure, which were effectively modelled *de novo*. These models were of high quality, with an average backbone RMSD value of 0.41Å over the set of 70 domain models. As expected from modelling domains in isolation, there are minor differences in side-chain orientation when compared to the experimental structure (all atom RMSD value = 0.98Å).

The 70 generated domain models were the basis for all subsequent modelling steps.

### 4.2.3 Generating whole protein models from individual domains

The full length protein model was generated by concatenating the two domain sequences, with each domain sequence aligned to the corresponding separate domain model, generated in the previous section.

In order to remove the possibility of the modelling procedure benefitting from the starting orientation of the domains, both domains were randomly rotated. To avoid clashes between the initial domain models caused by the random rotation step, after the models are randomly rotated, the two domains were separated. Each pair of randomly rotated domains was stored, so identical starting structures were used for each different application of restraints.

In the first instance, models were generated without specifying any interdomain restraints. 100 models of the full-length protein were produced for each of the 35 proteins specified in Section 4.2.2.1.

The procedure was replicated with additional refinement using Modeller's inbuilt Molecular MD/SA protocol. MD/SA refinement was performed using the "slow" option, performing thorough refinement.

## 4.2.4 Applying restraints

The Modeller program does not permit restraints to be assigned to non-specific atoms of an amino acid (i.e. those used in heavy-atom based contact definitions). Because of this, all restraints were added between the Cβ atom (Cα for Glycine) of each residue, with the upper bound for the interaction set to a distance of 8 ± 0.1Å, in line with the CASP definition of a contact (Monastyrskyy et al., 2014).

### 4.2.4.1 Experimentally observed contacts

To assess a best-case scenario, 10 interdomain contacts observed in the crystal structure were also used to restrain models. 10 contacts were randomly selected from the complete set of experimental structure-observed CB8A contacts. To avoid any effects of different sets of contacts biasing the modelling procedure, a single set of 10 randomly selected contacts were used for all models.

### 4.2.4.2 CCMpred predicted contacts

The same set of CCMpred predicted contacts from the best performing set of alignment parameters identified in Chapter 2 were used.

#### 4.2.4.3 Randomised CCMpred predicted contacts

In order to assess the added benefit of the specific pairings between contacts, the residues forming contacts were randomised. The aim of this procedure was to assess the contribution of the pairings between the two domains, whilst maintaining the same residues being evaluated. Contacts are specific pairings between two residues: $i$ and $j$. To randomise the contacts, the top 10 CCMpred predictions were taken, then split into two lists: $i_1$ to $i_{10}$, and $j_1$ to $j_{10}$. The order of the residues in both lists was then randomised, after which, pairings were generated from the random lists. This list was then compared to the starting list of predictions to ensure that no contact was regenerated during the randomisation procedure. If a contact was regenerated, the randomisation procedure was repeated. In this way, the same individual residues form contacts, but the specific interactions between each residue pair are broken. Of course, by randomising the predicted contacts, new correct contacts can be formed. The precision scores of the randomised contacts are shown in Table 4.1 which shows that the randomisation procedure was effective in breaking the specific links between residue pairs located at the domain interface, reducing the number of correct interdomain pairings. For target 1WJ9A, the list contains 1 correct contact both before and after the procedure.

A single set of randomised contacts was used simply as a starting points for experimentation. Of course, in future work, multiple random trials would present a fairer comparison for how the randomised interface contacts perform.

| | Number of correct contacts (HA5A threshold) | |
|---|---|---|
| **PDB ID** | **Before randomisation** | **After randomisation** |
| 1AF7A | 6 | 0 |
| 1AQTA | 7 | 3 |
| 1BL0A | 7 | 0 |
| 1EE8A | 8 | 1 |
| 1EH6A | 7 | 0 |
| 1GRJA | 7 | 4 |
| 1JDBF | 6 | 0 |
| 1KSLA | 9 | 1 |
| 1LI5A | 4 | 0 |
| 1MGPA | 8 | 0 |
| 1OI7A | 6 | 0 |
| 1PUJA | 3 | 0 |
| 1T6CA | 4 | 0 |
| 1U98A | 3 | 0 |
| 1V0BA | 9 | 1 |
| 1VHNA | 5 | 1 |
| 1VMAA | 10 | 0 |
| 1WF3A | 5 | 0 |
| 1WJ9A | 1 | 1 |
| 2B6CB | 7 | 0 |
| 2CGJA | 8 | 2 |
| 2DYIA | 5 | 3 |
| 2HIYC | 7 | 0 |
| 2QFLA | 9 | 2 |
| 2RA9A | 4 | 1 |
| 2WHYA | 6 | 3 |
| 3A4TA | 8 | 3 |
| 3CI0J | 6 | 2 |
| 3CWVA | 3 | 1 |
| 3FUXC | 8 | 0 |
| 3HP7A | 9 | 1 |
| 3NZKB | 7 | 1 |
| 3QCZA | 8 | 1 |
| 3VO8B | 7 | 0 |
| 3VRDA | 3 | 2 |
| **Mean** | 6.29 | 0.97 |

**Table 4.1: Number of correct contacts when compared to the experimental structure for the 10 CCMpred predictions and 10 randomised contacts.**

### 4.2.5 Model selection with the zDOPE score

Models were selected using the inbuilt DOPE scoring function (Shen and Sali, 2006) as the method has been demonstrated to outperform the alternative scoring functions offered by the Modeller suite (Shen and Sali, 2006). In order to facilitate comparisons between models and establish the efficacy of DOPE for identifying models with low iRMSD, the normalised DOPE score (zDOPE) was used. The normalised DOPE score converts the DOPE score into a Z-score. Both DOPE and zDOPE scores have been shown to be correlated with model error (Shen and Sali, 2006), where negative zDOPE scores generally represent good quality models, with the score increasing as the model error becomes larger. For each target, the model with the lowest zDOPE score from the 100 alternatives was taken as the predicted model.

In order to establish the efficacy of the zDOPE measure for identifying good quality models, the iRMSD values for the 3500 generated models were plotted against the calculated zDOPE scores (Figure 4.1). From Figure 4.1, it can be seen that negative zDOPE scores generally represent models with low iRMSD scores and that the zDOPE score is strongly correlated with model iRMSD value. However, there is also a dense area located in the bottom-left where many low iRMSD structures are assigned similar, negative zDOPE scores.

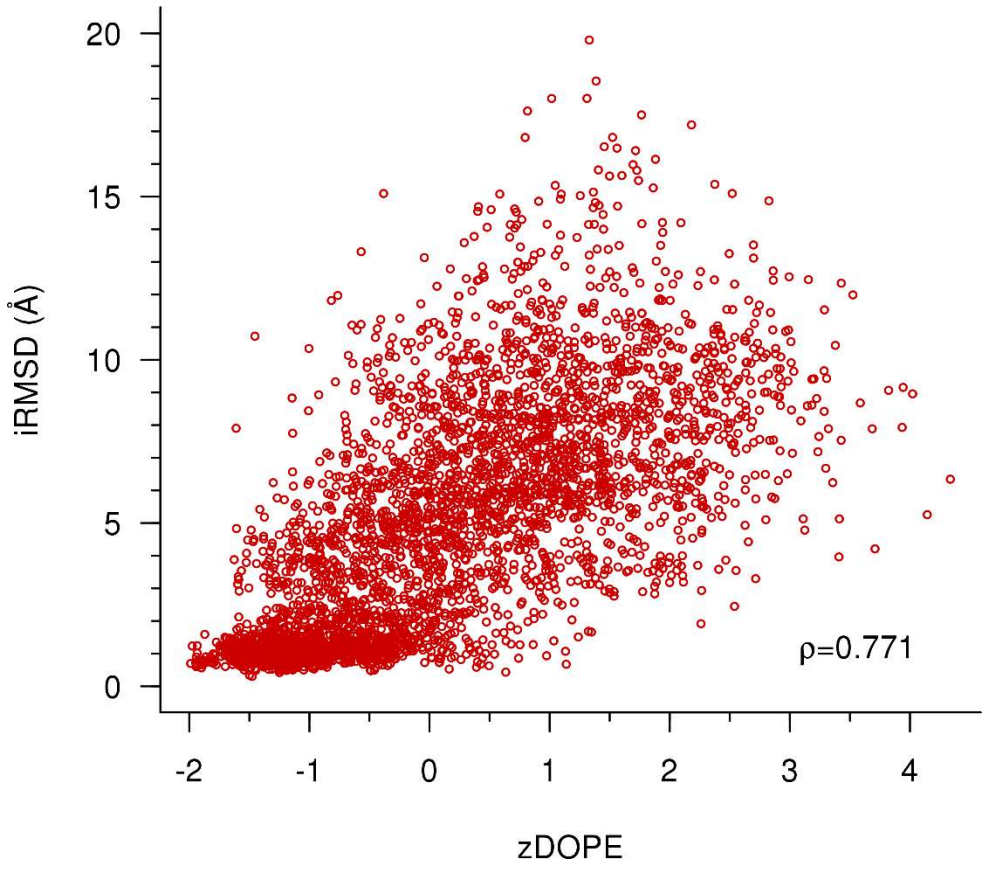



**Figure 4.1: iRMSD plotted against normalised DOPE (zDOPE) values.** Negative zDOPE scores typically relate to models with lower iRMSD values. Spearman's ρ = 0.771. The 3500 models shown are the output from the trial using 10 experimental structure-derived contacts, with MD/SA refinement.

### 4.2.6 Assessment of generated models

Similarly to the assessment of models in Chapter 3, models were assessed in terms of iRMSD and fNat scores (see Section 3.2.3). A new measure introduced in this section is the ΔiRMSD. The ΔiRMSD is the difference between iRMSD values of the model selected by the zDOPE score, and the lowest iRMSD model in the set of 100 alternative structures. If the zDOPE score successfully identified the lowest iRMSD model for a target protein,

this value would be 0. Only residues present in the experimental structure were used for analysis, i.e. any *de novo* modelled regions were not assessed.

## 4.3 Results and discussion

The use of predicted contacts for restraining the modelling of two domain proteins was conducted on a majority subset of the initial 37 proteins identified in Chapter 2. As the benefit of CCMpred contacts is being assessed, two targets were removed where CCMpred failed to identify any interdomain contacts correctly. The results for the remaining 35 proteins are presented below.

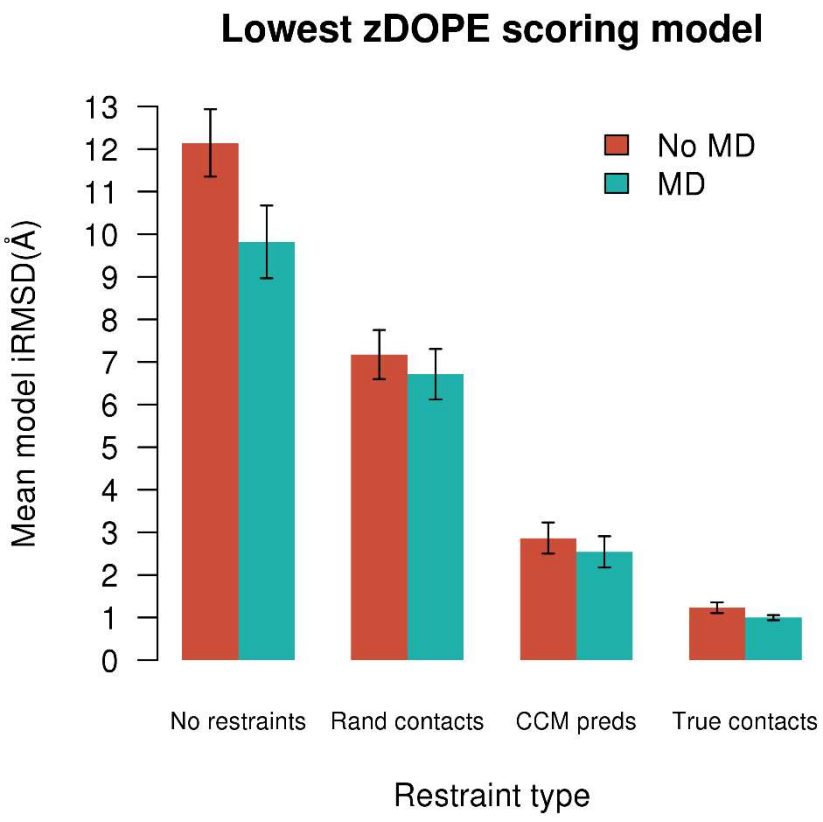



**Figure 4.2: Average model iRMSD for the set of 35 protein targets, where models were selected using the lowest observed zDOPE score.**

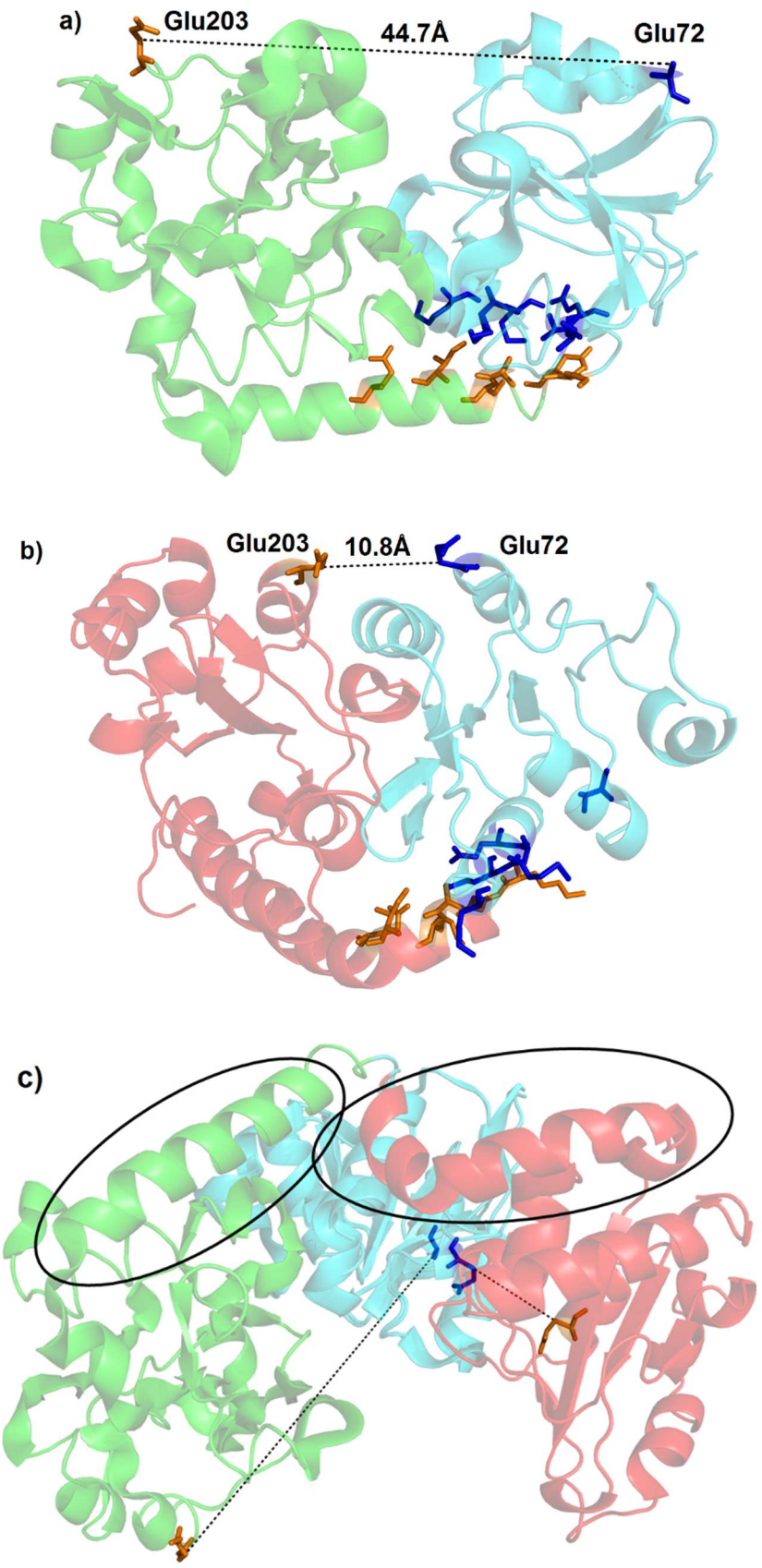
Figure 4.2 shows the effect of adding interdomain restraints prior to modelling. Without restraints, the two domains are merely linked and without cause for Modeller to bring the domains together, they remain distant, resulting in high iRMSD scores.

The randomised contact pairs are sufficient to bring the two domains together, in turn reducing the iRMSD scores in comparison to the trial with no restraints. However, in order to establish the effect of randomising the contacting pairs, one must compare the results with those in the non-randomised trial. By doing so, it is clear that the specific pairings, rather than merely the identification of residues located at the interdomain surface, are responsible for generating the low iRMSD values observed in the trial evaluating CCMpred predictions.

Using predicted contacts, Modeller generates CAPRI “acceptable models” on average across the 35 targets (iRMSD = 2.87Å without MD/SA refinement and iRMSD = 2.54Å with MD/SA refinement). Within the 35 targets, 6 have an iRMSD score ≤ 1Å, representing CAPRI “high” quality models, a further 13/35 with iRMSD scores between 1 and 2Å (CAPRI “medium” quality models) and 11/35 between 2 and 4Å (CAPRI “acceptable” quality models). However, 5 of the 35 models obtained iRMSD scores > 4Å, which would be deemed incorrect in the CAPRI assessment. In these 5 cases, generated models suffered due to Modeller attempting to minimise the distance between pairs of residues which are observed at long distances in the experimental structure. One such example is observed within target 2WHYA (Figure 4.3). The inter-residue Cβ distance between residues 72 and 203 in the experimental structure is 44.7Å (Figure 4.3a). In the provided restraints, the distance restraint between these residues was set with an upper-bound of 8 ± 0.1Å. During the generation of the molecular PDF, Modeller attempts to satisfy this contact, bringing these two residues together at a distance of 10.8Å in the lowest zDOPE model (Figure 4.3b). Modeller achieves this by breaking some of the restraints forming the domain structure (Figure 4.3c). This was not a unique phenomenon observed only within

the lowest zDOPE scoring model, with Modeller attempting to satisfy the contact between residues 72 and 203 in all 100 structures, resulting in similar inter-residue distances, and overall models as those seen in Figures 4.3b and 4.3c (mean distance = 10.6Å; standard deviation = 0.89 Å).

**Figure 4.3: Effect of assigning a long-distance covarying prediction as a CB8A restraint on the modelling of target 2WHYA.** a) The experimental crystal structure with the 10 CCMpred predictions shown as sticks. Residues located within domain 1 are shown in blue, and those in domain 2 shown in orange. The long-distance covarying residue pair of glutamic acids 72 and 203 is connected with a black line between Cβ atoms. b) Modeller generated model showing the location of the same 10 CCMpred predictions. Again, the contact between residues 72 and 203 is shown with a black line. This model achieves an iRMSD score of 11.3Å. c) Overall result of Modeller attempting to satisfy the long-distance contact between residues 72 and 203. The two proteins are aligned over domain 1 shown in cyan, with domain 2 of the experimental structure shown in green, and domain 2 of the model shown in red. The deformed helix in the model (red) and the corresponding experimentally observed structure (green) are circled.


a)
Glu203
44.7Å
Glu72
b)
Glu203
10.8Å
Glu72
c)

Looking at the level of contact satisfaction over the entire set of generated models, it can be seen that other long-distance contacts are also brought into close proximity in a similar manner to that observed above (Figure 4.4). However, the majority (177/220 or 80%) of true positive predictions (residue pairs observed to form CB8A contacts within the experimental structure) were formed by Modeller in the selected structures.

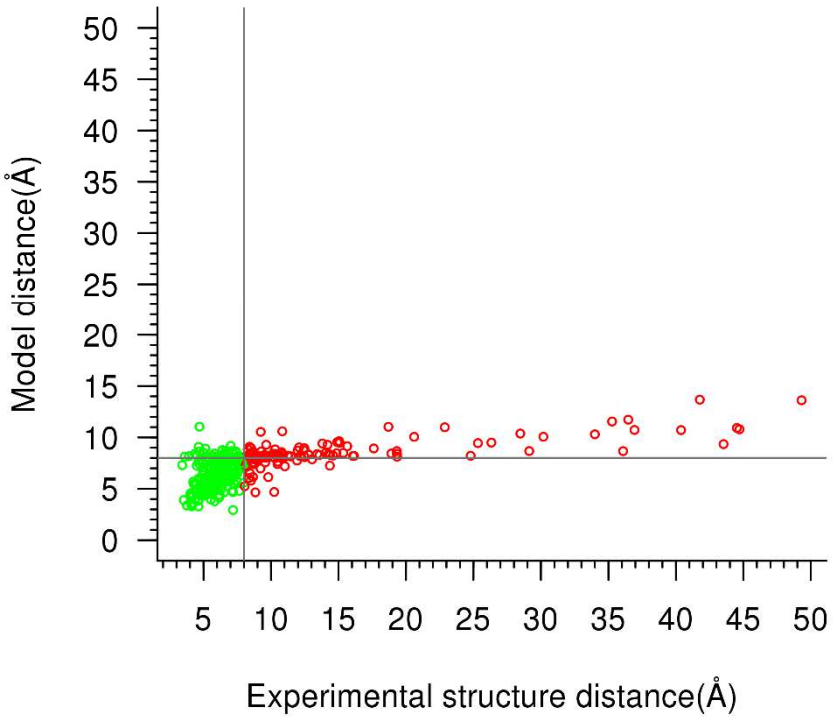



**Figure 4.4: Observed distances of the top-10 CCMpred predicted contacts in the experimental structure plotted against the distances of the residue pair in the Modeller-generated model.** Measurements are calculated between the Cβ atom of both residues (Cα in the case of Glycine). Contacts observed within 8Å in the experimental structure are shown in green, and those above this threshold shown in red. The 8Å threshold is shown in grey for both axes.

The final trial of Figure 4.2 simulated a best-case scenario if contacts could be predicted with perfect precision. Using 10 contacts retrieved from the experimental structure enabled Modeller to generate low iRMSD models (iRMSD =1.23Å without MD/SA and iRMSD = 0.99Å with MD/SA refinement), corresponding to “medium” and “high” quality CAPRI models, respectively.

The corresponding fNat scores for the zDOPE-selected models are shown in Figure 4.5.

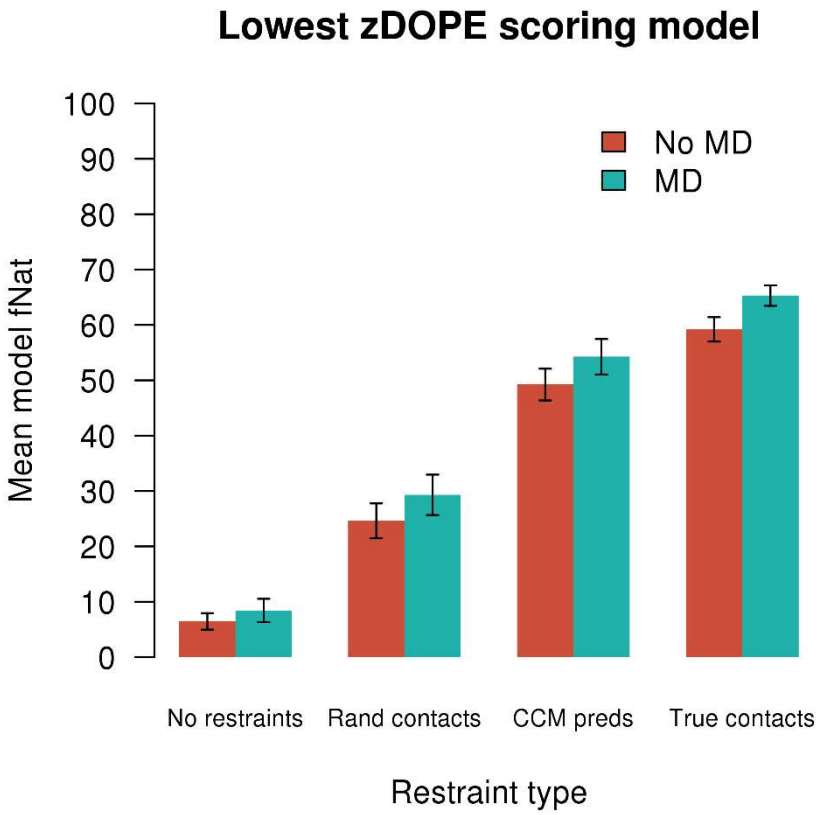



**Figure 4.5: Average model fNat for the set of 35 protein targets, where models were selected using the lowest observed zDOPE score.**

Figure 4.5 shows the average fNat scores achieved under the 8 trialled conditions, reflecting those shown in Figure 4.2. Without restraints the models generally recall a small percentage of the native interface contacts. Using CCMpred contacts, models are on

average able to recall over 50% of the native contacts, which would be equivalent to CAPRI “high” quality models.

However, the zDOPE score does not select the lowest-iRMSD model for every target. In the original paper (Shen and Sali, 2006), the authors note that the method is often less successful at identifying high-quality models when the general quality of the alternative models is low. In order to investigate whether this effect is observed in this study, we implemented the ΔiRMSD measure, and these results are shown in Figure 4.6.

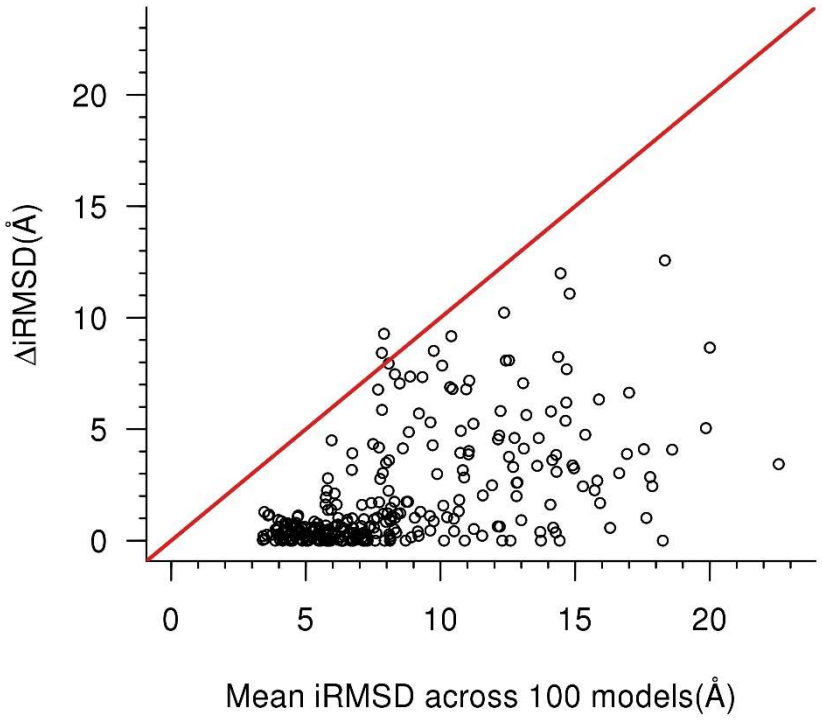



**Figure 4.6: ΔiRMSD values plotted against mean iRMSD values for the 100 alternative models for each of the 35 targets in the 8 trialled conditions.** Spearman’s ρ = 0.559. The red line shows $x = y$ for reference.

Figure 4.6 shows that the zDOPE score is generally more capable of identifying the lowest iRMSD structure when the set of alternative models are also of good quality, in agreement with the findings outlined in the original DOPE paper. When the set of alternative models have high iRMSD scores on average, the ΔiRMSD value, indicating the difference in iRMSD values between the zDOPE selected model and the lowest possible model, is often high. However, whilst the zDOPE score is imperfect, the score generally performs well. In only 2/280 cases (0.7%) is the model selected worse than the average iRMSD for the 100 alternatives. For 31/280 (11%) the ΔiRMSD score is equal to 0, indicating that zDOPE selects the optimal model. In over half of cases (155/280 or 55%), the selected models are within 1Å of the optimal model. Although the zDOPE score is sufficient to identify native-like models, improvements in model quality assessment would improve reported results. Figure 4.7 shows the results of the same modelling procedure if a perfect scoring function could be devised, where the lowest iRMSD model is selected in every instance.

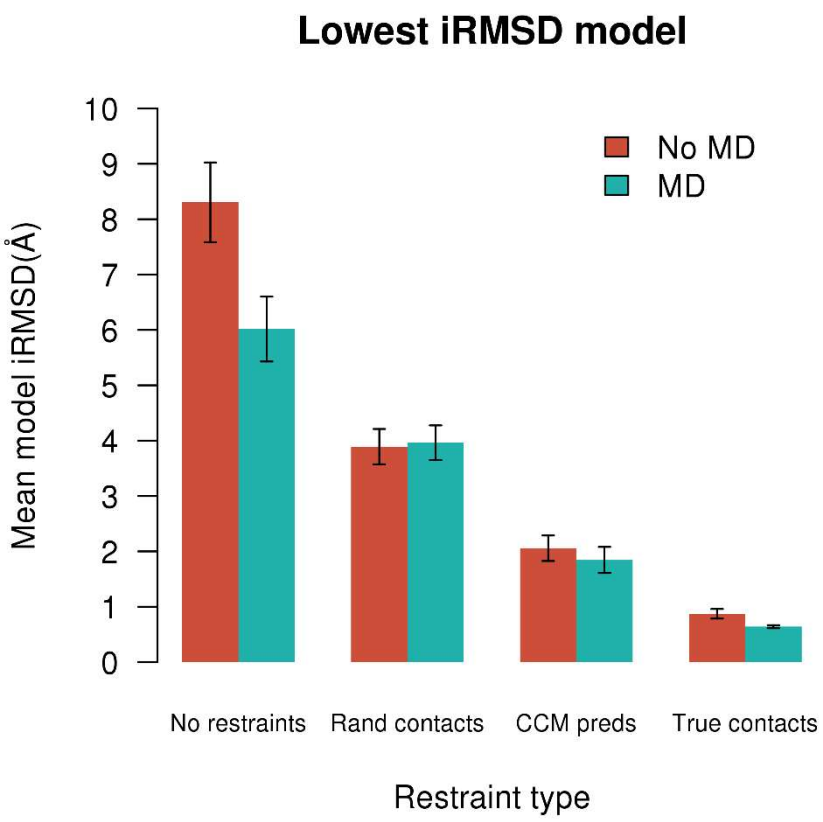



**Figure 4.7: Average model iRMSD for the set of 35 protein targets, where each model scores the lowest iRMSD value within the 100 alternatives.**

Figure 4.7 shows the average iRMSD for each target if the lowest iRMSD model could be selected without fail for every case. These results broadly resemble the results seen when models were selected using the zDOPE score. Interestingly, when considering the randomized contact pairs, the MD/SA procedure systematically reduces the iRMSD scores of the models. This is likely to stem from the MD/SA procedure improving the satisfaction of the provided restraints, which in the case of the randomised contact pairings, brings the model further away from the experimental structure.

Despite issues stemming from the attempted satisfaction of long-distance contacts, as shown in Figure 4.3, CCMpred contacts are able to generate low iRMSD structures, which

are not necessarily selected by the zDOPE score. With experimental contacts, showing an approximation of how well the approach could perform with perfect contacts, MD/SA refinement is able to generate models with an average iRMSD score of just 0.64Å. However, it should be noted that the experimental contacts were selected randomly, and other sets of contacts with a distribution across the domain interface may result in even higher quality models.

### 4.3.1 Areas for method development

Whilst in general the use of CCMpred predictions are able to generate models of reasonable quality (Figures 4.2, 4.5 and 4.7), the issues demonstrated in Figure 4.3 highlight a key problem with the current approach. While covarying pairs generally relate to nearby residues (Figures 2.6 and 4.4), there are instances of covariation identified at much longer distances (Figures 2.15, 2.17 and 4.4). Within the current approach, all covarying residue pairs are specified as small distance restraints. Modeller then attempts to satisfy these restraints, even if this requires altering the structure of the domain derived from the template (Figure 4.3c). There are a number of approaches which could be implemented in order to reduce, or remove, this effect.

Firstly, contacts could be filtered based on their location within a domain, as the domain assembly approach assumes that a structure of each domain is available. Considering only the green domain of Figure 4.3a, the orange residue located in the top left is clearly located in a different location to the other residues located towards the bottom-right of the structure. By identifying the location of the residues which are predicted by the covariation methods, it is reasonable to assume that methods could be developed to filter out residues likely to be involved in longer-range contacts. By doing this, the effect seen in Figure 4.3c is likely to be avoided. A number of different approaches could be employed to perform

this task. For example, it may be possible to use a measure termed the Solvent Accessible Distance (SAD), which is the shortest distance through the solvent between two residues without penetrating the protein surface (Kahraman et al., 2011). By calculating the SAD between residues identified by covariation approaches, residues within the same region could be clustered, and any residues which appear to be distant from the cluster could be removed. It may also be possible to exclude residues which are not solvent-accessible in this manner. Alternatively, instead of removing dubious contacts, the upper bound on the distance restraint could be set to a value scaled on confidence, in order to reduce their emphasis.

Secondly, generated models could be used to identify predicted contacts which may be long-distance. After an initial round of modelling has been conducted, predicted contacts which are not satisfied within the highest scoring models could be removed, and further rounds of modelling could be conducted with the revised set of restraints. Looking at Figure 4.4, 141 of the 350 (40%) predicted contacts are above the 8Å threshold in the selected models. Of these contacts, 98 of the 141 contacts (70%) also above this threshold in the experimental structure (i.e. they relate to false positive predictions), and 43 are within the contact threshold. If these 141 contacts were to be removed, the longest distance between predicted contacts would be reduced to 14.4Å. Whilst this may not entirely solve the issue of including long-distance covarying pairs as restraints, it would be sufficient to eliminate the longest covarying pairs observed, reducing the level of any distortion introduced into the generated models.

Another line of research could be undertaken to investigate the use of variable inter- Cβ distances based on the residues which are identified to be covarying. As mentioned in Section 2.3.6, the CB8A contact threshold relates to an approximation of the average distance between two residue side chains which are in contact. A more refined application of the inter-residue Cβ distance would be to consider the maximal Cβ distance between

two residues which permits a heavy atom contact. One such matrix of distances was presented by the group of David Baker during the CASP11 experiment (http://www.predictioncenter.org/casp11/doc/presentations/FM_Baker_expert.pdf, slide 8). While residue-specific distances may improve performance in cases where contacts are in close structural proximity, this would not solve the issues relating to long-distance covarying pairs.

Whilst the work conducted here focused on the application of 10 predicted contacts to restrain modelling, procedure refinement would be necessary in future work. No alternative numbers of restraints were trialled, and it would be reasonable to assume that smaller number of high precision contacts may benefit modelling, reducing the numbers of erroneous predictions which are introduced *a priori*. In an ideal scenario, large numbers of perfect predictions would lead to accurate models, as seen in Figures 4.2, 4.5 and 4.7. However, in a realistic scenario where perfect predictions are likely impossible, it is probable that an intermediate number of contacts would be optimal. Identifying where this point lies would be a key point for future investigations.

Although the zDOPE score generally performs well at selecting optimal or near-optimal models from the set of generated alternatives (Figure 4.6), further developments in scoring approaches would lead to better models being selected. Community-wide developments in model quality (MQ) assessment, like those in contact prediction, are also monitored in the biennial CASP experiment. Previous assessments of MQ have shown that consensus approaches, incorporating multiple separate estimations of MQ, generally outperform single methods (Kryshtafovych et al., 2015). Whilst the CASP assessment of models is performed at the domain level, the application of such consensus scoring approaches may also be of benefit for model selection for the work presented here.

## 4.4 Conclusions

The Modeller program generates models by satisfying distance restraints between amino acids. The majority of these restraints are generated from template structures, but can be supplemented with additional restraints from a variety of external sources including, but not limited to, experimental cross-linking, nuclear magnetic resonance (NMR) and site-directed mutagenesis experiments (Eswar et al., 2006). This chapter set out to establish whether covariation-based predictions of interdomain contacts were sufficiently accurate to supplement the modelling of proteins in the same manner.

This chapter describes a proof-of-principle study using experimentally elucidated structures as the basis for a domain assembly exercise. The results presented here show that in most cases, interdomain contacts predicted by CCMpred are sufficient to correctly orientate domains, resulting in low iRMSD models of two-domain proteins. These findings suggest that covariation-based contact prediction can act as an alternative approach to using experimentally-generated contact information.

However, modelling is complicated by the existence of covarying residues which are observed at long distance (Figures 4.3a and 4.4). Whilst the majority of contacts do correspond to close structural proximity (as shown in Chapter 2 and Figure 4.4), even a single long-distance contact is sufficient to detrimentally affect the quality of resultant models (Figures 4.3b and 4.3c). It is likely that these contacts can be identified and subsequently removed by considering the proximity of the identified residues to other predictions, and removing such contacts through the implementation of a clustering approach. Alternatively, modelling could be conducted in an iterative fashion, where the capability of Modeller to satisfy predictions is evaluated at each stage. Restraints which

are not satisfied in a model could be removed, under the assumption that they relate to long-distance covariation.

Whilst untested here, the use of Modeller for domain assembly should be applicable for real-world modelling problems, where template domains could be identified for a query sequence through sequence similarity searches to structures in the PDB. In such cases, the steps would be identical to those outlined here, but the domains of the query sequence would be modelled on a non-identical template structure, rather than the crystallographic structure.

# 5. General conclusions and outlook

The link between inter-residue covariation and their interaction has long been known and has spurred numerous studies investigating the reasoning behind - and methods to exploit - this promising link between sequence and structure. The field has recently had a resurgence of interest stemming from the seminal work of Weigt and colleagues (2009) who outlined a practical approach for dealing with coupling effects in a real-world example, building upon the foundational work outlined 10 years earlier (Lapedes et al., 1999). This breakthrough coincided with the explosion of available sequencing data, necessary for use with the proposed techniques. The co-occurrence of method developments and available sequence data helped accelerate progress, attracting a variety of groups stemming from different disciplines to work on this long-standing problem. As a result, in the short time since 2009, the field has progressed considerably. The work of Weigt and colleagues analysed just 60 residues, whereas many hundreds of positions are now routinely considered, including in the work presented here. Once full-length proteins were able to be studied, groups quickly put the novel developments to use, applying predicted contacts for a variety of modelling tasks. However, to date, these methods have not been explicitly applied to study covarying residues spanning domains. Whilst covariation between interdomain contacting residues has been investigated with MI-based approaches (Gomes et al., 2012), other studies have shown that “chaining-aware” methods offer substantially higher contact precision than the best MI-based approaches for intradomain contacts (Kamisetty et al., 2013; Jones et al., 2012). This formed the underlying rationale and motivation for the work conducted in this thesis.

The work presented here has shown that current state-of-the-art approaches are capable of identifying interdomain contacts from the MSAs of two-domain proteins.

In Chapter 2, a diverse dataset of two-domain proteins from the CATH database was constructed in order to develop a protocol for identifying interdomain contacts. From this dataset, two alternative HMM-based MSA programs were used to identify homologous sequences for each target. Parameter space was searched for the set of parameters which provided maximal performance from a number of covariation-based approaches employing different statistical models. In that chapter it was demonstrated that covarying residue pairs are frequently located in close proximity at the domain interface. Interestingly, a number of long-distance covarying residues were observed which are identified by all approaches, despite differences in the underlying approaches.

Based on these findings demonstrating that interdomain contacts can be identified with high precision, the same predictions were applied as an alternative scoring approach to rank docking-related decoy models in Chapter 3. Predicted contacts were used in order to re-rank docking models based on the number of predicted contacts observed. The results of this study showed that predicted contacts were capable of identifying optimal and near-optimal decoy models, and outperformed the PatchDock scoring function. The decoy selection procedure proposed in Chapter 3 is simple to implement, making it suitable for inclusion as an alternative decoy scoring approach by individual groups. Whilst imperfect, using covarying residues to rank models should act as a useful alternative to structure-based scoring functions.

Finally, in Chapter 4, predicted contacts were applied as distance restraints within the Modeller program. By adding restraints between the domains, the average quality of the models was improved, as measured by the iRMSD score. However, the long-distance covarying residues identified within Chapter 2 complicate the assignation of covarying pairs as short-distance restraints. In order to satisfy these long-distance instances of covariation provided as restraints, Modeller may distort the internal structure of the domain itself, resulting in lower quality models. These cases are likely to be improved by filtering

predictions based on their location in the unbound structures and removing these restraints based on observed contact satisfaction in generated models.

## 5.1 Limitations

The analysis of covarying residue pairs across domains, in the manner demonstrated in this thesis, suffers from a number of limitations.

Firstly, sufficiently large MSAs spanning both domains must be available for analysis. Previous work has suggested that at least 125-150 sequences should be analysed in order to reduce the effect of positional entropy on identified covariation (Martin et al., 2005). However, it has been observed both in this work (Figures 2.11 and 2.12) and other studies investigating intrachain covariation (Jones et al., 2012; Jones et al., 2015) that precision scores are correlated with the number and diversity of available sequences. Given this, it is generally suggested that hundreds, or ideally many thousands of sequences are used for analysis (Kosciolek and Jones, 2014; Mao et al., 2015; Marks et al., 2011), though large number of sequences are unlikely to produce high quality predictions if the sequences contain little diversity (Figure 2.12; Jones et al., 2012). Of course, vast numbers of sequences are not necessarily available for a target sequence, and in these cases reliable prediction of contacts is unlikely.

The approaches studied in this work identify covarying residues relating to both intra- and interdomain interactions. In order to isolate contacts relating to interdomain interactions, CATH domain definitions were used to filter these lists. To identify which of these covarying residues relate to interdomain interactions, the domain boundaries must be known, at least approximately. In real-world modelling cases, domain boundaries would likely come from resources such as Gene3D, SUPERFAMILY or Pfam. While the alignment of a target sequence and a domain superfamily may not provide a perfect

indication of the end of the target domain, an approximate indication of the domain boundary would be sufficient to identify the contacts required for interdomain modelling.

One major issue in the use of contacts for protein modelling arises from homooligomeric interactions. It is now well established that covariation between interacting residues occurs both within (Jones et al., 2012; Kosciolek and Jones, 2014; Nugent and Jones, 2012; Marks et al., 2011) and between protein chains (Weigt et al., 2009; Ovchinnikov et al., 2014; Hopf et al., 2014). In cases where a protein functions as a homooligomer, covariation will be observed relating to both intra and intermonomer contacts (dos Santos et al., 2015). If the protein structure is known, it is trivial to establish which covarying residue pairs do not relate to local intrachain contacts, and can be assumed to be involved in interchain interactions (dos Santos et al., 2015). However, if the structure is not known, mixed inter- and intrachain contacts add an additional layer of complication to *de novo* modelling. It may be possible to filter contacts using additional predictions relating to the protein structure, such as predicted solvent accessibility. An example of employing additional predictions to discriminate between intra and interchain contacts was demonstrated in a recent modelling study of homooligomeric transmembrane proteins (Wang and Barth, 2015). Here, the authors additionally considered predictions of lipid exposure in order to differentiate between intra- and intermolecular contacts, but more general approaches for use with globular proteins would be a great advancement.

## 5.2 Future directions for interdomain contact prediction

Whilst the approaches employed in this work are capable of generating good quality models, further work can be performed in order to improve results. This section outlines some potential areas which could lead to improved contact prediction in the future, along with improvements to the way predicted contacts as used for modelling.

### Improved alignment methods

As covariation-based approaches rely solely on a provided MSA in order to identify covariation between columns, the quality of the MSA is crucial. Work within the field, along with the work conducted in this thesis, has been predominantly based around the use of HMM-based approaches to identify and align homologous sequences. Recently a novel approach for sequence homology detection was published, based on Markov Random Fields (Ma et al., 2014). The method, MRFalign, was demonstrated to outperform PSSM and HMM-based approaches in both alignment accuracy and remote homologue detection. These qualities show great promise for the use of this method in contact prediction, though the alignments generated from this method have not yet been evaluated. Improvements to MSA methods are of particular importance as alignment errors have been shown to lead to erroneous observations of covariation (Dickson et al., 2010).

### Improvements to available sequence databases

Covariation-based contact prediction methods have been shown to be correlated with both the total number and effective number of sequences within alignments (Figures 2.11 and 2.12; Jones et al., 2012; Jones et al., 2014). In recent years there has been tremendous growth in the number of available protein sequences deposited in public databases. Whilst this has already benefitted the field of sequence-based contact prediction, additional sequences would likely improve results further, particularly for sequence families which are currently small. There have recently been a number of wider advancements which may permit the inclusion of novel diverse sequences, which are extremely important for this type of analysis. For example, it is exciting that it is now possible to grow soil-dwelling bacterial cultures *in situ*, which may result in a novel areas of previously unexplored sequence-space making their way into public repositories (Ling et al., 2015). Recent

efforts have also been sequencing the genomes of organisms from other exotic niches, such as the Arctic (Abraham and Thomas, 2015) and deep sea (Lauro et al., 2013). These remote and relatively unexplored locations may be host to a number of isolated and importantly divergent organisms, whose proteins have evolved in relative isolation from the last universal common ancestor. Proteins from such organisms may increase the diversity observed in sequence data banks, due to their requirement to function under such extreme conditions.

Whilst bacterial sequences are widely available with the aforementioned avenues to further increase their availability, eukaryotic sequences are less accessible. The current release of the UniProt-TrEMBL database is comprised of 61% bacterial sequences, in comparison to 31% from eukaryotes (http://www.ebi.ac.uk/uniprot/TrEMBLstats). Frustratingly, more eukaryotic proteins are multidomain than bacterial (Apic et al., 2001), and as such eukaryotic proteins would likely be the main beneficiaries from the type of analysis conducted here, but may be limited by the availability of sequence data.

**Improved handling of phylogenetic and entropic biases**

In this work and that performed by other groups investigating correlated mutations, the APC is routinely used (Dunn et al., 2008). However, there are alternative approaches (e.g. Little and Chen, 2009; Gloor et al., 2010) to handle phylogenetic and entropic biases shown to outperform the APC approach in studies using MI, which have been entirely neglected for contact prediction thus far. Trialling these alternatives to handle phylogenetic and entropic biases may improve contact prediction in general and should be considered in future studies.

**Improving contact prediction with machine learning**

Recently, machine learning approaches have been combined with covariation-based contact predictions in order to improve performance (refer to Section 1.5.6.5). Two leading methods, PconsC2 and MetaPSICOV, feed initial contact predictions from different covariation-based methods into another predictor which is trained on a series of protein contact maps in order to learn typical contact patterns relating to secondary structural elements and known protein folds. As general structural patterns are learnt in this way, this enables the methods to eliminate likely false positive predictions as well as adding plausible contacts based on the observed information, such as adjacent residues in beta sheets. Whilst these approaches have proven to be effective in raising the precision scores for intradomain contact prediction, whether they would improve the prediction of contacts between domains is currently unknown.

However, as these approaches were developed for intradomain prediction, they are unlikely to be optimised for the interdomain case. Developing an interdomain-specific approach is likely to improve the precision of predicted contacts, resulting in improved modelling over what has been demonstrated in this thesis. Where these methods may be of particular use is automating the procedure to remove likely false positive predictions. These more sophisticated approaches will almost certainly offer better performance than the simple consensus approach attempted in this work (see Section 2.2.8), which was unsuccessful at improving precision scores over any of the component methods (Figure 2.18).

The aforementioned areas of research are likely to offer improvements to the precision scores of predicted domain-domain contacts. In addition to these steps, research should also be conducted to devise approaches to improve the way these contacts are applied in modelling itself.

### Filtering unlikely contact predictions

Whilst it is interesting that long-distance instances of covariation exist, and may well indicate genuine evolutionary pressures between identified residues, they do further complicate the already difficult problem of protein modelling (as demonstrated in Chapter 4). In the knowledge of the individual structural components being combined, approaches could be developed in order to remove covarying residues which may not be involved at the interdomain interface (such as the proposed approach using SADs in Section 4.3.1, or based on solvent accessibility (Gomes et al., 2012)). Similarly, predicted contacts could be removed if they are impossible to project into three dimensions. This is likely to be implemented along the same lines as the machine learning approaches which learnt typical contact patterns from experimental contact maps, as mentioned in the previous section, or using various heuristic rules (Shao and Bystroff, 2003).

### Inclusion of other interdomain features

This thesis has only considered the application of covarying residue pairs for protein modelling. It may well be possible to incorporate additional features relevant to modelling to further improve performance. For instance, it has been shown that residues at the interdomain interface are more conserved than other regions of the solvent accessible surface (Littler and Hubbard, 2005). Including solvent accessible residues displaying high levels of conservation may help guide docking-type approaches, along with potentially indicating the location of the binding interface, which may help to filter predictions. In a similar manner, it may also be possible to incorporate predictions of “hot spots” – amino acids which contribute significantly to binding (Moreira et al., 2007). Using these residues in conjunction with covarying residues is likely to paint a fuller picture of the binding interface, resulting in improved models.

## 5.3 Wider study of interdomain contact prediction

The work presented in this thesis represents the first specific analysis of interdomain covariation using state-of-the-art approaches to distinguish between direct and indirect covarying residues. Excitingly, interdomain contact assessment has been evaluated (albeit briefly) as part of the last two CASP RR experiments (Monastyrskyy et al., 2014; Monastyrskyy et al., 2015). The CASP11 experiment was the first assessment which saw wide-scale participation of groups using current covariation-based approaches. In the final CASP evaluations, the assessors concluded that interdomain contacts were predicted with much lower precision scores than those within domains (reporting precision scores of 27% for intradomain prediction, and 4% for interdomain respectively, where both assessments are conducted assessing L/5 long range contacts, where L is the length of the protein, as measured in amino acids).

However, two things must be noted about the assessment. Firstly, the prediction of L/5 interdomain contacts is vast, and does not represent a realistic number of contacts required for modelling. In fact, only 60% of the proteins shown in Table 2.2 contain over L/5 interdomain contacts in total within the experimental structure. Assuming that every one of these observed contacts covaries with its partner is unrealistic. Secondly, the main contact assessment in the CASP experiment is conducted at the domain level, so groups are likely to have developed methods prioritising the prediction of intradomain contacts. This is likely to include MSA parameters (which typically involve high numbers of MSA iterations, shown to be suboptimal in Figure 2.4), along with any training of machine learning approaches. However, despite the flaws in the analysis of interdomain predictions in previous CASP experiments, it is encouraging to see the wider community take greater interest in the problem of interdomain contact prediction. Now that interdomain contacts

have been assessed in the previous two experiments, hopefully participating groups will develop interdomain-specific predictors alongside intradomain ones for the 12th CASP experiment.

## 5.4 Final conclusions

Predicting the structure of a protein from its amino acid sequence remains one of the "grand challenges" within bioinformatics (Dill and MacCallum, 2012). Although the analysis of covarying residue pairs by no means solves the problem, it does permit insight into structural and functional dependencies between amino acid pairs within a family-averaged structure of a protein.

Some have said that the field of covariation-based contact prediction is in a catch-22 situation (Kamisetty et al., 2013). As these approaches rely on the availability of hundreds or thousands of sequences in order to generate reliable contact predictions, the methods are inherently limited to abundant protein families. However, it is typically these families where a suitable template structure can be identified, effectively mooting the use of covariation-based methods for modelling when (typically more reliable) template-based modelling can be used in its place (Kamisetty et al., 2013). Whilst this may be true to some extent for intradomain modelling, covariation-based approaches have also been demonstrated to have great promise for more diverse challenges where experimental data are more limited, such as in the modelling of multidomain proteins (as demonstrated in this thesis), guiding the assembly of proteins into bound complexes (Hopf et al., 2014), or modelling transmembrane proteins (Hayat et al., 2015; Hopf et al., 2012; Nugent and Jones, 2012).

While long-distance covarying contacts complicate blinded inclusion of covarying residues for modelling (e.g. Figure 4.3), the underlying reason why such pairs covary is interesting.

Future work in collaboration with experimental groups would be of considerable mutual benefit. Experimental work to mutate one, or a pair of covarying residues and assessing the effect on structure and function is likely to paint a clearer picture of the roles of such long-distance pairings, if they are in fact genuine. Whilst the work in this thesis has been focused on using covarying residues to guide modelling, this type of analysis is also likely to be of use to experimental groups. If a protein structure is known, or can be reliably modelled, covariation-based analyses may be able to shed light on long-range effects which cannot be explained through local structural interactions, potentially revealing unobserved aspects of folding or function and act as a basis to guide exploratory research.

# 6. Appendices

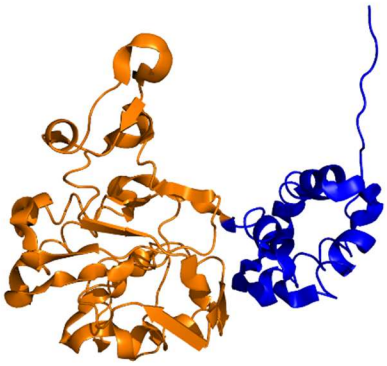
## 6.1 Table of abbreviations

| | |
|---|---|
| 3D | 3-Dimensional |
| Å | Ångström, equal to 0.1 nm or $10^{-10}$m |
| APC | Average Product Correction |
| BLAST | Basic Local Alignment Search Tool |
| CAPRI | Critical Assessment of PRediction of Interactions |
| CASP | Critical Assessment of protein Structure Prediction |
| CB8A | Contact defined as two residues with an inter-Cβ distance < 8Å |
| DNA | Deoxyribonucleic acid |
| DOPE | Discrete Optimized Protein Energy |
| fNat | Fraction of Native contacts |
| FP | False Positive |
| HA5A | Contact defined as two residues with an inter-heavy atom distance < 5Å |
| HA6A | Contact defined as two residues with an inter-heavy atom distance < 6Å |
| HMM | hidden Markov model |
| iRMSD | interface Root-Mean-Square Deviation |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| MD/SA | Molecular Dynamics with Simulated Annealing |
| MI | Mutual Information |
| MIp | Mutual Information with APC |
| MQ | Model Quality |
| MSA | Multiple Sequence Alignment |
| PDB | Protein Data Bank |
| PDF | Probability Density Function |
| PISA | Proteins, Interfaces, Structures and Assemblies |
| PSSM | Position-Specific Scoring Matrix |
| RMSD | Root-Mean-Square Deviation |
| SAD | Solvent Accessible Distance |
| SGI | Structural Genomics Initiative |
| TP | True Positive |
| vdW | van der Waals |
| zDOPE | Normalised DOPE |

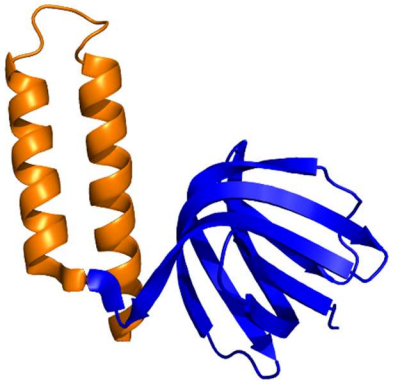
Table 6.1: Table of abbreviations.

## 6.2 Gallery of dataset experimental structures

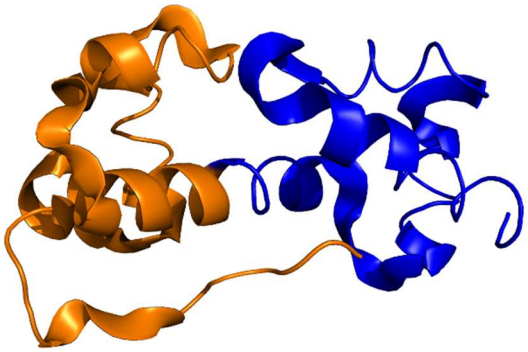
A gallery of the 37 proteins studied in this thesis with the N-terminal domain coloured blue, the C-terminal domain coloured orange and regions outside CATH-defined domains coloured grey, if present. The PDB code for each structure is given underneath.



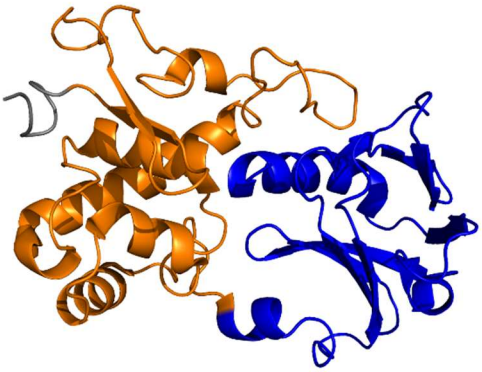
1AF7A



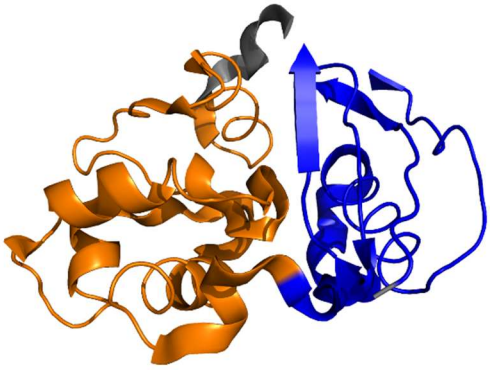
1AQTA



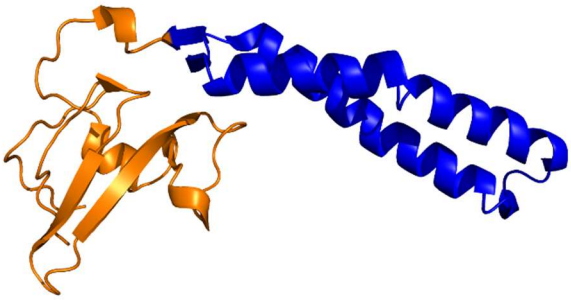
1BL0A



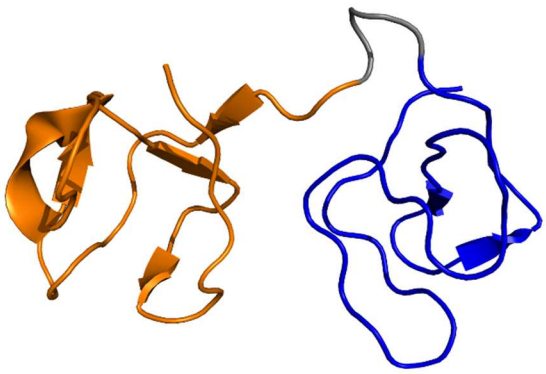
1EE8A



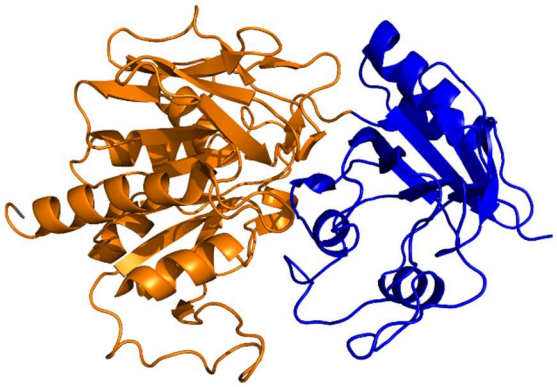
1EH6A



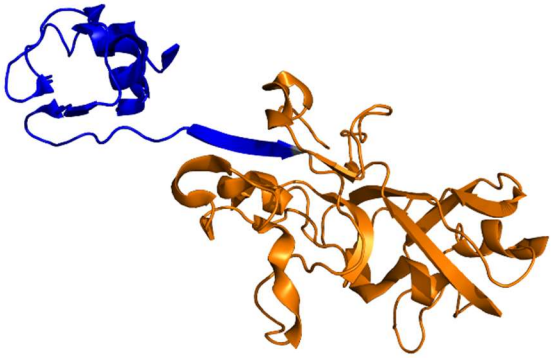
1GRJA



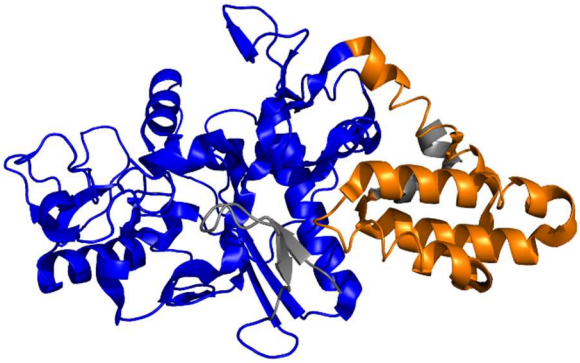
1H8PA



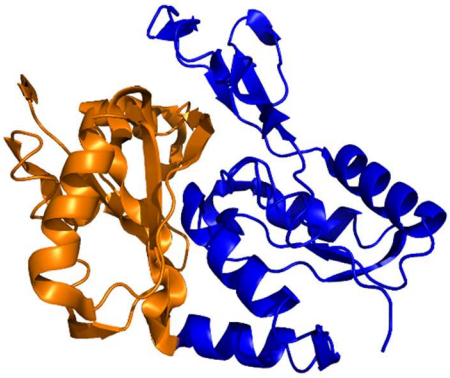
1JDBF



1KSLA

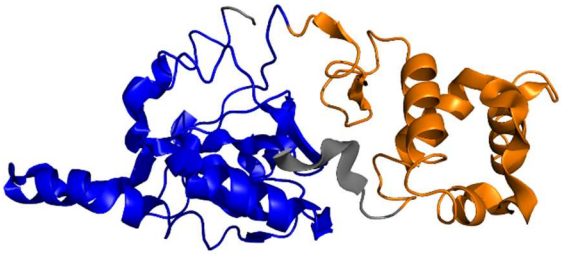


1LI5A

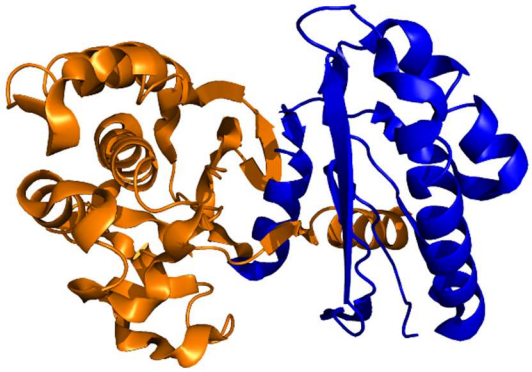


1MGPA

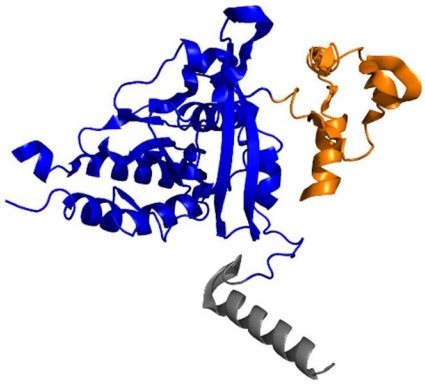
1OI7A



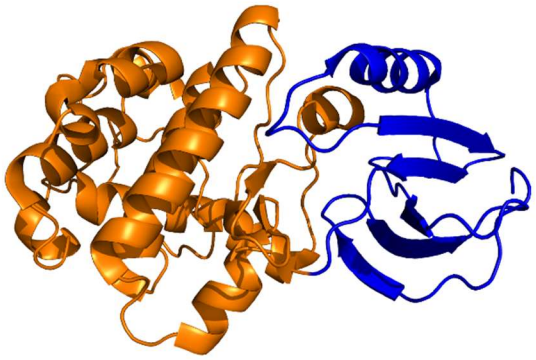
1PUJA



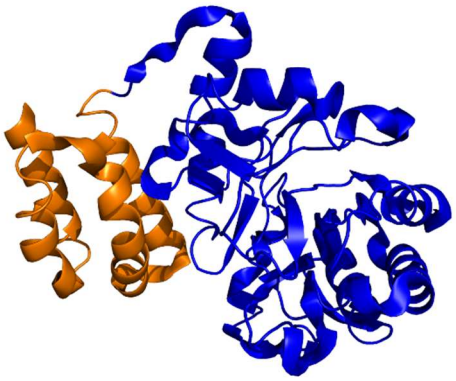
1T6CA



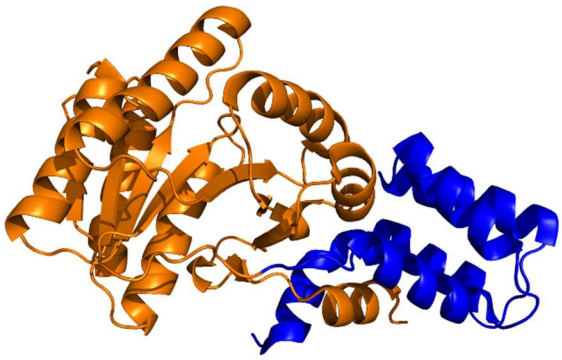
1U98A



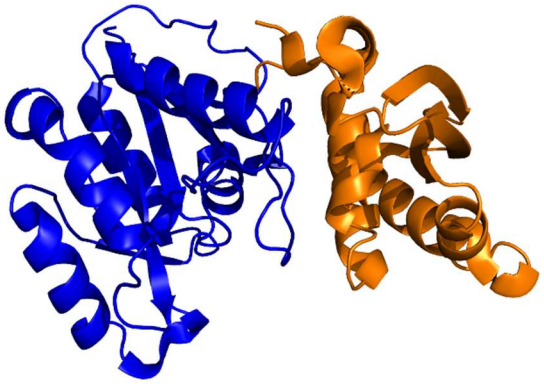
1V0BA



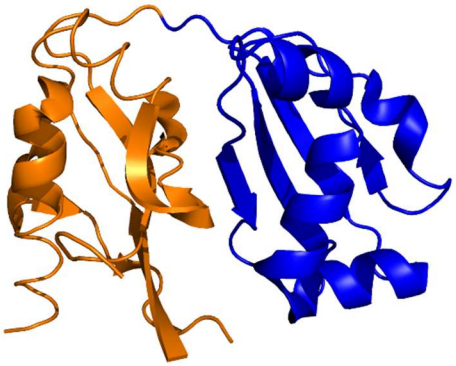
1VHNA



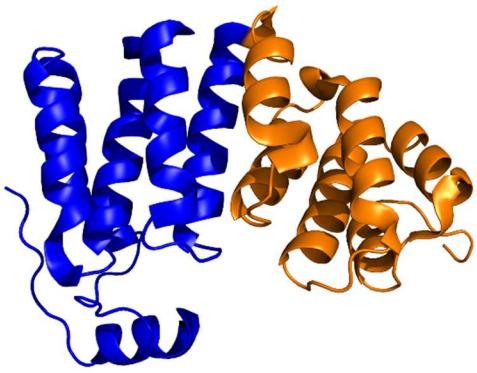
1VMAA



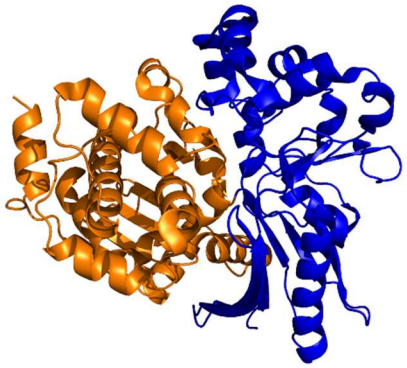
1WF3A



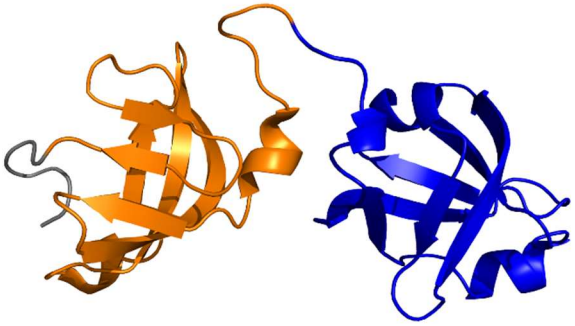
1WJ9A



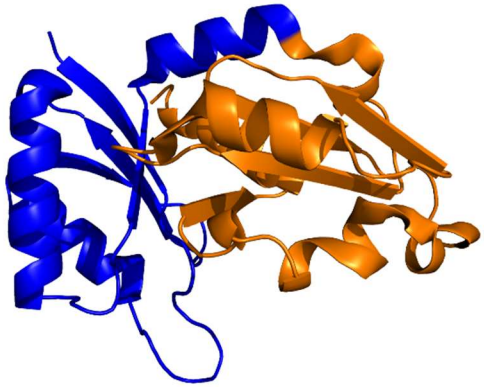
2B6CB



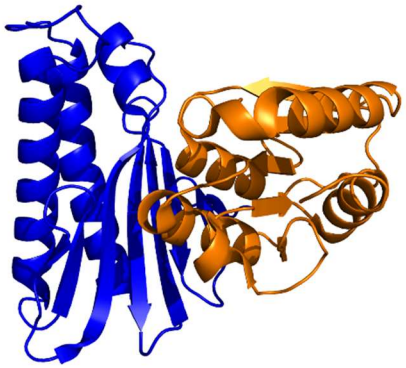
2CGJA



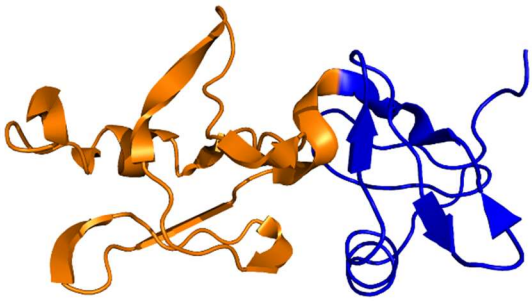
2DYIA



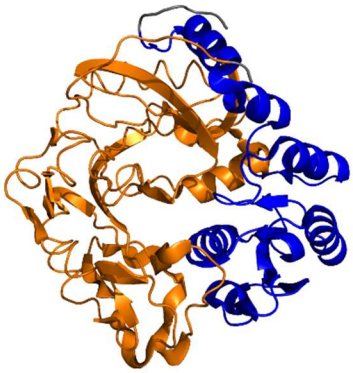
2HIYC



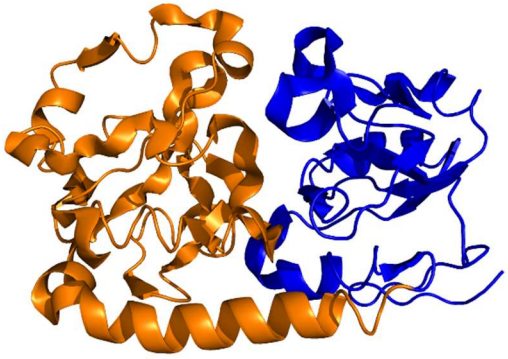
2QFLA



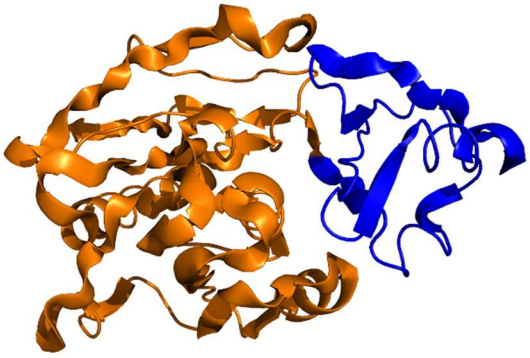
2RA9A



2W6PB

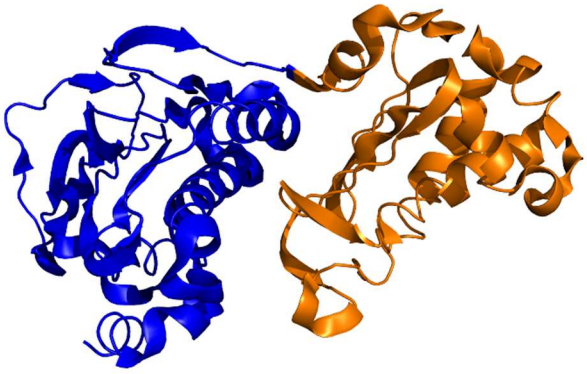


2WHYA

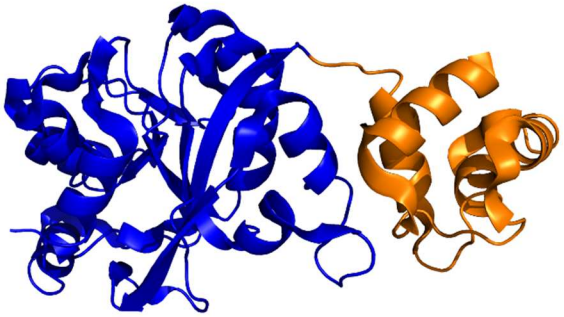


3A4TA

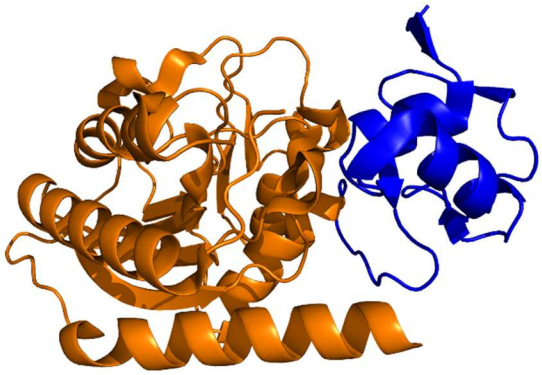
3CI0J



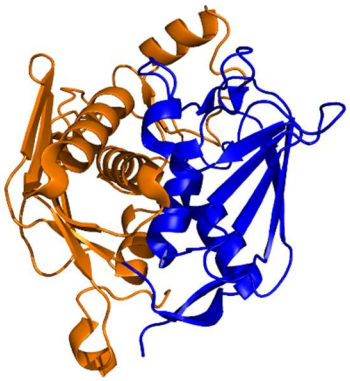
3CWVA



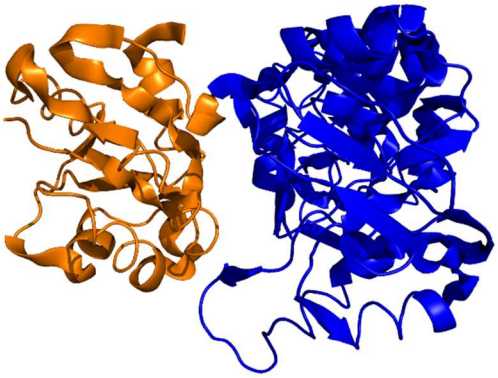
3FUXC



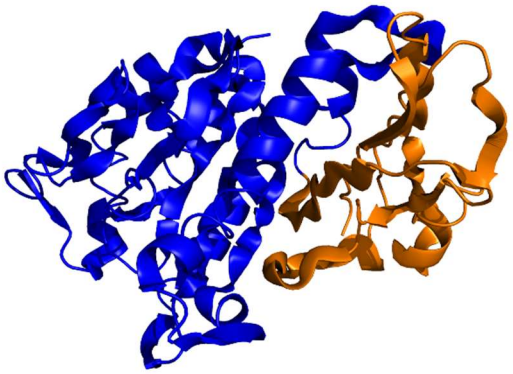
3HP7A



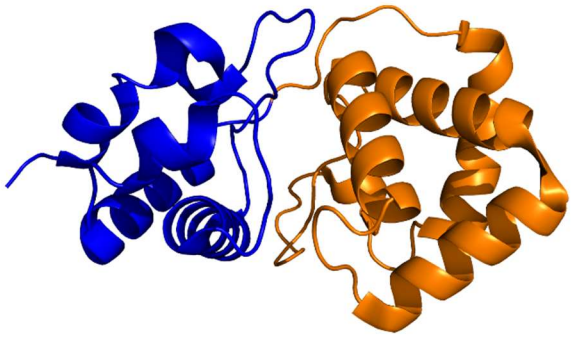
3NZKB



3QCZA



3VO8B



3VRDA

## 6.3 Bibliography


Abraham, W.P. & Thomas, S., 2015. Draft Genome Sequence of Pseudomonas psychrophila MTCC 12324 , Isolated from the Arctic at 79 ° N. *Genome announcements*, 3(3), pp.e00578–15.

Alberts, B., 1998. The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. *Cell*, 92(3), pp.291–294.

Aloy, P., Ceulemans, H., Stark, A. & Russell, R.B., 2003. The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, 332(5), pp.989–998.

Altschuh, D., Lesk, A.M., Bloomer, A.C. & Klug, A., 1987. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, 193(4), pp.693–707.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3), pp.403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17), pp.3389–3402.

Andreani, J., Faure, G. & Guerois, R., 2013. InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. *Bioinformatics Oxford England*, 29(14), pp.1742–9.

Andreatta, M., Laplagne, S., Li, S.S.C. & Smale, S., 2013. Prediction of residue-residue contacts from protein families using similarity kernels and least squares regularization. *arXiv preprint arXiv:1311.1301*, pp.1–8.

Apic, G., Gough, J. & Teichmann, S.A., 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of molecular biology*, 310(2), pp.311–325.

Apic, G. & Russell, R.B., 2010. Domain recombination: a workhorse for evolutionary innovation. *Science signaling*, 3(139), p.pe30.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. & Yeh, L.-S.L., 2004. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue), pp.D115–9.

Argos, P., 1988. An investigation of protein subunit and domain interfaces. *Protein Engineering, Design and Selection*, 2(2), pp.101–113.

Bai, X., McMullan, G. & Scheres, S.H.W., 2015. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences*, 40(1), pp.49–57.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.-I. & Langmead, C.J., 2011. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4), pp.1061–1078.

Banerjee, O., El Ghaoui, L. & D'Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9, pp.485–516.

Bashton, M. & Chothia, C., 2007. The Generation of New Protein Functions by the Combination of Domains. *Structure*, 15(1), pp.85–99.

Bashton, M. & Chothia, C., 2002. The geometry of domain combination in proteins. *Journal of molecular biology*, 315(4), pp.927–939.

Basu, M.K., Carmel, L., Rogozin, I.B. & Koonin, E. V, 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome research*, 18(3), pp.449–461.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. & Sonnhammer, E.L.L., 2000. The Pfam protein families database. *Nucleic Acids Res*, 28(1), pp.263–266.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., Studholme, D.J., Yeats, C. & Eddy, S.R., 2004. The Pfam protein families database. *Nucleic Acids Research*, 32(S1), pp.D138–D141.

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(1), pp.235–42.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M., 1978. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Archives of biochemistry and biophysics*, 185(2), pp.584–591.

Bhaskara, R.M., Padhi, A. & Srinivasan, N., 2014. Accurate prediction of interfacial residues in two-domain proteins using evolutionary information: Implications for three-dimensional modeling. *Proteins*, 82(7), pp.1219–1234.

Bondi, A., 1964. van der Waals Volumes and Radii. *The Journal of Physical Chemistry*, 68(3), pp.441–451.

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

Brenner, S.E., 2001. A tour of structural genomics. *Nature reviews. Genetics*, 2(October), pp.801–809.

Bryson, K., Cozzetto, D. & Jones, D.T., 2007. Computer-assisted protein domain boundary prediction using the Dom-Pred server. *Current Protein and Peptide Science*, 8(2), pp.181–188.

Buchan, D.W.A., Shepherd, A.J., Lee, D., Pearl, F.M.G., Rison, S.C.G., Thornton, J.M. & Orengo, C.A., 2002. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Research*, 12(3), pp.503–14.

Buslje, C. M., Santos, J., Delfino, J. M., & Nielsen, M. (2009). Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. *Bioinformatics*, *25*(9), pp.1125–1131.

Callaway, E., 2015. The Revolution Will Not Be Crystallized. *Nature*, 525, pp.172–174.

Caporaso, J.G., Smit, S., Easton, B.C., Hunter, L., Huttley, G.A. & Knight, R., 2008. Detecting coevolution without phylogenetic trees? Tree-ignorant metrics of coevolution perform as well as tree-aware metrics. *BMC evolutionary biology*, 8, p.327.

Cheng, T.M.K., Blundell, T.L. & Fernandez-Recio, J., 2008. Structural assembly of two-domain proteins by rigid-body docking. *BMC bioinformatics*, 9, p.441.

Chothia, C. & Gough, J., 2009. Genomic and structural aspects of protein evolution. *Biochemical Journal*, 419(1), pp.15–28.

Chothia, C. & Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4), pp.823–826.

Cover, T.M. & Thomas, J.A., 1991. Entropy, Relative Entropy and Mutual Information. In *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, pp. 12–15.

Csaba, G., Birzele, F. & Zimmer, R., 2009. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC structural biology*, 9, p.23.

Cunningham, B.A., Gottlieb, P.D., Pflumm, M.N. & Edelman, G.M., 1971. Immunoglobulin structure: diversity, gene duplication and domains. In B. Amos, ed. *Progress in Immunology*. New York: Academic Press, pp. 3–24.

DePristo, M.A., Weinreich, D.M. & Hartl, D.L., 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*, 6(9), pp.678–687.

Dickson, R.J., Wahl, L.M., Fernandes, A.D. & Gloor, G.B., 2010. Identifying and seeing beyond multiple sequence alignment errors using Intra-Molecular protein covariation. *PLoS ONE*, 5(6).

Dill, K. a & MacCallum, J.L., 2012. The protein-folding problem, 50 years on. *Science (New York, N.Y.)*, 338(6110), pp.1042–1046.

Dobson, C.M., 2003. Protein folding and misfolding. *Nature*, 426(6968), pp.884–890.

Duan, Y. & Reddy, B.V.B., 2005. Physicochemical and residue conservation calculations to improve the ranking of protein – protein docking solutions. *Protein science*, 14(2), pp.316–328.

Duhovny, D., Nussinov, R. & Wolfson, H.J., 2002. Efficient Unbound Docking of Rigid Molecules. In *Algorithms in Bioinformatics*. Springer, pp. 185–200.

Dumontier, M., Yao, R., Feldman, H.J. & Hogue, C.W.V., 2005. Armadillo: Domain Boundary Prediction by Amino Acid Composition. *Journal of Molecular Biology*, 350(5), pp.1061–1073.

Dunn, S.D., Wahl, L.M. & Gloor, G.B., 2008. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3), pp.333–340.

Durrant, J.D. & McCammon, J.A., 2011. Molecular dynamics simulations and drug discovery. *BMC Biology*, 9(1), p.71.

Eddy, S.R., 1996. Hidden Markov models. *Current opinion in structural biology*, 6(3), pp.361–365.

Edgar, R.C. & Batzoglou, S., 2006. Multiple sequence alignment. *Current opinion in structural biology*, 16(3), pp.368–73.

Ekeberg, M., Cecilia, L., Lan, Y., Weigt, M. & Aurell, E., 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1), p.012707.

Ekeberg, M., Hartonen, T. & Aurell, E., 2014. Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous. *Journal of Computational Physics*, 276, pp.341–356.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.-Y., Pieper, U. & Sali, A., 2006. Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics*.

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(5), pp.1356–1361.

Feinauer, C., Skwark, M.J., Pagnani, A., Aurell, E., Torino, P. & Science, C., 2014. Improving contact prediction along three dimensions. *PLoS Comput Biol*, 10(10), p.e1003847.

Felsenstein, J., 1985. Phylogenies and the comparative method. *The American Naturalist*, 125(1), pp.1–15.

Fodor, A.A. & Aldrich, R.W., 2004. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function and Genetics*, 56(2), pp.211–221.

Friedman, J., Hastie, T. & Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3), pp.432–441.

Giraud, B.G., Heumann, J.M. & Lapedes, a S., 1999. Superadditive correlation. *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, 59(5 Pt A), pp.4983–4991.

Glaser, F., Pupko, T., Paz, I. & Bell, R.E., 2003. ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information. *Bioinformatics (Oxford, England)*, 19(1), pp.163–164.

Gloor, G.B., Tyagi, G., Abrassart, D.M., Kingston, A.J., Fernandes, A.D., Dunn, S.D. & Brandl, C.J., 2010. Functionally compensating coevolving positions are neither homoplasic nor conserved in clades. *Molecular biology and evolution*, 27(5), pp.1181–1191.

Göbel, U., Sander, C., Schneider, R. & Valencia, A., 1994. Correlated Mutations and Residue Contacts in Proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4), pp.309–317.

Goh, C.S., Bogan, a a, Joachimiak, M., Walther, D. & Cohen, F.E., 2000. Co-evolution of proteins with their interaction partners. *Journal of molecular biology*, 299(2), pp.283–293.

Gomes, M., Hamer, R., Reinert, G. & Deane, C.M., 2012. Mutual information and variants for protein domain-domain contact prediction. *BMC research notes*, 5, p.472.

Gough, J., 2005. Convergent evolution of domain architectures (is rare). *Bioinformatics*, 21(8), pp.1464–1471.

Gough, J., Karplus, K., Hughey, R. & Chothia, C., 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4), pp.903–919.

Gromiha, M.M. & Selvaraj, S., 2004. Inter-residue interactions in protein folding and stability. *Progress in biophysics and molecular biology*, 86(2), pp.235–277.

Hadley, C. & Jones, D.T., 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, 7(9), pp.1099–1112.

Hamer, R., Luo, Q., Armitage, J. P., Reinert, G., & Deane, C. M. (2010). i-Patch: Interprotein contact prediction using local network information. *Proteins: Structure, Function, and Bioinformatics*, 78(13), pp.2781–2797.

Han, J.-H., Batey, S., Nickson, A. a, Teichmann, S. a & Clarke, J., 2007. The folding and evolution of multidomain proteins. *Nature reviews. Molecular cell biology*, 8(4), pp.319–330.

Han, J.H., Kerrison, N., Chothia, C. & Teichmann, S. a, 2006. Divergence of Interdomain Geometry in Two-Domain Proteins. *Structure (London, England : 1993)*, 14(5), pp.935–945.

Hayat, S., Sander, C., Marks, D.S. & Elofsson, A., 2015. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. *Proceedings of the National Academy of Sciences*, 112(17), pp.5413–5418.

Hendrickson, W. a, Horton, J.R. & LeMaster, D.M., 1990. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *The EMBO journal*, 9(5), pp.1665–1672.

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

Hertig, S., Goddard, T.D., Johnson, G.T. & Ferrin, T.E., 2015. Multidomain Assembler (MDA) Generates Models of Large Multidomain Proteins. *Biophysical Journal*, 108(9), pp.2097–2102.

Hopf, T. a., Colwell, L.J., Sheridan, R., Rost, B., Sander, C. & Marks, D.S., 2012. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7), pp.1607–1621.

Hopf, T. a., Morinaga, S., Ihara, S., Touhara, K., Marks, D.S. & Benton, R., 2015. Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nature Communications*, 6, p.6077.

Hopf, T. a., Schärfe, C.P.I., Rodrigues, J.P.G.L.M., Green, A.G., Sander, C., Bonvin, A.M.J.J. & Marks, D.S., 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3, p.e03430.

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

Huang, Y.J., Mao, B., Aramini, J.M. & Montelione, G.T., 2014. Assessment of template based protein structure predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 82(S2), pp.43–56.

Inbar, Y., Benyamini, H., Nussinov, R. & Wolfson, H.J., 2005. Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. *Physical biology*, 2(4), pp.S156–65.

Iyer, L.M., Leipe, D.D., Koonin, E. V & Aravind, L., 2004. Evolutionary history and higher order classification of AAA+ ATPases. *Journal of Structural Biology*, 146(1-2), pp.11–31.

Jana, B., Morcos, F. & Onuchic, J.N., 2014. From structure to function: the convergence of structure based models and co-evolutionary information. *Physical Chemistry Chemical Physics*, 16(14), pp.6496–6507.

Janin, J., 2010. Protein-protein docking tested in blind predictions: the CAPRI experiment. *Molecular BioSystems*, 6(12), pp.2351–2362.

Janin, J., Henrick, K., Moult, J., Eyck, L. Ten, Sternberg, M.J.E., Vajda, S., Vakser, I. & Wodak, S.J., 2003. CAPRI: A critical assessment of PRedicted interactions. *Proteins: Structure, Function and Genetics*, 52(1), pp.2–9.

Jardin, C., Stefani, A.G., Eberhardt, M., Huber, J.B. & Sticht, H., 2013. An information-theoretic classification of amino acids for the assessment of interfaces in protein-protein docking. *Journal of molecular modeling*, 19(9), pp.3901–3910.

Jaynes, E.T., 1957. Information Theory and Statistical Mechanics. *Physical review*, 106(4), pp.620–630.

Jefferson, E.R., Walsh, T.P. & Barton, G.J., 2008. A comparison of SCOP and CATH with respect to domain–domain interactions. *Proteins: Structure, Function, and Bioinformatics*, 70(1), pp.54–62.

Johnson, L.S., Eddy, S.R. & Portugaly, E., 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics*, 11, p.431.

Jones, D.T., 2001. Predicting novel protein folds by using FRAGFOLD. *Proteins*, 45(Suppl 5), pp.127–132.

Jones, D.T., 1999. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *Journal of molecular biology*, 292(2), pp.195–202.

Jones, D.T., 1997. Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins: Structure, Function and Genetics*, 29(S1), pp.185–191.

Jones, D.T., Buchan, D.W.A., Cozzetto, D. & Pontil, M., 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2), pp.184–90.

Jones, D.T., Singh, T., Kosciolek, T. & Tetchner, S., 2015. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics*, 31(7), pp.999–1006.

Jones, S., Marin, A. & Thornton, J.M., 2000. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Engineering Design and Selection*, 13(2), pp.77–82.

Jones, S., Stewart, M., Michie, a, Swindells, M.B., Orengo, C. & Thornton, J.M., 1998. Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Science*, 7(2), pp.233–242.

Kahraman, A., Herzog, F., Leitner, A., Rosenberger, G., Aebersold, R. & Malmström, L., 2013. Cross-Link Guided Molecular Modeling with ROSETTA. *PloS one*, 8(9), p.e73411.

Kahraman, A., Malmström, L. & Aebersold, R., 2011. Xwalk: Computing and visualizing distances in cross-linking experiments. *Bioinformatics*, 27(15), pp.2163–2164.

Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S. & Rost, B., 2014. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC bioinformatics*, 15, p.85.

Kamisetty, H., Ovchinnikov, S. & Baker, D., 2013. Assessing the utility of coevolution-based residue – residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39), pp.15674–15679.

Kass, I. & Horovitz, A., 2002. Mapping Pathways of Allosteric Communication in GroEL by Analysis of Correlated Mutations. *Proteins: Structure, Function, and Bioinformatics*, 48(4), pp.611–617.

Katchalski-Katzir, 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(6), pp.2195–2199.

Kendrew, J.C., Bodo, G., Dintzis, H.., Parrish, R.G. & Wyckoff, H., 1958. A three-dimensional model of the myoglobin molecule obtained by xray analysis. *Nature*, 181(4610), pp.662–666.

Khafizov, K., Madrid-Aliste, C., Almo, S.C. & Fiser, A., 2014. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proceedings of the National Academy of Sciences of the United States of America*, 111(10), pp.3733–3738.

Kim, D.E., Dimaio, F., Wang, R.Y., Song, Y. & Baker, D., 2013. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 82(S2), pp.208–218.

Konopka, B.M., Ciombor, M., Kurczynska, M. & Kotulska, M., 2014. Automated Procedure for Contact-Map-Based Protein Structure Reconstruction. *The Journal of Membrane Biology*, 247(5), pp.409–420.

Kosciolek, T. & Jones, D.T., 2015. Accurate contact predictions using covariation techniques and machine learning. *Proteins: Structure, Function, and Bioinformatics*, In Press.

Kosciolek, T. & Jones, D.T., 2014. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PloS one*, 9(3), p.e92197.

Krissinel, E. & Henrick, K., 2007. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology*, 372(3), pp.774–797.

Kryshtafovych, A., Barbato, A., Monastyrskyy, B., Fidelis, K., Schwede, T. & Tramontano, A., 2015. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. *Proteins: Structure, Function, and Bioinformatics*, In Press.

Kryshtafovych, A., Fidelis, K. & Moult, J., 2014. CASP10 results compared to those of previous CASP experiments. *Proteins*, 82(S2), pp.164–174.

Lapedes, A.S., Giraud, B.G., Liu, L. & Stormo, G.D., 1999. Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *Statistics in Molecular Biology*, 33, pp.236–256.

Lauro, F., Stratton, T., Chastain, R., Ferriera, S., Johnson, J., S, G., Yayanos, A. & Bartlett, D., 2013. Complete Genome Sequence of the Deep-Sea Bacterium Psychromonas strain CNPT3. *Genome announcements*, 1(3), pp.e00304–13.

Lees, J.G., Lee, D., Studer, R. a., Dawson, N.L., Sillitoe, I., Das, S., Yeats, C., Dessailly, B.H., Rentzsch, R. & Orengo, C. a., 2014. Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis. *Nucleic Acids Research*, 42(D1), pp.240–245.

Lensink, M.F. & Wodak, S.J., 2013. Docking, scoring, and affinity prediction in CAPRI. *Proteins*, 81(12), pp.2082–2095.

Levinthal, C., Wodak, S.J., Kahn, P. & Dadivanian, a K., 1975. Hemoglobin interaction in sickle cell fibers. I: Theoretical approaches to the molecular contacts. *Proceedings of the National Academy of Sciences of the United States of America*, 72(4), pp.1330–1334.

Levitt, M., 2009. Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, 106(27), pp.11079–11084.

Liang, S., Zhang, C., Liu, S., & Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. *Nucleic acids research*, 34(13), pp.3698–3707.

Ling, L.L., Schneider, T., Peoples, A.J., Spoering, A.L., Engels, I., Conlon, B.P., Mueller, A., Hughes, D.E., Epstein, S., Jones, M., Lazarides, L., Steadman, V. a, Cohen, D.R., Felix, C.R., Fetterman, K.A., Millett, W.P., Nitti, A.G., Zullo, A.M., Chen, C. & Lewis, K., 2015. A new antibiotic kills pathogens without detectable resistance. *Nature*, 517(7535), pp.455–459.

Lise, S., Walker-Taylor, A. & Jones, D.T., 2006. Docking protein domains in contact space. *BMC bioinformatics*, 7, p.310.

Little, D.Y. & Chen, L., 2009. Identification of coevolving residues and coevolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution. *PLoS ONE*, 4(3), p.e4762.

Littler, S.J. & Hubbard, S.J., 2005. Conservation of orientation and sequence in protein domain–domain interactions. *Journal of molecular biology*, 345(5), pp.1265–1279.

Ma, J., Wang, S., Wang, Z. & Xu, J., 2014. MRFalign: protein homology detection through alignment of Markov random fields. *PLoS computational biology*, 10(3), p.e1003500.

Ma, J., Wang, S., Wang, Z. & Xu, J., 2015. Protein Contact Prediction by Integrating Joint Evolutionary Coupling Analysis and Supervised Learning. *Bioinformatics*, In Press.

MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K.,

Lau, F.T., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D. & Karplus, M., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry. B*, 102(18), pp.3586–616.

Madaoui, H. & Guerois, R., 2008. Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proceedings of the National Academy of Sciences*, 105(22), pp.7708–7713.

Madhusudhan, M.S., Marti-Renom, M. a., Sanchez, R. & Sali, A., 2006. Variable gap penalty for protein sequence-structure alignment. *Protein Engineering, Design and Selection*, 19(3), pp.129–133.

Maisnier-Patin, S., Berg, O.G., Liljas, L. & Andersson, D.I., 2002. Compensatory adaptation to the deleterious effect of antibiotic resistance in Salmonella typhimurium. *Mol Microbiol*, 46(2), pp.355–366.

Mandal, S., Moudgil, M. & Mandal, S.K., 2009. Rational drug design. *European Journal of Pharmacology*, 625(1), pp.90–100.

Mao, W., Kaya, C., Dutta, A., Horovitz, A. & Bahar, I., 2015. Comparative Study of the Effectiveness and Limitations of Current Methods for Detecting Sequence Coevolution. *Bioinformatics*, 31(12), pp.1929–1937.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R. & Sander, C., 2011. Protein 3D structure computed from evolutionary sequence variation. *PloS one*, 6(12), p.e28766.

Marks, D.S., Hopf, T.A. & Sander, C., 2012. Protein structure prediction from sequence variation. *Nature Biotechnology*, 30(11), pp.1072–1080.

Martin, L.C., Gloor, G.B., Dunn, S.D. & Wahl, L.M., 2005. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22), pp.4116–4124.

Martin, R.B., 2001. Peptide bond characteristics. In *Metal ions in biological systems*. pp. 1–24.

Meinshausen, N. & Bühlmann, P., 2006. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3), pp.1436–1462.

Michel, M., Hayat, S., Skwark, M.J., Sander, C., Marks, D.S. & Elofsson, a., 2014. PconsFold: improved contact predictions improve protein models. *Bioinformatics*, 30(17), pp.i482–i488.

Mintseris, J. & Weng, Z., 2005. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, 102(31), pp.10930–10935.

Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A. & Kryshtafovych, A., 2014. Evaluation of residue-residue contact prediction in CASP10. *Proteins*, 82(S2), pp.138–153.

Monastyrskyy, B., D'Andrea, D., Fidelis, K., Tramontano, A. & Kryshtafovych, A., 2015. New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins: Structure, Function and Bioinformatics*, In Press.

Moore, A.D., Björklund, Å.K., Ekman, D., Bornberg-Bauer, E. & Elofsson, A., 2008. Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences*, 33(9), pp.444–451.

Morcos, F., Jana, B., Hwa, T. & Onuchic, J.N., 2013. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(51), pp.20533–20538.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T. & Weigt, M., 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49), pp.E1293–1301.

Morcos, F., Schafer, N.P., Cheng, R.R., Onuchic, J.N. & Wolynes, P.G., 2014. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, 111(34), pp.12408–12413.

Moreira, I.S., Fernandes, P.A. & Ramos, M.J., 2007. Hot spots—A review of the protein – protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, 68(4), pp.803–812.

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C., 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4), pp.536–540.

Neher, E., 1994. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*, 91(1), pp.98–102.

Nugent, T. & Jones, D.T., 2012. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(24), pp.E1540–1547.

Ohta, T., 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246, pp.96–98.

Oliva, R., Vangone, A. & Cavallo, L., 2013. Ranking multiple docking solutions based on the conservation of inter-residue contacts. *Proteins*, 81(9), pp.1571–1584.

Olmea, O., Rost, B. & Valencia, A., 1999. Effective use of sequence correlation and conservation in fold recognition. *Journal of molecular biology*, 293(5), pp.1221–1239.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. & Thornton, J., 1997. CATH - a hierarchic classification of protein domain structures. *Structure*, 5(8), pp.1093–1109.

Ovchinnikov, S., Kamisetty, H. & Baker, D., 2014. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 3, p.e02030.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C., 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of molecular biology*, 284(4), pp.1201–1210.

Pazos, F., Helmer-citterich, M., Ausiello, G., Valencia, A., Biologia, D., Roma, U. & Vergata, T., 1997. Correlated Mutations Contain Information About Protein-Protein Interaction. *Journal of molecular biology*, 271(4), pp.511–523.

Ponting, C.P. & Russell, R.R., 2002. The natural history of protein domains. *Annual review of biophysics and biomolecular structure*, 31, pp.45–71.

Poon, A. & Chao, L., 2005. The rate of compensatory mutation in the DNA bacteriophage φX174. *Genetics*, 170(3), pp.989–999.

Qian, J., Luscombe, N.M. & Gerstein, M., 2001. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology*, 313(4), pp.673–681.

Remmert, M., Biegert, A., Hauser, A. & Söding, J., 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, 9(2), pp.173–175.

Šali, A. & Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3), pp.779–815.

Sandler, I., Medalia, O. & Aharoni, A., 2013. Experimental analysis of co-evolution within protein complexes: The yeast exosome as a model. *Proteins: Structure, Function, and Bioinformatics*, 81(11), pp.1997–2006.

Dos Santos, R.N., Morcos, F., Jana, B., Andricopulo, A.D. & Onuchic, J.N., 2015. Dimeric interactions and complex formation using direct coevolutionary couplings. *Scientific Reports*, 5, p.13652.

Savitsky, P., Bray, J., Cooper, C.D., Marsden, B.D., Mahajan, P., Burgess-Brown, N.A. & Gileadi, O., 2010. High-throughput production of human proteins for crystallization: the SGC experience. *Journal of structural biology*, 172(1), pp.3–13.

Scheres, S.H., 2014. Beam-induced motion correction for sub-megadalton cryo-EM particles. *eLife*, 3, p.e03665.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H.J., 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research*, 33(S2), pp.W363–367.

Schrödinger LLC, The PyMOL Molecular Graphics System.

Schwede, T., 2013. Protein Modeling: What Happened to the “Protein Structure Gap”? *Structure*, 21(9), pp.1531–1540.

Seemayer, S., Gruber, M. & Söding, J., 2014. CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics (Oxford, England)*, 30(21), pp.3128–3130.

Shackelford, G., & Karplus, K. (2007). Contact prediction using mutual information and neural nets. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), pp.159–164

Shao, Y., & Bystroff, C. (2003). Predicting interresidue contacts using templates and pathways. *Proteins: Structure, Function, and Bioinformatics*, 53(S6), pp.497–502.

Shen, M., & Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11), pp.2507–2524.

Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., Yeats, C., Thornton, J.M. & Orengo, C.A., 2013. New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. *Nucleic Acids Research*, 41(D1), pp.490–498.

Simons, K.T., Kooperberg, C., Huang, E. & Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *Journal of molecular biology*, 268(1), pp.209–225.

Skwark, M.J., Abdel-Rehim, A. & Elofsson, A., 2013. PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics (Oxford, England)*, 29(14), pp.1815–1816.

Skwark, M.J., Raimondi, D., Michel, M. & Elofsson, A., 2014. Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns. *PLoS computational biology*, 10(11), p.e1003889.

Söding, J., 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7), pp.951–960.

Söding, J. & Remmert, M., 2011. Protein sequence comparison and fold recognition: Progress and good-practice benchmarking. *Current Opinion in Structural Biology*, 21(3), pp.404–411.

Stein, R.R., Marks, D.S. & Sander, C., 2015. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLOS Computational Biology*, 11(7), p.e1004182.

Strub, M.-P., Hoh, F., Sanchez, J.-F., Strub, J.M., Böck, A., Aumelas, A. & Dumas, C., 2003. Selenomethionine and Selenocysteine Double Labeling Strategy for Crystallographic Phasing. *Structure*, 11(11), pp.1359–1367.

Süel, G.M., Lockless, S.W., Wall, M.A. & Ranganathan, R., 2003. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural & Molecular Biology*, 10(1), pp.56–69.

Tanaka, S. & Scheraga, H. a, 1975. Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 72(10), pp.3802–3806.

Taylor, W.R. & Hatrick, K., 1994. Compensating changes in protein multiple sequence alignments. *Protein Engineering*, 7(3), pp.341–348.

Tetchner, S., Kosciolek, T. & Jones, D.T., 2014. Opportunities and limitations in applying coevolution-derived contacts to protein structure prediction. *Bio-Algorithms and Med-Systems*, 10(4), pp.243–254.

The Uniprot Consortium, 2015. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1), pp.D204–D212.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), pp.267–288.

Tress, M., de Juan, D., Graña, O., Gómez, M.J., Gómez-Puertas, P., González, J.M., López, G. & Valencia, A., 2005. Scoring docking models with evolutionary information. *Proteins: Structure, Function and Genetics*, 60(2), pp.275–280.

Urano, D., Dong, T., Bennetzen, J.L. & Jones, A.M., 2015. Adaptive Evolution of Signaling Partners. *Molecular Biology and Evolution*, 32(4), pp.998–1007.

Velankar, S., Dana, J.M., Jacobsen, J., Van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J. & Kleywegt, G.J., 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research*, 41(November 2012), pp.483–489.

Veretnik, S., Bourne, P. E., Alexandrov, N. N., & Shindyalov, I. N. (2004). Toward consistent assignment of structural domains in proteins. *Journal of Molecular Biology*, 339(3), pp.647–678.

Vernet, T., Tessier, D.C., Khouri, H.E. & Altschuh, D., 1992. Correlation of co-ordinated amino acid changes at the two-domain interface of cysteine proteases with protein stability. *Journal of molecular biology*, 224(2), pp.501–509.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C. & Teichmann, S.A., 2004. Structure, function and evolution of multidomain proteins. *Current opinion in structural biology*, 14(2), pp.208–216.

Vogel, C., Teichmann, S.A. & Pereira-Leal, J., 2005. The relationship between domain duplication and recombination. *Journal of Molecular Biology*, 346(1), pp.355–365.

Wang, Y. & Barth, P., 2015. Evolutionary-guided de novo structure prediction of self-associated transmembrane helical proteins with near-atomic accuracy. *Nature Communications*, 6.

Webb, B. & Sali, A., 2014. Protein Structure Modeling with MODELLER. In *Protein Structure Prediction*. New York: Springer, pp. 1–15.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A. & Hwa, T., 2009. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1), pp.67–72.

Wetlaufer, D.B., 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 70(3), pp.697–701.

Wheelan, S.J., Marchler-Bauer, A. & Bryant, S.H., 2000. Domain size distributions can predict domain boundaries. *Bioinformatics*, 16(7), pp.613–618.

Wodak, S.J. & Lensink, M.F., 2007. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins: Structure, Function, and Bioinformatics*, 69(4), pp.704–718.

Wodak, S.J. & Méndez, R., 2004. Prediction of protein-protein interactions: The CAPRI experiment, its evaluation and implications. *Current Opinion in Structural Biology*, 14(2), pp.242–249.

Wollacott, A.M., Zanghellini, A., Murphy, P. & Baker, D., 2007. Prediction of structures of multidomain proteins from structures of the individual domains. *Protein science*, 16(2), pp.165–175.

Wollenberg, K.R. & Atchley, W.R., 2000. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proceedings of the National Academy of Sciences of the United States of America*, 97(7), pp.3288–3291.

Xu, D., Jaroszewski, L., Li, Z. & Godzik, A., 2015. AIDA: Ab Initio Domain Assembly for Automated Multi-domain Protein Structure Prediction and Domain-Domain Interaction Prediction. *Bioinformatics*, 31(13), pp.2098–2105.

Xu, D., Jaroszewski, L., Li, Z. & Godzik, A., 2014. AIDA: Ab initio domain assembly server. *Nucleic Acids Research*, 42(W1), pp.W308–313.

Yanofsky, C., Horn, V. & Thorpe, D., 1964. Protein Structure Relationships Revealed By Mutational Analysis. *Science (New York, N.Y.)*, 146, pp.1593–1594.

Van Zundert, G.C.P., Rodrigues, J.P.G.L.M., Trellet, M., Schmitz, C., Kastritis, P.L., Karaca, E., Melquiond, A.S.J., van Dijk, M., de Vries, S.J. & Bonvin, A.M.J.J., 2015. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, In Press.