Role of Mutual Information for Predicting Contact Residues in Proteins

Mireille Gomes
Department of Statistics
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2012

For my parents, who travelled many miles and sacrificed their comforts and careers so that I could have mine.

Acknowledgements

This thesis is a culmination of the support of many people, without whom it would not have been possible.

The primary thank you goes to my supervisors, Prof. Gesine Reinert and Prof. Charlotte Deane, who enabled me to develop my scientific skills. Their direction, dedication and genuine interest over the last three years were the driving forces behind this dissertation. Special thanks also to Dr. Rebecca Hamer for her guidance and ever-present willingness to assist.

My OPIG colleagues who not only provided engaging scientific discussion and assistance with code, but also a warm and lively social circle. Thank you all for enlisting me in so-called "sports" tournaments, delivering cookies, cooking dinner, exploiting connections at Maxwells, buying not needed roll-on paint kits, lending me novels, agreeing to random travel expeditions and never failing to laugh "with" me. In no particular order, JP, Jamie, Eoin, Konrad, Seb, Leila, Hannah, James, Jean, Henry, Saulo, Waqar, Yoonjoo, Qiang, Rhodri, Anna Lewis, Sumeet, Markus, Faisal, Eleanor, Anthony and Anna Vangone. A heartfelt thank you to Beverley Lane whose caring and friendly disposition made the Medawar a lovely place to work.

My DTC family for the many laughs, nights out, camping trips, birthday drinks, bridge pictures, disparaging comments, overall support and high standards you set for me. To Pam Dawes and my MAK family for blindly believing in my potential, fully supporting every feather-brained idea I had and warmly encircling me into your lives. My pseudo-nuclear family in Oxford, Jennifer de Beyer, Tiago Rito, Eda Beyazit and Helen Curtis, thank you for answering calls about those many earth-shattering, colossal events that I just had to discuss, picking me up time after time after time, and giving me a "home" on this side of the ocean.

To my sister, Chantelle, for keeping me grounded, believing in me and always fighting my corner. My deepest gratitude to my parents who travelled many miles, and sacrificed their comforts and careers so that I could have mine. Mum and Dad this is for you.

Abstract

Mutual Information (MI) based methods are used to predict contact residues within proteins and between interacting proteins. There have been many high impact papers citing the successful use of MI for determining contact residues in a particular protein of interest, or in certain types of proteins, such as homotrimers. In this dissertation we have carried out a systematic study to assess if this popularly employed contact prediction tool is useful on a global scale.

After testing original MI and leading MI based methods on large, cross-species datasets we found that in general the performance of these methods for predicting contact residues both within (intra-protein) and between proteins (inter-protein) is weak. We observe that all MI variants have a bias towards surface residues, and therefore predict surface residues instead of contact residues. This finding is in contrast to the relatively good performance of i-Patch (Hamer *et al.* [2010]), a statistical scoring tool for inter-protein contact prediction. i-Patch uses as input surface residues only, groups amino acids by physiochemical properties, and assumes the existence of patches of contact residues on interacting proteins. We examine whether using these ideas would improve the performance of MI.

Since inter-protein contact residues are only on the surface of each protein, to disentangle surface from contact prediction we filtered out the confounding buried residues. We observed that considering surface residues only does indeed improve the inter-protein contact prediction ability of all tested MI methods. We examined a specific "successful" case study in the literature and demonstrated that here, even when considering surface residues only, the most accurate MI based inter-protein contact predictor, MIc, performs no better than random. We have developed two novel MI variants; the first groups amino acids by their physiochemical properties, and the second considers patches of residues on the interacting proteins. In our analyses these new variants highlight the delicate trade-off between signal and noise that must be achieved when using MI for inter-protein contact prediction.

The input for all tested MI methods is a multiple sequence alignment of homologous proteins. In a further attempt to understand why the MI methods perform poorly, we have investigated the influence of gaps in the alignment on intra-protein contact prediction. Our results suggest that depending on the evaluation criteria and the alignment construction algorithm employed, a gap cutoff of around 10% would maximise the performance of MI methods, whereas the popularly employed 0% gap cutoff may lead to predictions that are no better than random guesses.

Based on the insight we have gained through our analyses, we end this dissertation by identifying a number of ways in which the contact residue prediction ability of MI variants may be improved, including direct coupling analysis.

Contents

List of Figures

List of Tables

Glossary

3D	3-Dimensional
Å	Ångström
aMIc	Lee & Kim [2009] Mutual Information variant (Equation 1.18)
BLAST	Basic Local Alignment Search Tool (Altschul *et al.* [1990])
DCA	Direct Coupling Analysis (Section 1.5.2)
DDI	Domain-Domain Interaction
enrichment	Percent of contact residue pair scores in the subset of scores being considered
F-measure	The maximum F-measure may be considered as the optimal trade-off point between recall/sensitivity and specificity of a classifier (Liu *et al.* [2010], Equation 1.25)
FN	False Negative
FPR	False Positive Rate $= \frac{FP}{FP+TN}$
FP	False Positive
HK	Histidine Kinase
HMMER3	Hidden Markov Model based sequence alignment software (Eddy [2011])
HMM	Hidden Markov Model
i-Patch	Hamer *et al.* [2010] non-Mutual Information based algorithm to predict contacts between interacting proteins (Section 1.8)
inter-domain	between protein domains
inter-protein	between proteins

intra-domain within a protein domain

intra-protein within a protein

JOY protein sequence-structure analysis software (Mizuguchi *et al.* [1998])

MaxAlign sequence alignment software that attempts to maximise the number of amino acids in ungapped columns by selecting an optimal subset of sequences (Gouveia-Oliveira *et al.* [2007])

MCC curve Matthews Correlation Coefficient curve (Matthews [1975], Equation 1.23)

MI3DRA Our Mutual Information variant that considers triangles of residues and uses a reduced alphabet set of amino acids (Sections 4.3.1 and 4.3.2)

MI3D Our Mutual Information variant that considers triangles of residues (Equation 4.1)

MIcRA Our MIc variant that uses a reduced alphabet set of amino acids (Section 4.3.2)

MIc Lee & Kim [2009] Mutual Information variant (Equation 1.14)

MIp3DRA Our MIp variant that considers triangles of residues and uses a reduced alphabet set of amino acids (Sections 4.3.1 and 4.3.2)

MIp3D Our MIp variant that considers triangles of residues (Equation 4.2)

MIpRA Our MIp variant that uses a reduced alphabet set of amino acids (Section 4.3.2)

MIp Dunn *et al.* [2008] Mutual Information variant (Equation 1.10)

MIRA Our Mutual Information variant that uses a reduced alphabet set of amino acids (Section 4.3.2)

MI Mutual Information (Equation 1.7)

MSA Multiple Sequence Alignment

MUSCLE sequence alignment software (Edgar [2004])

N×MCC^2 Performance evaluation metric to determine if the Matthews Correlation Coefficient (MCC) value is significantly better than random (Baldi *et al.* [2000], Section 1.9.4)

P-ROC curve Precision Recall Operating Characteristic curve (Buckland & Gey [1994], Section 1.9)

PDB Protein Data Bank (Berman *et al.* [2000])

Pfam Protein family database (Punta *et al.* [2012])

PPI Protein-Protein Interaction

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

protein complex a group of two or more interacting proteins

PSICOV A Direct Coupling Analysis measure based on sparse inverse covariance estimation (Jones *et al.* [2012])

recall = sensitivity = true positive rate = $\frac{TP}{TP+FN}$

RR Response Regulator

SN Sensitivity = recall = true positive rate = $\frac{TP}{TP+FN}$

SP Specificity = $\frac{TN}{TN+FP}$

TN True Negative

TPR True Positive Rate = recall = sensitivity = $\frac{TP}{TP+FN}$

TP True Positive

ZNMI Z-scored-product Normalized Mutual Information (Brown & Brown [2010], Section 1.7.5)

Chapter 1

Introduction

1.1 Thesis Overview

Knowledge of protein structure and interactions is integral to our understanding of biological systems. Elucidation of residues in close proximity within a protein and between interacting proteins sheds light on the structure and function of the protein. To determine protein structure and interaction the concept of coevolving protein residues, *i.e.* "correlated mutations," has been exploited.

It is believed that residues that are physically proximal, in "contact," coevolve in order to preserve a protein's fold and/or maintain protein-protein interaction (Halperin *et al.* [2006]); such residues are thought to undergo "correlated mutations." Mutual Information (MI), a correlated mutation analysis measure, is commonly employed to detect contact residues within a protein (intra-protein) and between proteins (inter-protein) (Atchley *et al.* [2000]; Dunn *et al.* [2008]; Korber *et al.* [1993]; Martin *et al.* [2005]; Tillier & Lui [2003]; Wollenberg & Atchley [2000]).

However in previous inter- and intra-protein residue studies, high MI scores are associated with both contact and non-contact pairs (Dunn *et al.* [2008]; Halperin *et al.* [2006]; Martin *et al.* [2005]; Skerker *et al.* [2008]). Over the years limitations of MI

have been recognised and a number of techniques have been designed to overcome these. Nonetheless, the accuracy of MI variants for predicting inter- and intra-protein contacts remains low (Dunn *et al.* [2008]; Halperin *et al.* [2006]; Martin *et al.* [2005]).

In this dissertation we investigate MI based methods and find that their accuracy is surprisingly poor. We assess as possible explanations phylogenetic noise in the amino acid alphabet, the patch-like nature of residue interactions and gap cutoffs. None of these factors provide a satisfactory explanation for the low performance of MI based methods, if the assumption that correlated mutations are directly related to contact residue pairs is correct.

We begin in this chapter by discussing the biological motivation for this work and presenting background information on protein structure, protein-protein interaction and the theory of correlated mutations. As we are interested in assessing variants of MI for predicting structure and interaction we give an overview of the leading MI based methods and the evaluation measures used to appraise their performance.

Employing structural data of crystallised protein domains and associated multiple sequence alignments (MSAs), in Chapter 2 we find that the ability of all tested MI variants to predict both inter- and intra-protein contact residues is weak. We demonstrate that the MI variants are skewed towards predicting surface residues rather than contact residues; with some variation depending on the noise correction metrics of some MI algorithms (Chapter 3).

As residues involved in protein domain-domain interactions tend to be on the surface of the domains, we disentangle contact residue prediction from surface prediction. In Chapter 4 we filter out the confounding buried residues and assess the inter-domain contact residue determination abilities of the MI variants, using surface residues only. Motivated by the algorithm of i-Patch (Hamer *et al.* [2010]), a non-MI based tool for inter-protein contact prediction, we formulate two novel MI measures for this task. These new measures use a reduced alphabet amino acid set and the patch-like nature

of residue interactions.

In our study we evaluate MI and its variants on a large, general purpose, cross-species, inter-domain dataset. We show that MI algorithms have some predictive abilities for inter-domain contacts, but when considering a single "successful" test case even the recommended MI variant performs not significantly better than random.

Finally in Chapter 5, we closely examine the influence of gaps in the protein sequence on MI algorithms. We consider the various gap penalties imposed and speculate that in order to maximise MI performance, there is a balance that must be achieved between the exclusion of contact residue columns in the MSA and inclusion of noise in the data. We find that depending on the evaluation criteria and the alignment algorithm used to construct the MSA, a range of gap cutoffs may be suitable. As a rule of thumb, around the 10% gap cutoff appears to optimise performance. We observe that once a column in an MSA contains at least one gap, the number of gaps in the column is not strongly related to whether or not the corresponding residue is involved in a contact residue pair.

The Conclusions section of this dissertation details the significance of these findings and the future directions they offer to the field of MI based contact residue prediction.

1.2 Motivation

1.2.1 Biological significance of protein structure and interaction

Proteins constitute the majority of the dry mass in a cell (Alberts *et al.* [2007]). These organic compounds function not only as the building blocks of the cell, but are also responsible for executing most cellular processes. For example, proteins in the cytoskeleton maintain cell shape; membrane proteins transport small molecules in and out of the cell; signalling proteins transmit information within and between cells; and enzymes catalyse chemical transformations. Other proteins move cellular organelles,

translate DNA, function as antibodies, toxins, hormones and antifreeze molecules (Alberts *et al.* [2007]; Devlin [2005]). The structure of these versatile macromolecules dictates the specificity of their interactions and the diversity of their functions. Their chemical composition, size, shape and binding complementarity lie at the heart of our understanding of whole biological systems and the field of drug discovery.

1.2.2 The protein structure prediction problem

Recent developments in high-throughput sequencing technologies has led to an explosion of protein sequence data, though knowledge of structures is lagging far behind. As of August 2012 there were 16,393,342 sequences in the National Center for Biotechnology Information (NCBI) Non-Redundant RefSeq database (Pruitt *et al.* [2007]), while the Protein Data Bank (Berman *et al.* [2000]) contained only 47,453 protein structures of non-redundant sequences, *i.e.* sequences that are not 100% identical (83,983 total structures). Since the experimental determination of protein structures, via X-ray crystallography and Nuclear Magnetic Resonance (NMR), is laborious, time consuming and costly, there is great motivation for the development of computational structure prediction tools.

Computational structure prediction algorithms can be divided into two main classes: homology (template-based) modelling and *ab initio* (template-free) modelling (Gajda *et al.* [2011]). Template-based modelling aligns the target sequence to an evolutionarily related, experimentally solved protein structure. It is assumed that proteins with homologous sequences will share similar structure and function. Specifically it has been shown that proteins with sequence homology greater than 50% provide a good general model for 3-dimensional (3D) structure. Conversely, there will be large structural differences between proteins that have sequence homology less than 20% (Chothia & Lesk [1986]). However, proteins that have low sequence homology but share the same interacting partner, may share similar geometry at the interaction site (Read *et al.* [1984]).

The most widely used tool for this category of structure prediction is MODELLER (Sali & Blundell [1993]). Though the majority of the models generated by this method are approximately accurate, structural refinements are necessary in order to attain the atomic resolution achieved by experimental methods. Furthermore, the performance of this methodology is highly dependent on the availability of experimentally determined homologous structures. Hence we turn to template-free prediction.

Template-free or *ab initio* modelling includes a wide range of techniques. Some methods use chemical and physical information to simulate the folding process of a protein *in silico*. The underlying assumption of these algorithms is that a protein sequence folds to a native confirmation or set of confirmations that are at or near the global minimum free energy (Bonneau & Baker [2001]). Knowledge-based algorithms, such as ROSETTA (Simons *et al.* [1997]), utilise statistical potentials of recurrent patterns of known protein structures and sequences to determine the 3D shape of a protein of unknown structure. Still other algorithms employ MSAs (Section 1.3.6) and attempt to find correlated mutations (Section 1.5.1) in order to infer residues in a protein sequence that are in close proximity to each other (Halperin *et al.* [2006]; Jones [1997]; Jones *et al.* [2012]; Marks *et al.* [2011]; Weigt *et al.* [2009]).

Although the results of the most recent Critical Assessment of Methods of Protein Structure Prediction (CASP9) experiment indicate an improvement in both template-based and template-free modelling accuracy, there is still much room for advancement (Moult *et al.* [2011]). Better techniques for identifying the best model amongst a set of predicted models are necessary for the template-based prediction methods, and though template-free methods produce reasonable results for targets less than 120 residues they are not effective for larger proteins (Kryshtafovych *et al.* [2011]).

1.2.3 The protein-protein interaction prediction problem

The interaction(s) of a protein with other molecule(s) determine its biological function. The binding of a protein is believed to be very specific, such that a protein can only bind with one or a few molecules, compared to the thousands present in its environment (Alberts *et al.* [2007]). Almost all cellular processes involve proteins interacting with other proteins (Shenoy & Jayaram [2010]). It has been estimated that the human interactome contains approximately 650,000 protein interactions, while *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Caenorhabditis elegans* interactomes are estimated to have approximately 30,000, 100,000, and 240,000, protein interactions respectively (Stumpf *et al.* [2008]). These highly specific interactions pose attractive drug targets; by preventing certain proteins from interacting, the disease pathways they are involved in can be altered (Fuller *et al.* [2009]).

Several high-throughput screening experiments allow for the identification of interacting proteins. These methods include yeast two hybrid systems (Y2H) (Ito *et al.* [2001]; Uetz *et al.* [2000]), affinity purification (Rigaut *et al.* [1999]), phage display libraries (Smith [1985]) and protein-fragment complementation assays (PCA) (Pelletier *et al.* [1999]; Tarassov *et al.* [2008]). However these screens are often incomplete and have high false positive and false negative rates (Deane *et al.* [2002]; Gomez *et al.* [2003]; Guarracino *et al.* [2010]; Szilágyi *et al.* [2005]). Computational prediction methods attempt to augment the low amount of protein-protein interaction (PPI) data and do achieve some success (Jessulat *et al.* [2011]). These methods are based on gene neighbourhood conservation, phylogenetic profile comparison, interacting domain pair preservation, protein interaction network (PIN) inferences and high-throughput protein docking experiments (Jessulat *et al.* [2011]; Wass *et al.* [2011]).

Gene neighbourhood or co-localisation methods work on the premise that physically interacting genes or functionally related genes will be close to each other on the genome, and this co-localisation is especially significant if it is conserved across species.

In this manner physical or functional interactions between proteins are predicted across different organisms. Phylogenetic profile comparison proposes that genes with similar phylogenetic profiles co-evolve so that the interaction and function of the proteins are preserved. Another theory is that if pairs of protein domains are observed to interact in some organisms, it may be inferred that proteins with those domains interact in other organisms. Constructing a protein interaction network with nodes and edges, and then identifying within it dense subgraphs, one may conjecture unknown interactions between proteins with missing edges. A recent protein docking method compares docking scores of proteins with known interactions to those of non-interacting proteins, and uses the observed difference in the distributions of these scores to predict if a docked pair of proteins are likely to interact (Wass *et al.* [2011]).

None of these computational techniques, however, shed light on the pairs of residues in close contact in the interacting proteins; instead they only inform us of which proteins are interacting. Mutating residues in contact, with the aim of altering/disrupting the interactions of a protein, would allow for a better understanding of the biological pathway(s) the protein is involved in. Furthermore, knowledge of the residues in contact is necessary for the design of targeted drug molecules. The two experimental methods that provide this level of atomic detail, X-ray crystallography and NMR spectroscopy, are arduous and expensive; hence, the development of computational contact residue prediction tools has gained momentum.

Protein-protein contact residue prediction algorithms can be grouped into two categories: docking and sequence analysis. Typically docking aims to find the native binding state of two or more proteins by utilising physiochemical properties of the proteins and a repository of known binding interfaces (Tuncbag *et al.* [2009]). Sequence analysis methods on the other hand, apply a wide range of techniques, machine learning, sequence pattern recognition, and residue composition analysis, to an input multiple sequence alignment (MSA) in order to identify residues in contact (Jessulat *et al.* [2011]). The

main bottleneck in these prediction methods is a lack of biological understanding of protein-protein interactions, in addition to limited data such as, solved 3D structures of individual proteins and proteins in complex, experimentally verified true positive protein-protein interactions, and annotated protein interfaces. This in turn results in computational tools that yield low PPI prediction accuracy (Jessulat *et al.* [2011]).

1.3 Protein Structure

The overall framework of molecular biology is that genetic sequence encodes the sequence of a protein (Crick [1958]), the sequence of a protein determines its 3D structure (Anfinsen [1973]; Baker & Sali [2001]), the structure in turn dictates the interactions of the protein (Brinda & Vishveshwara [2005]; Williams & Lovell [2009]), and these interactions facilitate almost all biological processes (Alberts *et al.* [2007]). In this section we examine how the unique combination of residues in a protein sequence formulates its structural hierarchy and design. We also explore the databases housing protein sequences, structures and domains.

1.3.1 Primary structure

The sequence of a protein is composed of a unique arrangement of 20 types of residues, also known as amino acids. Each residue is linked to its sequence neighbour(s) within the protein via a covalent peptide bond making a polypeptide chain. The set of repeating atoms that lie along the core of the polypeptide chain is known as the protein backbone. The atoms attached to the C_α atoms of the backbone are called "side chains." The different chemical properties of these side chains, for example, aromatic, aliphatic, polar, non-polar, etc., are responsible for forming noncovalent bonds between residues (salt bridges, van der Waals forces and hydrogen bonds), which in turn may cause the protein chain to fold upon itself and form higher order structures.

1.3.2 Secondary structure

The amino acids in a protein sequence can noncovalently bond to assemble into one of two secondary structures: α-helices or β-sheets. α-helices involve hydrogen bonds between backbone atoms of every four residues, such that the polypeptide chain turns approximately every 3.6 amino acids giving rise to a cylinder. On the other hand, hydrogen bonds that connect strands of polypeptide chain result in rigid β-sheets. There are two varieties of β-sheets: parallel and anti-parallel. Parallel sheets are formed by strands running in the same direction, while anti-parallel sheets result from strands running in opposite directions.

1.3.3 Tertiary structure

Noncovalent bonds between secondary structures cause the polypeptide chain to pack into a compact, global, 3D, tertiary structure. The main driving force of protein folding is thought to be the hydrophobic effect; *i.e.* hydrophobic amino acids aggregate in the core of the protein, minimising exposure to the surrounding aqueous solution (Pace *et al.* [1996]).

Multiple tertiary structures assemble to form a quaternary structure. A quaternary structure consists of more than one polypeptide chain.

1.3.4 Domains

Proteins are often subdivided into domains which can be characterised as: a set of sequence residues common to many proteins (Uzman [2001]); structural regions of proteins that perform certain functions (Uzman [2001]); or parts of protein structure that can exist, function and evolve independently of the rest of the protein (Andreeva *et al.* [2004]; Richardson [1981]). For the purpose of this dissertation we employ the latter notion of protein domains.

We identify a domain as a set of residues that have a high number of contacts

within the set and a low number outside the set (Holm & Sander [1993]) (Figure 1.1). Our definition also specifies that the identified domains in a protein are connected to each other only by amino acid segments that have limited or no 3D structure (Devlin [2005]).

Defining structural domain boundaries in this manner is a subjective process and consequently domain definitions can vary in the number of leading and trailing residues (Richardson [1981]).

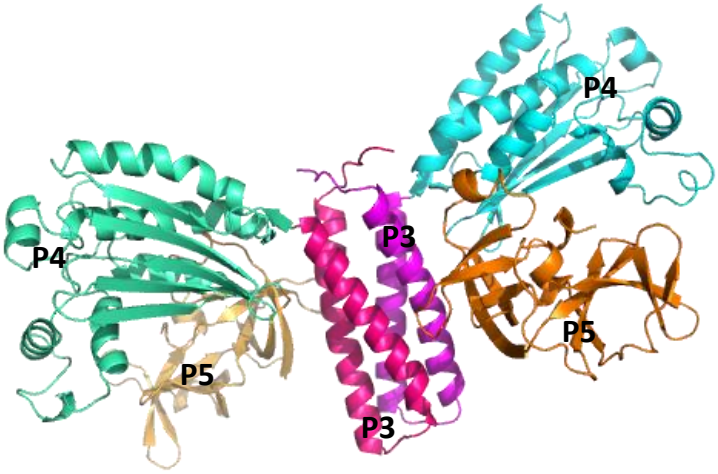

Figure 1.1: **Domains in a protein.** Depicted in the different colours are P3, P4 and P5 domains of *Thermotoga maritima* protein CheA (crystal structure 1B3Q.pdb (Bilwes *et al.* [1999])). Protein CheA exists as a homodimer (2 sets of identical domains). This image and subsequent protein structure images were created using the PyMOL graphics software (DeLano [2002]).

1.3.5 Surface and buried residues

Amino acids on the outside of a protein domain, "surface residues," tend to differ in chemical composition to their "buried residue" counterparts.

Amino acids with hydrophilic side chains are usually on the outside of the protein. Conversely, hydrophobic amino acids most often cluster in the core of the protein preventing accessibility to the surrounding aqueous solvent. The contact residues between

one protein or one domain and another will be surface residues. Although if the protein undergoes a conformational change upon binding, buried residues may be exposed to serve as a binding pocket for interacting molecules.

Given the 3D structure of the protein, the solvent-accessibility of a residue is usually measured by rolling a virtual water molecule over a protein, when not in complex, and summing the area of the residue that can be accessed by this molecule. In this thesis we use the definition that if this summed area is more than 7% of the maximally accessible area of the residue, the residue is identified as a "surface residue" (Mizuguchi *et al.* [1998]). The maximum surface area is calculated by placing the residue in a ALA-X-ALA (alanine-residue of interest-alanine) tripeptide structure. Conversely, those residues in the protein that are less than 7% solvent-accessible are termed "buried residues." In this work, the surface accessibility calculation of each residue in the structure of a protein is performed using JOY (Mizuguchi *et al.* [1998]). The surface and buried residues of a proxy histidine kinase and its response regulator can be visualised in Figure 1.2.

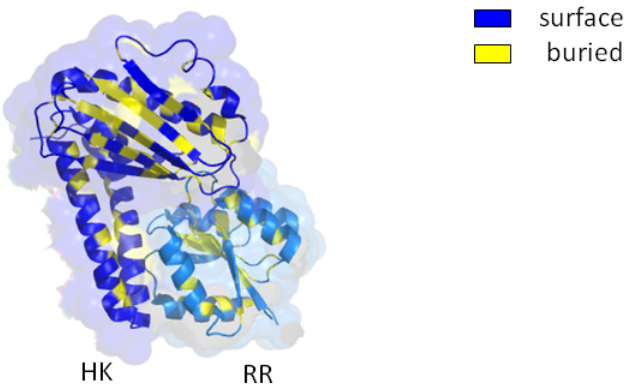

Figure 1.2: **Surface and buried residues in a pair of interacting domains** of a proxy histidine kinase (HK) and its response regulator (RR) (1F51.pdb (Zapf *et al.* [2000])). Surface and buried residues are determined for the HK and RR respectively, when each protein is not in complex. The cartoon ribbon depicts the backbones of the proteins, and the cloud surrounding the ribbon represents the surface area of the protein complex that is accessible to a solvent.

1.3.6 Multiple sequence alignments

A protein multiple sequence alignment (MSA) consists of protein sequences, at least three, which are homologs (Figure 1.6). A column of this alignment ideally consists of residues that superimpose structurally and share the same common ancestral residue. Achieving the "correct" alignment is not a trivial problem, for although X-ray crystallography and NMR determined protein structures can be superimposed, the ancestral history of residues in a sequence is not known from any primary source. Consequently, the quality of an MSA depends solely on the similarity of the sequences being aligned and the robustness of the alignment algorithm employed.

The most widely used alignment technique is 'progressive multiple alignment.' A set of 'distances' reflecting sequence similarity are calculated between every pair of sequences that need to be included in the alignment. Using these distances a tree is then built, such that more similar sequences are closer together and input sequences are at the end of every branch of the tree. Each time two sequences are aligned, they are compressed into a 'profile.' To build a multiple sequence alignment, the tree is then followed from branch ends to root; at each internal node profiles are aligned (Nuin *et al.* [2006]).

Alignment algorithms differ in how distances are calculated between pairs of sequences, how profiles are constructed, how the generated alignments are scored, as well as the iterative processes they use to refine these alignments. Some of the measures used to estimate the distance between pairs of sequences include expected alignment accuracy (ProbCons, Do *et al.* [2005]), the percentage of identical residues between each pair of sequences (T-Coffee (Notredame *et al.* [2000]), CLUSTALW (Thompson *et al.* [1994])), and the number of groups of residues within each sequence that are common between the two sequences (MUSCLE, Edgar [2004]).

After the alignments are built various refinement techniques can be used, such as making homologous sequences non-redundant above a certain sequence identity via

Cd-hit (Li & Godzik [2006]), or maximising the number of amino acids in ungapped columns by selecting an optimal subset of sequences using MaxAlign (Gouveia-Oliveira *et al.* [2007]).

1.3.7 Databases

1.3.7.1 Sequences

In this post-genomic era collating, curating, annotating and making protein sequence data accessible is no small task. There are a number of sequence databases that provide varying quantities of data and depth of associated details.

Basic sequence repositories such as NCBI's GenPept and Entrez Protein databases contain protein sequences translated from nucleotide sequences (Sayers *et al.* [2012]). Unlike GenPept, many of the entries in Entrez Protein contain additional information from curated databases such as SWISS-PROT (Gasteiger *et al.* [2001]) and TrEMBL (Boeckmann *et al.* [2003]), which are discussed later. NCBI's Reference Sequence (RefSeq) collection is more stringent, providing a non-redundant set of sequences by collapsing all reports of a protein sequence into a single record (Sayers *et al.* [2012]).

Curated protein databases such as PIR-PSD (Wu *et al.* [2003]), SWISS-PROT (Gasteiger *et al.* [2001]) and TrEMBL (Boeckmann *et al.* [2003]) contain manually or automatically verified sequences from all species, with annotations, including protein classification, known post-translational modifications, function, structure, etc. Entries in the PIR-PSD database are organised by superfamily and family, and annotated with structural, functional, bibliographic and genetic data. The annotations also include regions within the sequence of biological interest and the organism in which this protein naturally occurs. SWISS-PROT is scrutinously curated and annotated by biologists. Experimental evidence available in the literature is manually verified using a number of sequence analysis programs. In addition to the basic annotations, SWISS-PROT also includes pathways the protein is involved in, the tissue and developmental stages

in which the protein is expressed, diseases associated with the protein and similarities to other proteins. Since manually curating the SWISS-PROT database in this manner is laborious and time consuming, the TrEMBL database that does automated curation was built to incorporate new sequences quickly. TrEMBL includes unannotated protein sequences translated from nucleotide sequences, as well as sequences and annotations from SWISS-PROT and a number of other databases.

The builders of UniProt (UniProt Consortium [2011]) have attempted to combine SWISS-PROT, TrEMBL and PIR-PSD into a single database. UniProt aims to provide all of the mentioned database features under one umbrella, *i.e.* it provides an archive where new and updated sequences are deposited, an expertly curated sequence collection and a non-redundant reference set of protein sequences.

Still other sequence databases that are available may be specific to a species (*e.g.* Yeast Protein Database (YPD) - Hodges *et al.* [1998]) or type of protein (*e.g.* MPtopo for membrane proteins - Jayasinghe *et al.* [2001]).

1.3.7.2 Structures

Established in the 1970s, the Protein Data Bank (PDB) (Berman *et al.* [2000]) is a key resource in structural biology. This publically available repository contains structures of proteins and other biological macromolecules that have been determined via X-ray crystallography, Nuclear Magnetic Resonance (NMR), electron microscopy or a hybrid of these techniques. The primary data in the PDB are files listing the coordinates of each atom in a structure in 3D space. These files are also annotated with sequence information, chemistry and biology of the protein, experimental conditions, structure resolution, depositing author(s) and so on.

The protein structures from the PDB have been further annotated and classified in other specialised databases such as, Proteopedia (a wiki highlighting functional sites and ligands - Hodis *et al.* [2008]), SCOP (provides structural classification and evolu-

tionary information - Andreeva *et al.* [2008]) and PDB_TM (a database of transmembrane proteins - Tusnády *et al.* [2005]). For the purpose of this dissertation we obtain protein structures directly from the PDB.

1.3.7.3 Domains

Proteins sharing conserved regions are believed to perform similar functions (Mulder [2001]). Hence resources identifying proteins that have similar domains and subsequently grouping them into subfamilies, families and superfamilies, based on varying levels of their sequence, structural, evolutionary and functional relatedness are widely employed. These tools can be divided into three general categories: protein signature-based databases; sequence clustering databases; and amalgamated databases (Mulder [2001]).

Protein signatures are descriptors of a protein family or domain. They describe identified similarities of proteins in the family using various mathematical approaches. Examples of such approaches include regular expressions (defining patterns of conserved residues), profiles (tables stating probabilities of finding particular amino acids at a given sequence position) and Hidden Markov Models (HMM - a series of states and transitions encompassing the residue likelihood at and between positions). Unlike sequence clustering algorithms, protein signature methods usually begin with manually curated MSAs of known members of the protein family and are hence considered more reliable. Popularly used protein signature-based databases are Pfam (Punta *et al.* [2012]), PROSITE (Hulo *et al.* [2006]), PRINTS (Attwood *et al.* [2003]), SMART (Letunic *et al.* [2012]), TIGRFAMs (Haft *et al.* [2003]), PANTHER (Mi *et al.* [2005]), EVEREST (Portugaly *et al.* [2007]), Gen3D (Yeats *et al.* [2006]) and SUPERFAMILY (Wilson *et al.* [2007]). The HMMs used in EVEREST, Gen3D and SUPERFAMILY databanks are based on superfamily classifications formulated in the SCOP (Andreeva *et al.* [2008]) and CATH (Cuff *et al.* [2011]) databases, which in turn are derived from

proteins of known structure.

The sequence clustering-based databases usually use fully automated algorithms to group protein sequences that have high similarity. These processes struggle to recognise distant relationships between new and existing members of the same protein family. Additionally, biological annotation is sparse (Mulder [2001]). Examples of these type of databases are ProDom (Bru *et al.* [2005]), ProtoMap (Yona *et al.* [2000]), SYSTERS (Meinel *et al.* [2005]) and CluSTr (Petryszak *et al.* [2005]).

Since most of the mentioned protein family databases use protein sequences found in the UniProt Knowledgebase resource (UniProtKB - UniProt Consortium [2011]), there is a vast amount of redundancy and overlap in the defined families (Mulder [2001]). In order to allow for comparison between databases and easy access to all available information, several tools import data from these primary resources to form amalgamated databases. Widely employed amalgamated databases include CDD (Marchler-Bauer *et al.* [2011]), InterPro (Hunter *et al.* [2012]) and BLOCKS+ (Hunter *et al.* [2012]).

SCOP and CATH

The SCOP (Andreeva *et al.* [2008]) and CATH (Cuff *et al.* [2011]) databases mentioned previously use hierarchical clustering techniques to classify protein domains. A majority of the classification in CATH is carried out by automatic methods, while SCOP emphasises manual classification through visual inspection and comparison to observed 3D structures.

The CATH hierarchical groupings are as follows, from highest to lowest levels of classification:

1. Class: major constituents of secondary structure (α, β, $\alpha\beta$, other)
2. Architecture: gross spatial arrangement of secondary structures irrespective of their connectivity

3. Topology: the number, arrangement and connectivity of secondary structures (fold)

4. Homologous superfamily: evident evolutionary relationship based on a sequence identity >35%, or else having significant structural similarity, or if less than 20% sequence identity evidence of structural and functional parity

The levels of SCOP, from highest to lowest hierarchical grouping:

1. Class: major constituents of secondary structure (α, β, α/β, $\alpha+\beta$, other)

2. Fold: major secondary structure elements with the same arrangement and topological connections

3. Superfamily: domains that have low sequence identity, but whose structures and in many cases functional features show evidence of common evolutionary origin

4. Family: sequence identity $\geq$30%, or else having functional or structural similarity

1.4 Protein-protein Interactions

Cellular processes are largely determined by protein-protein interactions (PPIs). In this section we consider the residue composition of protein interfaces, the limited availability of experimentally determined structures of interacting proteins (complexes), as well as the residues in "contact" in these complexes.

1.4.1 Binding interfaces

Protein interaction (Figure 1.3) is determined by the size, shape and electrostatic complementarity of the residues in the binding interface of each protein (Argos [1988]; Janin [1995]; Jones & Thornton [1995]; Miller [1989]; Moreira *et al.* [2007]).

One investigation states that the standard interface size ranges from 1,200 to 2,000Å (Horton & Lewis [1992]); small interfaces, 1,150-1,200Å, constitute transient, low-stability complexes; and large interfaces, 2,000-4,660Å, can mostly be observed in the signal transduction system (Horton & Lewis [1992]; Lo Conte *et al.* [1999]). Residues in the interface are commonly hydrophobic. These hydrophobic interface residues drive the protein into a complex so that a favourable free energy change occurs as the hydrophobic residues move from an aqueous to a non-polar environment (Korn & Burnett [1991]; Young *et al.* [1994]). Electrostatic complementarity of protein interfaces is also known to promote binding (Camacho *et al.* [1999]; Ivanov *et al.* [2001]; Sheinerman *et al.* [2000]; Stevens *et al.* [2000]; Vijayakumar *et al.* [1998]; Xu *et al.* [1997a,b]).

In the binding interface, a small subset of amino acids are hypothesised to be the main contributors to the binding affinity; this residue subset is commonly referred to as "hot spots." The main hot spot amino acids are tryptophan (21%), arginine (13.3%) and tyrosine (12.3%) (Moreira *et al.* [2007]). Hot spot residues have been shown to be structurally conserved (Keskin *et al.* [2005]; Li *et al.* [1998]). It has also been observed that these residues frequently occur in protein interfaces that are used to bind to multiple interacting partners (Barnett *et al.* [2000]; Harris *et al.* [2001]; Luque & Freire [2000]; Ma *et al.* [2002]; Thornton [2001]).

There is evidence to suggest that transient interactions, brief contact, between proteins causes an increased mutation rate of interface residues that show no signs of coevolution. In contrast, residues in the interface of obligate complexes, proteins that may fold and bind simultaneously, evolve at a slower rate. This is perhaps because the interface residues of proteins in obligatory complexes coevolve with their interacting partners (Mintseris & Weng [2005]). A study has shown that multi–partner interfaces are not more conserved than other interface sites (Tyagi *et al.* [2009]). Further analysis of multi–partner protein interfaces suggests that the interfaces used to bind more than one protein are smaller and less packed than the interfaces of complexes with specific

partners (Keskin & Nussinov [2007]).

An atomic perception of protein-protein interfaces is necessary to further our understanding of the chemical and physical interactions that occur between residues in the interface. This information could be used to develop highly specific, small molecule, pharmaceutical interventions that disrupt or modify these interactions. Knowledge of these chemical and physical residue-residue interactions is also widely employed in the field of Synthetic Biology (Pleiss [2011]), and is used to predict the interacting partner(s) of a protein (Tuncbag *et al.* [2011]; Valencia & Pazos [2002]). Structures of protein complexes offer this level of atomic detail.

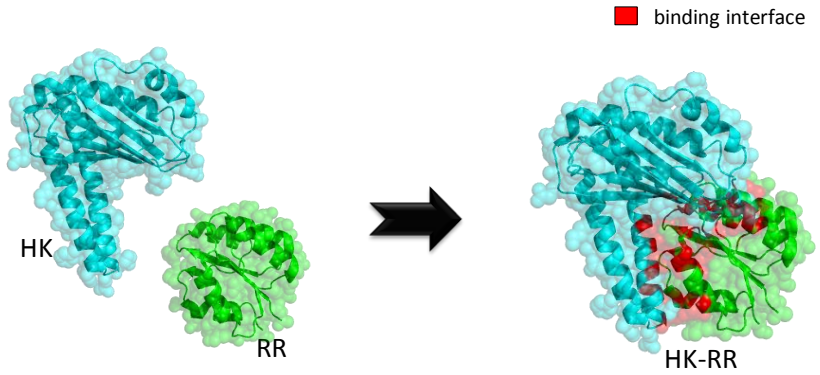

Figure 1.3: **Protein complex formation.** Physical interaction between a proxy histidine kinase (HK) and its response regulator (RR) to form an HK-RR complex (1F51.pdb (Zapf *et al.* [2000])). The area denoted in red constitutes the binding interface of this complex. In each protein, the cartoon ribbon depicts the backbone of the protein, and the cloud surrounding the ribbon represents the surface area of the protein that is accessible to solvent.

1.4.2 Crystallography complexes

In the PDB 89% of protein structures are experimentally determined via X-ray crystallography (29 August 2012) (Berman *et al.* [2000]). Figure 1.3 presents an example of a crystallographic complex. Protein crystallography, however, is extremely difficult, as highlighted by the fact that less than 1% of known sequences have resolved structures

(Section 1.2.2). The challenges associated with crystallising a protein are compounded when attempting to crystallise interacting proteins, protein complexes. The main limitations of this process are: (i) producing soluble protein for each of the proteins in the complex; (ii) making a soluble complex; (iii) generating sufficient soluble complexes at the high concentration levels necessary for the crystallisation pipeline (Shen *et al.* [2005]); (iv) determining the highly specific conditions suitable for growing crystals (Rhodes [2006]). Here "soluble" refers to proteins dissolved in solvent, such that they do not aggregate and precipitate out of solution. This in turn prevents the formation of well-ordered crystal lattices, which are necessary for X-ray diffraction.

Most proteins are insoluble when expressed in a non-native environment or when purified at high concentration levels, thus tailored protocols are required for each of the target proteins in complex. Additionally, PPIs are difficult to induce *in vitro* from individual protein components, and generation of sufficient protein complexes *in vivo* is almost impossible. Furthermore, PPIs are dynamic with the half-life of protein complexes varying greatly. The half-life of a protein complex in some cases is dependent on post-translational modifications made to the proteins involved. The post-translational modifications and the transient state of the complexes are difficult to support in the highly stable environment required to generate crystals (Shen *et al.* [2005]). Moreover, several conditions have to be tested to determine an environment in which crystals will grow. These crystallisation conditions are specific to the protein complex under consideration and include pH, purity and concentration of the complex, precipitants, additives, temperature, etc. (Rhodes [2006]). Achieving a suitable combination of conditions that yield crystals which diffract well is an extremely difficult task. It is also not possible to crystallise some protein complexes for various reasons, such as, disordered regions within the 3D structure of a protein (Bracken *et al.* [2004]). Consequently the amount of available X-ray crystallography protein complexes is low, which in turn limits our insight into the residue composition of binding interfaces of protein complexes.

1.4.3 NMR spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy is the second most popular form of protein structure elucidation, contributing 10.7% of all PDB protein structures (29 October 2012) (Berman *et al.* [2000]). Unlike crystallography it requires only a small volume of concentrated protein solution. This protein solution is then placed in a strong magnetic field and probed with radio waves. A major advantage of NMR over crystallography is that it provides information about the protein when in solution, rather than a rigid crystal lattice. Consequently, using this technique we can closely examine confirmation changes of proteins, as well as the flexible regions of a protein. However, NMR requires large amounts of experimental data per atom in a protein, making this method infeasible for structure elucidation of large proteins or protein complexes.

In this dissertation we use only crystallography structures to ensure analysis of a representative set of proteins of all sizes.

1.4.4 Contact residues

Residues within the binding interface (Figure 1.3) of a pair of interacting protein domains are labelled as "contact" residues. There is not one particular accepted definition for contact residues (Camacho *et al.* [1999]; Carugo & Argos [1997]; Halperin *et al.* [2006]). Most definitions consider pairs of residues to be in contact if the distance between all, or a particular pair of atoms in the respective amino acids (*e.g.* C_β-C_β), is less than a given distance, usually in the range of 4.5 to 9Å.

For the purpose of this dissertation we define two residues, one from each interacting protein domain, to be a "contact pair" if:

1. They are both on the surface of the individual protein domains.

2. The distance between any pair of atoms, one from each residue, is less than 4.5Å

(Carugo & Argos [1997]).

3. The solvent accessibility of the residues changes upon binding of the protein domains.

If the above criteria are not met the residues are denoted as a "non-contact pair" (Figure 1.4).

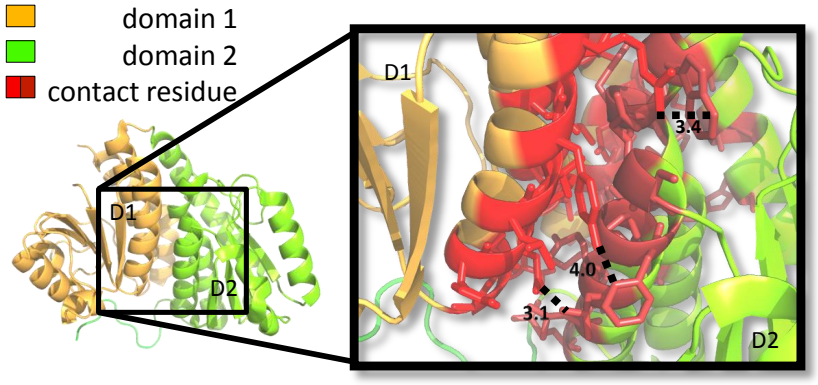

Figure 1.4: **Contact residues in a pair of interacting domains.** Test case 1J5X.pdb (Zapf *et al.* [2000]). The two structurally defined domains are depicted in orange (residue 2 to 169) and green (residue 170 to 319) respectively. In the magnified frame, residues in red denote contact residues. Dotted lines and corresponding numbers indicate the ångström distance between a pair of atoms in the connected residues.

When examining residues within a domain (intra-domain), we define "contact" residue pairs as surface or buried residues within the protein domain that are less than 4.5Å from each other (all residue atoms considered). Conversely, intra-domain "non-contacts" are those residue pairs that are at least 4.5Å distance from each other (taking all atoms into account). Neighbouring residues in the sequence will always be less than 4.5Å apart and are trivial intra-domain contacts. It has also been shown that adjacent residue pairs in sequence have high correlated mutation scores (Halperin *et al.* [2006]). Hence, like Lee & Kim [2009] we consider only residue pairs that are four or more residues apart in sequence.

1.5 Correlated Mutations

In this section we describe the theory of "correlated mutations," which lies at the heart of many structure and protein-protein interaction prediction algorithms. We go on to explain the newly popular "direct coupling analysis" methods that aim to improve correlated mutation-based predictions. We conclude by explaining why it should be plausible to use domain-domain interactions (DDIs) as proxies for protein-protein interactions (PPIs) in correlated mutation studies.

1.5.1 Correlated mutations: theory and analysis

It is theorised that the residues necessary for preserving the structure of a protein or protein complex, contact residues, coevolve in order to maintain the residue-residue interaction. That is, if a residue mutates and contact with an interacting residue partner is perturbed, the partner will more likely mutate to allow for physiochemical complementarity in order to maintain the residue-residue interaction. In this manner, the structure of the protein or protein complex is preserved (Figure 1.5). Turning this around, residues observed to coevolve are likely to form contacts. This theory is popularly known as "correlated mutations" (Halperin *et al.* [2006]; Jones *et al.* [2012]).

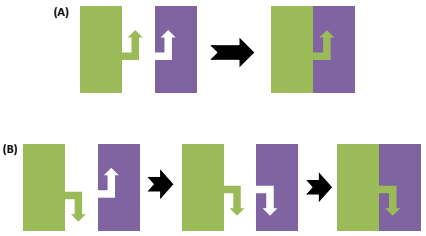

Figure 1.5: **Schematic of correlated mutations in interacting proteins.** (A) The green protein binds with the purple protein in a lock and key fit to form a protein complex. (B) The green protein in A undergoes a mutation. This prevents the purple protein from binding to it. The structure of the purple protein experiences a compensatory mutation to allow the green and purple proteins to once again form a complex.

Several techniques have been developed and have had some success in using the

idea of correlated mutations to predict intra- and inter- protein contact sites. These algorithms attempt to identify columns of an MSA that exhibit correlated amino acid compositions (Figure 1.6), and subsequently infer contact residues. Examples of such methods include Statistical Coupling Analysis (SCA) (Lockless & Ranganathan [1999]; Suel *et al.* [2003]), Explicit Likelihood of Subset Covariation (ELSC) (Dekker *et al.* [2004]), Observed Minus Expected Squared (OMES) (Fodor & Aldrich [2004]; Kass & Horovitz [2002]), McLachlan Based Substitution Correlation (McBASC) (Göbel *et al.* [1994]; Pazos *et al.* [1997]), H2r (Merkl & Zwick [2008]) and Mutual Information (MI) (Atchley *et al.* [2000]; Clarke [1995]; Dunn *et al.* [2008]; Martin *et al.* [2005]; Tillier & Lui [2003]). These methods are reviewed by Halperin *et al.* [2006]. To give some examples of how these methods work, OMES performs a chi-square test on every pair of columns, searching for amino acids that occur more frequently than expected. The null hypothesis is that there is no correlation between columns. ELSC on the other hand calculates the probability that a random subset of the alignment has the observed amino acid profile at a given site. As a final example, SCA calculates a conservation weighted covariance matrix of pairs of columns in the alignment.

It should be noted that some correlated mutation analysis measures employed by the various algorithms have been given the name "correlated mutation score" (Chakrabarti & Panchenko [2009]; Halperin *et al.* [2006]).

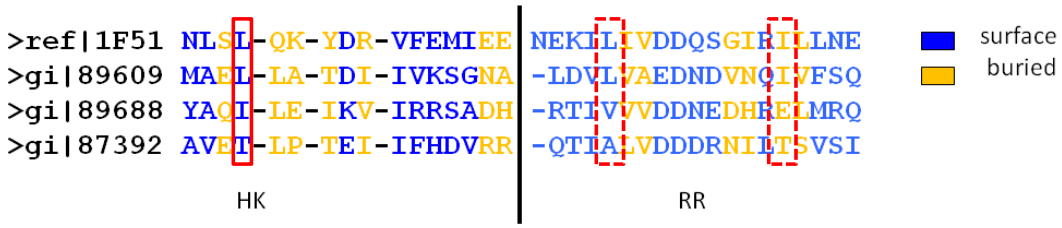

Figure 1.6: **Depiction of correlated changes in a multiple sequence alignment.** Excerpt of a histidine kinase (HK) response regulator (RR) MSA, taken from Hamer *et al.* [2010]. Homologous HKs are retrieved and aligned, in the same manner as the RRs. The two alignments are then concatenated such that each HK is joined to its interacting RR. The blue and yellow colours denote surface and buried residues respectively, as determined from protein structure 1F51.pdb (Zapf *et al.* [2000]). The amino acid compositions of the selected HK residue column (solid box) and the indicated RR columns (dashed boxes) are correlated. All three marked columns have identical residues in the first two sequences and dissimilar residues in sequences three and four.

1.5.2 Direct coupling analysis

It should be noted that not all residues that undergo correlated mutations are necessarily in contact. For example, residue A mutates, residue B which is in contact with residue A then mutates in order to maintain physiochemical complementarity. Residue C which is in close proximity to B also mutates as a result. Therefore direct correlated mutation coupling effects in contact residue pair AB will result in indirect coupling effects in the non-contact residue pair AC. Disentangling the direct from the indirect coupling has been likened to the inverse Ising problem in statistical physics by Lapedes *et al.* [1999]. That problem can be solved using maximisation of entropy.

Recently, methods attempting to distinguish direct from indirectly coupled correlated mutation scores have received a lot of attention. Weigt *et al.* developed a novel algorithm based on entropy maximisation entitled, "message-passing Direct Coupling Analysis (mpDCA)" (Lunt *et al.* [2010]; Weigt *et al.* [2009]). These authors used message-passing to estimate the marginal distributions for single and multiple positions in the MSA (Lunt *et al.* [2010]; Weigt *et al.* [2009]). The estimated distributions are then used in their formulated Direct Information (DI) measure.

Unlike MI which is a pairwise measure, DI seeks to model the joint distribution of all

residues, $P(A_1, ..., A_N)$, via maximising the entropy for residues A_1 to A_N. The entropy maximisation can be formulated by a Boltzmann distribution with pairwise couplings $e_{ij}(A_i, A_j)$ and local biases $h_i(A_i)$ (the preference of amino acid A_i at position i):

$$P(A_1, ..., A_N) = \frac{1}{Z} \prod_{i<j} exp\{-e_{ij}(A_i, A_j)\} \prod_i exp\{h_i(A_i)\}. \tag{1.1}$$

Z in the above equation is the partition function, which is described as

$$Z = \sum_{\{A_i\}} \prod_{i<j} exp\{-e_{ij}(A_i, A_j)\} \prod_i exp\{h_i(A_i)\}. \tag{1.2}$$

Here $e_{ij}(A_i, A_j)$ and $h_i(A_i)$ are once again the pairwise couplings and local biases, respectively, of amino acid A_i at position i (Weigt *et al.* [2009]).

Parameters for this model are determined iteratively as follows:

1. For a given energy function formulated in Equation 1.1, marginal distributions for single positions and pairs of positions are calculated. These are approximated by message-passing.

2. Using gradient descent, parameters in Equation 1.1 are updated according to the difference of the estimated marginals and frequencies of amino acid occurrences calculated from the data.

When starting this iterative procedure $e_{ij}(A_i, A_j)$ is set to 0.

$DI(i, j)$, the direct information, measures only the contribution introduced by the direct link (i, j) to the correlation between the corresponding amino acids i and j. It is essentially similar to using the norm $||e_{ij}||^2 = \sum_{A_i,A_j} [e_{ij}(A_i, A_j)]^2$ to measure the coupling strength.

In order to assess the ability of mpDCA to identify directly coupled correlated mutations, Weigt *et al.* tested their method on a bacterial two component signalling protein-protein interaction, namely histidine kinase with its interacting response reg-

ulator (Weigt *et al.* [2009]). Due to the high computational complexity of mpDCA, the less computationally expensive correlated mutation algorithm, Mutual Information (MI), was used to limit the number of MSA column pairs that should be analysed with mpDCA. In the histidine kinase - response regulator MSA 60 columns were identified in the top 140 MI scores; *i.e.* 60 MSA columns contributed to the top 140 MI pair scores. When analysis is confined to these 60 columns, MI falsely assigns the second highest score to a non-contact pair, conversely the top 10 mpDCA scores (of 408 scores) relate to contact residue pairs.

In order to reduce the computational complexity of mpDCA, Weigt and colleagues replaced the slow converging message-passing approach by a heuristic algorithm based on a mean-field approach, and renamed the pipeline "mfDCA" (Morcos *et al.* [2011]). For a protein with the same number of residues in its sequence, mfDCA is approximately 10^3 to 10^4 times faster than mpDCA (Morcos *et al.* [2011]). Hence mfDCA does not have to be restricted to just 60 columns and one test case, as in the previous mpDCA study. When using a test set of 131 predominantly bacterial MSAs (Morcos *et al.* [2011]), the new mfDCA algorithm successfully predicted residue pairs between 3-5Å and 7-8Å within the top 10 scores (Morcos *et al.* [2011]).

Around the same time as mfDCA's release, Jones *et al.* [2012] published PSICOV, another correlated mutations decoupling technique. Unlike mpDCA and mfDCA, which are based on entropy maximisation, PSICOV is formulated on sparse inverse covariance estimation. Jones *et al.* borrowed the phylogenetic and entropic noise normalisation metric used in the Mutual Information variant MIp (Dunn *et al.* [2008]) and incorporated it in PSICOV. This novel method evaluates an MSA in a median of 30 minutes, with the evaluation time for their 150 protein domain test cases varying from 1 to 240 minutes (Jones *et al.* [2012]). For 44% of the test cases PSICOV correctly predicted more than 50% contacts per residue. Jones *et al.* showed that PSICOV outperforms a Bayesian network approach to decoupling (Burger & van Nimwegen [2010]), as well

as MIp (Dunn *et al.* [2008]), in intra-protein domain contact prediction (Jones *et al.* [2012]).

In the last year these direct coupling analysis methods have started being used for protein structure prediction.

Marks *et al.* [2011] used mfDCA within their EVfold pipeline to determine the structure of proteins from sequence alone. They use the contact residues predicted by mfDCA as constraints when using distance geometry and simulated annealing to fold the protein sequence. For 15 test proteins, with different types of folds, they are able to determine the 3D structure of the protein to within 2.7-4.8Å C_α-rmsd of the known structure, over at least two-thirds of the protein. The developers of this algorithm then extended EVfold to predict the structures of 11 membrane proteins that have been experimentally determined, and called this extension EVfold_membrane (Hopf *et al.* [2012]). They achieved an overall C_α-rmsd of 4-5Å across all residues in the 11 proteins, suggesting that although the protein folds were correctly identified, the predicted packing of side chains and loops need improvement (Hopf *et al.* [2012]).

Additionally, mfDCA has been used within the novel DCA-fold method that also aims to determine protein structure from sequence. This time the mfDCA predicted contacts are used as constraints when attempting to fold the protein using a modified structure-based model (Sułkowska *et al.* [2012]). This methodology claims to have the capacity to predict the structure of proteins, up to approximately 200 amino acids in length, to within 3Å of their native structures.

The caveat of employing these correlated mutations decoupling algorithms is requiring MSAs that have more than 1,000 sequences as input in order to estimate the model parameters. Even in this post-genomic era having more than 1,000 homologous sequences for a target protein is rare. For instance, in the Pfam-A protein domain family database (version 26.0; downloaded August 2012; Punta *et al.* [2012]) only 22.7% domain families have more than 1,000 sequences in their MSA (*i.e.* 3,110 out of 13,672

families). If we were to extend these direct coupling analysis measures to assess higher correlations, between three or more residue columns in the MSA, the minimum number of sequences that would be required would be 10,000 and higher, making contact prediction infeasible given the amount of data we currently have (Weigt *et al.* [2009]).

In Chapter 6 of this dissertation we provide a comparison of the contact residue prediction ability of a DCA measure, namely PSICOV, against the MI based method that performs best in our analyses, MIc (Lee & Kim [2009]).

1.5.3 Domain-domain interactions as a proxy for protein-protein interactions

Unfortunately the seemingly straightforward approach of building MSAs by separately finding homologous sequences of the two proteins known to be in complex, and then pairing the sequences originating from the same species, is not feasible for the following reasons. Inferring protein-protein interactions (PPIs) across species based on sequence homology has a low level of accuracy, requiring a sequence identity of far higher than 70% (Lewis *et al.* [2010]; Mika & Rost [2006]). Using only sequences with more than 70% identity would result in MSAs with a low number of sequences and few amino acid changes, not sufficient enough to yield statistically significant correlated mutation scores (Martin *et al.* [2005]). Furthermore, within a species there may be multiple homologs of the interacting proteins, and selecting the correct pairs out of this is an unsolved problem (Mika & Rost [2006]). Hence previous protein-protein contact residue prediction investigations have chosen to use domain-domain interactions (DDIs) as a proxy for PPIs (Hamer *et al.* [2010]; Pazos *et al.* [1997]).

Proteins are composed of one or more domains. A multidomain protein, protein A, will often use one of its domains to bind to protein B and another to bind to protein C, thus allowing protein A to perform multiple functions. Consequently, in reality PPIs are often DDIs (Pagel *et al.* [2004]). Thus using DDIs within proteins as a representative

for PPIs ensures that interacting "proteins" are accurately paired in each MSA, while capturing interaction mechanisms.

Hamer *et al.* [2010] have shown that the propensities of amino acids to occur as contact residues between two domains in a protein and between two proteins in complex are highly similar. While it is plausible to use DDIs as a proxy for inter-protein interactions, it is however possible that protein-protein interfaces may indeed differ from domain-domain interfaces.

In this dissertation we employ proteins of known structure, with two domains, that have several homologous sequences available to build MSAs. We hope that our findings will shed light on both DDIs and PPIs.

1.6 Shannon Entropy

In order to quantify residue conservation in columns of an MSA, *i.e.* the amount of disorder in a column, variations of Shannon Entropy are often employed (Merkl & Zwick [2008]; Pirovano *et al.* [2006]; Schneider & Stephens [1990]; Valdar [2002]; Yeo & Burge [2004]). The basic form of Shannon entropy is denoted by equation 1.3.

$$H_{unstandardised}(J) = -\sum_{j=1}^{n} P(J=j)\log P(J=j). \tag{1.3}$$

In this equation J is a random variable with probabilities $P(J=j)$ for a discrete set of n events $j_1, ..., j_n$. The product $P(J=j)\log P(J=j)$ is taken to be zero if $P(J=j)=0$. The entropy is maximum when all j are equally likely to occur, *i.e.* $P(J=j)=1/n$ and $H_{unstandardised}(J) = -\sum \frac{1}{n}\log\frac{1}{n} = \log n$ (Durbin *et al.* [1998]). Entropy thus measures disorder. Commonly log base 2 is used making the unit of entropy a 'bit'.

In order to compare the entropies from different MSAs we standardise the entropy

score as follows:

$$H(J) = \frac{H_{unstandardised}(J) - \overline{H}_{unstandardised}}{\sigma_{H_{unstandardised}}}, \tag{1.4}$$

where $H_{unstandardised}(J)$ is the entropy of column J in the MSA, and $\overline{H}_{unstandardised}$ and $\sigma_{H_{unstandardised}}$ are the average entropy and estimated standard deviation, respectively, over all columns in the MSA combined.

The Shannon Entropy for an excerpt of the HK-RR MSA taken from Hamer *et al.* [2010] is graphically displayed in the logo (Schneider & Stephens [1990]) in Figure 1.7. The residue position with the highest entropy is the shortest stack and the shortest amino acid in a stack is the least conserved residue in that MSA column. The colours denote seven biochemical groupings of amino acids as defined by Hamer *et al.* [2010]. Let us consider stack 7 in Figure 1.7. R is the most conserved residue in the corresponding MSA column followed by L. The MSA column corresponding to stack 7 is more conserved than the column corresponding to stack 12 in Figure 1.7, as indicated by the greater height of stack 7 in comparison to stack 12.

The joint Shannon Entropy of two columns J and K is defined as:

$$H(J;K) = -\sum_{j=1}^{n}\sum_{k=1}^{m} P(J=j, K=k)\log P(J=j, K=k), \tag{1.5}$$

where column J has n different residues, and column K has m different residues.

1.7 Mutual Information

In this dissertation we analyse Shannon Entropy-based MI methods for detecting contact residues in proteins.

The general MI formula is:

$$MI(J;K)_{unstandardised} = H_{unstandardised}(J) + H_{unstandardised}(K) - H(J;K), \tag{1.6}$$

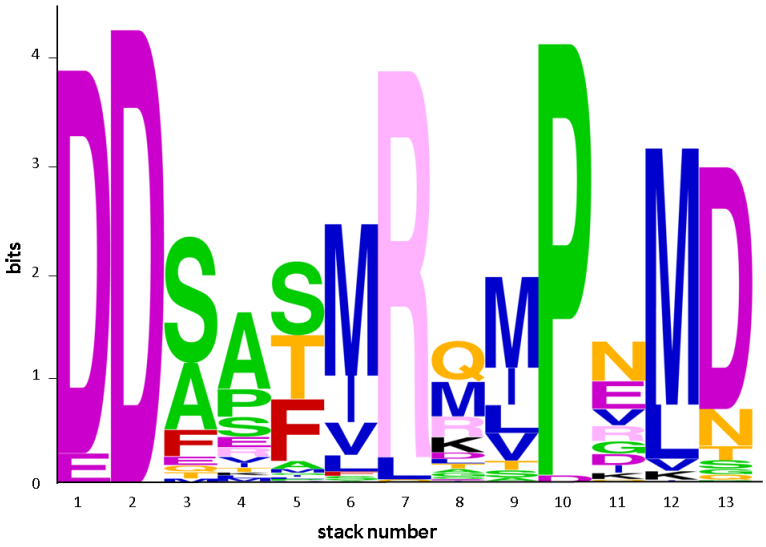

Figure 1.7: **Logo of an MSA of a subset of residues in a response regulator.** The graph contains one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position ($\log_2 20 - H_{unstandardised}(J)$), while the height of symbols within the stack indicates the relative frequency of each amino acid at that position. The amino acids are colour coded according to seven physiochemical categories as defined by Hamer *et al.* [2010]: small (S,G,A,P) - green, hydrophobic (V,M,I,L,C) - blue, negatively charged (D,E) - purple, aromatic (F,Y,W) - red, polar (Q,T,N) - orange, favoured positively-charged (R,H) - pink, disfavoured positively charged (K) - black. Hamer *et al.* introduced the disfavoured and favoured positively charged categories as Lysine (K) was found to be rare in protein/domain interfaces, while Arginine (R) and Histidine (H) were far more common. This sequence logo (Schneider & Stephens [1990]) was generated using an online tool developed by Pirovano *et al.* [2006].

where $H_{unstandardised}(J)$ and $H_{unstandardised}(K)$ are calculated as described in Equation 1.3, and $H(J;K)$ is computed as denoted in Equation 1.5.

MI is used to measure how related two random variables are. It asks the question *how* dependent two random variables J and K are, by measuring the relative entropy 'distance' between the distributions $P(J = j, K = k)$ and $P(J = j)P(K = k)$ (Durbin *et al.* [1998]). The applications of MI are far-reaching; from measuring the information carrying capacity of communication channels, to solving data analysis and estimation problems in image processing, speech recognition and signal processing (Togneri & DeSilva [2003]). For example, in signal processing MI has been successfully employed to separate mutually interfering signals (Togneri & DeSilva [2003]). MI has been used for several bioinformatics applications such as, RNA secondary structure prediction (Bindewald & Shapiro [2006]), phylogenetics (Atchley *et al.* [2000]; Clarke [1995]; Korber *et al.* [1993]), and transcription factor binding site analysis (Tomovic & Oakeley [2007]). It has also been employed to analyse protein coevolution (Dunn *et al.* [2008]; Fernandes & Gloor [2010]; Martin *et al.* [2005]; Wollenberg & Atchley [2000]). In these bioinformatics applications MI is used to identify columns in an MSA that may contain correlated mutations.

MI was first applied to sequence alignments by Korber *et al.* to identify covarying positions in a viral peptide (Korber *et al.* [1993]). We hypothesise that MI has since gained popularity because it is non-parametrised, *i.e.* the scores of MI are solely dependent on an MSA and no additional information, such as a phylogeny propensity table (Hamer *et al.* [2010]), a similarity matrix (Göbel *et al.* [1994]; Pazos *et al.* [1997]) or other is required. Furthermore, unlike other algorithms that predict contact "patches" (Bradford & Westhead [2005]; Xu & Tillier [2010]), or individual contact residues (Davis [2011]; Zhang *et al.* [2010]), MI attempts to predict specific pairs of residues that are in contact with one another.

Horner *et al.* [2008] have collated the accuracies of intra-protein contact residue pre-

diction from several publications that employ correlated mutation analysis algorithms, and showed that MI has an accuracy between 2 and 18%. Accuracy here refers to the percentage of predictions that are correct. The low accuracy of MI has in turn precipitated many variants that attempt to improve its performance. These variants specifically attempt to correct for the following three recognised limitations of MI: highly variable MSA columns, phylogenetic relationships and insufficient sequences in the MSA. There is evidence that columns in the MSA that have a high variability contribute to random and non-random high MI scores (Fodor & Aldrich [2004]; Martin *et al.* [2005]), while phylogenetic relationships (Wollenberg & Atchley [2000]) and insufficient number of sequences in the MSA (Martin *et al.* [2005]) weaken the signal detection ability of MI. In 1995, Clarke corrected the MI score by a measure relating to the number of amino acid pairs occurring at each position to negate the influence of highly diverged sequences that may be inappropriately aligned in the MSA (Clarke [1995]). Later, Wollenberg and Atchley used parametric bootstrapping to adjust for evolutionary relationships (Wollenberg & Atchley [2000]). Tillier and Lui designed a tool which removes columns in an MSA that carry a high phylogenetic signal and then employs MI to try to identify positions in the resulting MSA that coevolve with each other, but do not coevolve significantly with other positions (Tillier & Lui [2003]). As performance was still disappointing, Martin *et al.* attempted to remove the noise caused by entropy, by dividing the MI score of a pair of columns by their joint entropy (Martin *et al.* [2005]). These authors also suggested that a minimum of 125 sequences should be used in an MSA to reduce stochastic noise. Dunn *et al.* improved on this score by introducing MIp, which modified the MI value by a measure that aims to eliminate phylogenetic and entropic effects (Dunn *et al.* [2008]). Subsequently, Lee and Kim introduced two other powerful phylogenetic noise reduction MI measures, MIc and aMIc (Lee & Kim [2009]). In 2010 Brown and Brown suggested yet another MI measure, ZNMI, that accounts for different alphabet sizes among columns in the MSA (Brown

& Brown [2010]). These authors also proposed a pipeline to yield highly reproducible scores. Despite all of these efforts, to date no single MI measure has achieved general utility or wide acceptance for predicting intra-protein contact sites.

MI has begun to be extended to predict inter-protein contact residues. Halperin *et al.* carried out a small study of original MI and other correlation algorithms on 15 bacteria and archea fusion protein families (Halperin *et al.* [2006]), and Lee and Kim evaluated their MI measures on a specialised dataset of 27 homo-trimers (Lee & Kim [2009]). There have also been several high profile case studies on small datasets (one to three cases), such as Brown & Brown [2010]; Dunn *et al.* [2008]; Little & Chen [2009]; Martin *et al.* [2005] and Skerker *et al.* [2008]. Recently, we published the first systematic study on a large, general purpose, cross-species dataset of the performance of MI and its latest variants on inter-protein contact residue prediction, using domain-domain interactions as a proxy for protein-protein interactions (Gomes *et al.* [2012]).

In our work we evaluate MI measures that do not require any additional information and rely solely on the sequence alignment itself; focusing on the original MI and its most recent extensions MIp, MIc, aMIc, as well as ZNMI, alongside our own 3D and reduced alphabet MI variants.

1.7.1 Original MI

MI is calculated as described in Equation 1.6. The maximum MI occurs when residues in columns J and K always covary, *i.e.* $P(J = K) = 1$ making the $MI = -\sum_{j=1}^{n} P(J = j) \log P(J = j)$. The maximal MI that can be achieved for protein sequences, which have 20 varying residues, is $log_2 20 \simeq 4.32$ (Durbin *et al.* [1998]).

Note that $P(J = j)$ and $P(K = k)$, which are used to calculate $H_{unstandardised}(J)$ and $H_{unstandardised}(K)$ (Equation 1.3) components of the MI formula (Equation 1.6), and $P(J = j, K = k)$ that is used to calculate the $H(J; K)$ component (Equation 1.5), are the relative observed frequencies in a finite data set, the MSA being considered.

We could interpret them as probabilities of randomly drawn sites to take on these amino acids. Different methods of estimating these probabilities, which may or may not account for sampling variance, result in dissimilar estimated MI scores that only converge on extremely large data sets (Fernandes & Gloor [2010]). However Fernandes & Gloor [2010] have shown that despite the discrepancy, all methods are biologically relevant estimates of MI, for each method is based on equally plausible hypotheses and is effectively the Kullback-Leibler divergence between two random variables. Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions on the same set (Cover & Thomas [1991]).

We calculate "standardised MI scores" for each protein, so that MI values of different proteins can be compared. In order to do this the average and estimated standard deviation of the MI of all residue pairs in the protein are calculated. The "standardised MI score" is formulated as

$$MI(J;K) = \frac{MI_{unstandardised}(J;K) - \overline{MI}_{unstandardised}}{\sigma_{MI_{unstandardised}}}. \tag{1.7}$$

Here $MI_{unstandardised}(J;K)$ is the MI of columns J and K in the MSA, and the average and estimated standard deviation of MI for all considered column pairs in the MSA are denoted as $\overline{MI}_{unstandardised}$ and $\sigma_{MI_{unstandardised}}$ respectively.

If the pair of columns under consideration contains only gapped pairs, *i.e.* one or two gaps ('-') in the residue pair, we assign a "not a number" (nan) MI score (Equation 1.6) to that column pair. In other words, if a pair of columns are only made up of residue pairs (X,-), (-,X) and/or (-,-), where X represents any amino acid, the MI score assigned to that column pair is nan. These nan scores are omitted in further calculations. We extend this method of handling pairs of columns with only gapped pairs to our calculations of MIp, MIc and aMIc which follow (Equations 1.10, 1.14 and 1.18).

1.7.2 MIp

Dunn *et al.* designed an MI variant that aims to correct for background (random and phylogenetic) noise, MIp (Dunn *et al.* [2008]). This MI correction is denoted by the equation

$$MIp_{unstandardised}(J;K) = MI_{unstandardised}(J;K) - APC(J;K), \tag{1.8}$$

where $MI_{unstandardised}(J;K)$ is evaluated as denoted in Equation 1.6. $APC(J;K)$, the average product correction, is an adjustment term for columns J and K in the MSA, calculated as follows,

$$APC(J;K) = \frac{\overline{MI}_{unstandardised}(J) \times \overline{MI}_{unstandardised}(K)}{\overline{MI}_{unstandardised}}. \tag{1.9}$$

Here the average mutual information for column J is denoted by $\overline{MI}_{unstandardised}(J)$, the average mutual information for column K is denoted by $\overline{MI}_{unstandardised}(K)$, and $\overline{MI}_{unstandardised}$ is the overall average mutual information.

MIp scores for each protein are standardised in a manner similar to MI (Equations 1.7), so that MIp values from different proteins can be compared,

$$MIp(J;K) = \frac{MIp_{unstandardised}(J;K) - \overline{MIp}_{unstandardised}}{\sigma_{MIp_{unstandardised}}}. \tag{1.10}$$

Here $MIp_{unstandardised}(J;K)$ is the MIp of columns J and K in the MSA, whereas $\overline{MIp}_{unstandardised}$ and $\sigma_{MIp_{unstandardised}}$ are the average MIp and estimated standard deviation respectively, over all column pairs being considered in the protein.

1.7.3 MIc

Lee and Kim formulated normalising measures that also attempt to reduce phylogenetic noise in MI scores (Lee & Kim [2009]). Their first metric, the coevolutionary pattern

similarity score (CPS), measures the similarity between the MI scores of the two residues being considered:

$$CPS(J;K) = \frac{1}{n-2} \sum_{L \neq J;K} MI_{unstandardised}(J;L) MI_{unstandardised}(K;L). \tag{1.11}$$

In this equation $MI_{unstandardised}(J;L)$ is the mutual information score of the columns J and L, and is calculated as described in Equation 1.6. The number of columns being considered in the MSA are denoted by n.

Since CPS is the product of two MI scores, a normalising term is necessary. The authors use the square root of the mean of all CPS scores,

$$NCPS(J;K) = \frac{CPS(J;K)}{\sqrt{\frac{1}{n(n-1)} \sum_{J,K} CPS(J;K)}}. \tag{1.12}$$

The original MI pair score is then corrected by the corresponding NCPS score to yield Lee and Kim's noise reduced MI variant, MIc,

$$MIc_{unstandardised}(J;K) = MI_{unstandardised}(J;K) - NCPS(J;K). \tag{1.13}$$

As done for previous MI variants (Equations 1.7 and 1.10), MIc scores for each protein are standardised to allow for comparison of scores of the different proteins.

$$MIc(J;K) = \frac{MIc_{unstandardised}(J;K) - \overline{MIc}_{unstandardised}}{\sigma_{MIc_{unstandardised}}}, \tag{1.14}$$

where $MIc_{unstandardised}(J;K)$ is the MIc of columns J and K in the MSA, and the average of MIc and its estimated standard deviation for all calculated column pairs in the protein are denoted by $\overline{MIc}_{unstandardised}$ and $\sigma_{MIc_{unstandardised}}$ respectively.

The code made available by Lee and Kim (Lee & Kim [2009]) includes the nan MI scores, resulting from columns with all gapped pairs, in their CPS calculations

(Equation 1.11). This in turn produces nan CPS values, which causes the denominator of NCPS to be nan (Equation 1.11). Subsequently all NCPS scores for the alignment will be nan (Equation 1.11), and consequently all MIc values for the MSA will also result in nan (Equation 1.13). To avoid this loss of information we wrote a version of the code that ignores all nan MI values when calculating the CPS (Equation 1.11) thus yielding a greater number of valid MIc scores. We only observed this loss of information in our final piece of work (Chapter 5). Hence all the work described in Chapters 2, 3 and 4 employ the original code provided by Lee & Kim [2009], while the work in Chapter 5 uses our revised code.

1.7.4 aMIc

In the same study Lee and Kim attempted to further remove background noise by accounting for column entropy in their additionally normalised measure, aMIc (Lee & Kim [2009]). They begin by calculating the entropic factor $E(J;K)$ as follows

$$E(J;K) = H_{unstandardised}(J)H_{unstandardised}(K)(1 - H_{unstandardised}(J)H_{unstandardised}(K)), \tag{1.15}$$

where $H_{unstandardised}(J)$ and $H_{unstandardised}(K)$ are the entropies of columns J and K calculated as described in Equation 1.3. This entropic factor $E(J;K)$ is then used to adjust the $MIc_{unstandardised}(J;K)$ score of columns that have extreme entropy, as described in the equation below.

$$eMIc(J;K) = E(J;K)MIc_{unstandardised}(J;K), \tag{1.16}$$

where $E(J;K)$ is the entropic factor (Equation 1.15) and $MIc_{unstandardised}(J;K)$ is the MIc score (Equation 1.13) for columns J and K in the MSA. The aMIc score then normalises the unstandardised MIc and eMIc scores of the column pair, by the maximum unstandardised MIc and eMIc scores of all considered column pairs in the

MSA. This is denoted by

$$aMIc_{unstandardised}(J;K) = \frac{1}{2}\left[\frac{MIc_{unstandardised}(J;K)}{max(MIc_{unstandardised})} + \frac{eMIc(J;K)}{max(eMIc)}\right]. \quad (1.17)$$

Here $MIc_{unstandardised}(J;K)$ and $eMIc(J;K)$ are calculated as described by Equations 1.13 and 1.16, respectively, for MSA columns J and K.

The aMIc scores are standardised in a manner similar to MI, MIp and MIc (Equations 1.7, 1.10 and 1.14), to allow for score comparison across proteins.

$$aMIc(J;K) = \frac{aMIc_{unstandardised}(J;K) - \overline{aMIc}_{unstandardised}}{\sigma_{aMIc_{unstandardised}}}, \quad (1.18)$$

where $aMIc_{unstandardised}(J;K)$ is the aMIc of columns J and K in the MSA and $\overline{aMIc}_{unstandardised}$ and $\sigma_{aMIc_{unstandardised}}$ are the average aMIc and estimated standard deviation respectively, over all column pairs being considered in the protein.

As with MIc (Section 1.7.3), in Chapter 5 we employ our revised version of the code that calculates MIc ignoring the nan scores from CPS calculations. This results in a greater number of valid aMIc scores than generated by the authors' original code for the same test cases.

1.7.5 ZNMI

Since the MI score for a pair of residues is highly variable depending on the sequences included in an MSA, Brown and Brown designed a novel MI measure, ZNMI, as well as a methodology to yield reproducible and accurate contact pair prediction scores (Brown & Brown [2010]). Their suggested algorithm repeatedly partitions the MSAs into 50% sub-alignments, calculates the pair scores, retains significant scoring pairs for each partition and subsequently compares all partitions to acquire consensus pair scores. Through personal correspondence the authors provided us code for MI (Durbin *et al.* [1998]), MIp (Dunn *et al.* [2008]), OMES (Fodor & Aldrich [2004]; Kass & Horovitz

[2002], SCA (Halabi *et al.* [2009]), ZNMI (Brown & Brown [2010]) and ZRES (Little & Chen [2009]), the correlated mutation measures wrapped within their proposed pipeline, but unfortunately not for MIc and aMIc.

1.8 i-Patch

i-Patch is a leading inter-protein contact predictor, which unlike MI uses pre-calculated residue propensities. A brief overview of i-Patch follows, a more detailed explanation can be found in Hamer *et al.* [2010].

The propensities of residues, pairs of residues and triangles of residues that are in contact were calculated using a set of multidomain proteins and complexes with known structure. In order to reduce the size of these propensity tables the residues were grouped by physiochemical properties into seven categories: small (S,G,A,P), hydrophobic (V,M,I,L,C), negatively charged (D,E), aromatic (F,Y,W), polar (Q,T,N), favoured positively-charged (R,H), and disfavoured positively charged (K). Hamer *et al.* introduced the disfavoured and favoured positively charged categories as Lysine (K) was found to be rare in protein/domain interfaces, while Arginine (R) and Histidine (H) were far more common.

These pre-calculated propensities are incorporated into the i-Patch score. i-Patch considers only residues on the surface of the two interacting proteins. Each surface residue is scored using its propensity to be in contact weighted by its (intra-protein) neighbouring surface residues. Additionally, i-Patch employs a triangle score that is based on the idea that residue interactions occur in patches and not simply between pairs of residues (Madaoui & Guerois [2008]). In an assessment on a blind dataset of 31 inter-protein test cases i-Patch was found to outperform all other tested predictions achieving a precision of 59% at 20% recall.

Like all MI variants, the original version of i-Patch uses a multiple sequence align-

ment (MSA) containing homologous sequences of the interacting proteins to predict contact residues. However, it has been observed that the precision of i-Patch is approximately the same with or without the MSA. The use of the pre-calculated propensity tables and the sequences of the two target proteins, results in predictions of similar accuracy to the propensity tables used with an MSA of the target protein complex (personal communication with the developers of i-Patch, 20 September 2012).

In this dissertation we compare the contact prediction abilities of the MI variants to i-Patch (Hamer *et al.* [2010]). We observe that i-Patch, a purely statistical measure having no biological underpinnings, outperforms all MI variants. Hence we attempt to improve MI based inter-protein contact prediction by integrating with the MI calculations some of the heuristics and assumptions that have contributed to the success of i-Patch; specifically, grouping residues into physiochemical categories, only considering residues on the surface of the proteins, and incorporating the idea that residue interactions occur in patches.

1.9 Performance Evaluation Measures

In order to assess how well a classifier such as MI is correctly identifying contact residues, metrics based on a confusion matrix are employed. This matrix records the correctly and incorrectly classified cases. A confusion matrix for binary classification is depicted in Table 1.1. Here TP denotes the number of true positives, FP the number of false positives, TN the number of true negatives and FN the number of false negatives counts.

	Actual Positive	**Actual Negative**
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 1.1: **Confusion matrix for binary classification.** The number of correctly and incorrectly identified cases by a classifier. TP, FP, TN and FN indicate the number of true positive, false positive, true negative and false negative counts, respectively.

In our investigation the MI scores are ranked, and all scores over a certain percentile are considered to be positive, a contact residue pair, while the lower scores are classified as negative, a non-contact pair. Since each of the MSAs contains a representative structure, we can correctly identify the TP, FP, TN and FN scores.

Popular performance evaluation measures include accuracy, sensitivity, specificity and ROC-curves (Liu *et al.* [2010]). However when the number of positive and negative cases is disproportionate, like the ratio of contact to non-contact residue pairs in protein structure, P-ROC (Precision Recall Operating Characteristic) (Buckland & Gey [1994]) and MCC (Matthews Correlation Coefficient) (Matthews [1975]) curves provide an alternative to ROC (Receiver Operating Characteristic) (Fawcett [2006]) curves for assessing the performance of a classifier. In this section we describe the widely employed ROC-curve, and explore some of the performance evaluation measures that assess classifiers that work with disproportionate binary classes.

1.9.1 ROC-curves

As the discrimination percentile cutoff is varied, the true positive rate (TPR), or sensitivity (SN), is plotted against the false positive rate (FPR) in a receiver operating characteristic (ROC) curve (Fawcett [2006]). Here

$$TPR = SN = \frac{TP}{TP + FN}, \tag{1.19}$$

and

$$FPR = \frac{FP}{FP + TN}. \tag{1.20}$$

TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negatives, respectively.

In a ROC-curve a perfect classifier would produce a point at (0,1) indicating no

false positives and no false negatives. A random prediction would fall on the diagonal from (0,0) to (1,1). A good classifier would score above the diagonal, while a poor classifier would score below the diagonal (Fawcett [2006]).

In cases when the classifier identifies very few true positives and false positives, such as when discriminating between the low ratio of contact to non-contact protein residue pairs, the TPR and FPR are not very informative. Hence we turn to P-ROC curves (Buckland & Gey [1994]).

1.9.2 P-ROC curves

P-ROC (Precision Recall Operating Characteristic) curves plot *precision* against *recall* (Buckland & Gey [1994]), where

$$precision = \frac{TP}{TP + FP}, \tag{1.21}$$

and

$$recall = TPR = SN = \frac{TP}{TP + FN}. \tag{1.22}$$

A flat horizontal line in a P-ROC plot at $\frac{TP}{total\ scores}$ denotes the probability of randomly discriminating positive *versus* negative cases; $total\ scores = TP + TN + FP + FN$.

1.9.3 MCC curves

The values in the MCC (Matthews Correlation Coefficient) curve always range from -1 to +1, where -1 signifies total disagreement between predicted and actual classification and +1 indicates total agreement. An MCC of 0 denotes random prediction. MCC is

calculated as follows,

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}. \tag{1.23}$$

1.9.4 N×MCC^2 curves

MCC may be unable to accurately assess a classifier when the prediction algorithm indicates both few TPs and FPs (Baldi *et al.* [2000]); for example, at the 100^{th} percentile only one score, the highest score, can be allocated as either a TP or FP. In order to determine whether the MCC value is significantly better than random, 0, Baldi *et al.* [2000] recommend using a chi-squared test applied to the 2×2 confusion matrix containing TP, FP, TN and FN (Table 1.1). This test statistic is formalised such that,

$$\chi^2 = N \times MCC^2, \tag{1.24}$$

where N is the total number of scores being considered and MCC is calculated as described in Equation 1.23. 3.84 is the 95^{th} percentile of a χ^2 distribution with one degree of freedom. In the N×MCC^2 plots we draw a horizontal line at 3.84.

1.9.5 F-measure values

Often the maximum F-measure is selected as the optimal trade-off point between the sensitivity (SN) and specificity (SP) of a classifier (Liu *et al.* [2010]). The F-measure is calculated as

$$F - measure = \frac{2 \times SN \times SP}{SN + SP}. \tag{1.25}$$

Sensitivity is evaluated as described in Equation 1.19. Specificity on the other hand is described as,

$$SP = \frac{TN}{TN + FP}. \tag{1.26}$$

TN and FP represent true negative and false positive, respectively (Liu *et al.* [2010]). Specificity is equal to $1 - FPR$ (Equation 1.20).

Chapter 2

Preliminary Assessment of Mutual Information Based Methods for Contact Prediction

2.1 Chapter Overview

This chapter summarises the predictive power of original MI and the current leading MI variants, specifically their ability to predict contact residues within domains and between domains. We find that all MI based methods predict contact residues poorly.

2.2 Introduction

We begin our investigation by evaluating original MI and current leading variants, MIp (Dunn *et al.* [2008]), MIc (Lee & Kim [2009]) and aMIc (Lee & Kim [2009]) on 40 interacting domain pairs (inter-domains). The contacts between interacting domains serve as a proxy for contacts between interacting proteins (Pazos *et al.* [1997]). Since the MI based methods calculate scores for pairs of residue columns in the alignment,

we assign each residue column the maximum score it achieves with any other column in the alignment.

In our analysis MIc outperformed the other MI variants achieving a precision of 34.7% at 20% recall. In comparison, a non-MI based contact predictor, i-Patch (Hamer *et al.* [2010]), attains a precision of 48.9% at 20% recall on the same dataset.

We also evaluated the performance of these MI based methods for contact residue pair prediction within a domain (intra-domain). In order to do so we split our 40 domain-domain cases into 80 single domains.

In intra-domains there is no distinct subset of residues that are in contact, such as the residues in the interface of interacting domains. Hence instead of predicting contact residues, we set ourselves the harder task of predicting pairs of residues that are in contact with each other. We find that the ability of all MI measures to predict pairs of residues that are in contact within a domain to be poor. Yet again, MIc outperforms the other methods; attaining a precision of 12.8% at 20% recall for intra-domain contact pair prediction. In comparison, when attempting to predict pairs of residues that are in contact in inter-domains, MIc, once again the lead predictor, achieves a precision of 2.26% at 20% recall .

2.3 Materials and Methods

The MI, MIp, MIc and aMIc calculations used in this chapter are performed as described in Section 1.7. Entropy is also measured (Section 1.6). As in previous work (Dunn *et al.* [2008]; Lee & Kim [2009]), only ungapped columns in the alignments are considered in the analysis; 33.8% of 11,846 total columns in our dataset have one or more gaps and are not included in calculations for all MI variants. Refer to Chapter 5 for a fuller examination of the effect of gaps.

Our definition of contact and non-contact residues can be found in Section 1.4.4.

The P-ROC (Precision Recall Operating Characteristic) curve (Buckland & Gey [1994]) detailed in Section 1.9 is used to assess the contact prediction abilities of the MI variants. This evaluation metric is selected because in interacting proteins the ratio of contact to non-contact residues is disproportionate. For example, there are 1,342 domain-domain contact and 6,505 non-contact residues in our 40 inter-domain analysis (Appendix Tables 7 and 8), and 14,967 contact and 474,440 non-contact residue pairs in the 80 intra-domain study (Appendix Table 9).

2.3.1 Inter-domain and intra-domain datasets

Inter-domain

For the inter-domain investigation we use proteins that have two domains, rather than protein complexes, and treat each domain as a separate protein. The multiple sequence alignments (MSAs) are taken from Hamer *et al.* [2010], available at www.stats.ox.ac.uk/research/bioinfo/resources, which in turn are based on datasets in Holm & Sander [1994]; Pazos *et al.* [1997]; Siddiqui & Barton [1995] and Sowdhamini & Blundell [1995]. One protein in each MSA has a known PDB structure of X-ray resolution 2.5Å or better, and well annotated domain boundaries. This structure is henceforth referred to as the "reference structure" and is used to identify surface, buried, contact and non-contact residue columns within the MSA. The MSA was generated using the structural protein as a BLAST query (Altschul *et al.* [1990, 1997]) against the NCBI-NR database (Sayers *et al.* [2012]). The homologs identified were made non-redundant at the 90% level using Cd-hit (Li & Godzik [2006]). The final alignment was generated using MUSCLE (Edgar [2004]) and MaxAlign (Gouveia-Oliveira *et al.* [2007]). All non-standard amino acid entries, such as B, Z, X, * and ? are treated as gaps.

Amongst the set of 67 protein cases available from Hamer *et al.* [2010], proteins that contain a single domain that interacts with more than one other domain in the set are disregarded. We choose to omit these proteins as domains interacting with

multiple domains may have undergone correlated mutations not pertaining to the pair of domains being presently considered. We thus lose 15 of the 67 cases. In order to aid statistical analysis of the results we select only those domain pairs that have at least 20 contact and 20 non-contact residues on each domain, and the corresponding MSA columns of these residues must be ungapped and have an entropy greater than 0. Therefore a further 9 test cases are lost. We also remove 1 test case that has less than 20 surface and buried residues respectively, and 2 cases that have poorly annotated secondary structures in their reference PDB structure file. This leaves us with 40 inter-domain MSAs (Table 2.1).

Intra-domain

For assessing the ability of MI variants to predict intra-domain contact residue pairs we split each MSA in our 40 inter-domain dataset (Table 2.1) into the two separate domains they correspond to. The lengths of the resulting 80 single domains range from 60 to 376 residues.

protein	name	species	D1	D2	sequences	species
1A45	gamma-F-Crystallin	*Bos taurus*	1 82	83 173	160	E(146)N(14)
1BIB	BirA	*Escherichia coli*	67 270	271 317	236	A(12)B(201)N(23)
1BKS	tryptophan synthase	*Salmonella typhimurium*	1 188	189 268	478	A(21)B(401)E(10)N(46)
1FNB	ferredoxin-NADP+-oxidoreductase	*Spinacia oleracea*	19 152	153 314	58	B(22)E(34)N(2)
1G8A	fibrillarin-like pre-RNA processor	*Pyrococcus horikoshii*	1 51	52 227	75	A(47)E(20)N(8)
1G8P	magnesium-chelatase 38kda subunit	*Rhodobacter capsulatus*	18 216	261 350	230	A(10)B(143)E(49)N(28)
1I39	ribonuclease HII	*Archaeoglobus fulgidus*	1 158	159 200	688	A(32)B(538)E(7)V(1)U(1)N(109)
1J5X	glucosamine-6-phosphate deaminase	*Thermotoga maritima*	2 169	170 319	252	A(9)B(183)E(5)N(55)
1LAP	cytosol aminopeptidase	*Bos taurus*	1 147	148 484	454	A(2)B(331)E(84)N(37)
1LLD	L-lactate dehydrogenase	*Bifidobacterium longum*	7 148	149 319	709	A(33)B(389)E(221)N(66)
1MRI	alpha-momorcharin	*Momordica charantia*	1 162	163 246	68	B(2)E(65)N(1)
1PII	phosphoribosylanthranilate isomerase	*Escherichia coli*	1 255	256 452	75	B(65)N(10)
1RHD	rhodanese	*Bos taurus*	1 156	157 293	505	A(26)B(365)E(57)U(1)N(56)
1THM	thermitase	*Thermoactinomyces vulgaris*	1 127	128 208	106	A(1)B(62)E(34)N(9)
1W98	cyclin E1/CDK2	*Homo sapiens*	88 227	228 357	70	E(64)N(6)
1WRU	43 kDa tail protein	*Enterobacteria phage Mu*	3 176	177 346	64	B(58)V(2)N(4)
1X2G	lipoate-protein ligase A	*Escherichia coli*	1 246	247 337	224	A(2)B(155)E(42)N(25)
2AAA	alpha-amylase	*Aspergillus niger*	1 376	377 484	245	B(141)E(74)N(30)
2AHE	chloride intracellular channel protein 4	*Homo sapiens*	16 108	109 253	144	B(25)E(100)N(19)
2D3V	leukocyte Ig-like receptor A5	*Homo sapiens*	3 95	96 195	77	E(71)N(6)
2D8N	recoverin	*Homo sapiens*	16 97	102 189	240	E(195)N(45)
2E64	biotin-[acetyl-CoA-carboxylase] ligase	*Pyrococcus horikoshii OT3*	1 188	189 235	294	A(9)B(231)E(4)U(1)N(49)
2I00	Acetyltransferase	*Enterococcus faecalis V583*	10 300	301 406	116	A(2)B(80)N(34)
2IU5	Dihydroxyacetone operon	*Lactococcus lactis*	1 71	72 180	65	B(56)N(9)
2NPO	Acetyltransferase	*Campylobacter jejuni*	3 76	77 188	224	A(3)B(182)U(1)N(38)
2NRC	Succinyl-CoA:3-ketoacid-coenzyme A transferase 1	*Sus scrofa*	1 247	261 480	188	A(9)B(96)E(68)N(15)
2OF7	Putative tetR-family transcriptional regulator	*Streptomyces coelicolor*	17 67	68 207	204	B(135)N(69)
2OI8	Putative regulator SCO4313	*Streptomyces coelicolor*	8 86	87 216	215	B(151)N(64)
2PGD	6-phosphogluconate dehydrogenase	*Ovis aries*	1 172	178 433	317	B(211)E(78)N(28)
2PGE	MenC	*Desulfotalea psychrophila*	3 136	137 368	138	A(6)B(102)E(1)N(29)
2PGX	UPF0341 protein yhiQ	*Escherichia coli*	2 56	57 250	102	B(87)N(15)
2PHZ	FeuA	*Bacillus subtilis*	20 142	143 296	420	A(4)B(343)N(73)
2QY9	ftsY	*Escherichia coli*	201 284	285 495	471	A(32)B(344)E(15)N(80)
2REB	recA	*Escherichia coli*	23 268	269 328	482	B(434)E(12)N(36)
2TS1	tyrosyl-tRNA synthetase	*Geobacillus stearothermophilus*	1 220	248 319	598	B(512)E(34)N(52)
4ENL	enolase	*Saccharomyces cerevisiae*	1 126	127 436	649	A(32)B(448)E(122)N(47)
4MDH	cytoplasmic malate dehydrogenase	*Sus scrofa*	1 154	155 333	339	A(6)B(173)E(134)N(26)
5FBP	fructose-1,6-bisphosphatase	*Sus scrofa*	1 201	202 335	355	A(3)B(213)E(112)N(27)
6GST	mu glutathione s-transferase of isoenzyme 3-3	*Rattus rattus*	1 82	90 217	374	B(10)E(312)N(52)
8TLN	thermolysin	*Bacillus thermoproteolyticus*	1 135	136 316	44	A(1)B(36)E(2)N(5)

Table 2.1: **Summary of the Hamer dataset.** This dataset is taken from Hamer *et al.* [2010]. The "protein" column contains a list of pdb identifiers (Berman *et al.* [2000]). The "name" and "species" columns refer to the name of the crystallised protein identified in column 1 and the species it came from. D1 and D2 columns denote the start and end pdb residues of domains 1 and 2, respectively. For all pdbs listed, the start and end residues are located in chain A of the structure, except for pdb 1W98 where the mentioned domains are in chain B, and pdb 8TLN in chain E. The "sequences" column indicates the number of sequences present in the multiple sequence alignment (MSA). The final column states the distribution of sequences in each MSA taken from the various species' domains: eukaryotes (E); archea (A); bacteria (B); viruses (V); unclassified (U); and not found (N), *i.e.* those sequences that could not be found in the NCBI Taxonomy Database (Wheeler *et al.* [2006]).

2.3.2 MSA columns with 0 entropy

Pairing any MSA column with a fully conserved column, *i.e.* a column with an entropy of 0, results in a joint entropy equivalent to the entropy of the non-fully conserved column and subsequently an MI score of 0 for that pair. Since conserved columns do not give any indication of correlated mutations, MI scores involving these columns are ignored. This is standard procedure; for example, Tillier & Lui [2003]. The relationship between percent MI scores of 0 and percent of columns in an MSA with an entropy of 0 is shown in Figure 2.1.

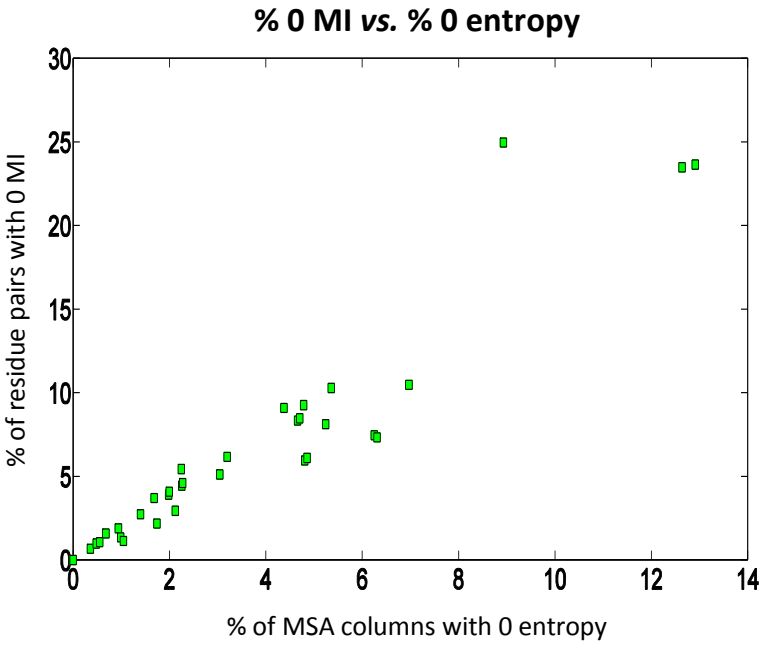

Figure 2.1: **Effect of entropies of 0 on MI scores.** The percent of columns in an MSA that have an entropy of 0 is plotted against the percent of all inter-domain residue pairs in the corresponding complex that have an MI value of 0. Only those columns in the MSA that correspond to a residue in the reference structure are used. Columns that have one or more gaps are ignored. Each point on the plot represents a single case study in our inter-domain dataset.

2.3.3 Identifying the contact *versus* non-contact residue pairs

For each reference structure protein in the dataset, contact and non-contact residue pairs are identified based on the criterion described in Section 1.4.4.

Inter-domain

For the inter-domain analysis, once a residue is identified as being involved in a contact pair it is labelled as a "contact" residue. All residues that do not participate in any contact pair are labelled as "non-contact." This information is then annotated to the entire MSA column to which the residue belongs. After eliminating residue columns that have an entropy of 0 or contain a gap, we are left with 1,342 domain-domain contact and 6,505 non-contact residues in our 40 inter-domain dataset (Appendix Tables 7 and 8).

Intra-domain

In the intra-domain analysis we consider contact and non-contact residue pairs, for there is no clear subset of contact residues; all residues will be in contact, within 4.5Å, of some other residues in the domain. Our 80 single domain dataset comprises of 14,967 contact pairs and 474,440 non-contact residue pairs, after residue columns that have an entropy of 0 or contain a gap are eliminated (Appendix Table 9). In order to compare the performance of MI variants for intra-domain *versus* inter-domain contact prediction, we also examine inter-domain contact residue pairs. After discarding 0 entropy and gapped columns, we are left with 1,301 contact and 362,399 non-contact pairs in our 40 inter-domain dataset (Appendix Table 7).

2.4 Results and Discussion

2.4.1 Inter-domain MI analysis

We would like to predict whether a residue is a contact residue or not using MI based scores on our domain-domain MSAs. As MI assigns scores to pairs of columns in an MSA, first we calculate the MI score for all pairs of columns. To obtain a score for individual columns, each residue in all 40 test cases is assigned the maximum MI score achieved by that residue column. We also tested assigning the average score of each residue column, but this resulted in a significant decrease in performance of the MI variants.

For the 40 inter-domain test cases employed, the probability of randomly selecting contacts, *i.e.* correctly picking a contact residue from the total set of residues, without any information about the proteins involved, is 17.1%. MI, MIp, MIc and aMIc achieved precisions of 21.5%, 31.8%, 34.7% and 34.5% respectively, at 20% recall (Figure 2.2 and Table 2.2). Running i-Patch (Hamer *et al.* [2010], Section 1.8), a non-MI based domain-domain contact predictor, on our 40 inter-domain dataset resulted in a precision of 48.9% at 20% recall. Thus the performance of all MI methods is below that of the parametrised method i-Patch, which uses additional information such as pre-calculated propensity tables of residues in contact and surface residues only. Since i-Patch uses surface residues only, the number of i-Patch scores compared to MI, MIp, MIc and aMIc is less, therefore the P-ROC curve of i-Patch in Figure 2.2 does not end at the same location as the MI variant curves.

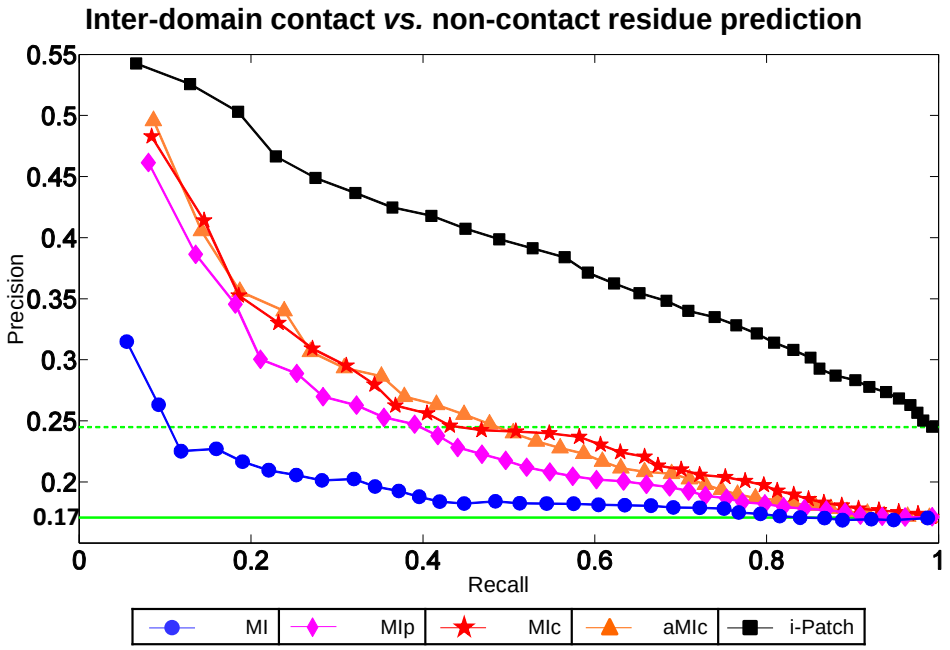

Figure 2.2: **Contact *versus* non-contact prediction P-ROC curves for MI variants and i-Patch on the 40 inter-domain test cases.** The performance of MI, MIp, MIc and aMIc variants when distinguishing contact from non-contact residues. The solid green line at 0.171 depicts the chance of randomly selecting a contact residue from the entire set of residues, surface and buried. The dashed green line at 0.245 depicts the chance of randomly selecting a contact residue from the set of surface residues only.

MI variant	inter-domain contact precision at 20% recall
i-Patch	48.9%
MIc	34.7%
aMIc	34.5%
MIp	31.8%
MI	21.5%
Random	17.1%

Table 2.2: **Precision for detecting contact *versus* non-contact residues at 20% recall for inter-domains.** Results are given for the 40 inter-domain test cases. MI variants and i-Patch are listed in descending order of contact *versus* non-contact precision, *i.e.* best to worst classifier of contact residues. The probability of randomly selecting a contact residue from all residues is 17.1%.

2.4.2 Intra-domain MI analysis

A similar analysis is carried out on the 80 single domains in our dataset (Table 2.1). When assessing the ability of the MI variants to predict intra-domain contacts however, we do not assign each residue its maximum MI score as we did previously for inter-

domain contact prediction. Unlike inter-domains that have a distinct subset of interface residues that form contacts, all residues within a domain are in close proximity with some residues and are far from others. Hence here we consider pair scores, and evaluate the ability of MI variants to identify pairs of residues that are in close proximity. To allow for comparison with inter-domain prediction we also examine inter-domain contact residue pair scores.

We find that the intra-domain contact pair prediction abilities of all MI variants is also poor (Figure 2.3A and Table 2.3). In our dataset, the probability of randomly selecting a contact pair from all residue pairs is 3.05%. On our test cases, MIc performs best, achieving a precision of only 12.8% at 20% recall.

The predictive power of all MI variants for inter-domain contact residue pair prediction is also weak (Figure 2.3B and Table 2.3). Once again MIc outperforms the other MI based methods, attaining a precision of 2.26% at 20% recall. The chance of randomly selecting a contact residue pair in the inter-domain dataset is 0.358%.

We recognise that the precision at 20% recall of MIc is approximately six times above random for inter-domain contact pair prediction, and only approximately four times above random for intra-domain prediction. Nevertheless, in terms of absolute prediction a precision of 2.26% is less useful than a precision of 12.8%. Neither precisions however are above 50%, the precision threshold accepted as necessary for an algorithm to be useful for protein structure determination (Jones *et al.* [2012]).

	intra-domain pair contact precision at 20% recall	**inter-domain pair contact precision at 20% recall**
MIc	12.8%	2.26%
aMIc	12.6%	2.26%
MIp	12.2%	1.70%
MI	4.15%	0.600%
Random	3.05%	0.358%

Table 2.3: **Precision for detecting contact *versus* non-contact residue pairs at 20% recall.** Results are given for the 80 intra-domain and 40 inter-domain test cases. MI variants are listed in descending order of contact *versus* non-contact precision, *i.e.* best to worst classifier of contact residue pairs. The probability of randomly selecting a contact residue pair from all residue pairs in the intra-domain test set is 3.05% and in the inter-domain test set it is 0.358%.

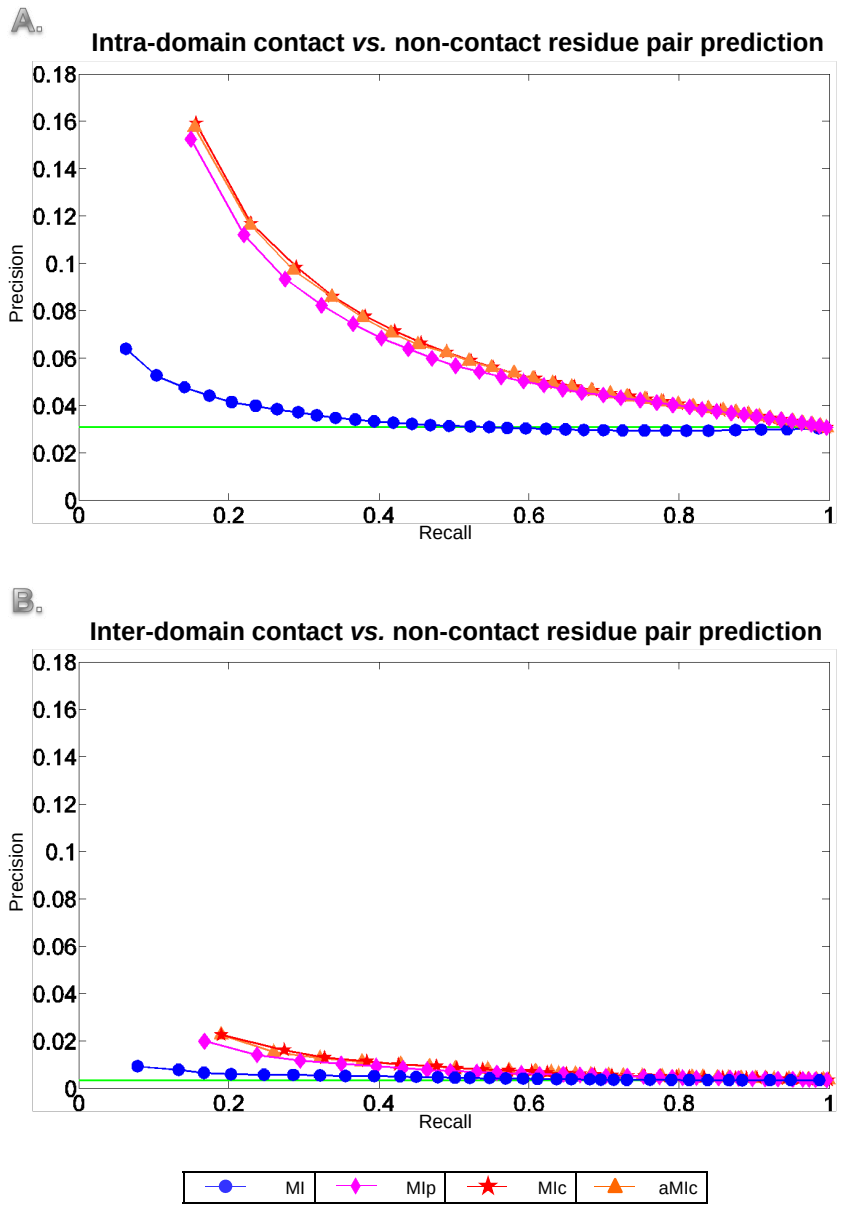

Figure 2.3: **Contact *versus* non-contact residue pair prediction P-ROC curves for MI variants on the 80 intra-and 40 inter-domain test cases.** A and B illustrate the performance of MI, MIp, MIc and aMIc on the 80 intra- and 40 inter-domain datasets, respectively, when distinguishing contact from non-contact residue pairs. The solid green line at 0.0305 in (A) and 0.00358 in (B) depicts the chance of randomly selecting a contact residue pair in each dataset.

A previous study found that the average precision (Equation 1.21) for intra-protein contact prediction using original MI is 10 times higher than that of inter-proteins (Halperin *et al.* [2006]). To assess if the latest leading variants of MI also exhibit this behaviour we calculate average precisions for both intra- and inter- datasets. In agreement with Halperin *et al.* [2006], all MI variants have higher average precisions in their top 100 scores in intra-domain than inter-domain contact pair prediction (Figure 2.4). For the highest score, $n = 1$, the average precision for all MI variants is approximately twice as high for intra-domains. For example when considering MIc, in Figure 2.4B when $n = 1$ along the x-axis, the average precision for the inter-domain cases is approximately 0.3, while in Figure 2.4A when $n = 1$ the average precision for the intra-domain cases is approximately 0.6.

In our analysis the average precision for original MI is not 10 times higher as determined by Halperin *et al.* [2006], perhaps because these authors use a different dataset and employ a less strict definition for "contact residue pairs;" we require that any two atoms, one from each residue, be within 4.5Å from each other, while they allow 6Å.

The observation that MI variants have a greater ability to predict contact residue pairs in intra-protein domains than inter-protein domains could suggest that:

1. From an evolutionary viewpoint residues have a higher tendency to preserve protein folding rather than protein-protein binding. This may be because single protein units are more stable than bound protein complexes.

2. The evolutionary rate of a single protein's sequence is more conserved than that of two interacting proteins (Kim *et al.* [2004]).

3. Interacting proteins are under the same evolutionary pressure and subsequently they have equal rates of evolution across the entire length of their sequences (Hakes *et al.* [2007]). Therefore, residue pairs may not be undergoing correlated mutations and consequently we cannot infer residues in contact in this manner.

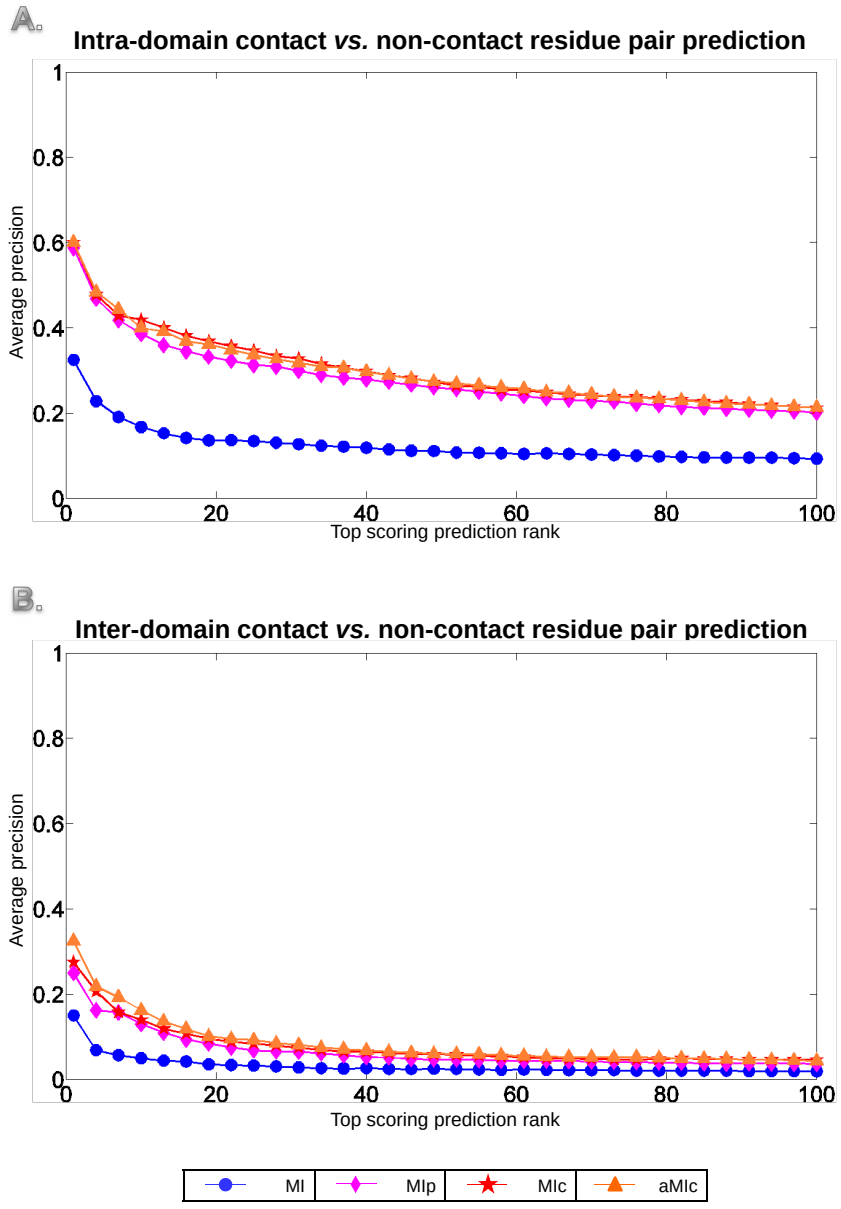

Figure 2.4: **Average precision for top n ranked contact *versus* non-contact pair predictions on 80 intra- and 40 inter-domains.** A and B illustrate the performance of MI, MIp, MIc and aMIc on the 80 intra- and 40 inter-domain datasets, respectively, when distinguishing contact from non-contact residue pairs. For each test case in a dataset, the precision (Equation 1.21) when considering the top n MI scores is calculated; n varies from 1 to 100. For each n, the average precision over all test cases in the dataset is then determined and plotted. This process is repeated for MIp, MIc and aMIc scores respectively.

4. MI based methods are not robust enough in detecting correlated mutations and subsequently residues in contact (Halperin *et al.* [2006]).

2.5 Conclusions

This chapter examines the properties of MI when applied to contact *versus* non-contact residues in domains, as well as domain-domain complexes. For our analysis we used structural data of crystallised proteins and associated MSAs of 40 inter-domain and 80 intra-domain test cases.

We find that the ability of the MI variants to distinguish between contacts and non-contacts in both inter- and intra-domains is weak. MIc achieved the highest performance in both inter- and intra-domain contact prediction, followed closely by aMIc, then MIp and finally original MI. The precision at 20% recall of MIc for inter- and intra- domain contact residue pair prediction is 2.26% and 12.8% respectively. The average precisions when considering the top 100 scores are higher for intra-domain than for inter-domain contact pairs for all tested MI variants. These discrepancies between intra- and inter-proteins MI based contact prediction may be related to the different evolutionary pressures on protein fold *versus* protein-protein interactions.

When attempting to predict inter-domain contact residues, i-Patch (Hamer *et al.* [2010], Section 1.8), a non-MI based inter-domain contact predictor, achieved a precision of 48.9% at 20% recall on our 40 domain-domain test cases. In contrast, MIc, the MI contact residue predictor that performed best, attained a precision of 34.7% at 20% recall. We conjecture that i-Patch may be more successful than the tested MI variants because unlike MI, i-Patch uses pre-calculated contact residue propensities, and considers only residues on the surface of a protein when attempting to predict contacts between proteins. In the next chapter we examine the effect of surface and buried residues on inter- and intra-domain MI contact prediction.

It is believed that a contact prediction algorithm should attain a precision of over 50% to be considered sufficiently useful for determining protein structure (Jones *et al.* [2012]). Our assessment of MI algorithms suggests that in their current state MI based methods are not very useful for protein residue contact prediction.

Chapter 3

Mutual Information Based Methods Exhibit Bias Towards Surface Residues

3.1 Chapter Overview

In this chapter we examine the abilities of original MI and its variants to predict contacts, considering surface and buried residues separately. Our analysis shows that all MI methods carry a signal for surface residues. We hypothesise that it is because surface residues tend to receive high MI scores that previous inter-protein contact prediction investigations appear to be successful.

Some of the work discussed in this chapter is presented in the published article Gomes *et al.* [2012].

3.2 Introduction

In the previous chapter we observed that the contact prediction ability of original MI and the current leading variants, MIp (Dunn *et al.* [2008]), MIc (Lee & Kim [2009]) and aMIc (Lee & Kim [2009]) on 40 interacting domain pairs (inter-domains) and the corresponding 80 single domains (intra-domain) is weak. We speculate that, i-Patch (Hamer *et al.* [2010], Section 1.8), a non-MI based, inter-domain contact predictor outperforms these methods because it uses pre-calculated contact residue propensities, and limits contact residue prediction between domains to surface residues only. In this chapter we analyse the influence of surface and buried residues on MI scores.

Similar to inter-domain residues in the previous chapter, each surface or buried residue is assigned the maximum score it attains when paired with any other residue column in the alignment.

For inter-domains we find that original MI distinguishes between surface and buried residues, with surface residues tending to have slightly higher MI values. We speculate that the moderately higher entropy detected in surface residue alignment columns contributes to the high MI scores of these residues. We observe that the adjusted MIp, MIc and aMIc scores weakens the signal between surface and buried residues, but in most cases surface residues still tend to have a higher score than buried residues.

A distinction between surface and buried MI values is also observed when analysing intra-domains, with surface residues once again tending to have higher MI values. This applies to all tested MI variants with the exception of MIc, which we observed in the previous chapter to be the best intra-domain contact pair predictor for the dataset used. The bias towards surface residues is slightly less pronounced in intra-domains as compared to inter-domains. For example, the scores of original MI are most skewed towards surface residues in both inter- and intra-domains. When attempting to distinguish surface from buried residues, at 20% recall the precision is 86.9% for inter-domains

compared to 83.8% for intra-domains.

Our analysis of two high-profile studies that have successfully used MI to find inter- and intra- protein contacts, respectively, suggests that the employed MI algorithms mostly select surface residues in their top scoring MI pairs. These results further support our hypothesis that high MI scores are biased towards the highly entropic surface residues, rather than residues that are in contact. We conjecture that it is because contact sites in interacting proteins are on the surface of each protein, that MI based methods appear to be successful for inter-protein contact prediction in previous studies.

3.3 Materials and Methods

The 40 inter-domain and 80 intra-domain dataset employed in the previous chapter is used for analysis in this chapter (Table 2.1). The MI, MIp, MIc and aMIc calculations are performed as described in Section 1.7 and entropy is measured as outlined in Section 1.6. In keeping with the methodology employed in the previous chapter, columns that have an entropy of 0, or contain one or more gaps are excluded.

P-ROC (Precision Recall Operating Characteristic) curves (Buckland & Gey [1994]) (Section 1.9) are used to assess the surface residue bias of the MI variants.

Identifying the surface *versus* buried residue pairs

For a given reference structure protein in the dataset, we calculate the solvent accessibility of the residues using JOY (Mizuguchi *et al.* [1998]); each domain is treated as a separate entity. In the reference structure, residues that are more than 7% accessible to a 1.4Å radius water molecule are denoted as "surface" residues (Mizuguchi *et al.* [1998]). Those that do not meet this criterion are termed "buried." This information about a residue is then annotated to the entire MSA column to which it belongs.

Employing this criterion on our 80 domains, along with eliminating residue columns that have an entropy of 0 or contain a gap, leaves us with 5,483 surface residues and 2,364 buried residues (Appendix Tables 7, 8 and 9).

3.4 Results and Discussion

3.4.1 Inter-domain MI analysis

After calculating the MI score for all pairs of columns, each residue in all 40 test cases is assigned the maximum MI score that the residue column achieved with any other residue column in its MSA. For our 40 inter-domains the probability of randomly selecting a surface residue from all residues without any information of the proteins involved is 69.9%. Using original MI, MIp, MIc and aMIc on our dataset as surface residue predictors (is the highest scoring residue on the surface?), we observed that each of the measures surpassed this random classification and attained precisions of 86.9%, 75.5%, 74.1% and 80.8% respectively at 20% recall (Figure 3.1 and Table 3.1). Thus it appears that high scores of all four variants of MI are skewed towards surface residues.

	inter-domain surface precision at 20% recall	**intra-domain surface precision at 20% recall**
MI	86.9%	83.8%
aMIc	80.8%	72.5%
MIp	75.5%	71.1%
MIc	74.1%	66.2%
Random	69.9%	69.9%

Table 3.1: **Precision for detecting surface *versus* buried residues at 20% recall.** Results are given for the 40 inter-domain and 80 intra-domain test cases. MI variants are listed in descending order of surface *versus* buried precision, *i.e.* best to worst classifier of surface residues. The probability of randomly selecting a surface residue from all residues in both inter- and intra-domain datasets is 69.9%.

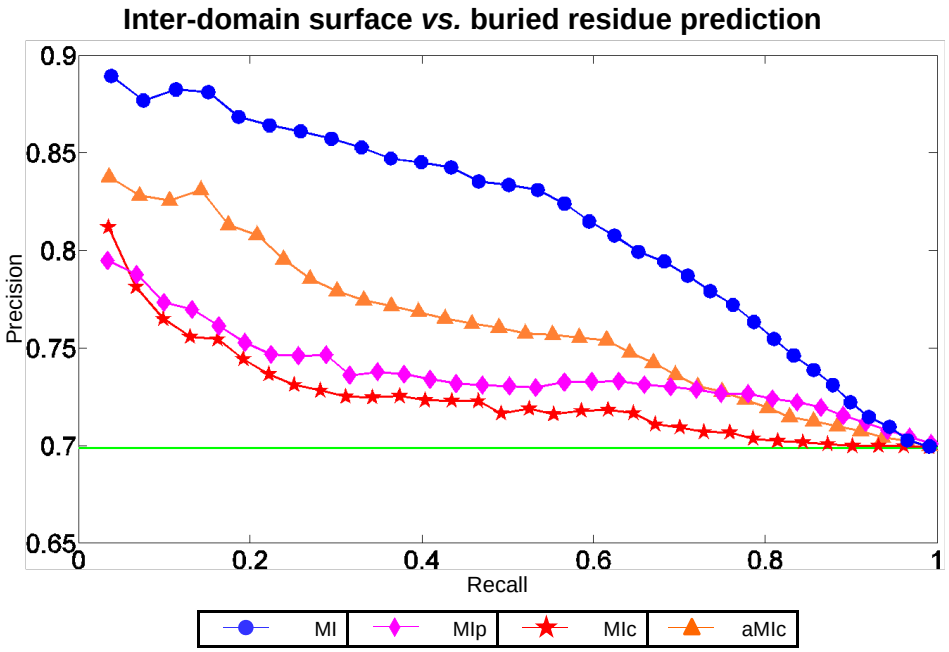

Figure 3.1: **Surface *versus* buried prediction P-ROC curves for MI variants on the 40 inter-domain test cases.** The performance of MI, MIp, MIc and aMIc variants when distinguishing surface from buried residues. The solid green line at 0.699 depicts the chance of randomly selecting a surface residue.

A possible explanation for the ability of MI based methods to predict surface residues is that the observed higher entropy of surface residue columns (Figure 3.2) contribute to the higher MI scores of these residues. Prior investigations have shown that MI scores strongly correlate with the entropy of the columns involved (Fodor & Aldrich [2004]; Martin *et al.* [2005]). Figure 3.2 shows that MSA columns corresponding to surface residues tend to have a higher entropy than those associated with buried residues. The observed lower column entropy for buried residues is consistent with previous studies that have indicated that buried residues are under greater evolutionary constraints than solvent-accessible surface residues (Bustamante *et al.* [2000]; Goldman *et al.* [1998]; Lin *et al.* [2007]; Overington *et al.* [1992]). A slower rate of evolution of these residues is unsurprising since buried residues often play a crucial role in maintaining the 3D structure of a protein. We hypothesise that this skewness of MI based methods towards surface residues in turn perturbs the ability of these measures

to predict contact residues. Hence in the next chapter we eliminate buried residues and re-evaluate the performance of original MI, MIp and MIc for inter-domain contact prediction when only surface residues are considered.

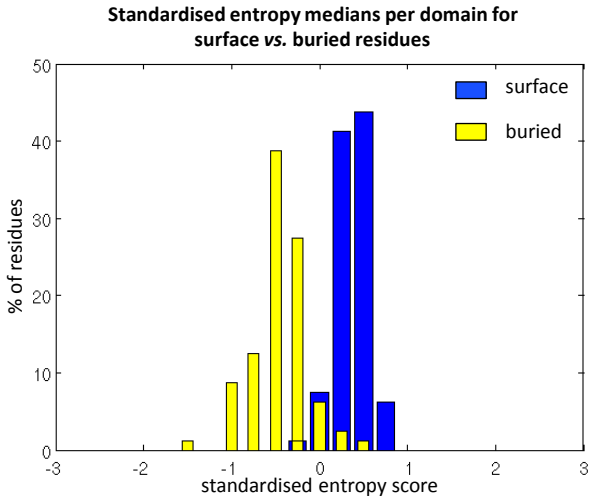

Figure 3.2: **Standardised entropy medians of surface *versus* buried residue columns for all 80 domains in the dataset.** Comparing the medians of the standardised entropy scores of the surface residue columns of each domain (blue) against the medians of the buried residue columns of each domain (yellow). Residue columns containing one or more gaps, or having an entropy score of 0 are not included in the median calculation.

3.4.2 Inter-domain case study

A paper by Skerker *et al.* [2008] has received a lot of attention for successfully determining inter-protein contact specificity residues with the aid of MI. The authors used original MI (Equation (1.6)) to determine a subset of contact residues that allow for specific binding of a histidine kinase (HK) with its interacting response regulator (RR) (Figure 3.3). The MSA provided by these authors does not contain the sequence of the structure used in their analysis. Hence we ran MI on the HK-RR MSA provided by Hamer *et al.* [2010], which does include the sequence of this reference structure. Similar to Skerker *et al.* for these MI calculations we used log_e. We observe 23 HK and 28 RR residues in MI pairs above 0.35, the MI score cutoff imposed in the Skerker

investigation. All but 1 residue are surface, and this exception is from the RR.

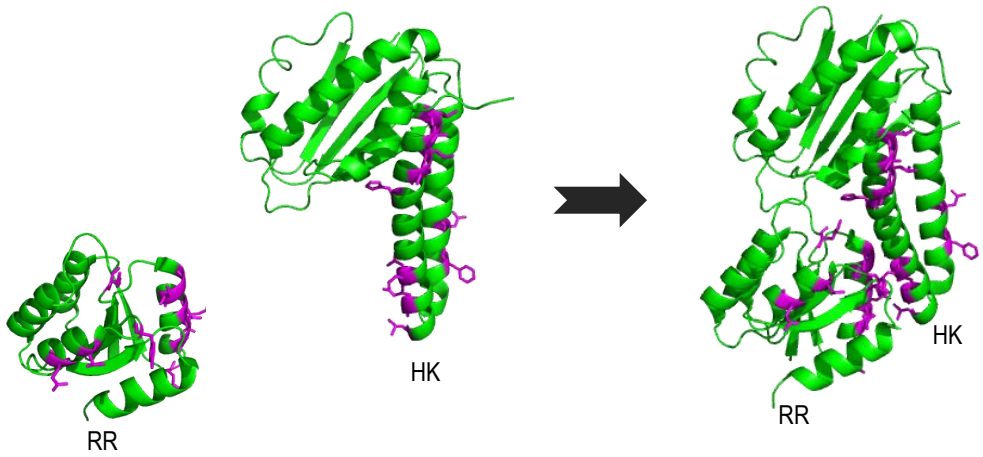

Figure 3.3: **The Skerker *et al.* [2008] high MI scoring residues.** Histidine kinase (HK) and its interacting response regulator (RR) in and out of complex. All magenta coloured residues with visible side chains have been explicitly identified by Skerker *et al.* [2008], in Figure 2 of their paper, as residues having contributed to high MI scores (greater than 0.35). Crystal structure 1F51.pdb (Zapf *et al.* [2000]) is used in this illustration.

3.4.3 Intra-domain MI analysis

The analysis is repeated on the 80 single domains in our dataset (Table 2.1). We observe that higher MI scores are preferentially biased towards surface residues even for intra-domains. As can be seen in Figure 3.4 and Table 3.1 this holds true for all MI variants, except MIc, which is also the leading intra-domain contact predictor in our investigation (Chapter 2, Figure 2.3A and Table 2.3). This observed bias towards surface residues is slightly weaker in intra-domains as compared to inter-domains (Table 3.1). We speculate that this reduced bias may partly explain the improved contact prediction abilities of MI based methods in intra-domains over inter-domains (Chapter 2).

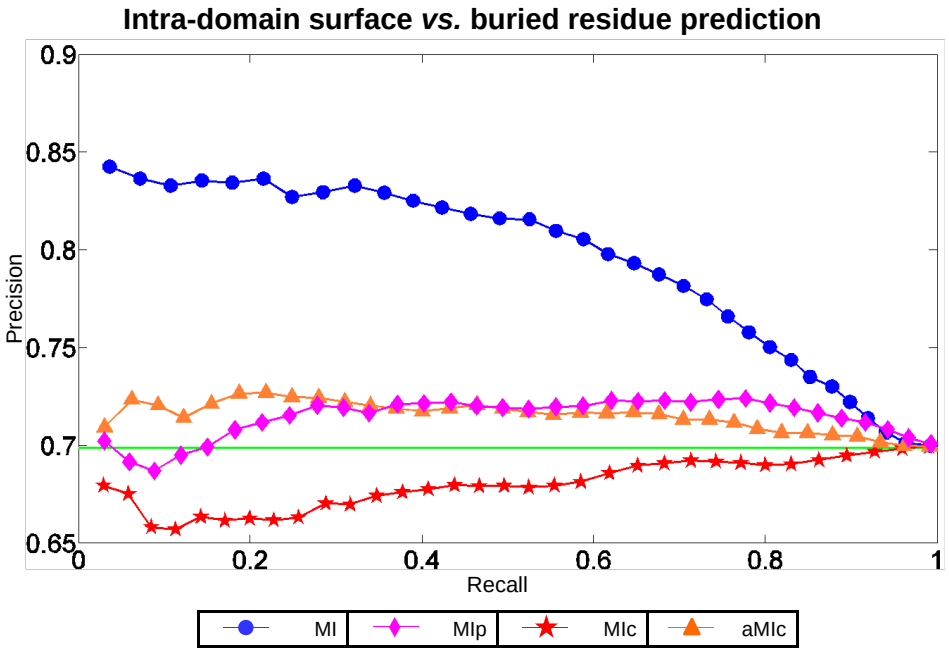

Figure 3.4: **Surface *versus* buried prediction P-ROC curves for MI variants on the 80 intra-domain test cases.** The performance of MI, MIp, MIc and aMIc variants when distinguishing surface from buried residues. The solid green line at 0.699 depicts the chance of randomly selecting a surface residue.

3.4.4 Intra-domain case study

Dunn *et al.* [2008] used triosephosphate isomerase to demonstrate the performance of their MI variant MIp, and a previous variant also designed by the same group, MIr (Martin *et al.* [2005]).

We observe that both MIr and MIp select surface-surface residue pairs in their top three scores (Table 3.2). Out of the 30 total residues involved in their identified 15 significantly scoring MIp residue pairs listed in Table 3.2, 22 residues are on the surface of the protein and 8 are buried. This may be owing to the strong bias MI has for surface residues. We also note that there is no relationship between the distance of two residues in structure or sequence (Table 3.2 columns 3 and 4), and the MIr or MIp score ranking of the residue pair.

These observations are consistent with our results which have shown that the signal

between surface and buried residues is stronger than the signal between contact and non-contact residue pairs in intra-proteins, as illustrated by the higher precisions the MI methods attain for surface residue prediction rather than contact (Tables 3.1, 2.2 and 2.3, and Figures 3.1, 3.4, 2.2 and 2.3).

residue i	residue j	structure distance (Å)	sequence distance	MIr	MIp
184	227	2.8	43	7.5	10.5
68	70	3.4	2	5.3	7.9
135	176	5.6	41	4.8	7.7
187	219	10.7	32	4.5	7.5
216	241	3.1	25	4.2	7
6	9	7.3	3	2.1	6
138	141	3.3	3	-	5.7
139	186	5.7	47	-	5.7
140	189	4.5	49	-	5.0
10	235	8	225	-	4.9
221	224	3.1	3	-	4.2
179	181	3.1	2	-	4.0
39	246	3.9	207	-	3.7
15	71	23.8*	56	-	3.4
142	145	3.1	3	-	3.1

surface pair buried pair mixed pair

Table 3.2: **The Dunn *et al.* [2008] high scoring residue pairs in triosephosphate isomerase.** Entries in the table are arranged in descending order of MIp values. Rows in blue depict a surface residue pair, while those in yellow signify a buried pair in structure 1IIH.pdb (Noble *et al.* [1991]). A line in pink denotes a pair with one surface and one buried residue. Taken from Dunn *et al.* [2008]. A '-' indicates that data does not exist in the Dunn *et al.* [2008] paper for that entry.

3.5 Conclusions

Original MI scores carry a signal distinguishing surface from buried residues, with surface residues tending to have slightly higher MI values for both inter- and intra-domains. We hypothesise that the generally higher surface residue scores of original MI

can be attributed to the observed higher entropy of columns in an MSA corresponding to the surface residues. Two of the three tested variants of MI, namely MIp and aMIc, give rise to the same trends. However, the distinction in signal between surface and buried residues is reduced in both inter- and intra-domains due to the noise correction metrics utilised by MIp and aMIc. MIc, although biased towards surface residues in inter-domains, does not exhibit a bias for surface residues in intra-domains.

An analysis of two studies that have successfully used MI to find inter- and intra-protein contacts respectively, suggests that that the employed MI algorithms mostly select surface residues in their top scoring MI pairs. This supports our hypothesis that MI is preferentially biased towards highly entropic surface residues instead of contact residues that may have undergone correlated mutations.

In summary, MI is conjectured to predict contact sites, but we find that MI instead predicts surface residues. When there is an available protein structure, surface residues, can be determined more easily and accurately via protein sequence-structure analysis software, such as JOY (Mizuguchi *et al.* [1998]), GETAREA (Fraczkiewicz & Braun [1998]) and POPS (Cavallo *et al.* [2003]). In the case where there is no available representative protein structure, algorithms such as SANN (Joo *et al.* [2012]), RSARF (Pugalenthi *et al.* [2012]) and Jnet (Cuff & Barton [2000]) can be used to predict residues that have 5% solvent accessibility with 85.5%, 78.3% and 79.8% accuracy, respectively. Here accuracy is defined as the percentage total number of residues assigned to the correct category, *i.e.* surface or buried. It is the fact that contact sites between proteins are surface residues, that make previous MI based methods appear to be successful.

Since the contact residues of a pair of interacting domains are only on the surface of the domains, in the next chapter we consider only surface residues and evaluate whether the inter-domain contact prediction ability of original MI, MIp and MIc is enhanced when contact residue prediction is disentangled from surface prediction.

Chapter 4

Mutual Information Based Methods for Protein Inter-domain Contact Prediction

4.1 Chapter Overview

We have observed in the previous chapter that most MI variants give higher scores to surface residues. Since only residues on the surface of a domain are involved in domain-domain binding, we now consider just surface residues when attempting to predict inter-domain contacts. Original MI, MIp, MIc and ZNMI are tested on 40 inter-domain cases. We also formulate and assess two new versions of MI. These two novel MI measures are founded on assumptions and heuristics incorporated into the non-MI based, successful inter-protein contact predictor, i-Patch (Hamer *et al.* [2010]). The first is based on the idea that interactions occur in patches; hence we extend the MI variants to consider triangles of residues rather than pairs. The second novel MI approach we propose accounts for the physiochemical properties of the amino acids by employing a reduced alphabet residue set.

Our analysis reveals that eliminating buried residues improves the performance of MI, MIp and MIc. After considering only surface residues, MIc still outperforms the other tested MI algorithms. Our triangle and reduced alphabet variants of MI are not as successful, but are useful in calling attention to the trade-off between signal and noise when employing MI for contact prediction. A closer examination of a "successful" contact prediction case study, showed that even when considering surface residues only, the most accurate inter-domain contact predictor, MIc, does not perform significantly better than random.

If methods based on MI cannot be improved in their predictive power for contact pairs, then this fact may shed doubt on the theory that residues in contact undergo correlated mutations.

A majority of the content of this chapter is presented in the published article Gomes *et al.* [2012].

4.2 Introduction

In this chapter we eliminate buried residues in the 40 inter-domain test cases used in the previous chapters, and perform a systematic study of MI and its most recent extensions. We evaluate original MI, and variants MIp (Dunn *et al.* [2008]), MIc (Lee & Kim [2009]) and ZNMI (Brown & Brown [2010]) for inter-protein contact residue prediction.

We do not include aMIc (Lee & Kim [2009]) in this investigation because Lee & Kim [2009] found that their MIc measure outperformed their additionally normalised aMIc score, when both methods were assessed for inter-domain contact prediction on a set of 27 homo-trimers. Additionally, we observed in the previous chapters that the contact prediction performance of aMIc is very similar to MIc on our 40 inter-domain test set (Chapter 2), and aMIc has a slightly greater bias towards surface residues than

MIc (Chapter 3).

The ZNMI algorithm (Brown & Brown [2010]) evaluated in this investigation accounts for different alphabet sizes among columns in the multiple sequence alignment (MSA). Unlike the other measures, the ZNMI score is embedded in an iterative pipeline that aims to yield highly reproducible scores (Brown & Brown [2010]). Hence we have to analyse the performance of ZNMI differently to its competitors.

We have also attempted to strengthen the predictive capabilities of MI by introducing two new MI variants, both of which are motivated by the framework of a leading inter-protein contact predictor, i-Patch (Hamer *et al.* [2010], Section 1.8). Like the i-Patch score, the first variant considers triangles of residues rather than pairs, with the aim of enhancing the signal for contacts. This variant is referred to as MI3D and MIp3D. As MIc already considers a third column in its normalising term, it is not extended to triangle scores. Our second variant is designed to reduce noise by grouping residues in the MSA into seven physiochemical categories and subsequently calculating MI. This modification is indicated by the suffix RA (reduced alphabet), and the resulting five variants are: MIRA, MI3DRA, MIpRA, MIp3DRA and MIcRA. The i-Patch study introduced the seven physiochemical categories we employ and used this residue grouping technique to reduce the size of the propensity tables incorporated into the i-Patch score. Thus altogether we examine the inter-domain contact prediction ability of 10 MI measures: MI, MI3D, MIRA, MI3DRA, MIp, MIp3D, MIpRA, MIp3DRA, MIc and MIcRA, alongside the pipeline ZNMI.

Amongst the 10 tested MI variants and ZNMI, we find that MIc is the leading MI inter-domain contact predictor. After eliminating buried residues, the performance of all variants improve, the precision of MIc increases from 34.7% to 44.9% at 20% recall. Nevertheless, all versions of MI do not perform as well as the non-MI based inter-domain contact predictor, i-Patch (Hamer *et al.* [2010]), which achieves a precision of 48.9% at 20% recall on the same 40 test cases. The enhanced predictive ability of i-Patch may

arise from its use of surface residues only and residue propensity scores, in conjunction with the sequences of the two interacting proteins. Conversely, MI variants rely solely on the MSA.

We also revisit the highly popular Skerker *et al.* [2008] case study, and demonstrate that here, when considering surface residues, even the most accurate inter-domain contact predictor, MIc, performs no better than random.

4.3 Materials and Methods

For this inter-domain MI investigation we used the same 40 test cases described in the previous chapter (Section 2.3.1) and detailed in Table 2.1. Once again, we calculate the MI score for all pairs or triangles of columns, and then assign each residue the maximum MI score that its residue column achieved with any other residue column in the MSA. As previously, columns that have an entropy of 0, or one or more gaps are ignored. Contact and non-contact residue pairs are determined based on the criterion described in Section 1.4.4, while surface and buried residues are identified as described in Section 3.3.

Employing this criterion on our 40 test cases leaves us with 5,482 surface residues and 2,364 buried residues (Appendix Table 8). These numbers decline to 5,362 and 2,174 respectively when employing the 40 reduced alphabet MSAs, as the reduced alphabet MSAs have a greater number of columns with 0 entropy (Equation 1.3). The ratios of surface to buried residues in the reduced and non-reduced alphabet sets are 2.47 and 2.32 respectively.

Similarly, there are 1,342 contact and 4,141 non-contact surface residues (Appendix Table 8), over all 40 test cases. These numbers decline to 1,306 and 4,056 respectively when employing the reduced alphabet on the 40 test cases. The ratio of contact to non-contact residues is 0.322 and 0.324 in the reduced and non-reduced alphabet sets

respectively.

Entropy was measured as described in Section 1.6. The MI, MIp, MIc and ZNMI calculations used in this chapter are performed as outlined in Section 1.7. The number of 0 entropy columns in the reduced and non-reduced alphabet set are 668 and 326 respectively, while the number of columns with one or more gaps in both sets are 3,479.

P-ROC (Precision Recall Operating Characteristic) (Buckland & Gey [1994]) and MCC (Matthews Correlation Coefficient) (Matthews [1975]) curves (Section 1.9) are used to assess the contact prediction capabilities of each of the MI methods and i-Patch (Hamer *et al.* [2010]).

4.3.1 3-dimensional (3D) MI and MIp

Original MI and MIp (Dunn *et al.* [2008]) were adapted to consider triangles of residues;

$$MI3D_{unstandardised}(J;K;L) = \sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{l=1}^{s} P(J=j,K=k,L=l) \times \log \frac{P(J=j,K=k,L=l)}{P(J=j)P(K=k)P(L=l)}, \quad (4.1)$$

where MSA column J from domain 1 has n different residues, column K from domain 2 has m different residues, and column L from domain 2 has s different residues. Residues in the representative protein structure, corresponding to columns K and L, should be less than 4.5Å from each other in order to be considered as being on the same patch in the domain.

MIp3D is defined as

$$MIp3D_{unstandardised}(J;K;L) = MI3D_{unstandardised}(J;K;L) - APC3D(J;K;L). \quad (4.2)$$

In this equation $MI3D_{unstandardised}(J;K;L)$ is calculated as denoted in Equation 4.1

and $APC3D(J;K;L)$ is calculated as

$$APC3D(J;K;L) = \frac{\overline{MI3D}_{unstandardised}(J)\ \overline{MI3D}_{unstandardised}(K)\ \overline{MI3D}_{unstandardised}(L)}{\overline{MI3D}_{unstandardised}}, \quad (4.3)$$

where $\overline{MI3D}_{unstandardised}(J)$ is the average 3D mutual information for column J, $\overline{MI3D}_{unstandardised}(K)$ is the average 3D mutual information for column K, $\overline{MI3D}_{unstandardised}(L)$ is the average 3D mutual information for column L, and $\overline{MI3D}_{unstandardised}$ is the overall average 3D mutual information.

In order to compare the 3D mutual information scores between test cases, MI3D and MIp3D scores were standardised in a manner similar to those described in Equations 1.7 and 1.10, respectively. Once again 0 entropy columns and columns containing one or more gaps were ignored.

4.3.2 Reduced alphabet MI scores

We grouped the 20 amino acids into the same seven physiochemical categories employed by Hamer *et al.* in their inter-domain contact predictor, i-Patch (Hamer *et al.* [2010]). These seven categories include: Small (S,G,A,P), Hydrophobic (V,M,I,L,C), Negatively charged (D,E), Aromatic (F,Y,W), Polar (Q,T,N), Favoured Positively-charged (R,H), and Disfavoured Positively-charged (K). These physiochemical groups are abbreviated to S, H, N, A, P, F and D respectively. Hamer *et al.* introduced Disfavoured and Favoured Positively-charged categories because Lysine (K) was found to be rare in protein/domain interfaces (propensity 0.66), while Arginine (R) and Histidine (H) were far more common (propensities of 1.05 and 1.11, respectively) (Hamer *et al.* [2010]).

We replaced the amino acid alphabets in each MSA by their corresponding category abbreviation and recalculated MI, MIp, MIc, MI3D and MIp3D as described above. The five new MI variant scores are referred to as MIRA, MIpRA, MIcRA, MI3DRA and MIp3DRA.

We choose to employ this particular set of seven physiochemical categories as it was successfully used by i-Patch (Hamer *et al.* [2010]) in inter-domain contact prediction. We do not expect another grouping to dramatically improve the predictive capabilities of MI and its variants further.

4.3.3 Sub-sampling to test stability of MI scores

To test the stability of the 10 MI variant scores under minor changes in the MSA, for each test case 70% of sequences in the MSA are randomly selected and all 10 MI scores are recalculated and 10 respective P-ROC curves are plotted. This sub-sampling and calculation process is repeated 100 times per test case for every MI variant. Then the average and standard error of the precision values for the 100 P-ROC curves are calculated for each MI variant.

This sub-alignment creation and MI recalculation process is only carried out on those 24 test cases that have at least 200 sequences to ensure that a minimum of 125 sequences are retained in each sub-alignment, the suggested minimum number of sequences required to reduce the stochastic noise in the MSA (Martin *et al.* [2005]).

4.4 Results and Discussion

4.4.1 Prediction capability of MI variants for contact *versus* non-contact surface residues

After filtering out buried residues in the 40 test cases, the precision of MIc increases from 34.7% (Figure 2.2) to 44.9% at 20% recall (Figure 4.1). The probability of randomly selecting a contact residue is now 24.5%, as opposed to 17.1% when buried residues were included. Excluding buried residues therefore clearly has a considerable effect. As can be observed in Figure 4.1, Table 4.1 and Appendix Figure 5, MIc still outperforms the other MI variants. MI and MIp achieve a precision of 24.4% and 42.3% respectively

at 20% recall (Figures 4.1 A and B, and Table 4.1).

MI variant	**precision contact *vs.* non-contact**
MIc	44.9
MIp	42.3
MIcRA	36.9
MIpRA	35.8
MIp3D	31.8
MIp3DRA	29.4
MIRA	28.4
Random	24.5
MI	24.4
RandomRA	24.4
MI3DRA	23.5
MI3D	19.9

Table 4.1: **Precision for detecting contact *versus* non-contact residues at 20% recall for inter-domains, when only surface residues are considered.** Results are given for the 40 test cases. MI variants are listed in descending order of contact *versus* non-contact precision, *i.e.* best to worst classifier of contact residues. The probability of randomly selecting a contact residue from all surface residues is 24.5%. This probability changes to 24.4% when using the reduced alphabet amino acid set because residues are lost as the entropy of their corresponding MSA column reduces to 0.

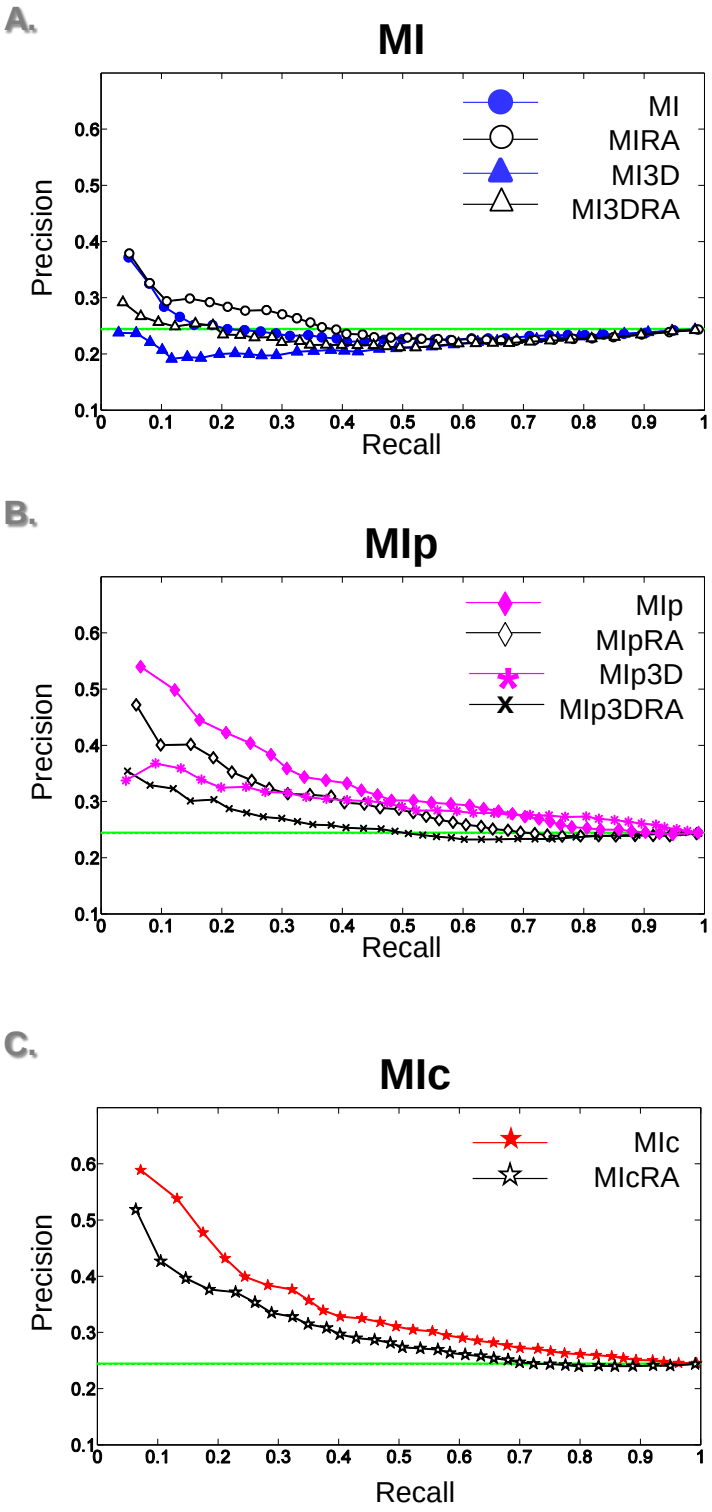

Figure 4.1: **Contact *versus* non-contact prediction P-ROC curves for MI variants on the 40 test cases, when only surface residues are considered.** A, B and C illustrate the performance of MI, MIp and MIc variants respectively when distinguishing contact from non-contact surface residues. The solid green line in all plots depicts the chance of randomly selecting a contact residue, while the dashed green line indicates the probability of randomly selecting a contact residue when employing the reduced alphabet amino acid set.

Using our sub-sampling procedure to test stability, we observed that the rank order of the top five MI variant scores is maintained (Table 4.2). The results in Table 4.1 refer to all 40 test cases, while those in Table 4.2 pertain to the subset of 24 cases with at least 200 sequences in the MSA. Based on two-sample t-tests, with a sample size of 24, the differences between the top four scores in Table 4.2 are highly significant at the 0.1% level.

Additionally, it is worth noting that the performance of all non-3D MI variants improve when using MSAs that have at least 200 sequences (Tables 4.1 and 4.2).

MI variant	**70% AVG**	**70% STD DEV**	**100%**
MIc	52.5*	2.1	54.8
MIp	46.0*	2.1	47.4
MIcRA	41.9*	1.8	41.4
MIpRA	38.2*	1.5	38.5
MIp3D	30.6	1.3	28.5
MIRA	30.2*	1.2	31.4
MIp3DRA	28.0*	1.2	30.9
MI	25.8*	0.8	27.6
MI3DRA	23.2*	1.0	25.5
Random	-	-	24.4
RandomRA	-	-	24.4
MI3D	20.0	0.6	21.8

Table 4.2: **Precision for detecting contact *versus* non-contact surface residues at 20% recall, for sub-alignments of 70%.** 70% of sequences in an MSA were randomly selected and the 10 MI variant scores based on the new sub-alignment were calculated. This subset selection and calculation procedure was repeated 100 times for those test cases that had $\geq$200 sequences to ensure $\geq$125 sequences in each sub-alignment (Martin *et al.* [2005]). Thus 24 test cases were used. For each of the 100 iterations a P-ROC curve similar to Figure 2 was plotted for the 24 test cases (figures not shown), and the precision at 20% recall recorded. Columns one and two, respectively, contain the averages and standard deviations of these 100 precision values. Column three indicates the precision attained at 20% recall when all sequences in the 24 original MSAs were used. When considering this set of 24 cases, the probability of randomly selecting a contact residue from all surface residues is 24.4% generally and when using the reduced alphabet MSAs. The MI variants are listed in descending order of the average precision of the 100 70% sub-alignments. The presence of an '*' at a MI variant indicates that the difference between the precision of this MI variant and the next lowest is significant at the 0.1% level, when using two-sample t-tests with a sample size of 24.

To account for the mentioned variability in scores due to changes in the MSA, Brown and Brown designed a novel MI measure, ZNMI, as well as a methodology to yield highly reproducible and accurate contact pair prediction scores (Brown & Brown [2010]). Their suggested algorithm repeatedly partitions the MSAs into 50% sub-alignments, calculates the pair scores, retains significant scoring pairs for each

partition and subsequently compares all partitions to acquire consensus pair scores. It should be noted that unlike our methodology, this pipeline does not filter out buried residues. The authors provided us with code for original MI, MIp (Dunn *et al.* [2008]), OMES (Fodor & Aldrich [2004]; Kass & Horovitz [2002]), SCA (Halabi *et al.* [2009]), ZNMI (Brown & Brown [2010]) and ZRES (Little & Chen [2009]) measures wrapped within their proposed pipeline, but unfortunately not for MIc. Having run this code on our 40 inter-domain test cases we find that using ZNMI in conjunction with their algorithm does improve on the performance of original MI; at 20% recall the precision of ZNMI is 30.5% (Table 4.3), as opposed to the 24.4% precision of original MI (Figure 4.1A and Table 4.1). ZNMI within the Brown and Brown pipeline even outperforms MIp, when MIp is incorporated into the same pipeline (27.1% precision at 20% recall; Table 4.3). However, the performance of MIp independent of the pipeline, after filtering out buried residues and columns with one or more gaps, supersedes ZNMI and all other coevolving residue algorithms tested by the authors, as illustrated by its precision of 42.3% at 20% recall (Table 4.3).

algorithms	**precision contact *vs.* non-contact**
MIp - original, minus buried	42.3
SCA	31
ZNMI	30.5
ZRES	28.9
MIp	27.1
MI	25.7
OMES	25.7
MI - original, minus buried	24.4

Table 4.3: **Precision at 20% recall of contact prediction algorithms used within Brown & Brown [2010] pipeline.** Results are given for the 40 test cases. The Brown & Brown [2010] pipeline was applied to the contact residue prediction algorithms listed in column one, with the exceptions of MIp and MI "original, minus buried." As in Table 4.1, these two algorithms were run independently of the pipeline and buried residue columns, residue columns with one ore more gaps or an entropy of 0 were filtered out. The table is arranged in descending order of precision.

4.4.2 3-dimensional (3D) and reduced alphabet MI adjustments

To investigate methods that might further enhance the predictive power of MI variants we designed two adjustments that are motivated by the implementation of i-Patch (Hamer *et al.* [2010], Section 1.8). The first adjustment considers triangles of columns rather than pairs, based on the idea that interactions occur in patches (Hamer *et al.* [2010]; Madaoui & Guerois [2008]). This variant is denoted by the suffix 3D. The second adjustment, suffixed RA, reduces the 20 amino acids to seven categories based on their physical and chemical properties, with the aim of reducing noise.

The idea behind the 3D version is that protein binding involves patches of residues in contact. This idea has been previously used to predict contact residues (Hamer *et al.* [2010]; Madaoui & Guerois [2008]). Furthermore, the success of MIc lies in its normalising factor, the coevolutionary pattern similarity (CPS) score, which estimates the coevolutionary relationship between the pair of residues currently under consideration and all other residues in the MSA (Lee & Kim [2009]). We thus speculated that adding additional residue information to MI and MIp pair scores may enhance their inter-domain contact predictive capabilities. Hence we created new versions of MI and MIp that consider triangles rather than pairs of columns to identify contacts (MI3D (Equation 4.1) and MIp3D (Equation 4.2)). Increasing the dimensionality of MI and MIp in this manner surprisingly worsened performance in both cases; the precision at 20% recall of MI3D and MIp3D are 19.9% and 31.8% respectively as compared to precision of MI and MIp of 24.4% and 42.3% (Figure 4.1 A and B, and Table 4.1). We conjecture that adding an extra dimension to MI and MIp magnifies the noise in the MSA more than it boosts the signal.

Assuming that contact residues mutate in a correlated manner in order to maintain their interaction, it is not evident how much of a change a residue can undergo while still maintaining its contacts. Using a reduced alphabet residue set addresses this point; as it groups residues by their physiochemical properties, under the assumption that

residues with the same physiochemical properties will maintain similar interactions. Grouping the 20 amino acids into seven categories only improved the performance of basic MI and MI3D, which rose in precision at 20% recall from 24.4 to 28.4% and 19.9 to 23.5% respectively (Figure 4.1A and Table 4.1). In all other cases the reduced alphabet (RA) appeared to reduce noise as well as signal (Figure 4.1, Table 4.1 and Appendix Figure 5).

All 3D and RA variants also exhibit a bias towards surface residues (Appendix Figure 6). As speculated in Chapter 3, this is probably owing to the mathematical relationship between MI (Equation 1.6) and entropy (Equation 1.3), and subsequently the observed higher entropy of MSA columns corresponding to surface residues than those associated with buried residues (Figure 3.2).

4.4.3 Case study

We revisit the Skerker *et al.* [2008] case study discussed in the previous chapter, as it has received a lot of attention for successfully determining inter-protein contact specificity residues with the aid of MI. To reiterate, the authors used original MI (Equation 1.6) to determine a subset of contact residues that allow for specific binding of a histidine kinase (HK) with its interacting response regulator (RR). In Chapter 3 we found that all but one residue involved in MI scores above the Skerker *et al.* imposed score cutoff are surface residues.

The MSA provided by the authors does not contain the sequence of the structure used in their analysis. Hence we now compute MI and MIc on the HK-RR MSA provided by Hamer *et al.* [2010], which does include the sequence of this reference structure. As Skerker *et al.* were interested in residue pairs only between the DHp domain (four helix bundle) of the HK and its interacting RR, only these MI and MIc scores are considered when examining performance. In accordance with our evaluation method on the 40 test cases in this chapter, all buried residues are eliminated, as are

residues corresponding to columns that have one or more gaps, or an entropy of 0. This leaves us with 46 DHp residues, nine of which are contacts, and 68 RR residues, amongst which 24 are contacts.

We check the number of correct predictions among the top nine predictions for DHp, since DHp has 9 contact residues. If there was no relationship between the MI scores and contact sites, then the number of correct predictions would follow a Binomial distribution with sample size nine and probability of success 9/46. Under this model we would expect 1.76 correct predictions.

For RR there are 24 contact residues. We check the number of correct predictions among the top 24 predictions. If there was no relationship between the MI scores and contact sites, then the number of correct predictions would follow a Binomial distribution with sample size 24 and probability of success 24/68. Under this model we would expect 8.47 correct predictions.

The results are recorded in Table 4.4. The p-value is the probability of seeing a number this large or larger under the corresponding Binomial model. None of the p-values are below 5%. Therefore at the 5% level there is no statistical evidence to reject the null hypothesis that in this case study random guess does as well as MI and MIc.

	DHp: 9 contacts out of 46		**RR: 24 contacts out of 68**	
	contacts among top 9	**p-value**	**contacts among top 24**	**p-value**
MI	3	0.2503	9	0.4864929
MIc	4	0.0800	9	0.4864929

Table 4.4: **Performance of MI and MIc on a histidine kinase (HK) - response regulator (RR) complex.** MI and MIc are run on the HK-RR MSA provided by Hamer *et al.* [2010]. Each surface, ungapped DHp residue column, having a column entropy greater than 0, is assigned the maximum score it achieves when paired with the RR residue columns. These DHp residues are then ranked according to score and the number of true contact residues amongst the top nine scores are recorded for MI and MIc respectively. The same steps are applied to residues in the RR and the number of true contact residues in the top 24 scores are counted for MI and MIc. The p-value refers to the probability of seeing a number this large or larger under the corresponding Binomial model.

4.5 Conclusions

Eliminating buried residues improves inter-domain contact prediction. Amongst the MI variants we tested, MIc is the best inter-domain contact predictor. Its predictive capabilities, however, are not as high as i-Patch (Hamer *et al.* [2010]), a non-MI based inter-domain contact predictor, but unlike this algorithm MIc relies solely on sequence information in an MSA. Our 3D and reduced alphabet variants of MI did not improve prediction, but illustrate the delicate trade-off between signal to noise in the use of MI for inter-domain contact prediction. Examining a "successful" contact prediction case study revealed that, after buried residues are eliminated, even MIc does not perform significantly better than random.

Chapter 5

Gap Cutoffs for Alignment Columns when Using Mutual Information Based Methods for Intra-domain Contact Prediction

5.1 Chapter Overview

During the work undertaken in previous chapters we became aware that the choice of gap cutoffs for alignment columns influences the contact prediction abilities of the MI based methods. To our knowledge there is no systematic study analysing the influence of gaps on MI. An understanding of the appropriate gap cutoff to use may strengthen the structure prediction capabilities of MI based algorithms. This chapter details our investigation on this topic.

5.2 Introduction

When executing MI algorithms there are several seemingly ad hoc choices made that may greatly influence the quality of prediction. For example, in Martin *et al.*'s multiple sequence alignment (MSA) simulation studies they found that MSAs with at least 125 sequences yield more accurate MI contact predictions. They use this MSA sequence count cutoff in conjunction with their MI variant, MIr (Martin *et al.* [2005]). The same group later designed MIp and continued to use MSAs with at least 125 sequences (Dunn *et al.* [2008]). Conversely, Lee and Kim used MSAs with more than 100 sequences when investigating their novel MI based contact prediction tools, MIc and aMIc (Lee & Kim [2009]). This discrepancy in the minimum of number of sequences that can be used also carries over to gap cutoffs.

Gaps represent insertions and deletions in homologous sequences in the MSA. It is theorised that conserved residues in the MSA of a protein family are integral to maintaining the structure and function of proteins belonging to that family (Barton [1990]; Benner *et al.* [1994]; Crawford *et al.* [1987]). This implies that residues in gapped columns in an MSA may be less essential for protein structure and interaction. This is perhaps why most studies on MI based algorithms have opted to consider only ungapped alignment columns in their analyses (Chakrabarti & Panchenko [2009]; Dunn *et al.* [2008]; Lee & Kim [2009]), while some allow columns with up to 10% (Brown & Brown [2010]) or 50% (Buslje *et al.* [2009]) gaps.

In this chapter we consider 2,144 Pfam domain MSAs (Punta *et al.* [2012]) and assess the performance of original MI and the current leading MI variants, MIp (Dunn *et al.* [2008]), MIc (Lee & Kim [2009]) and aMIc (Lee & Kim [2009]) for intra-domain contact prediction with varying gap cutoffs. Contrary to popular belief we do not find that the percent of contact pairs out of total pairs steadily declines in columns with increasing number of gaps. In other words, residues involved in contacts are not

restricted to conserved columns. At the 0% gap penalty only 2.02% of contact pairs are contained in this dataset; suggesting that this gap cutoff employed by MIp, MIc and aMIc is too stringent, and results in a significant loss of information. We find that the highest proportion of contacts is correctly identified by all variants around the 10% gap cutoff, and 48.8% of contact pairs in our dataset are included in the MI analyses at this gap penalty. However, when total number of residue pairs evaluated is of importance, the performance of all variants is optimal around the 40% gap cutoff.

In order to assess whether the results are partial to the algorithm used to build the alignments in the Pfam database, we re-run all our tests on the 80 domains employed in Chapters 2 and 3 (Table 2.1), that were originally acquired from Hamer *et al.* [2010]. Unlike Pfam that uses a Hidden Markov Model to build its alignments, the Hamer *et al.* MSAs were generated using the alignment software MUSCLE (Edgar [2004])and MaxAlign (Gouveia-Oliveira *et al.* [2007]). Hamer *et al.* then checked the MSAs by eye, and those that did not appear to be correct were excluded from the dataset. Once again we observe that residues involved in contacts are not restricted to conserved columns. When using this dataset, the highest proportion of correctly identified contacts by MIp, MIc and aMIc occurs when the 0% gap cutoff is used. The 10% gap cutoff is best for original MI. When the total number of residue pairs evaluated is important, a gap cutoff of 20, 30 or 40% optimises performance. Significant improvement in performance of MIp, MIc and aMIc is only made by residue pairs added between the 0 and 10% gap cutoff interval, suggesting that a 0 or 10% cutoff would generally optimise performance of the MI variants when alignments from the Hamer dataset are used.

Our analysis has thus shown that depending on the evaluation criteria a range of gap cutoffs are suitable. The desired proportion of correctly predicted contact pairs, and the alignment method used, dictates the gap cutoff that should be employed. As a rule of thumb, a 10% gap cutoff appears to maximise performance. At the 10% gap cutoff, MIc is the leading intra-protein contact predictor, when either our Pfam or

Hamer dataset are used.

5.3 Materials and Methods

Contact *versus* non-contact residues in this chapter are determined using the criterion listed in Section 1.4.4.

5.3.1 Datasets

Pfam dataset

Unlike previous investigations (Dunn *et al.* [2008]; Lee & Kim [2009]; Martin *et al.* [2005]) we do not use MSAs from the Conserved Domain Database (CDD) (Marchler-Bauer *et al.* [2011]) because these alignments are curated such that all gaps are removed from the conserved structural core motifs across all rows of the MSA, leaving gapped, unaligned regions between structural motifs.

Instead we use the Pfam-A database. We choose Pfam-A, rather than Pfam-B, as it contains seed alignments that are manually curated, resulting in better quality alignments (Punta *et al.* [2012]).

The 2,144 MSAs used in this study were obtained from the Pfam-A database, version 26.0, downloaded April 2012. Pfam-A uses a curated seed alignment, which is made up of a small set of representative members in a domain family, to build a profile Hidden Markov Model (HMM). This HMM is then used to automatically generate a full alignment containing all detectable sequences in the Pfamseq database belonging to that family.

In order to assess the contact prediction capabilities of the various MI algorithms, Pfam-A MSAs that include a sequence of known structure are required. Amongst the 13,672 Pfam-A entries, 4,132 MSAs have one or more sequences with a known Protein Data Bank structure (PDB, Berman *et al.* [2000]) having a resolution of less than

3Å. Of these 4,132 alignments, 2,451 MSAs contain a sequence of known structure that covers at least 75% of the alignment. This allows us to test a minimum of 75% of columns in the MSA for correct contact and non-contact prediction. In order to acquire a representative set of MSAs and remove MSA outliers that have too few or too many sequences, we rank the 2,451 entries by number of sequences and eliminate the top and bottom 5% of entries. The top 5% of the 2,451 alignments have between 11,496 to 288,250 sequences, whilst the bottom 5% have between 2 to 27 sequences. Eliminating these leaves us with 2,205 entries. To diminish the possibility that we are analysing fragments of protein domains, rather than entire domains, we then discard 53 entries that have at most 20 columns in the MSA. The PDB structure files of the sequences of known structure for eight of the remaining 2,152 entries contain errors that prevent us from determining contact, non-contact, surface and buried residues of these eight MSAs. Subsequently, we are left with 2,144 Pfam-A alignments in our final dataset. These data acquisition steps are outlined in Figure 5.1.

In each of the 2,144 MSAs, the sequence of known structure with the highest alignment coverage, *i.e.* least number of gaps, and lowest resolution is selected as the "reference structure" for that alignment. We then truncate each MSA to begin and end at the first and last position of the reference structure sequence. We also only consider columns in the MSA that do not contain a gap in the reference sequence. These choices are made because there is no available structure information for columns beyond the length of the reference sequence, or for gapped positions.

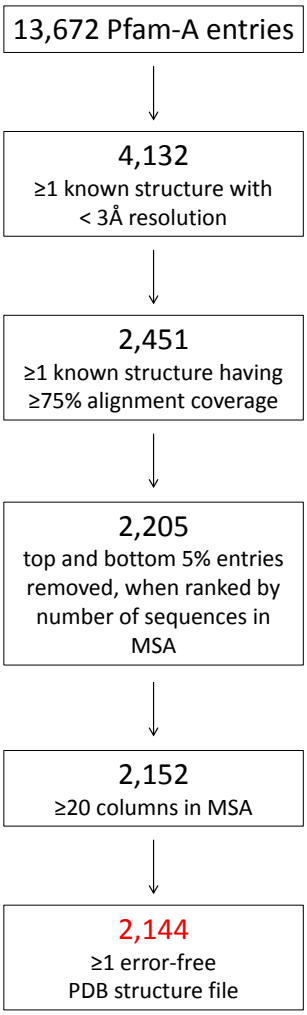

Figure 5.1: **Data acquisition for gap cutoff analysis.** Present in the Pfam-A database (version 26.0, downloaded April 2012, Punta *et al.* [2012]) are 13,672 entries, each with an associated multiple sequence alignment. Of these, 4,132 MSAs have one or more sequences with a known PDB structure having a resolution of less than 3Å. Out of the 4,132 MSAs, 2,451 MSAs contain a sequence of known structure that covers at least 75% of the alignment. The 2,451 entries are ranked by number of sequences and the top and bottom 5% of entries are eliminated, leaving 2,205 MSAs. Entries that have at most 20 alignment columns are discarded next. The PDB structure files of the sequences of known structure for eight of the remaining 2,152 entries contain errors. This leaves us with 2,144 Pfam MSAs.

An overview of the properties of the alignments in our Pfam dataset can be found in Table 5.1. Any non-standard amino acid entries in the 2,144 MSAs, such as B, Z, X, * and ?, are treated as a gap as there is no established method of processing these;

0.0014% of the residue positions over all 2,144 MSAs are dealt with in this manner.

	Pfam dataset			Hamer dataset		
	minimum	**maximum**	**median**	**minimum**	**maximum**	**median**
sequences	28	11,485	809	44	709	227
columns	21	1,262	148	57	873	225
percent gaps	0	95.5	9.94	4.84	72.0	37.5

Table 5.1: **Properties of the alignments in the datasets.** The minimum, maximum and median number of sequences, columns and percent of gaps in the alignments of the Pfam and Hamer datasets.

There are 852,289 contact and 46,504,119 non-contact residue pairs over all 2,144 test cases. Contact residue pairs constitute 1.80% of all pairs. The 2,144 Pfam entries contain a total of 98,282 buried residues and 283,918 surface residues. The ranges of the MI, MIp, MIc and aMIc values and their corresponding standardised scores for our Pfam dataset are given in Table 5.2.

	Pfam dataset		Hamer dataset	
	non-standardised	**standardised**	**non-standardised**	**standardised**
MI	0 to 0.750	-3.40 to 59.8	0 to 0.700	-2.80 to 11.6
MIp	-1.38 to 0.473	-84.1 to 62.1	-0.857 to 0.419	-24.6 to 19.0
MIc	-1.47 to 0.477	-86.3 to 26.1	-0.924 to 0.482	-18.3 to 11.0
aMIc	-3.99×10^3 to 1.00	-108 to 39.3	-2.33 to 1.00	-9.23 to 11.6

Table 5.2: **Range of MI scores in the datasets.** The range of values of the MI based methods when all 2,144 Pfam test cases and 80 Hamer test cases, respectively, are considered. The range of corresponding standardised scores are also included.

Hamer dataset

The 80 domains listed in Table 2.1 and used in this investigation are taken from the Hamer *et al.* [2010] study. To create each MSA the authors used a protein of known structure with less than 2.5Å resolution and well defined domain boundaries as a BLAST query (Altschul *et al.* [1990, 1997]) against the NCBI-NR database (Sayers *et al.* [2012]). Homologs identified were made non-redundant at the 90% level using Cd-hit (Li & Godzik [2006]). The final alignment was generated using MUSCLE (Edgar [2004]) and MaxAlign (Gouveia-Oliveira *et al.* [2007]). The authors discarded alignments that were obviously flawed.

When all residue columns in these alignments are considered, there are a total of 25,704 contact and 951,532 non-contact residue pairs across all 80 test cases; therefore 2.63% of all residue pairs are contacts. The 80 test cases in the Hamer dataset consist of a total of 3,023 buried residues and 8,500 surface residues. Across the 80 MSAs, 0.0025% of residue positions are non-standard amino acid entries and are treated as gaps. An overview of the properties of these 80 alignments and the ranges of the MI variant scores attained for these test cases can be found in Tables 5.1 and 5.2 respectively.

5.3.2 Varying the gap cutoff

In this investigation we evaluate the ability of MI algorithms to predict contact pairs, when MSA columns of varying gap percentages are retained in the analysis. Hence we calculate the percent of gaps in each column. At a given gap percent cutoff, only those columns in the MSA for which the percent of gaps are less than or equal to the specified cutoff are used in the MI analysis. For example, at a gap penalty of 0% MI, MIp, MIc and aMIc are calculated for ungapped columns only. Similarly at 50% gap cutoff MI, MIp, MIc and aMIc are determined for all columns that are composed of at most 50% gaps. At a 100% gap penalty all columns are included in the MI analysis.

5.3.3 Performance evaluation metrics

The N×MCC^2, MCC and F-measure evaluation metrics, outlined in Section 1.9, are used to assess the contact prediction abilities of the MI variants.

At each incrementing gap cutoff a higher number of columns in the alignments are considered, and subsequently a greater number of MI pair scores are calculated. The N×MCC^2 measure takes into account the total number of MI scores being considered at each gap cutoff in its N term (Equation 1.24). It assesses how well the MI variant can differentiate contact from non-contact residue pairs given the total number of residue

columns included in the analysis. Therefore it offers a standardised scale for comparison of the performance of an MI variant when different gap cutoffs are used and the total number of MI scores calculated varies. The N×MCC^2 measure also allows us to identify if the overall prediction of an MI variant is better than random at a particular gap cutoff, given the number of column pairs evaluated at that cutoff.

The MCC measure, which does not account for the number of scores being considered (Equation 1.23), instead enables us to identify the gap cutoff at which an MI variant achieves the highest recall / sensitivity. Recall, also known as sensitivity, is the proportion of positives that are correctly identified as such.

The best F-measure is often selected as the optimal trade-off point between recall/sensitivity and specificity. Specificity is the proportion of negatives that are correctly identified as such.

In this study positives and negatives are contact and non-contact residue pairs, respectively.

5.3.4 Calculating the MI variants

The MI, MIp, MIc and aMIc calculations used in this chapter are performed as described in Section 1.7.

As mentioned in Section 1.7, for this chapter alone we did not use the code made available by Lee & Kim [2009] to calculate their MIc and aMIc measures, but instead wrote our own code. We observed a significant loss of information with incrementing gap cutoffs when using the code provided by the authors. As only ungapped columns were used in previous chapters this problem was not observed previously.

The Lee & Kim [2009] code includes the "not a number" (nan) MI scores in their CPS calculations (Equation 1.11). These nan MI scores result from columns with all gapped pairs. Including these nan scores in the CPS calculations produces nan CPS values, which in turn causes the denominator of NCPS to be nan (Equation 1.11).

Subsequently all NCPS scores for the alignment will be nan, and consequently all MIc and aMIc values for that particular test case will also result in nan (Equation 1.13 and 1.17). To avoid this loss of information we wrote a version of the code that ignores all nan MI values when calculating CPS (Equation 1.11), therefore producing a greater number of valid MIc and aMIc scores, and successfully using MIc and aMIc on a larger number of test cases. This is evident in Table 5.3. For the gap cutoffs from 0 to 50% on which the original code also yields valid MIc and aMIc scores for all 80 Hamer test cases, the results of both versions of the code are consistent. This consistency can be observed in Figures 5.2 and 5.3; to ease visualisation the MCC scores of only the 0, 20 and 50% gap cutoffs are shown.

gap cutoff	**Original Code no. of test cases**		**Our Code no. of test cases**	
	MIc	**aMIc**	**MIc**	**aMIc**
0	80	80	80	80
10	80	80	80	80
20	80	80	80	80
30	80	80	80	80
40	80	80	80	80
50	80	80	80	80
60	79	79	80	80
70	76	76	80	80
80	61	61	80	80
90	33	33	80	80
100	0	0	80	80

Table 5.3: **Number of Hamer dataset test cases successfully calculated by the original MIc and aMIc code *versus* our code.** Out of the 80 test cases in our Hamer dataset, the number of test cases for which the code provided by Lee & Kim [2009] did not produce nan MIc and aMIc scores, respectively, as compared to our code to calculate the same.

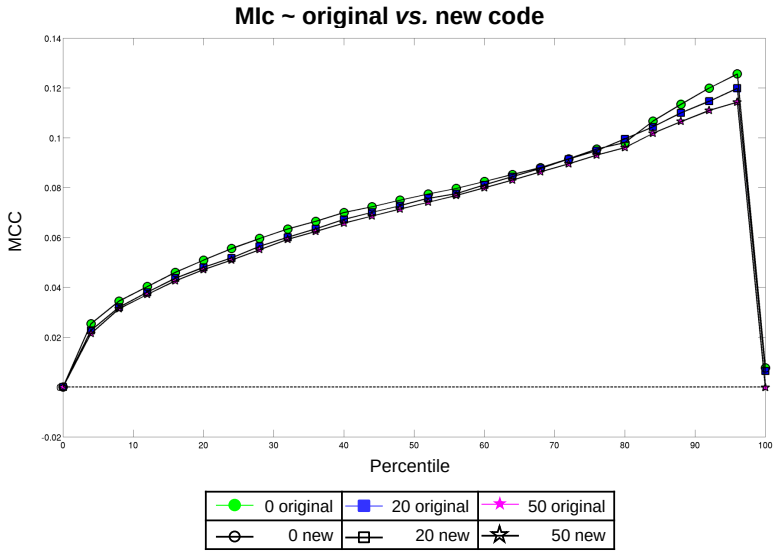

Figure 5.2: **Comparing the MCC curves at the 0%, 20% and 50% gap cutoffs, using the original *versus* new MIc code on the Hamer dataset.** A curve illustrates the performance of MIc when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue. The results of the MIc calculations performed by code provided by Lee & Kim [2009] are depicted by the solid colour curves, while the results produced by our MIc code are depicted by a black outline. The results of both versions of the code overlap exactly.

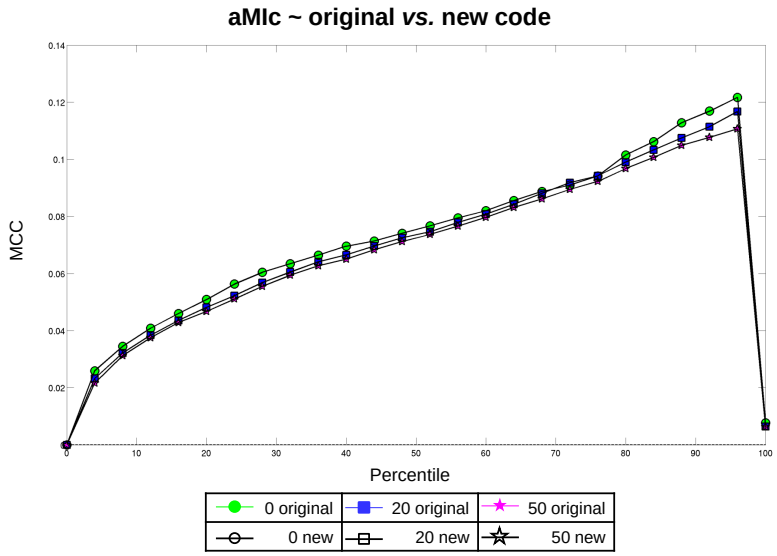

Figure 5.3: **Comparing the MCC curves at the 0%, 20% and 50% gap cutoffs, using the original *versus* new aMIc code on the Hamer dataset.** A curve illustrates the performance of aMIc when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue. The results of the aMIc calculations performed by code provided by Lee & Kim [2009] are depicted by the solid colour curves, while the results produced by our aMIc code are depicted by a black outline. The results of both versions of the code overlap exactly.

5.4 Results and Discussion

5.4.1 The Pfam dataset

5.4.1.1 Contacts lost with varying gap cutoffs

It is believed that residues integral to preserving the structure of a protein are conserved in proteins belonging to the same family (Barton [1990]; Benner *et al.* [1994]; Crawford *et al.* [1987]). Hence we would expect the columns in the MSA pertaining to contact residues to contain no or few gaps. This is perhaps why the MIp, MIc and aMIc studies decided to use only ungapped columns (Dunn *et al.* [2008]; Lee & Kim [2009]). In our Pfam dataset, however, we find that just 2.02% of contact pairs are considered when using only ungapped columns, *i.e.* at the 0% gap cutoff. Additionally, we do not observe a trend that suggests that the percent of contact pairs introduced steadily declines with increasing gap cutoffs (Figure 5.4). Therefore we cannot assume that residues in columns with more gaps are less important for contacts.

It is true however that we are more likely to find contacts in columns that have up to 10% gaps, 48.8% of contacts (Table 5.4 and Figure 5.5), but we speculate that this could simply be because MSA algorithms build alignments on the premise that contact residues are more preserved, and not because in evolutionary biology contact residues are more conserved.

These findings would suggest that either contacts are no more preserved than other residues or that the quality of the Pfam MSAs employed in this study are not very good, or both.

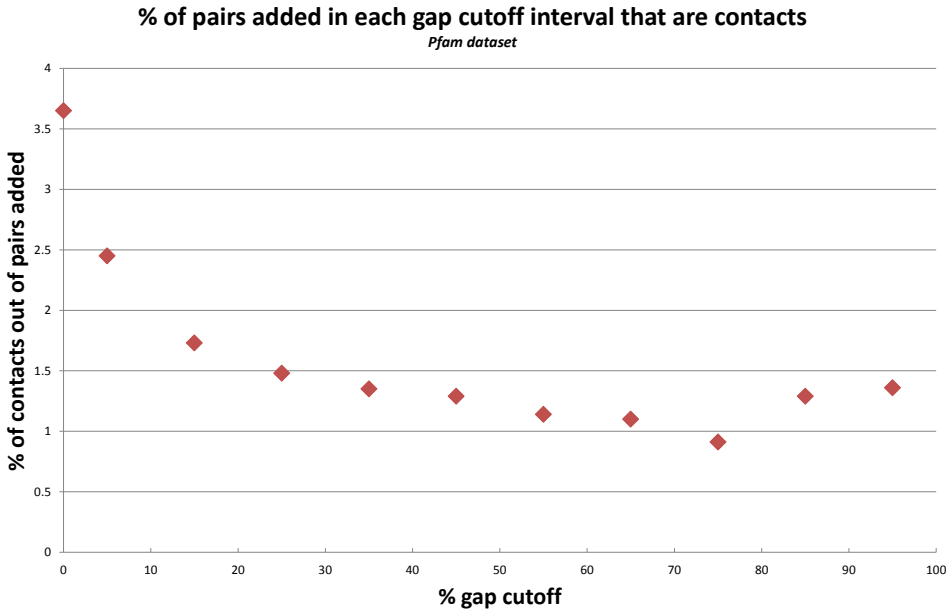

Figure 5.4: **Percent of pairs added in each gap cutoff interval that are contacts, using the Pfam dataset.** The percent of total residue pairs added in consecutive gap cutoff intervals that are contact pairs. To take one example, of the 2,926,813 total pairs introduced at the 50% cutoff that were not included at the 40% gap cutoff, 1.29% are contacts. When considering only ungapped columns, *i.e.* at 0% gap cutoff, 3.65% of the 470,765 total pairs present at this gap penalty are contacts.

Gap Cutoff Interval	% of Total Contact Pairs Added
0%	2.02
0%-10%	46.8
10%-20%	19.4
20%-30%	10.5
30%-40%	6.14
40%-50%	4.42
50%-60%	3.10
60%-70%	2.52
70%-80%	1.97
80%-90%	1.47
90%-100%	1.73

Table 5.4: **Contact pairs added in each gap cutoff interval, using the Pfam dataset.** When considering only ungapped columns, *i.e.* at the 0% gap cutoff, 2.02% of the 852,289 total contact pairs are present. At the 10% gap cutoff, 46.8% of the 852,289 total contact pairs are introduced that were not included at the 0% gap cutoff, at the 20% gap cutoff 19.4% of contact pairs are newly introduced, at the 30% cutoff 10.5%, and so on.

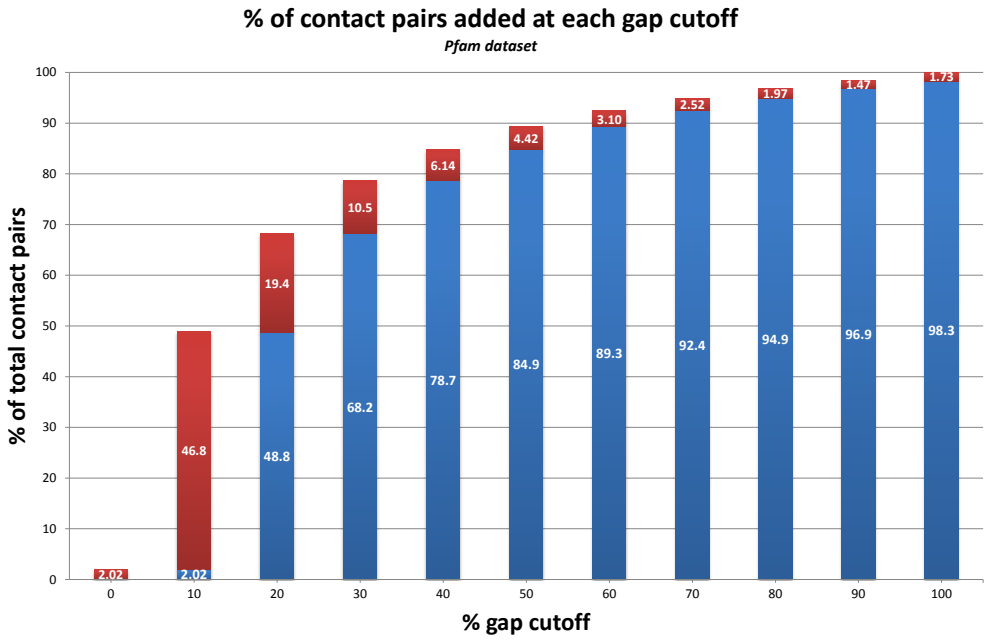

Figure 5.5: **Contact pairs added at each gap cutoff, using the Pfam dataset.** The red bars represent the percent of contact pairs newly introduced at each gap cutoff, while the blue bars denote the percent of contact pairs present at the preceding gap cutoff. The number in white within each red and blue bar, indicate the percent of contact residue pairs accounted for in each red and blue bar. For example, when considering only ungapped columns, *i.e.* at the 0% gap cutoff, 2.02% of the 852,289 total contact pairs are present. At the 10% gap cutoff, 46.8% of the 852,289 total contact pairs are newly introduced, making the percent of total contact pairs included at the 10% gap cutoff 48.8%.

5.4.1.2 The effect of gap cutoffs on MI

The figures for this section can be found on pages 103 to 111.

By the 40% gap cutoff 84.9% of contact pairs are included (Table 5.4 and Figure 5.5), and all are included at the 100% gap cutoff, therefore for ease of visualisation we only plot 0%, 10%, 20%, 30%, 40% and 100% gap cutoff curves in Figures 5.6, 5.7 and 5.8.

Plotting MCC curves we observe that the MI algorithms do not attain the highest proportion of correctly classified contact residue pairs at the 0% gap cutoff, nor at the 100% limit, but rather between a 5 to 20% gap cutoff (Figure 5.6). The cutoff used in

ZNMI (Brown & Brown [2010]) is 10% and this appears to be approximately optimal.

On the other hand, when keeping in mind the number of column pairs evaluated, the performance of all MI variants is generally best at the 40% gap cutoff (Figure 5.7 and Table 5.5).

Therefore if a higher recall is desired, between a 5 to 20% gap cutoff appears to maximise performance. However if the total number of residue pairs considered is more important than a high recall, the performance of the MI measures is optimal around the 40% gap cutoff.

	cutoff	**percentile**	**MCC**	**cutoff**	**percentile**	**N×MCC^2**
MIc	10	97	0.104	40	97	2.62×10^5
aMIc	10	98	0.103	30	98	2.46×10^5
MIp	10	95	0.0926	40	96	2.13×10^5
MI	10	96	0.0263	40	97	1.90×10^4

Table 5.5: **Overall highest MCC and N×MCC^2 achieved, using the Pfam dataset.** MI, MIp, MIc and aMIc were calculated for the 2,144 test cases for 11 gap cutoffs ranging from 0 to 100%, with 10% cutoff increments. Subsequently MCC and N×MCC^2 curves were plotted. The gap penalty and corresponding percentile at which each of the MI variants achieve the highest MCC and N×MCC^2, respectively, are recorded along with the MCC and N×MCC^2 values.

Upon closely examining the plots around N×MCC$^2 = 3.84$, the 95^{th} percentile of a χ^2 distribution with one degree of freedom, we find that MIp, MIc and aMIc all score significantly better than random when either 0, 10, 20, 30, 40 or 100% gap cutoffs are employed (Figure 5.8). However, as can be seen in Figure 5.8 all gap cutoff curves of original MI fall below the horizontal line at 3.84 for some percentiles, illustrating that the performance of original MI is not always significantly better than random.

Since 46.8% of contact pairs in the Pfam dataset are introduced between the 0 and 10% gap cutoffs, the largest increase (Table 5.4 and Figure 5.5), it is unsurprising that the recall of the MI variants is highest around 10% cutoff. In order to further verify this observation we consider only the residue pairs that are introduced with each gap cutoff increment and evaluate the performance of the MI variants on these subsets of residue pairs. As illustrated in Figures 5.9 and 5.10, MIp, MIc and aMIc perform best on the residue pairs added between 0 and 10% gap cutoffs, and consistently score

significantly better than random in this interval (Figure 5.11). Conversely, no significant improvement is made by the residue pairs added in in any of the gap cutoff intervals for some percentiles when using original MI, or for the 50-100% interval when using MIp, MIc and aMIc (Figure 5.11). We now take a closer look at the performance of the MI variants around the 10% cutoff (Figures 5.12, 5.13 and 5.14).

We find that the MCC of each MI measure is slightly better at the 8% gap cutoff, while the highest N×MCC^2 is attained at the 12% gap cutoff (Figures 5.12, 5.13 and 5.14). Therefore we recommend the in-between 10% gap cutoff.

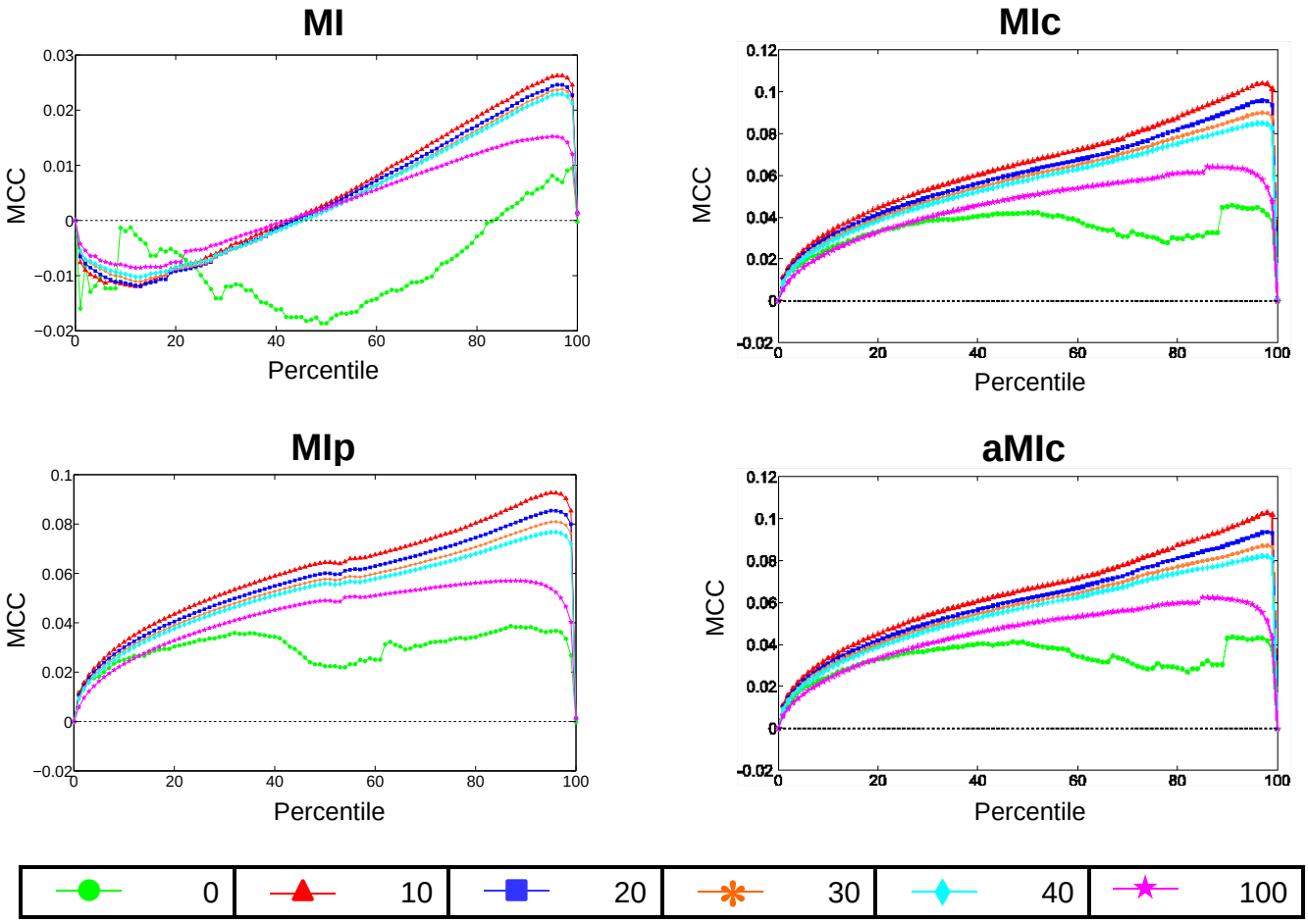

Figure 5.6: **Contact *versus* non-contact prediction MCC curves at the 0%, 10%, 20%, 30%, 40% and 100% gap cutoff for MI variants, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue. Percentile on the x-axis of this and subsequent figures indicates the percentage ranking of the MI score equal to and above which residue pairs are predicted to be contacts. For example, at the 50^{th} percentile all residue pairs that have an MI score equal to or above the median score are predicted to be contacts, while all other residue pairs are predicted to be non-contacts.

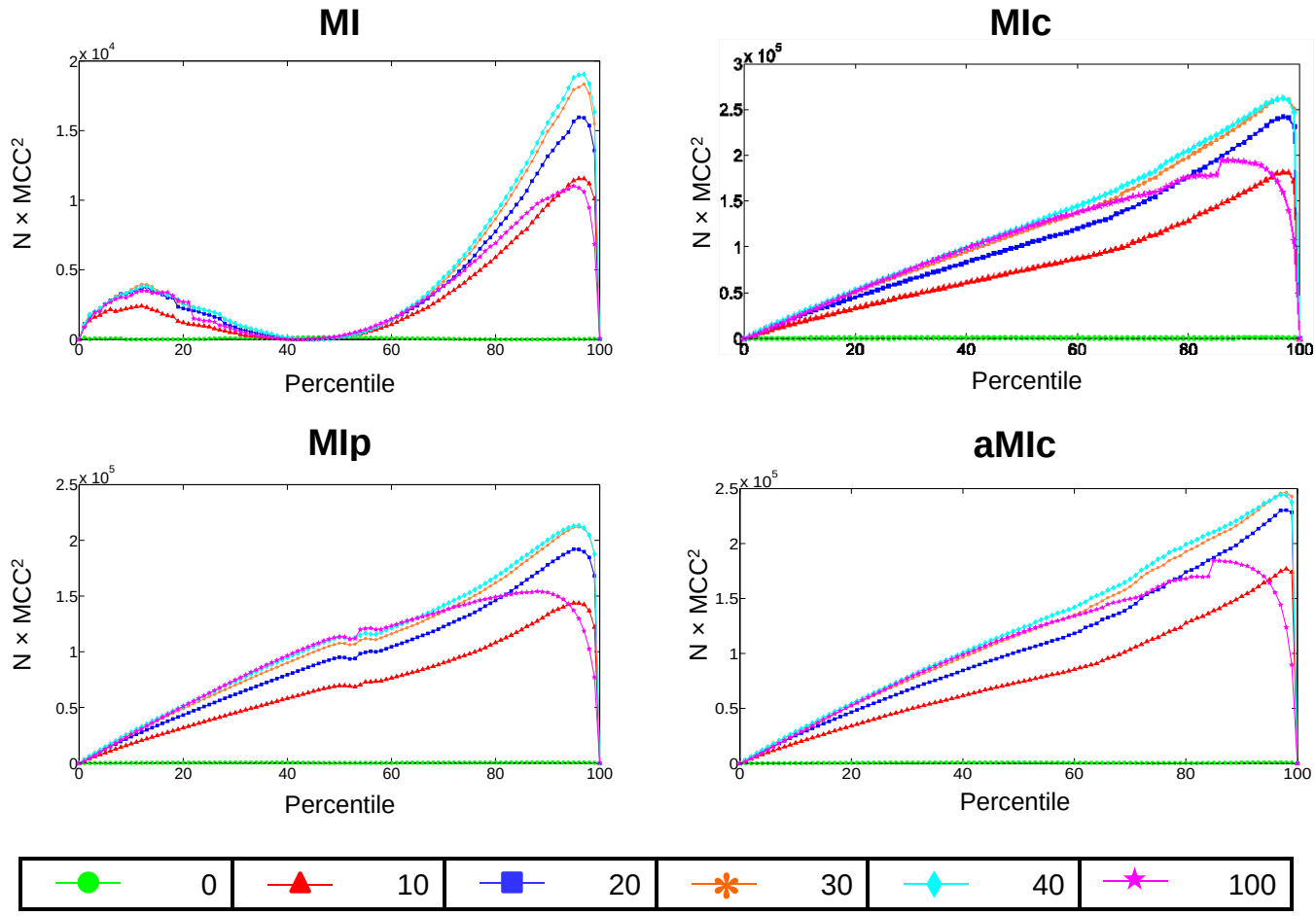

Figure 5.7: **Contact *versus* non-contact prediction N×MCC^2 curves at the 0%, 10%, 20%, 30%, 40% and 100% gap cutoff for MI variants, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

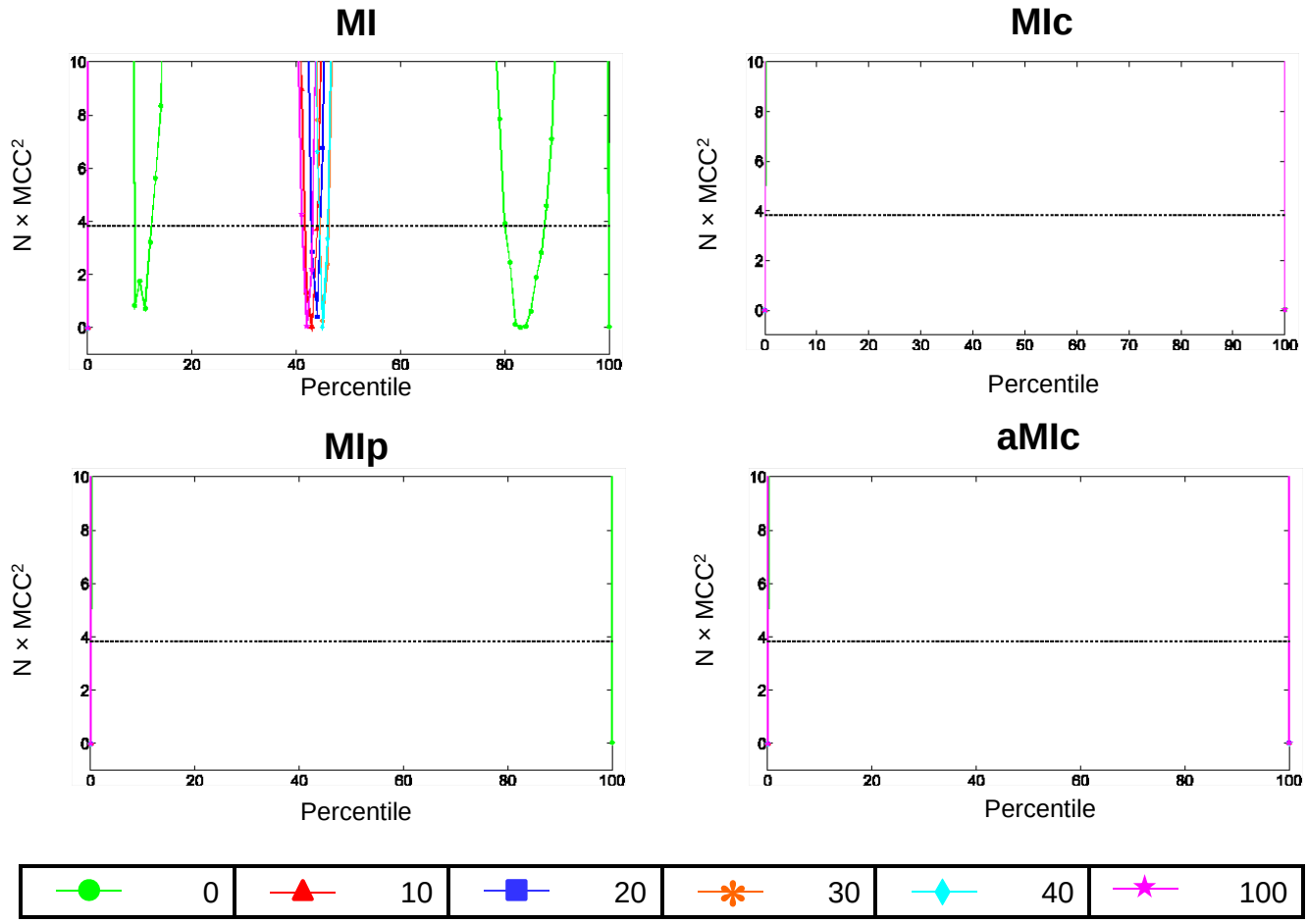

Figure 5.8: **Contact *versus* non-contact prediction N×MCC^2 curves, around N×MCC^2 = 3.84, at the 0%, 10%, 20%, 30%, 40% and 100% gap cutoff for MI variants, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

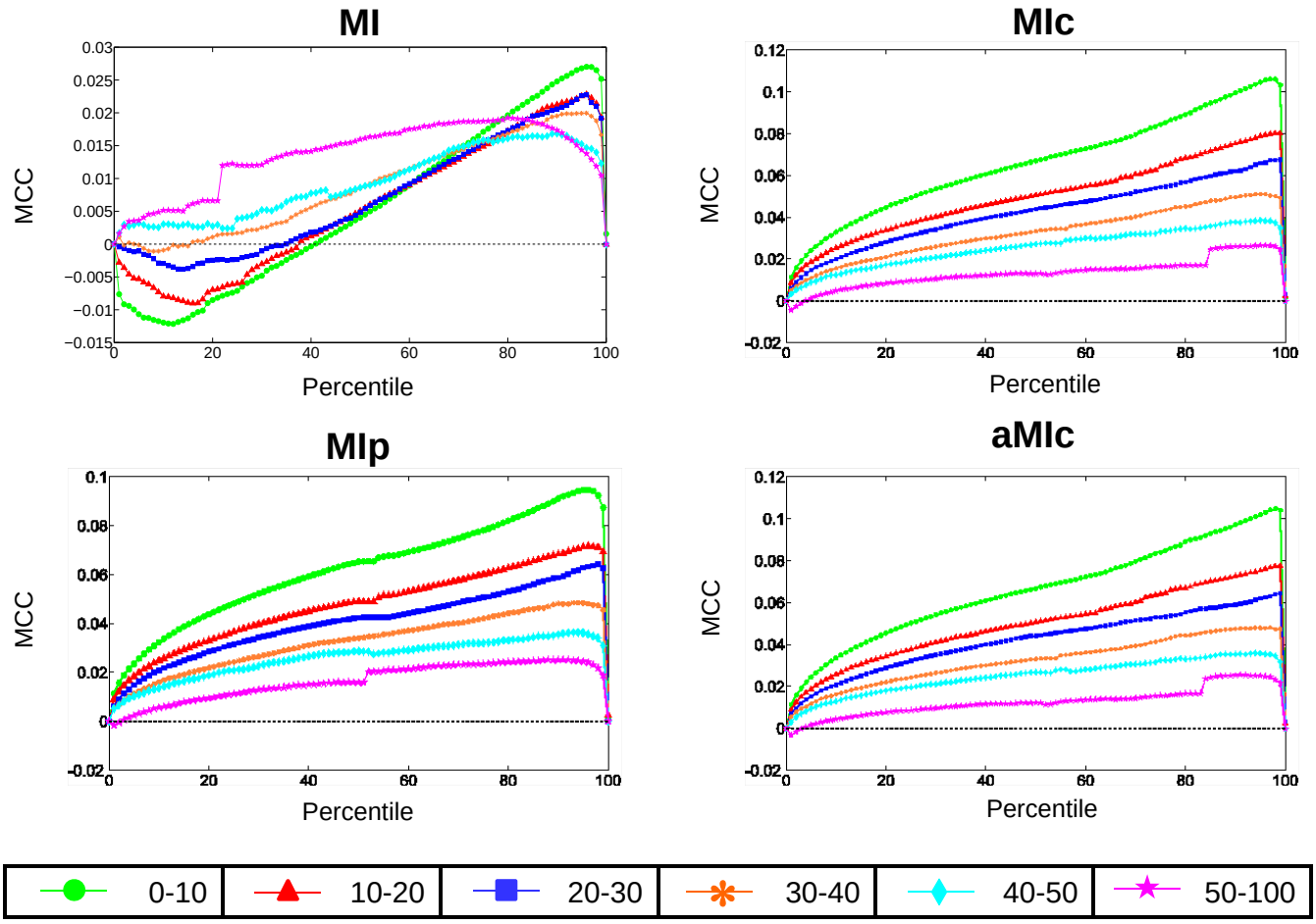

Figure 5.9: **Contact *versus* non-contact prediction MCC curves considering only the residue pairs that are introduced with each gap cutoff increment for MI variants, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant, when distinguishing the newly included contact and non-contact residue pairs for the specified gap cutoff increment. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue.

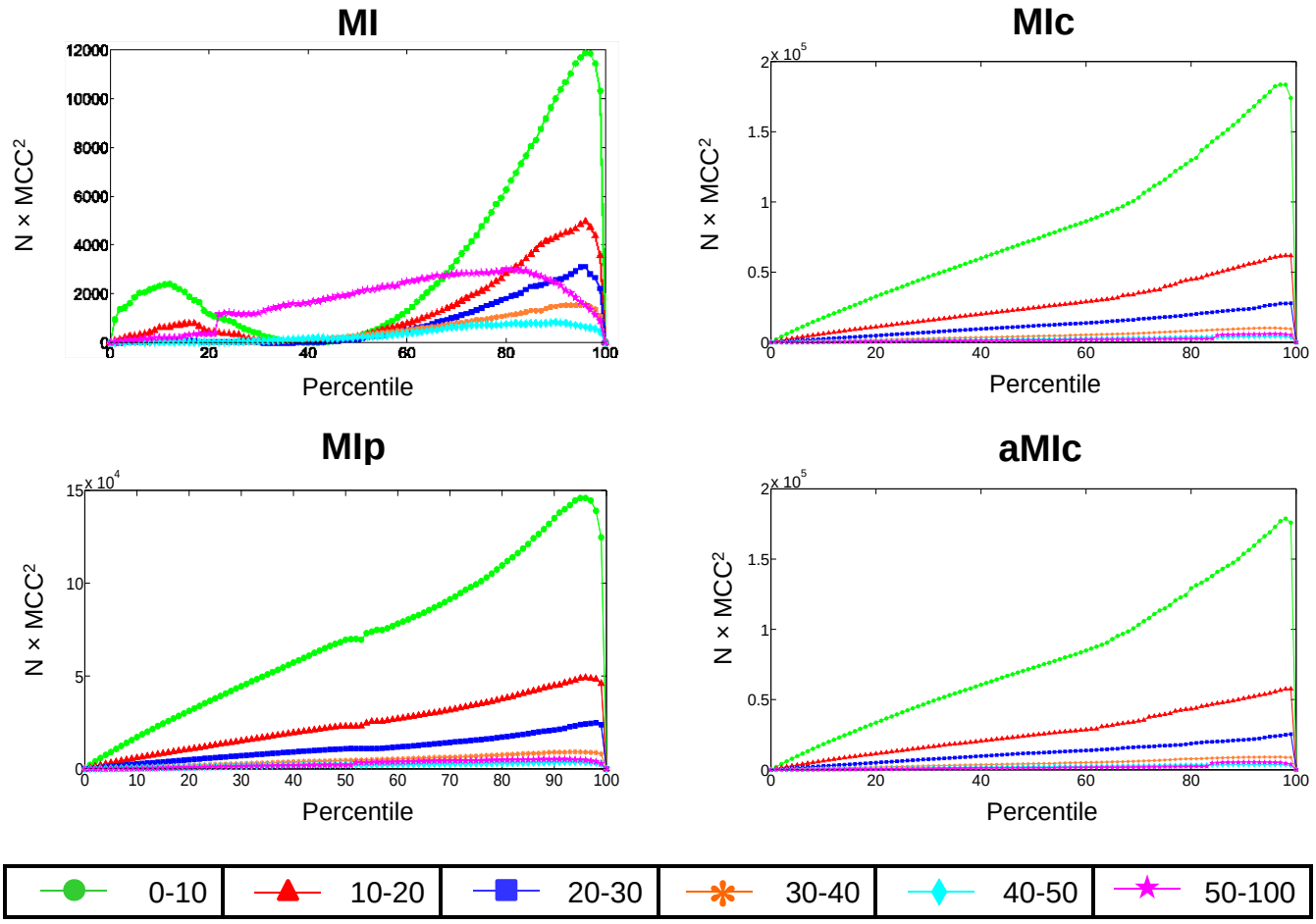

Figure 5.10: **Contact *versus* non-contact prediction N×MCC^2 curves considering only the residue pairs that are introduced with each gap cutoff increment for MI variants, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant, when distinguishing the newly included contact and non-contact residue pairs for the specified gap cutoff increment. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

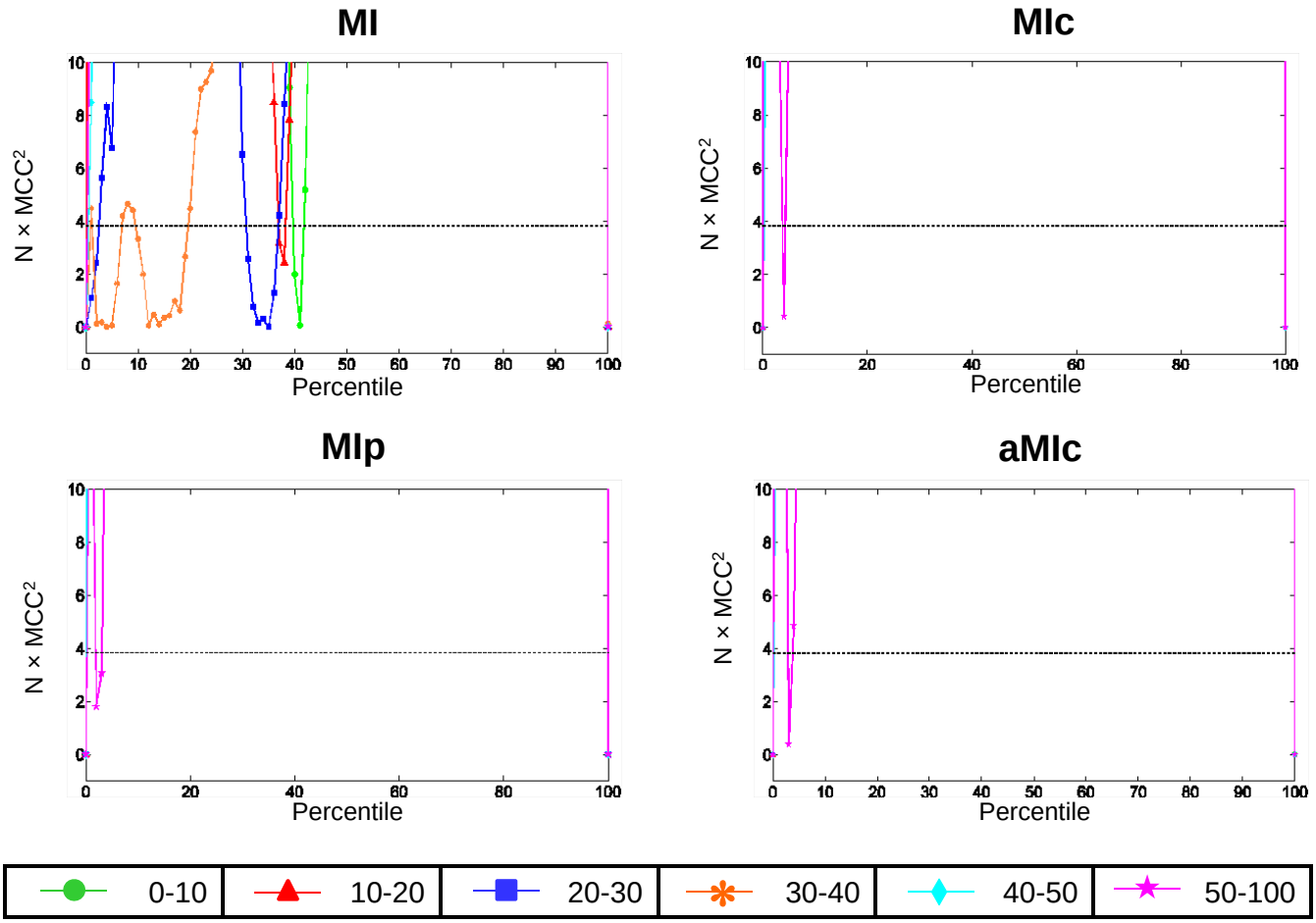

Figure 5.11: **Contact *versus* non-contact prediction N×MCC^2 curves, around N×MCC^2 = 3.84, considering only the residue pairs that are introduced with each gap cutoff increment for MI variants, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant, when distinguishing the newly included contact and non-contact residue pairs for the specified gap cutoff increment. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

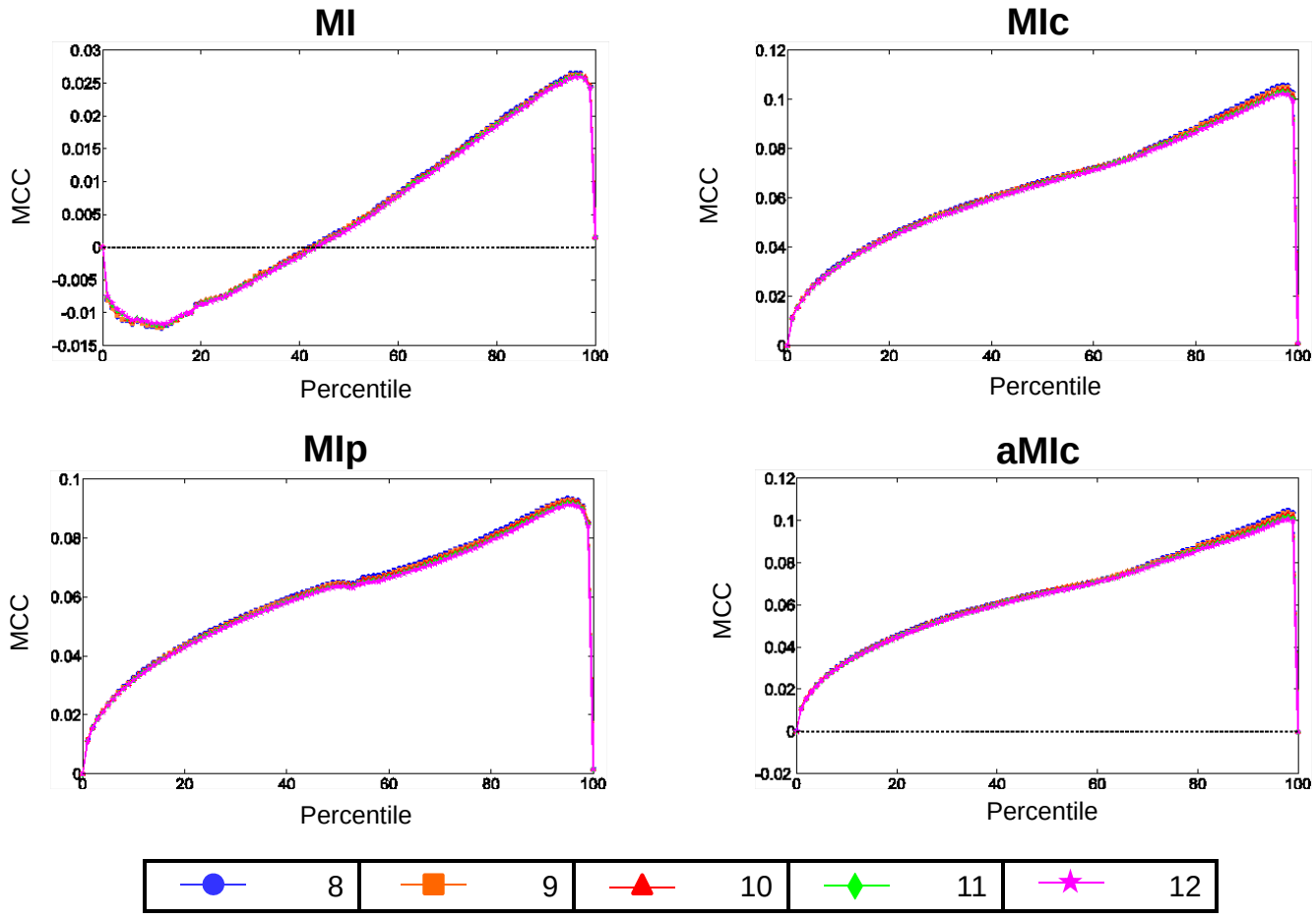

Figure 5.12: **Contact *versus* non-contact prediction MCC curves when varying the gap cutoff from 8% to 12% for MI variants, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue.

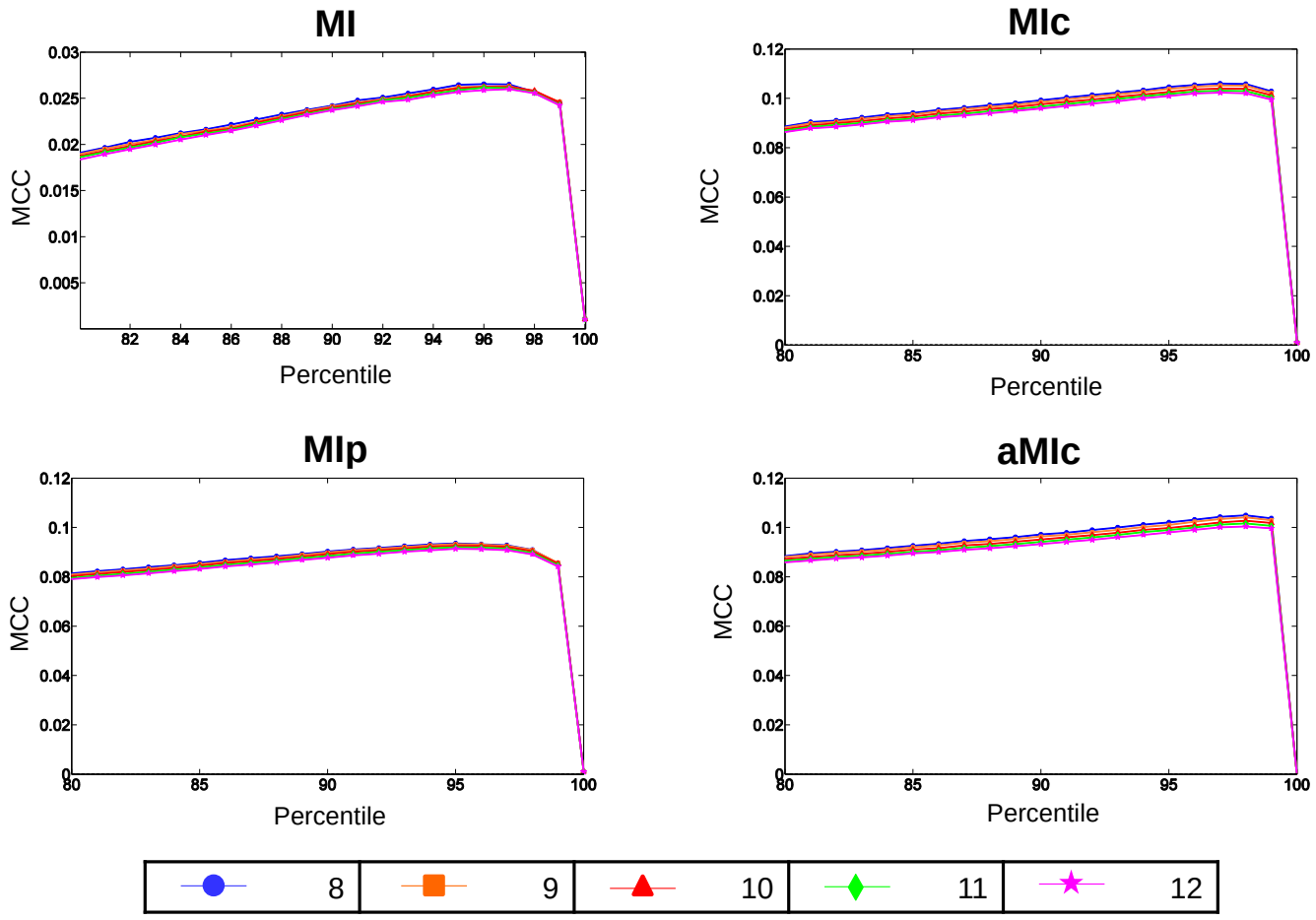

Figure 5.13: **Contact *versus* non-contact prediction MCC curves when varying the gap cutoff from 8% to 12% for MI variants, between 80 to 100 precentile, using the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff.

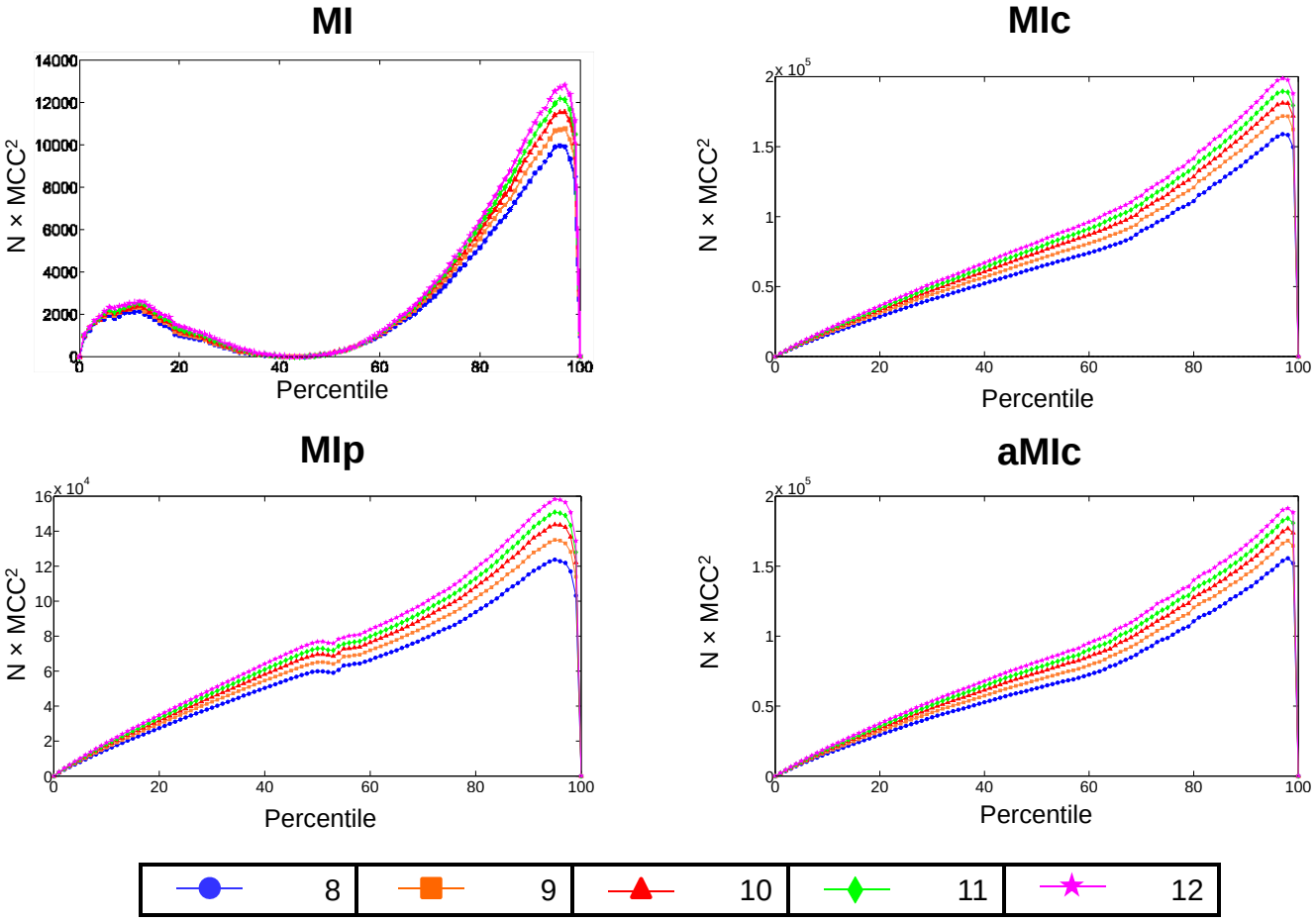

Figure 5.14: **Contact *versus* non-contact prediction N×MCC^2 curves when varying the gap cutoff from 8% to 12% for MI variants on the 2,144 Pfam test cases.** In each subplot a curve illustrates the performance of the MI variant when classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

5.4.1.3 Comparing the performance of MI based methods at the 10% gap cutoff

Next we examine the ability of the MI variants to predict contact pairs across different percentiles at a 10% gap cutoff. Percentile performance means, for example, at the 80^{th} percentile we are predicting that the residue pairs with the top 20% of scores are contact pairs and the remaining 80% of residue pairs are non-contacts. Similarly, at the 99^{th} percentile the residue pairs with the top 1% of scores are predicted to be contacts.

If we were to employ a gap cutoff of 10%, which appears to optimise performance, we find that for the highest percentiles, 80 to 100, MIc is the best intra-protein contact residue predictor among those considered, for lower percentiles the performance of MIc, aMIc and even MIp are approximately equal (Figures 5.15 and 5.16). Typically when using an MI measure to predict contact residue pairs of a protein of unknown structure, one would assume that the highest scores are more likely to be associated with true contact residue pairs. The results in Figures 5.15 and 5.16 are summarised in Table 5.6, which shows that MIc consistently achieves the highest MCC, N×MCC^2 and F-measure, followed by aMIc, MIp and finally original MI.

	percentile	**MCC**	**percentile**	**N×MCC^2**	**percentile**	**F-measure**
MIc	97	0.104	97	1.81×10^5	62	0.614
aMIc	98	0.103	98	1.77×10^5	61	0.613
MIp	95	0.0926	95	1.44×10^5	61	0.607
MI	96	0.0263	96	1.16×10^4	52	0.506

Table 5.6: **Highest MCC, N×MCC^2 and F-measure achieved at 10% gap cutoff, using the Pfam dataset.** MI, MIp, MIc and aMIc were calculated for the 2,144 test cases at a 10% gap cutoff. Subsequently MCC and N×MCC^2 curves were plotted (Figures 5.15 and 5.16), and F-measures calculated. The percentiles at which each of the MI variants achieve the highest MCC, N×MCC^2 and F-measure, respectively, are recorded along with their corresponding values.

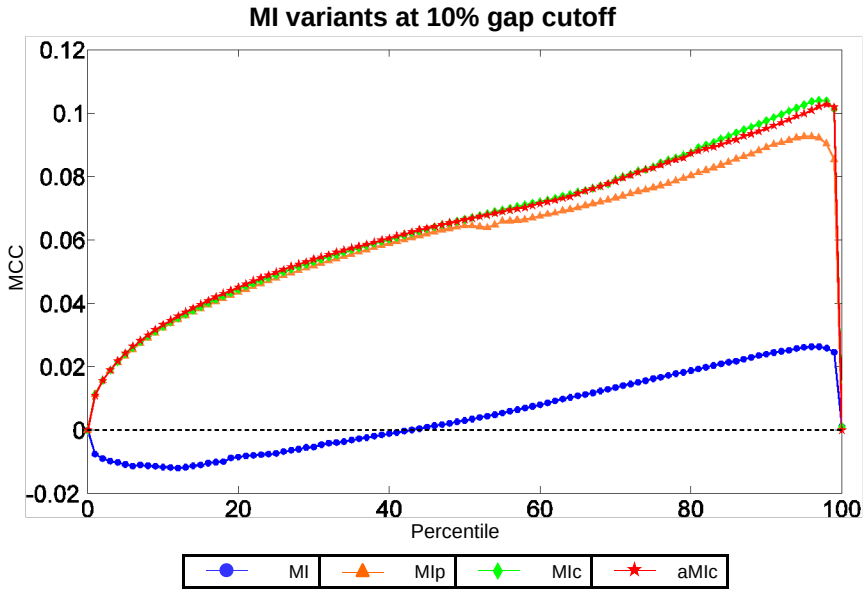

Figure 5.15: **Contact *versus* non-contact prediction MCC curves at the 10% gap cutoff for MI variants, using the 2,144 Pfam test cases.** Each line on the curve illustrates the performance of MI, MIp, MIc and aMIc respectively when distinguishing contact from non-contact residue pairs at the 10% gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue.

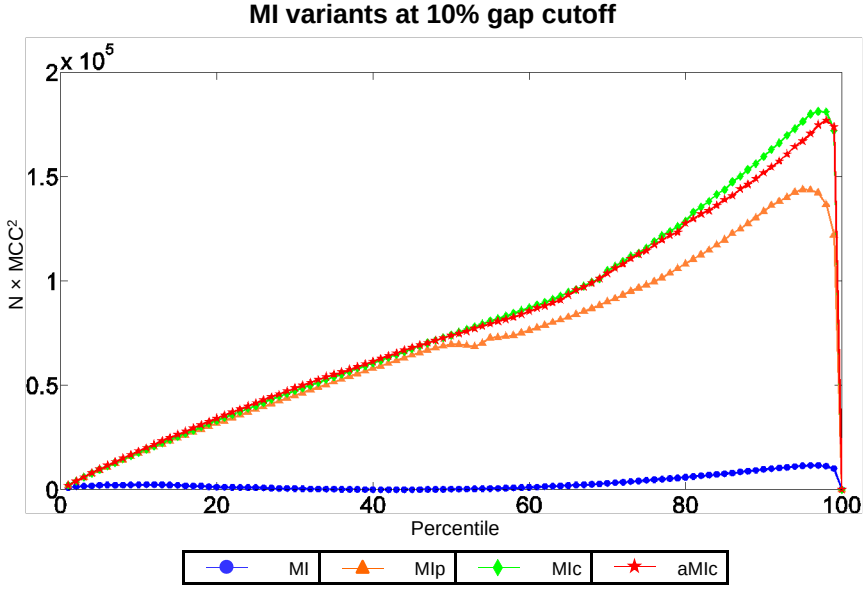

Figure 5.16: **Contact *versus* non-contact prediction N×MCC^2 curves at the 10% gap cutoff for MI variants, using the 2,144 Pfam test cases.** Each line on the curve illustrates the performance of MI, MIp, MIc and aMIc respectively when distinguishing contact from non-contact residue pairs at the 10% gap cutoff. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

5.4.2 The Hamer dataset

We run all tests discussed in this chapter on the 80 alignments from Hamer *et al.* [2010] (Table 2.1). The comparable figures and tables can be found in Appendix B. We find that the different algorithms used to construct the multiple sequence alignments, specifically the HMMER3 (Eddy [2011]) Hidden Markov Model based alignment software employed by Pfam, and the MUSCLE (Edgar [2004]) and MaxAlign (Gouveia-Oliveira *et al.* [2007]) alignment software used by Hamer *et al.*, lead to dissimilar results when assessing the influence of gaps on MI based methods.

At the 0% gap penalty, 63.3% of contact pairs are included in the Hamer dataset, as opposed to the mere 2.02% of contact pairs included at this gap cutoff in the Pfam dataset (Table 5.4 and Figure 5.5). This high contact pair retention in ungapped columns in the Hamer dataset could be a direct result of using MaxAlign which attempts to maximise the number of amino acids in ungapped columns by selecting an optimal subset of sequences. We would expect the fraction of contact pairs added in each incrementing gap cutoff interval to be monotone decreasing. However, like the Pfam dataset, this is not the case.

When using the MCC measure, the performance of MIp, MIc and aMIc appear to be best at the 0% gap cutoff, unlike on our Pfam dataset for which these MI variants achieved the highest MCC around the 10% gap cutoff. When examining the $N \times MCC^2$ curves the 20, 30 or 40% gap cutoffs seem to perform equally well. We speculate that when using the Hamer dataset there is minimal difference in performance between the 20 and 40% gap cutoffs because in this interval only 2.52% of new contact pairs are introduced, as opposed to the Pfam dataset in which 16.6% of contact pairs are added in the same interval (Table 5.4).

Similar to the Pfam dataset MIp, MIc and aMIc perform best on the residue pairs added between 0 and 10% gap cutoffs, and consistently score significantly better than random in this interval. On the other hand, no significant improvement is made by

the residue pairs added in the 20 to 30, 30 to 40, 40 to 50 and 50 to 100% gap cutoff intervals for some percentiles. Thus suggesting that for the Hamer test cases, a gap cutoff of 0% or around 10% would maximise the performance of the MI variants.

For the Hamer dataset we observe almost no difference in performance between the 8 to 12% gap cutoffs. Yet again MIc outperforms the other MI methods at the 10% gap cutoff.

At the 10% gap cutoff we compare the enrichment, percent of contact residues, in the highest 10% of scores for each MI variant, when evaluated on the different datasets (Table 5.7). We find that there is a higher percent of contacts in the top scores of all MI variants when evaluated on the Hamer dataset, as compared to the Pfam set. This is despite the percent of contact pairs out of total pairs at the 10% gap cutoff being approximately the same in both datasets, 2.78 and 2.48 in the Hamer and Pfam sets respectively. Thus suggesting that the MI methods are more partial to the alignment construction algorithm employed by Hamer *et al.*, rather than the one used by Pfam.

	Pfam	**Hamer**
MI variant	**highest 10%**	**highest 10%**
MI	3.60	4.32
MIp	6.64	8.06
MIc	7.03	8.40
aMIc	6.92	8.27

Table 5.7: **Enrichment for contacts in the highest 10% of MI scores.** Each MI variant is run on the Pfam and Hamer datasets using the 10% gap cutoff. The percent of scores relating to contact pairs in each MI variant's highest 10% of scores is recorded. At the 10% gap cutoff, the percent of contact pairs out of total pairs is 2.48 and 2.78, in the Pfam and Hamer sets respectively.

5.4.3 Relationship between alignment gaps and biological properties

One would expect contact residue pairs, integral to preserving the structure of the protein, to belong to conserved columns in an alignment. For both datasets we find that once there is a gap in the alignment columns there is no strong relation with whether or not it is a contact residue column. It is true however that in the Hamer dataset, 63.3% of contact pairs are included at the 0% gap penalty, but in the Pfam

dataset only 2.02% of contact pairs are considered at the this gap cutoff (Table 5.4 and Figure 5.5). We hypothesise that this may be due to the difference in quality of the alignments we are using. We now investigate this further taking into account the properties of surface and buried residues.

Previous studies have shown that buried residues are under greater evolutionary constraints than solvent-accessible surface residues (Bustamante *et al.* [2000]; Goldman *et al.* [1998]; Lin *et al.* [2007]; Overington *et al.* [1992]). A slower rate of evolution of these residues is unsurprising since buried residues often play a crucial role in maintaining the 3D structure of a protein. Therefore, in theory, the more conserved buried residue columns should contain fewer gaps than columns pertaining to surface residues. Once again we observe this to be the case for the Hamer dataset, but not for the Pfam test set, even though the ratio of surface to buried residues in both sets is approximately equal, 2.81 for the Hamer and 2.89 for the Pfam test set.

In the Hamer dataset 85.4% of residue columns pertaining to buried residues have 0% gaps, while 71.5% of surface columns are ungapped (Figure 5.17B). Conversely in the Pfam dataset only 4.81% of buried residue columns are ungapped, compared to the 6.22% ungapped surface columns (Figure 5.17A). This discrepancy in alignment properties may be due to the different rules and penalties each of the MSA construction algorithms employed.

Pfam is more permissive than the measures employed by Hamer *et al.* For example, when attempting to find homologous sequences for an alignment Hamer *et al.* required that the accepted BLAST (Altschul *et al.* [1990, 1997]) hits were between 75 and 125% of the length of the sequence of known structure, and the hit covered the central 50% of the query sequence (Hamer *et al.* [2010]). Only those alignments that had more than 50 and less than 1,000 BLAST hits were retained. The hits were initially aligned using MUSCLE, and then MaxAlign was employed to filter out sequences that augmented the percentage of gaps in the alignment. Pfam on the other hand imposes a “gathering

threshold" that a sequence must have or surpass in order to be considered for a Pfam family and subsequently be included in the family's MSA. Unlike Hamer *et al.*, Pfam does not limit the number of sequences that can be included in an MSA. The maximum number of sequences in the Hamer dataset is 709, while the maximum in our Pfam set is 11,485 (Table 5.1). Typically the greater the number of sequences, the greater the percentage of gaps in an alignment in order to accommodate all sequences. This is especially true for the Pfam dataset (Figure 5.18). Additionally, Pfam does not retrieve the optimal subset of sequences that minimises the number of gaps as done by MaxAlign. Furthermore, the Pfam MSAs are not tailored primarily to the sequence(s) of known structure.

More extensive evaluation of the performance of MI based methods using MSAs generated by different alignment algorithms would be beneficial. There are several studies in the literature evaluating the quality of MSAs (Aniba *et al.* [2010]; Edgar [2010]; Golubchik *et al.* [2007]; Thompson *et al.* [2011]), but none specific to the performance of MI.

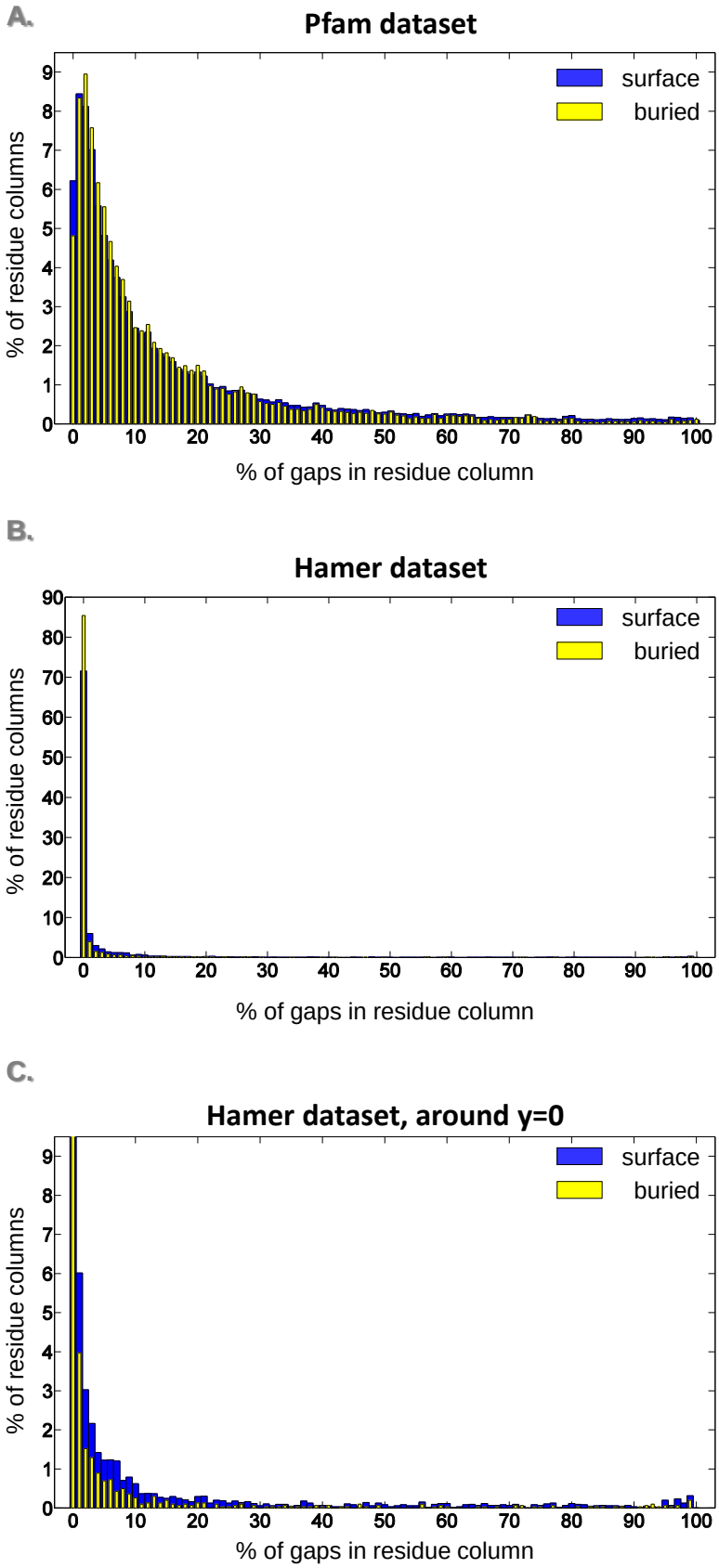

Figure 5.17: **Percent of gaps in surface and buried residue columns.** The percent of surface and buried residue columns, respectively, containing varying percents of gaps. Plot A illustrates the distribution of surface and buried columns in the 2,144 Pfam test cases, while B and C depict the distribution in the 80 Hamer test cases. Plot C is a magnification of plot B around y = 0.

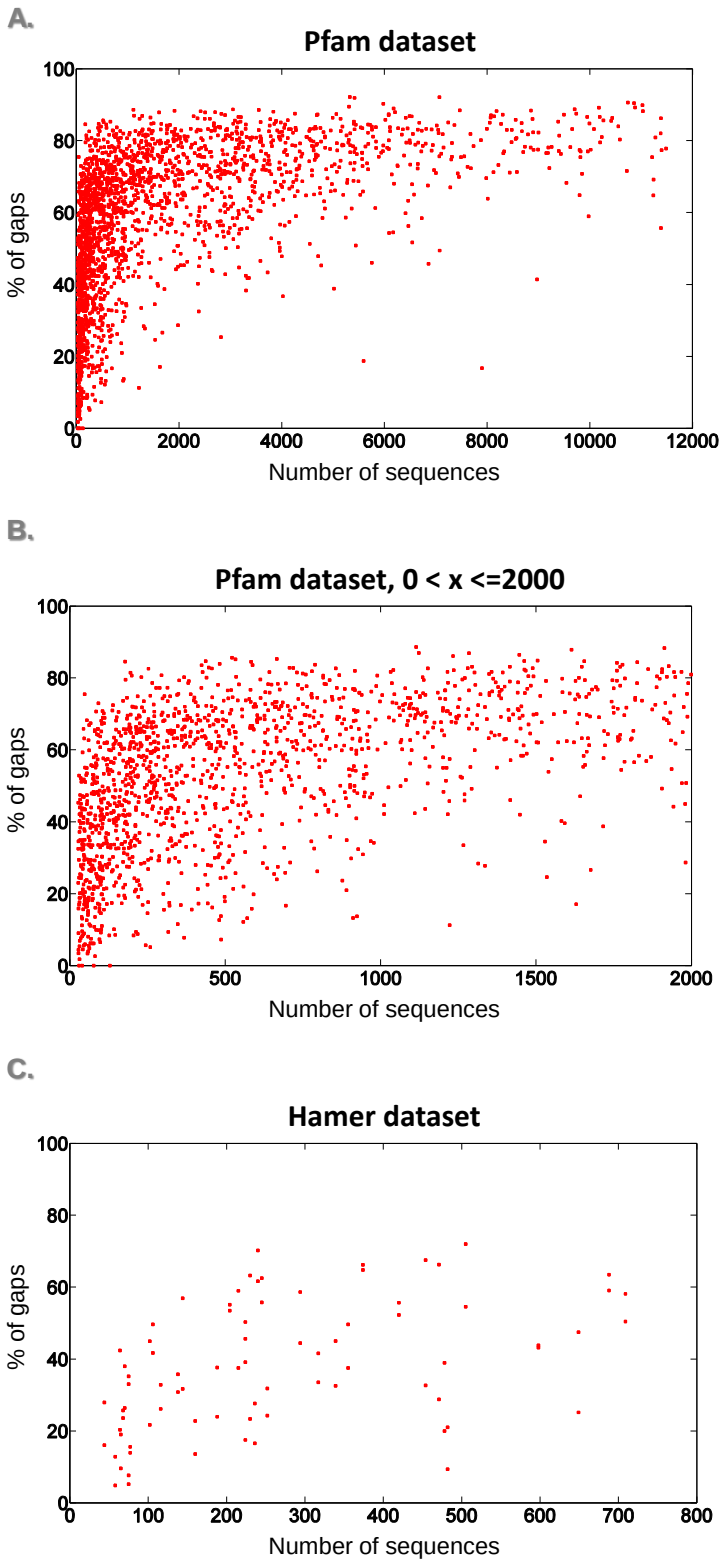

Figure 5.18: **Relationship between number of sequences and percent of gaps in the alignment.** The number of sequences in each multiple sequence alignment *versus* the percent of gaps, when all columns in the alignment are considered. Plots A and B illustrate the relationship in the 2,144 Pfam test cases, while C depicts the distribution in the 80 Hamer test cases. Plot B is a magnification of plot A for alignments that have 0 to 2,000 sequences.

5.5 Conclusions

Contrary to popular belief we do not observe that contact residue pairs only belong to conserved columns in the MSA. We surprisingly see a non-monotone relationship between the percent of contact pairs added and the percent of gaps in the alignment columns, when either our Pfam or Hamer datasets are used. This suggests that once there is a gap in the alignment column there is no strong relation with whether or not it is a contact residue column.

Using our Pfam dataset we find that all tested MI algorithms do not perform optimally when only columns with no gaps are allowed (0% gap cutoff) or when columns with all gaps are permitted (100% gap cutoff). At these extremes we speculate that there is either too much loss of information, 98.0% contact pairs are lost at 0% gap penalty, or too much noise is included in the data, inhibiting the signal-detection ability of the MI based methods. We use the MCC performance evaluation measure and observe that for all tested MI measures the highest proportion of correctly predicted contacts is attained when using a gap cutoff of around 10%, and not at the 0% cutoff frequently used by MIp, MIc and aMIc studies. When total number of residue pairs considered is of importance, the performance of the MI measures is generally optimal around 40% gap cutoff as deduced from the $N \times MCC^2$ plots.

For the Hamer dataset in which 63.3% of contact pairs are already included at the 0% gap penalty, MIp, MIc and aMIc achieve the highest proportion of correctly predicted contacts when using the 0% gap cutoff. Original MI does best around the 10% gap cutoff. Significant improvement in performance is only observed for contact pairs introduced between the 0 and 10% gap cutoff interval, suggesting that a cutoff around 0 or 10% would generally maximise recall for test cases in the Hamer set. If total number of residue pairs evaluated is important, the 20, 30 and 40% gap cutoffs attain close to optimal performance.

When attempting to predict intra-domain contacts using original MI and the three leading variants MIp, MIc and aMIc, we thus find that a range of gap cutoffs are suitable and the precise choice depends on the evaluation criteria. The recall desired and the alignment method used dictate the gap cutoff that should be employed. As a rule of thumb, around the 10% cutoff appears to optimise performance. At the 10% gap cutoff MIc is the leading intra-protein contact pair predictor among the ones tested, when either dataset is used.

We also observe that the four tested MI algorithms perform better on the Hamer set than on the Pfam dataset. For example, at the 10% gap cutoff the highest MCC achieved by MIc on the Hamer set is 0.122 at the 96^{th} percentile, while on the Pfam set is 0.104 at the 97^{th}. Likewise the enrichment for contacts in the top 10% of scores for MIc, using the 10% gap cutoff, is 8.39% for the Hamer set and 7.03% for the Pfam set. Further investigation into the dependency of MI performance on different alignment algorithms may significantly help improve the contact prediction capabilities of MI based methods.

Chapter 6

Conclusions and Future Directions

6.1 Conclusions

Proteins serve as the building blocks of cells, and are also responsible for executing most cellular processes. Although recent developments in high-throughput technologies have led to a an explosion of protein sequence data, knowledge of protein structure and interaction is lagging far behind. Experimental methods to determine the same are laborious, time consuming and costly. Subsequently the ability to accurately predict residues within and between proteins that are in contact *in silico* is necessary to further our understanding of protein functionality. This in turn will shed light on the mechanistics of whole biological systems and help advance drug discovery.

In this dissertation we examined the capacity of Mutual Information (MI) based methods to predict contact residues within (intra-) and between (inter-) proteins. The use of MI for contact residue prediction is founded on the widely accepted theory of "correlated mutations." This theory postulates that if two residues are in close proximity it is likely that a change in the size, shape or chemistry of one will need to be compensated for by a change in the other, if the contact is to remain energetically favourable (Fitch & Markowitz [1970]; Lockless & Ranganathan [1999]; Poon & Chao

[2005]; Yanofsky *et al.* [1964]). MI based methods attempt to identify these compensatory/correlated residue mutations by measuring the dependence between two residue columns in a multiple sequence alignment (MSA) of homologous sequences of the target protein(s). Under the theory of correlated mutations it is assumed that contact residue column pairs will have a greater dependency and therefore yield a higher MI score.

Through this dissertation we have gained new insight into how MI based methods work. We have identified factors that influence its low accuracy in both intra- and inter-protein domain contact residue prediction, specifically the properties of surface and buried residues, and the alignment gap cutoffs used.

We began our investigation by assembling a dataset of 40 interacting protein domains of known structure to assess MI methods for inter-protein contact prediction. We then split each of the 40 domain pairs into 80 single domains for the intra-protein MI analysis. These 40 domain-domain test cases were acquired from Hamer *et al.* [2010] and serve as a proxy for inter-protein contacts (Pazos *et al.* [1997]). To our knowledge this is the first study assessing the predictive ability of MI variants for inter-protein domain contact prediction on a large, general purpose, cross-species dataset.

Our preliminary assessment of the ability of MI based methods to predict contact residues within and between protein domains, Chapter 2, revealed that the capacity of all MI variants to distinguish contact from non-contact residues is weak. MIc (Lee & Kim [2009], Equation 1.14) outperformed the other MI methods for both inter- and intra-domain contact residue pair prediction achieving a precision of 2.26% and 12.8%, respectively, at 20% recall. The average precision when considering the top 100 scores is higher for intra-domain than inter-domain contact pairs, which may be attributed to the different evolutionary pressures acting on protein fold *versus* protein interaction. When attempting to predict contact residues in interacting protein domains, i-Patch (Hamer *et al.* [2010]) a purely statistical measure, attained a precision of 48.9%, while MIc, the leading MI inter-domain contact residue predictor in our investigation, founded

on the evolutionary biology theory of correlated mutations, only achieved a precision of 34.7% at 20% recall.

As i-Patch uses surface residues only in its calculation, in Chapter 3 we investigated the influence of surface and buried residues on MI contact prediction. We observed that MI variants have a bias towards surface residues in both inter- and intra-domains, such that surface residues attain higher MI scores than buried residues. This bias is most evident in original MI (Equation 1.7), and though it exists, is less prominent in MIp (Dunn *et al.* [2008], Equation 1.10) and aMIc (Lee & Kim [2009], Equation 1.18) due to the noise correction metrics employed by these MI measures. MIc although biased towards surface residues in inter-domains does not exhibit this bias in intra-domains. An analysis of two studies that have "successfully" employed the MI variants to predict contacts in inter- and intra- proteins respectively, suggests that the MI methods used predict surface residues in their top scoring residues rather than contacts. Therefore although MI is conjectured to predict contact residues we find that MI based algorithms instead predict surface residues. Since contact residues between protein domain are only on the surface, in the Chapter 4 we eliminated buried residues and reassessed the inter-domain contact prediction capacity of MI based methods.

When using surface residues only, we observed that the performance of all tested MI variants improved (Gomes *et al.* [2012]). After disentangling surface from contact prediction by eliminating buried residues, we found that MIc continues to be the leading inter-domain contact predictor on our test cases, with its precision increasing from 34.7% to 44.9% at 20% recall. However the abilities of MIc still fall shy of i-Patch which achieved a precision of 48.9% at 20% recall. A closer look at the popular Skerker *et al.* [2008] "successful" MI contact prediction test case revealed that MIc in fact, performs no better than random (Gomes *et al.* [2012]).

In Chapter 4 and Gomes *et al.* [2012] we introduced two novel MI variants that are based on the heuristics and assumptions incorporated in i-Patch. We conjecture that

the reduced alphabet based MI variant, which considers the physiochemical properties of the residues in the MSA, underperforms because while reducing the noise it also reduces the amount of signal available in the inter-domain data. Conversely, the additional alignment column in the 3D MI variant we formulated, that uses triangles of residues rather than pairs, perhaps enhances the noise in the inter-domain data while boosting the signal, and therefore also underperforms.

An investigation into the influence of gap penalties on MI algorithms in Chapter 5 suggests that the optimal gap cutoff is dependent on the recall desired and the multiple sequence alignment construction algorithm employed. As a rule of thumb around a 10% gap cutoff appears to optimise performance. In our study, at the 10% gap cutoff, we found MIc to be the best MI based intra-domain contact residue predictor among the ones considered.

It is commonly assumed that residues integral to preserving the structure of proteins in a family are conserved (Barton [1990]; Benner *et al.* [1994]; Crawford *et al.* [1987]). Although this holds true for the Hamer *et al.* [2010] dataset, for the Pfam (Punta *et al.* [2012]) alignments we found that the percent of contact and buried residues are not highest at the 0% gap cutoff as one would expect. This may be attributed to the different assumptions, penalties and alignment algorithms used by Hamer *et al.*, as opposed to those employed by Pfam. The different heuristics used to build the two sets of alignments resulted in better performance of all MI tested measures on the Hamer set. Further evaluation of the performance of MI based methods using MSAs generated by different alignment algorithms would be of great benefit to the field of contact residue prediction.

Based on these findings we recommend that, for intra- and inter-protein contact prediction when relying solely on sequence information, amongst the MI methods tested MIc (Lee & Kim [2009]) performs best. Surface residues only should be considered when attempting to predict inter-domain contacts and a gap cutoff around 10% is generally

optimal for maximising the performance of MI variants.

Our assessment of MI algorithms suggests that in their current state MI based methods are not very useful for contact prediction.

There are a number of hypothetical explanations for the poor performance of MI, such as:

- The theory of correlated mutations does not hold. We have shown that the inter-domain contact predictor i-Patch (Hamer *et al.* [2010]), a purely statistical measure compiled using a database of structures, not founded on any biological underpinnings, outperforms the correlated mutations based MI variants. Other statistical measures besides i-Patch have also achieved success in contact prediction (Geppert *et al.* [2011]; Liang *et al.* [2006]; Neuvirth *et al.* [2004]).

- MI's low accuracy may be a result of low input quality, the multiple sequence alignment. It has been observed that the accuracy of i-Patch with or without an MSA is approximately the same (personal communication with the developers of i-Patch, 20 September 2012). However unlike i-Patch, MI cannot be calculated independently of the alignment, and subsequently its performance is highly dependent on the quality of the input MSA.

- The protein crystal structures used to assess correct / incorrect contact residue pair predictions of MI present only one conformation of these dynamic macromolecules. Thus a non-contact residue pair in one protein conformation could in fact be a contact residue pair in another. Hence the prediction ability of MI based methods may be underestimated.

- Additionally, the protein structure used to assess the performance of MI methods on an MSA may not be representative of the whole MSA. Therefore contact residues for some sequences may differ from the contact residues of sequences that are homologous to the sequence of known structure.

- Recent studies have proposed that MI methods fail for contact residue prediction because these methods are not powerful enough to distinguish correlated mutations that arise from direct *versus* indirect couplings (Hopf *et al.* [2012]; Jones *et al.* [2012]; Marks *et al.* [2011]; Morcos *et al.* [2011]; Sułkowska *et al.* [2012]; Weigt *et al.* [2009]). For example, if residue A mutates, residue B which is in contact with residue A then mutates in order to maintain physiochemical complementarity. Residue C which is in close proximity to B also mutates as a result. Therefore direct coupling effects in contact residue pair AB will result in indirect coupling effects in the non-contact residue pair AC. These studies suggest that MI methods cannot distinguish between these two types of residue pairs.

Given all of the above, we would like to suggest a few ways in which the contact prediction ability of MI based methods could be improved.

6.2 Future Directions

6.2.1 Composition of protein interfaces

Previous studies have done extensive work characterising the composition of protein interfaces using protein complexes of known structure (Bogan & Thorn [1998]; Jones [1997]; Jones & Thornton [1996]; Sengupta & Kundu [2012]; Xia *et al.* [2010]). For example, Jones [1997] found that interfaces between proteins that only exist in complex are hydrophobic, whilst interfaces of proteins that exist both independently and in complex are more hydrophilic. Additionally, the interfaces of the later group of proteins can be the most polar patches on the proteins' surface. These authors also found that while most interfaces are planar, in some hetero-complexes the smaller protein has a protrusion that fits into the surface cleft of its binding partner. Another study observed that the free binding energy is not evenly distributed across the interface, but is localised in areas enriched in tryptophan, tyrosine and arginine (Bogan & Thorn [1998]). These

localised areas, "hot spots," have larger relative side-chain accessible surface areas (Xia *et al.* [2010]). Performance of MI methods may be enhanced by accounting for these characteristics of protein interfaces.

6.2.2 Assessing the quality of the alignment

Some steps can be taken to evaluate the quality of an alignment, and subsequently determine if it is suitable for MI analysis.

For alignments containing two or more sequences of known structure, the structures could be aligned using structural alignment software such as TM-align (Zhang & Skolnick [2005]). If the structures do not align well then there is reason to doubt the quality of the alignment. If the structures do align well, but the fraction of residues pairs in the structure alignment that are also aligned in the sequence alignment, the QDeveloper score (Wang & Dunbrack [2004]), is small, the alignment quality can also be judged to be poor.

Five Pfam test cases with varying number of sequences were picked randomly from our dataset to demonstrate these suggestions. Each of the Pfam alignments selected have two or more sequences of known structure. The "reference structure" used in our MI assessment was structurally aligned to another PDB structure associated with the MSA that has the most similar number of residues. TM-align was used to perform this structural alignment. Out of the residue pairs that structurally aligned, the number of residue pairs that shared this alignment in the MSA were counted and the QDeveloper score calculated. The sequence identity of the two PDB structures considered in PF04135 is low, the root mean square deviation (RMSD) of the aligned structures is high and the QDeveloper score is 0 (Table 6.1). We therefore believe that this alignment is of poor quality. The sequence identity *versus* QDeveloper score of the other four Pfam alignments are as expected according to the Wang & Dunbrack [2004] study.

Other measures can also be used to assess the quality of the alignment. As il-

test case	no. of sequences	PDB 1	PDB 2	sequence identity	no. of structurally aligned residues	RMSD	QDeveloper
PF02294	29	1WTX;A;2-62	1SSO;A;1-61	74.6%	59	2.12	0.869
PF04135	310	3LWO;B;1-53	1Y2Y;A;3-53	12.5%	24	3.69	0.00
PF03301	780	2NOX;J;28-299	2NW9;A;9-178	52.2%	157	1.05	0.767
PF00318	4,890	2VQE;B;10-226	3KC4;B;8-224	43.5%	216	3	0.972
PF00013	11,485	2PQU;C;15-75	2QND;A;5-64	22.2%	54	1.8	0.897

Table 6.1: **Assessing Pfam MSAs using structure alignments.** Five Pfam test cases with varying number of sequences were selected. The number of sequences in each Pfam MSA is recorded in column 2. 28, 11,485 and 780 are the minimum, maximum and median number of sequences in the alignments in our 2,144 Pfam dataset. Each of the five Pfam's selected had at least two sequences of known structure; their corresponding PDB structures are recorded in column 3 and 4 (PDB ID; chain ID; residue range). TM-align (Zhang & Skolnick [2005]) is used to align the two PDB structures. The TM-align calculated sequence identity, number of structurally aligned residues and RMSD of the two PDB structures are specified in columns 5, 6 and 7 respectively. The QDeveloper score, listed in the last column, is the fraction of residue pairs in the structure alignment that are also aligned in the Pfam sequence alignment.

lustrated in Figure 6.1A the alignment can be viewed in programs such as Jalview (Waterhouse *et al.* [2009]) to ensure that the physiochemical properties of residues in a column are consistent. These properties include for example, residue size, charge and hydrophobicity. The residue columns in Figure 6.1A are coloured by the physiochemical groupings used by the sequence alignment software ClustalW. Jalview also provides a "conservation" score for each column. A '*' in place of the column's conservation score indicates that the residues in a column are fully conserved, a '+' denotes that the all residues in a column belong to the same physiochemical grouping, whilst a number less than or equal to 9 signifies the amount of deviation of the mutations from the majority physiochemical residue category. The "quality" score specified for each column is inversely proportional to the average cost of all pairs of mutations in the column. A high alignment quality score implies that there are no mutations, or most mutations observed in the column are favourable. The "consensus" score of the column is the percentage of observations of the most frequently occurring residue in the column.

Pfam provides a "heatmap" for its alignments, which depicts the alignment uncertainty (Figure 6.1C). The greener the residue highlight, the higher the posterior probability that the alignment of the amino acid to the match/insert state in the Hidden Markov Model profile used to build the alignment is correct. The lower the posterior probability, the lower the certainty, the closer the colour to red.

Jalview also allows users to calculate phylogenetic trees from the alignment (Figure 6.1D). If the distances in the phylogenetic tree suggests that the alignment is too diverse, *i.e.* the alignment includes highly dissimilar sequences, the quality of the alignment should be questioned.

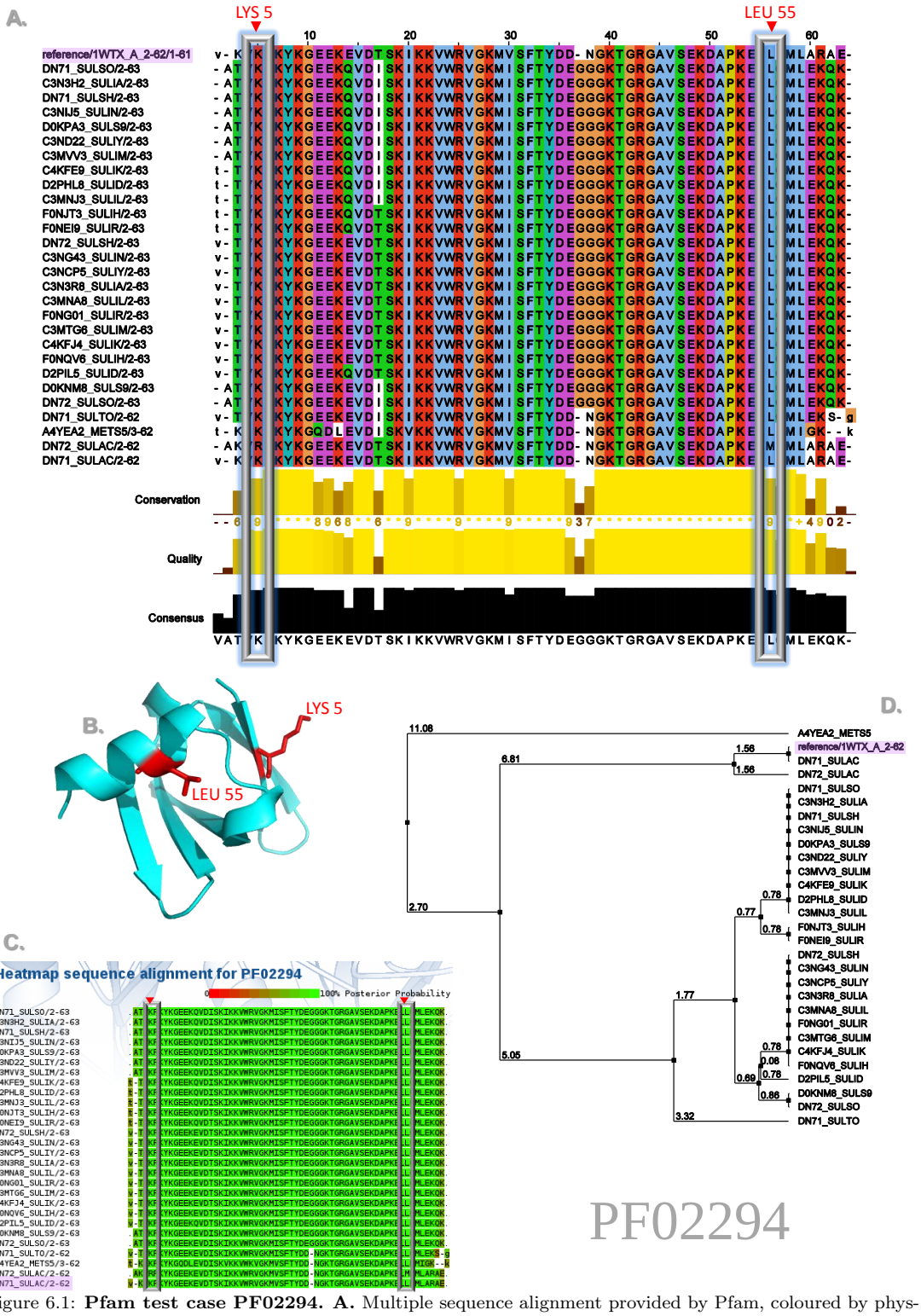

Figure 6.1: **Pfam test case PF02294. A.** Multiple sequence alignment provided by Pfam, coloured by physiochemical residue groupings used by alignment software ClustalW. **B.** PDB structure associated with sequence DN71_SULAC/2-62 (1WTX.pdb, Chain A, residues 2 to 62). **C.** Heatmap reflecting alignment uncertainty. The greener the residue highlight, the higher the posterior probability that the alignment of the amino acid to the match/insert state in the Hidden Markov Model profile used to build the alignment is correct. The lower the posterior probability, the lower the certainty, the closer the colour to red. **D.** Phylogenetic tree inferred from the percent identity of the sequences, average distance labelled on each branch. Sequence DN71_SULAC/2-62, used as the "reference" sequence in our MI assessment, is highlighted in purple in A, C and D. Grey boxes and red triangles in A and D indicate the residue column pair that achieved the highest MIc and aMIc scores (Lysine 5 and Leucine 55, 1WTX.pdb Chain A). These two residues are emphasised in red in the PDB structure in B. Jalview (Waterhouse *et al.* [2009]), PyMOL (DeLano [2002]) and the Pfam web application (Punta *et al.* [2012]) were used to generate these figures.

6.2.3 Relationship between number of sequences in the alignment and the performance of MI

In our analyses MIc outperformed the other tested MI variants. A preliminary investigation into the influence of the number of sequences in the alignment on the performance of MIc suggests that alignments with larger numbers of sequences carry a stronger signal for correlated mutations and subsequently the contact prediction ability of MIc is improved. Since the number of sequences in the alignments in our Pfam dataset shows greater variation than in our Hamer dataset, we consider the top 20% and bottom 20% of our 2,144 Pfam alignments, when ranked by number of sequences. The range of total sequences in each alignment subset are 3,223 to 11,485, and 28 to 188, respectively. We calculate MIc at the 10% gap cutoff, which was found to be among the best for contact prediction on the Pfam test set. We find that although the percent of contact pairs in the top and bottom subsets are approximately equal, 2.57 and 2.83 respectively, the MCC score, N×MCC^2 score, precision at 20% recall and enrichment for contacts in the highest 10% of scores are much higher for the subset of alignments with greater number of sequences (Figure 6.2 and Table 6.2).

Based on the difference in performance we observe here, we believe that in addition to Martin *et al.* [2005]'s initial study, further investigation into the number of sequences in an MSA that optimise the contact prediction capabilities of MI is necessary. We recommend that this analysis be carried out in conjunction with appraisal of the performance of MI when using MSAs from different alignment construction algorithms.

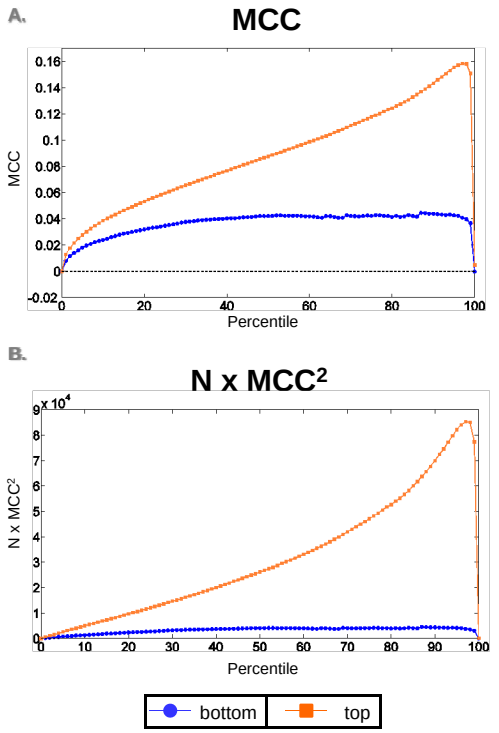

Figure 6.2: **Performance of MIc when considering alignments in our Pfam dataset that have the highest and lowest number of sequences.** At the 10% gap cutoff, MIc was calculated on the top 20% and bottom 20% of our 2,144 Pfam alignments, when the alignments are ranked by number of sequences. Plots A and B are the MCC and N×MCC^2 curves, respectively, for MIc on both subsets of alignments.

	% of contacts	percentile	MCC	percentile	N×MCC^2	precision at 20% recall	enrichment highest 10%
top	2.57	97	0.158	97	8.53×10^4	16.8%	9.37%
bottom	2.83	87	0.0445	87	4.52×10^3	4.83%	5.01%

Table 6.2: **Summary of the performance of MIc when considering alignments in our Pfam dataset that have the highest and lowest number of sequences.** At the 10% gap cutoff, MIc was calculated on the top 20% and bottom 20% of our 2,144 Pfam alignments, when the alignments are ranked by number of sequences. The percent of contact residue pairs out of total pairs is listed for each alignment subset. The highest MCC and N×MCC^2 scores achieved by MIc on each subset are recorded, along with the percentile at which each of these scores were attained. The precision and recall for the MIc scores were calculated and the precisions achieved at 20% recall are listed. The percent of contact pair scores in the highest 10% of MIc scores is also noted for each alignment subset.

6.2.4 Relationship between number of columns in the alignment and the performance of MI

In order to assess if a smaller search space, *i.e.* fewer columns in an MSA, results in improved contact prediction by MI based methods, we consider the top and bottom

20% of alignments in our 2,144 Pfam dataset, when ranked by number of columns in the alignment. The range of total columns in each alignment subset are 255 to 1,262, and 21 to 86, respectively. We find that MIc does indeed perform better on MSAs with fewer columns (Figure 6.3 and Table 6.3). The MCC, precision and enrichment is higher for alignments with fewer number of columns. Conversely, the N×MCC^2 score is greater for alignments with more columns. Since the number of residue column pairs and subsequently N is larger for alignments with greater number of columns, a higher N×MCC^2 score is unsurprising.

We speculate that the improved performance of MIc on alignments with fewer columns could be a result of the greater percentage of contact pairs in this subset of alignments, 5.98%, as compared to the 1.65% of contact pairs in the subset of alignments with greater number of columns (Table 6.3). We also hypothesise that by eliminating buried residues we in effect reduce the search space in inter-domain contact prediction, which partly contributes to the improved performance of the MI variants. Therefore finding additional ways in which columns can be eliminated before MI analysis is carried out may yield more accurate results.

	% of contacts	**percentile**	**MCC**	**percentile**	**N×MCC^2**	**precision at 20% recall**	**enrichment highest 10%**
top	1.65	98	0.101	98	8.80×10^4	7.84%	5.15%
bottom	5.98	95	0.113	95	6.94×10^3	15.0%	13.8%

Table 6.3: **Summary of the performance of MIc when considering alignments in our Pfam dataset that have the highest and lowest number of columns.** At the 10% gap cutoff, MIc was calculated on the top 20% and bottom 20% of our 2,144 Pfam alignments, when the alignments are ranked by number of columns. The percent of contact residue pairs out of total pairs is listed for each alignment subset. The highest MCC and N×MCC^2 scores achieved by MIc on each subset are recorded, along with the percentile at which each of these scores were attained. The precision and recall for the MIc scores were calculated and the precisions achieved at 20% recall are listed. The percent of contact pair scores in the highest 10% of MIc scores is also noted for each alignment subset.

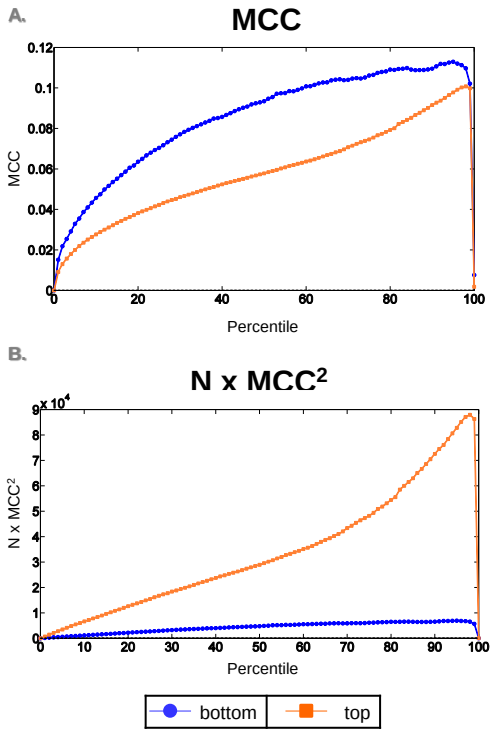

Figure 6.3: **Performance of MIc when considering alignments in our Pfam dataset that have the highest and lowest number of columns.** At the 10% gap cutoff, MIc was calculated on the top 20% and bottom 20% of our 2,144 Pfam alignments, when the alignments are ranked by number of columns. Plots A and B are the MCC and N×MCC^2 curves, respectively, for MIc on both subsets of alignments.

6.2.5 Entropy of domain-domain contact *versus* non-contact surface residue columns

A plausible option for reducing the number of columns in the alignment is eliminating residue columns that have a low entropy. We have observed that non-contact surface residue columns in our 40 inter-domain dataset have a higher entropy than contact residue columns in the same set (Figure 6.4). We believe that identifying the appropriate entropy cutoff would be beneficial to the field of MI based contact prediction.

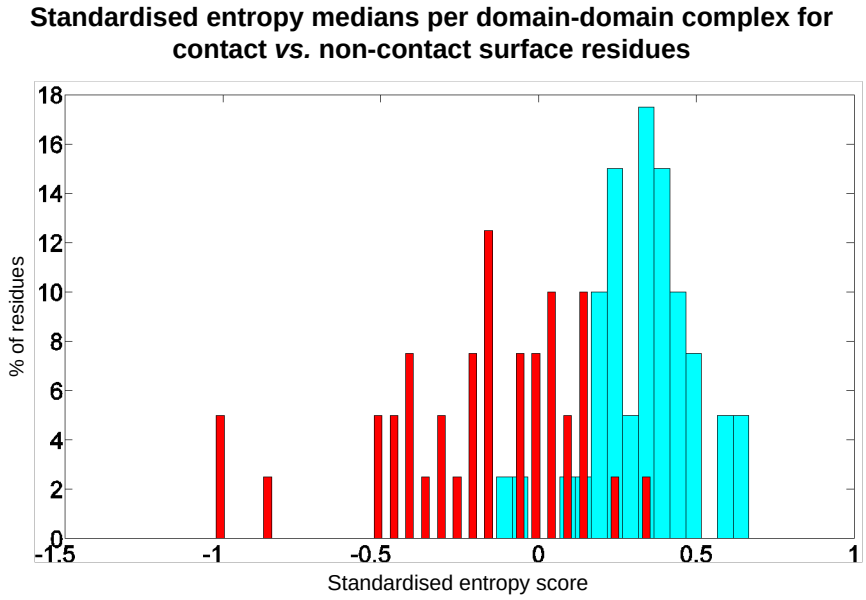

Figure 6.4: **Standardised entropy medians of contact *versus* non-contact surface residue columns for all 40 inter-domains in the Hamer dataset** (Table 2.1). Comparing the medians of the standardised entropy scores of the contact residue columns of each domain-domain complex (red) against the medians of the surface non-contact residue columns of each complex (cyan). Residue columns containing one or more gaps, having an entropy score of 0, or belonging to buried residue columns are not included in the median calculation.

6.2.6 Top scoring residue pairs

The top scoring residue pairs of each of the four leading MI based methods, original MI, MIp, MIc and aMIc were examined for five Pfam and three Hamer intra-domain test cases. Each MI variant was calculated at the 10% gap cutoff. We attempted to use the same Pfam test cases for which pairs of PDB structures were aligned (Table 6.1), however PF03301 and PF00318 do not have sufficient columns in the MSA with at most 10% gaps for the MI analysis to be performed. Subsequently PF02462 and PF01575 that have similar number of sequences in their alignments are used.

We find no obvious pattern between the residue pairs assigned the highest score by each of the different MI based methods (Tables 6.4 and 6.5, and Figure 6.1B), but a further systematic study could be undertaken.

test case	number of sequences	MI	MIp	MIc	aMIc
PF02294	29	36, 61	36, 61	5, 55	5, 55
PF04135	310	7, 11	8, 23	33, 51	7, 11
PF02462	780	90, 95	90, 95	90, 95	90, 95
PF01575	4,869	95, 99	44, 67	44, 67	44, 67
PF00013	11,485	22, 27	38, 42	38, 42	8, 42

Table 6.4: **Each MI variant's top scoring residue pair for five Pfam test cases.** Five Pfam test cases with varying number of sequences were selected. 28, 11,485 and 780 are the minimum, maximum and median number of sequences in the alignments in our 2,144 Pfam dataset. Each MI variant was calculated at the 10% gap cutoff and the residue pair that achieved the highest score recorded. The residue numbers in this table correspond to the PDB structure residue number. Residue pairs highlighted in light red are contact pairs, *i.e.* four or more residues apart in sequence and less than 4.5Å from each other.

test case	number of sequences	MI	MIp	MIc	aMIc
8TLN_1	44	19, 31	66, 135	102, 122	72, 132
1G8P_1	230	51, 212	54, 170	20, 214	20, 214
1LLD_1	709	48, 148	87, 91	87, 91	11, 76

Table 6.5: **Each MI variant's top scoring residue pair for three Hamer dataset cases.** Three Hamer dataset cases with varying number of sequences were selected. 44, 709 and 227 are the minimum, maximum and median number of sequences for the alignments in our Hamer dataset. Each MI variant was calculated at the 10% gap cutoff and the residue pair that achieved the highest score recorded. The residue numbers in this table correspond to the PDB structure residue number. Residue pairs highlighted in light red are contact pairs, *i.e.* four or more residues apart in sequence and less than 4.5Å from each other.

6.2.7 Direct coupling analysis

The recent Direct Coupling Analysis (DCA) measures, which also use MSAs as input and are founded on the theory of correlated mutations, claim to more accurately determine contact residues than MI based methods, by disentangling direct from indirect coupling (Hopf *et al.* [2012]; Jones *et al.* [2012]; Marks *et al.* [2011]; Morcos *et al.* [2011]; Sułkowska *et al.* [2012]; Weigt *et al.* [2009]). Our exploratory analysis of a leading Direct Coupling Analysis measure, PSICOV (Jones *et al.* [2012]), suggests that at a small recall its intra-domain contact prediction ability is better than that of MIc, which in our investigation outperformed the other tested MI variants. It is not possible with PSICOV to obtain predictions beyond the recall generated by the algorithm, because PSICOV does not assign scores to all pairs of columns in the alignment, but just to a limited set of column pairs that PSICOV determines are important. For this analysis we used the 977 Pfam test cases in our 2,144 Pfam dataset that have more than 1,000 sequences in their alignments, as this is the minimum number of sequences required to

run DCA measures. We evaluated the performance of PSICOV using the 100, 10 and 0% gap cutoffs and compared its performance to MIc. The definition of intra-domain "contact residues pairs" employed in this analysis, is consistent with the definition used throughout this dissertation; *i.e.* residue pairs that are four or more residues apart in sequence and are less than 4.5Å from each other are identified as "contact residue pairs."

PSICOV requires a large number of diverse homologous sequences in order to calculate a covariance matrix and subsequently PSICOV scores for each residue pair. Despite using alignments with greater than 1,000 sequences there may be insufficient number of columns in the alignment or limited diversity between sequences resulting in PSICOV calculations not converging and PSICOV failing to produce scores for some test cases. Out of the 977 Pfam test cases PSICOV computed scores for 908, 908 and 0 cases when the 100, 10 and 0% gap cutoffs respectively were used. We speculate that there were insufficient number of columns when the 0% gap cutoff was employed.

A PSICOV study considers scores 0.5 or greater to be a positive prediction, *i.e.* a contact residue pair (Nugent & Jones [2012]). We calculate the precision and recall attained by PSICOV accordingly and compare it to the precision achieved by MIc on the test cases for which PSICOV computed scores (Table 6.6). In Table 6.6 the precision of MIc is specified at the recall attained by PSICOV, as well as at 20% recall, in order to allow for comparison and fair appraisal of both contact prediction methods. PSICOV did not achieve a recall as high as 20%, hence the precision at 20% recall is not recorded for PSICOV.

When using the 100 and 10% gap cutoff we find that the precision attained by MIc is lower than that of PSICOV at the PSICOV recall (Table 6.6). This may suggest that PSICOV does indeed successfully disentangle direct from indirect couplings and subsequently better predicts intra-domain contact residue pairs than MIc. However unlike PSICOV, MIc is not limited by the number of sequences, number of columns or

	PSICOV				MIc	
gap cutoff	**no. of computed test cases**	**no. of positive predictions (TP+FP)**	**recall**	**precision**	**precision at PSICOV recall**	**precision at 20% recall**
100%	908	5.58×10^4	8.81%	57.0%	8.67%	6.25%
10%	908	3.51×10^4	8.87%	62.1%	24.5%	13.9%
0%	0	0	N/A	N/A	N/A	N/A

Table 6.6: **PSICOV and MIc analysis on Pfam test cases with greater than 1,000 sequences.** PSICOV and MIc were run on the 977 alignments in our 2,144 Pfam test set that have more than 1,000 sequences. PSICOV successfully converged on 908, 908 and 0 test cases when using the 100, 10 and 0% gap cutoffs respectively. The number of positive predictions, scores greater than or equal to 0.5, are recorded in column 3. The corresponding PSICOV recalls and precisions are listed in columns 4 and 5. The precisions and recalls of MIc were calculated for the Pfam domains on which PSICOV successfully converged. The precisions achieved by MIc on these test cases at the recalls attained by PSICOV, and at 20% recall, are recorded in columns 6 and 7 respectively. At the 0% gap cutoff, PSICOV did not produce scores for any test case, hence the precisions and recalls are not recorded.

the diversity of the sequences in the alignment and is able to make predictions for all alignments. Therefore if a 0% gap cutoff is necessary, the similarity of the sequences are high, the lengths of the sequences are short, or less than 1,000 homologous sequences are available, MIc and not PSICOV should be used. Alternatively, PSICOV and other Direct Coupling Analysis methods could be refined to be more permissive.

To date these direct coupling metrics have only been used on one inter-protein test case (Weigt *et al.* [2009]). Since MSAs with a minimum of 1,000 sequences are required for these methods, building such inter-protein MSAs will be extremely difficult.

As previously observed for the MI methods tested, the performance of PSICOV is best not at the 100% gap cutoff or at the 0% cutoff, but at 10% gap penalty. Further investigation on the influence of gaps on PSICOV and other Direct Coupling Analysis measures should be carried out.

As suggested for MI, we believe these metrics would also benefit from evaluation of performance using MSAs generated by different alignment algorithms.

Perhaps in these ways the use of correlated mutation analysis for contact residue prediction may be better understood.

Appendix A: Chapters 2, 3 and 4 Supplementary Tables and Figures

In Chapters 2, 3 and 4 we use 40 interacting domain pairs of known structure to evaluate the inter-domain contact prediction ability of MI based methods. These 40 inter-domain test cases were taken from Hamer *et al.* [2010]. We split each of the 40 domain pairs into 80 single domains to test the intra-domain contact prediction ability of the MI variants. Tables and figures for these inter- and intra-domain datasets, that are not included in Chapters 2, 3 and 4, are presented here.

residue type	**number in dataset**
surface residues	5,483
buried residues	2,364
contact residues	1,342
non-contact residues	6,505
contact residue pairs	1,301
non-contact residue pairs	362,399

Table 7: **Hamer dataset inter-domain summary.** The number of residues or residue pairs in each category across all 40 inter-domain test cases, after eliminating residue columns that have an entropy of 0 or contain a gap.

	surface	**buried**
contact	1,342	0
non-contact	4,141	2,364

Table 8: **Hamer dataset inter-domain break down.** The number of residues in each category across all 40 inter-domain test cases, after eliminating residue columns that have an entropy of 0 or contain a gap.

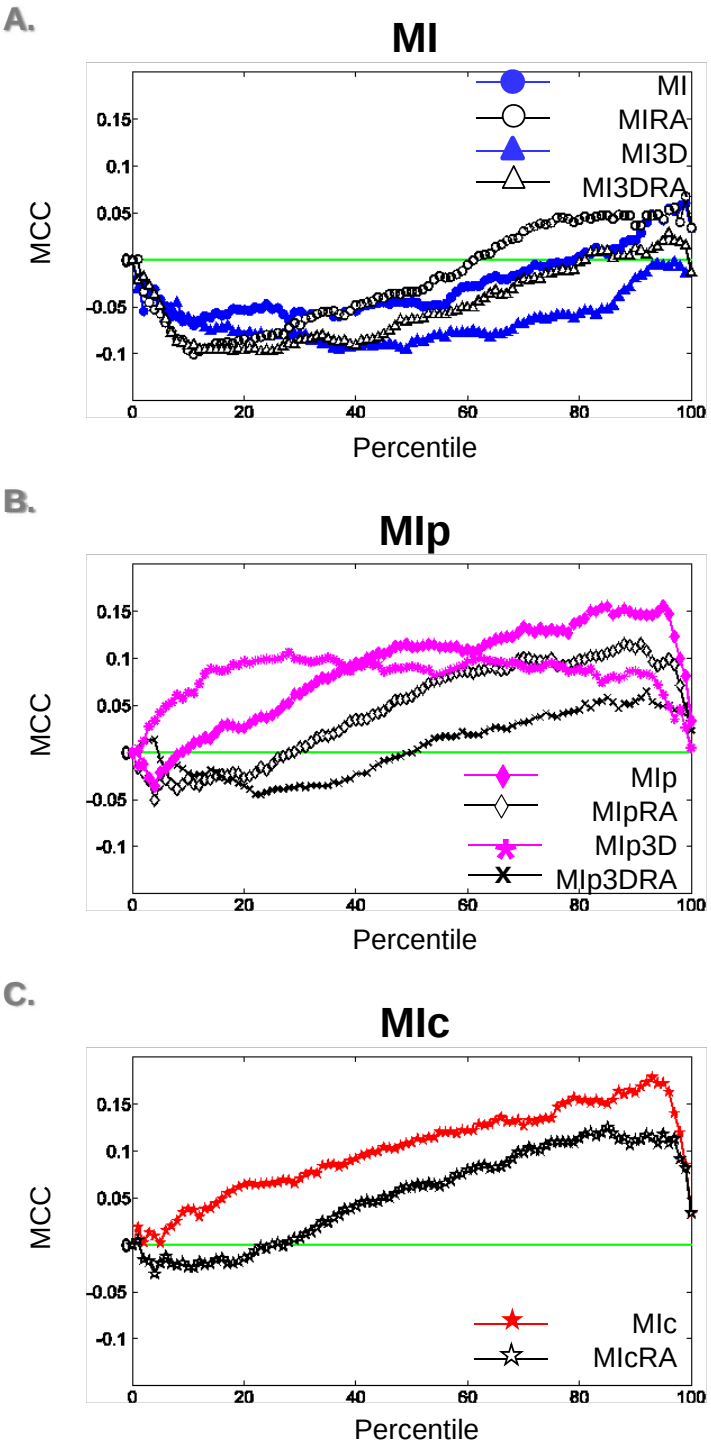

Figure 5: **Contact *versus* non-contact prediction MCC curves for MI variants on 40 inter-domain test cases taken from Hamer *et al.* [2010], when only surface residues are considered.** Performance evaluation of the predictive power of MI, MIp and MIc using the Matthews Correlation Coefficient (MCC) score (Matthews [1975]). A, B and C illustrate the performance of MI, MIp and MIc variants respectively when distinguishing contact from non-contact surface residues. The solid green line at 0 in all plots depicts the chance of randomly selecting a contact residue. An MCC score of $+1$ indicates a perfect prediction, while a score of -1 represents total disagreement between prediction and observation.

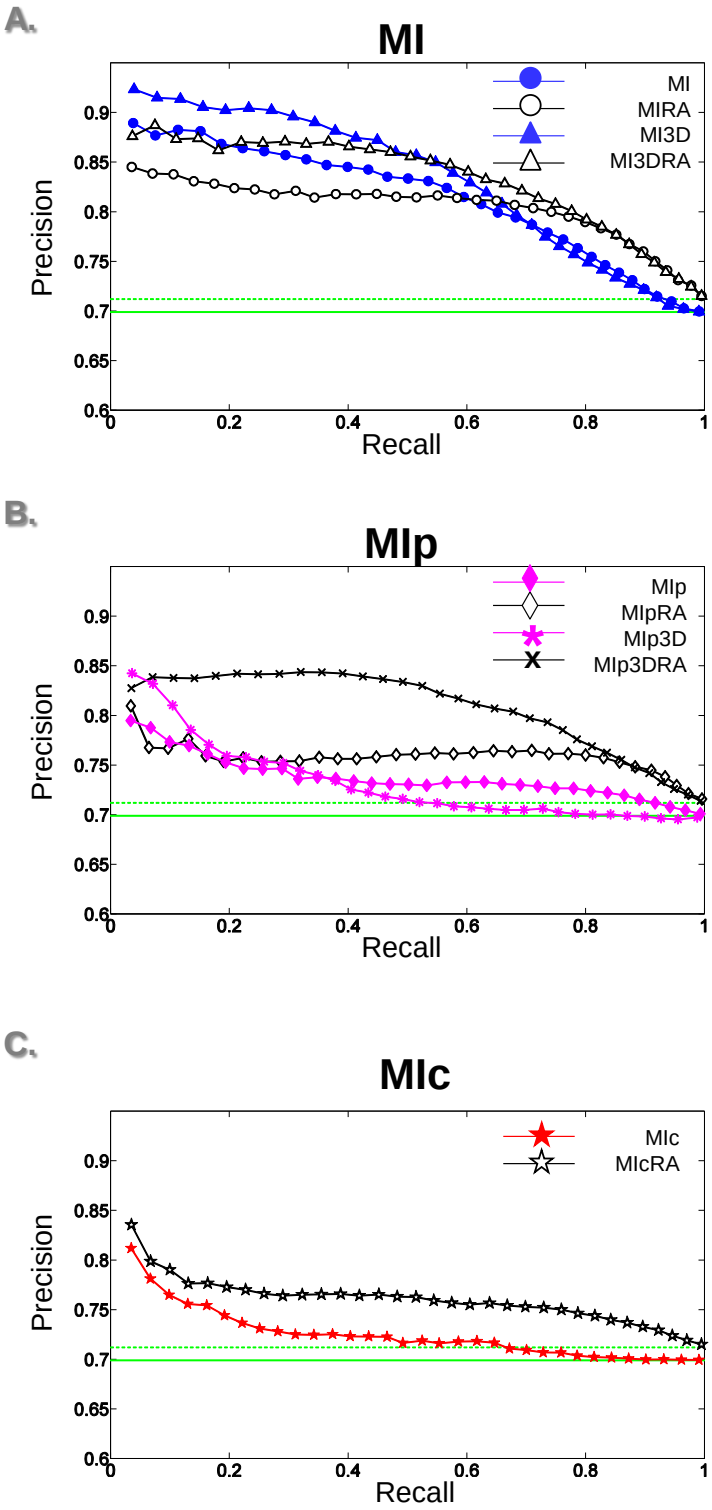

Figure 6: **Surface *versus* buried prediction P-ROC curves for MI variants on 40 inter-domain test cases taken from Hamer *et al.* [2010].** A, B and C illustrate the performance of MI, MIp and MIc variants respectively when distinguishing surface from buried residues. The solid green line in all plots depicts the chance of randomly selecting surface residues, while the dashed green line indicates the probability of randomly selecting a surface residue when employing the reduced alphabet amino acid set.

APPENDIX A

residue type	number in dataset
surface residues	5,483
buried residues	2,364
contact residue pairs	14,967
non-contact residue pairs	474,440

Table 9: **Hamer dataset intra-domain summary.** The number of residues or residue pairs in each category across all 80 domains, after eliminating residue columns that have an entropy of 0 or contain a gap.

Appendix B: Results of the Gaps Investigation Using the Hamer Dataset

Chapter 5 is an assessment of the influence of gaps in the sequence alignment on the contact prediction ability of MI based methods. All tests performed on the 2,144 Pfam domains in Chapter 5 were also carried out on the 80 Hamer dataset domains, which were used in previous chapters. The resulting tables and figures for the gap investigation on the Hamer dataset are provided here.

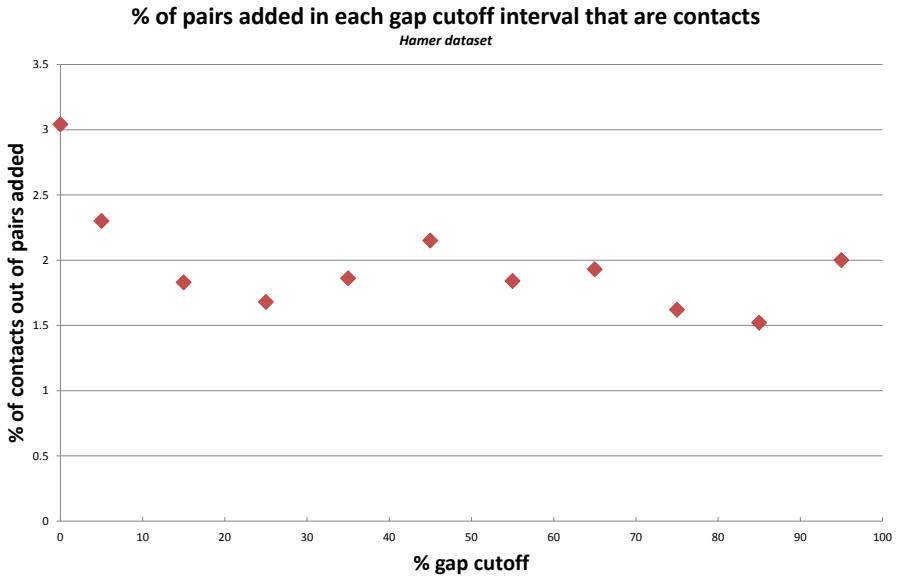

Figure 7: **Percent of pairs added in each gap cutoff interval that are contacts, using the Hamer dataset.** The percent of total residue pairs added in consecutive gap cutoff intervals that are contact pairs. To take one example, of the 11,900 total pairs introduced at the 50% cutoff that were not included at the 40% gap cutoff, 2.15% are contacts. When considering only ungapped columns, *i.e.* at 0% gap cutoff, 3.04% of the 534,849 total pairs present at this gap penalty are contacts.

Gap Cutoff Interval	% of Total Contact Pairs Added
0%	63.3
0%-10%	26.0
10%-20%	3.02
20%-30%	1.54
30%-40%	0.977
40%-50%	0.996
50%-60%	0.755
60%-70%	0.704
70%-80%	0.805
80%-90%	0.619
90%-100%	1.32

Table 10: **Contact pairs added in each gap cutoff interval, using the Hamer dataset.** When considering only ungapped columns, *i.e.* at the 0% gap cutoff, 63.3% of the 25,704 total contact pairs are present. At the 10% gap cutoff, 26.0% of the 25,704 total contact pairs are introduced that were not included at the 0% gap cutoff, at the 20% gap cutoff 3.02% of contact pairs are newly introduced, at the 30% cutoff 1.54%, and so on.

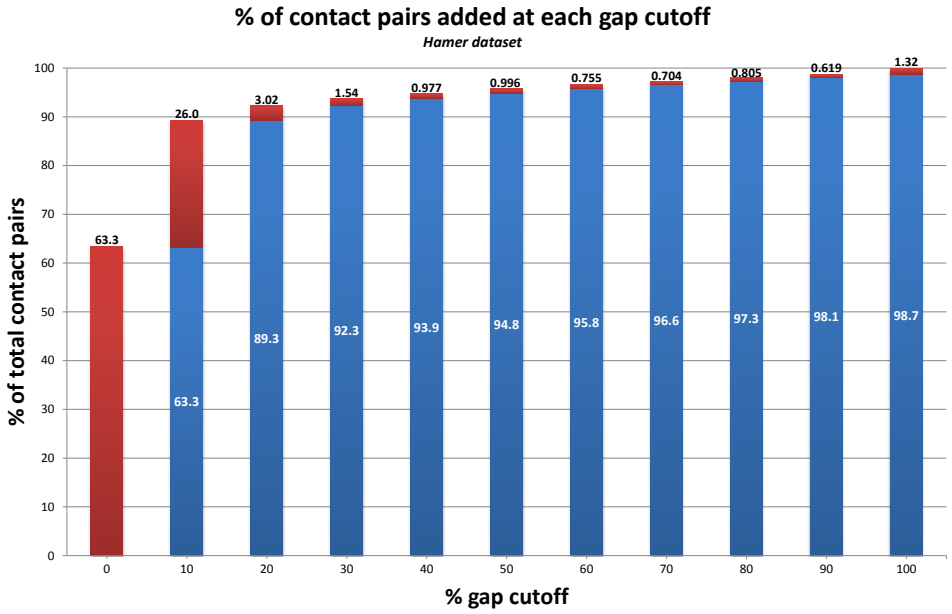

Figure 8: **Contact pairs added at each gap cutoff, using the Hamer dataset.** The red bars represent the percent of contact pairs newly introduced at each gap cutoff, while the blue bars denote the percent of contact pairs present at the preceding gap cutoff. The number in black above each red bar, and in white within each blue bar, indicate the percent of contact residue pairs accounted for in each red and blue bar respectively. For example, when considering only ungapped columns, *i.e.* at the 0% gap cutoff, 63.3% of the 25,704 total contact pairs are present. At the 10% gap cutoff, 26.0% of the 25,704 total contact pairs are newly introduced, making the percent of total contact pairs included at the 10% gap cutoff 89.3%.

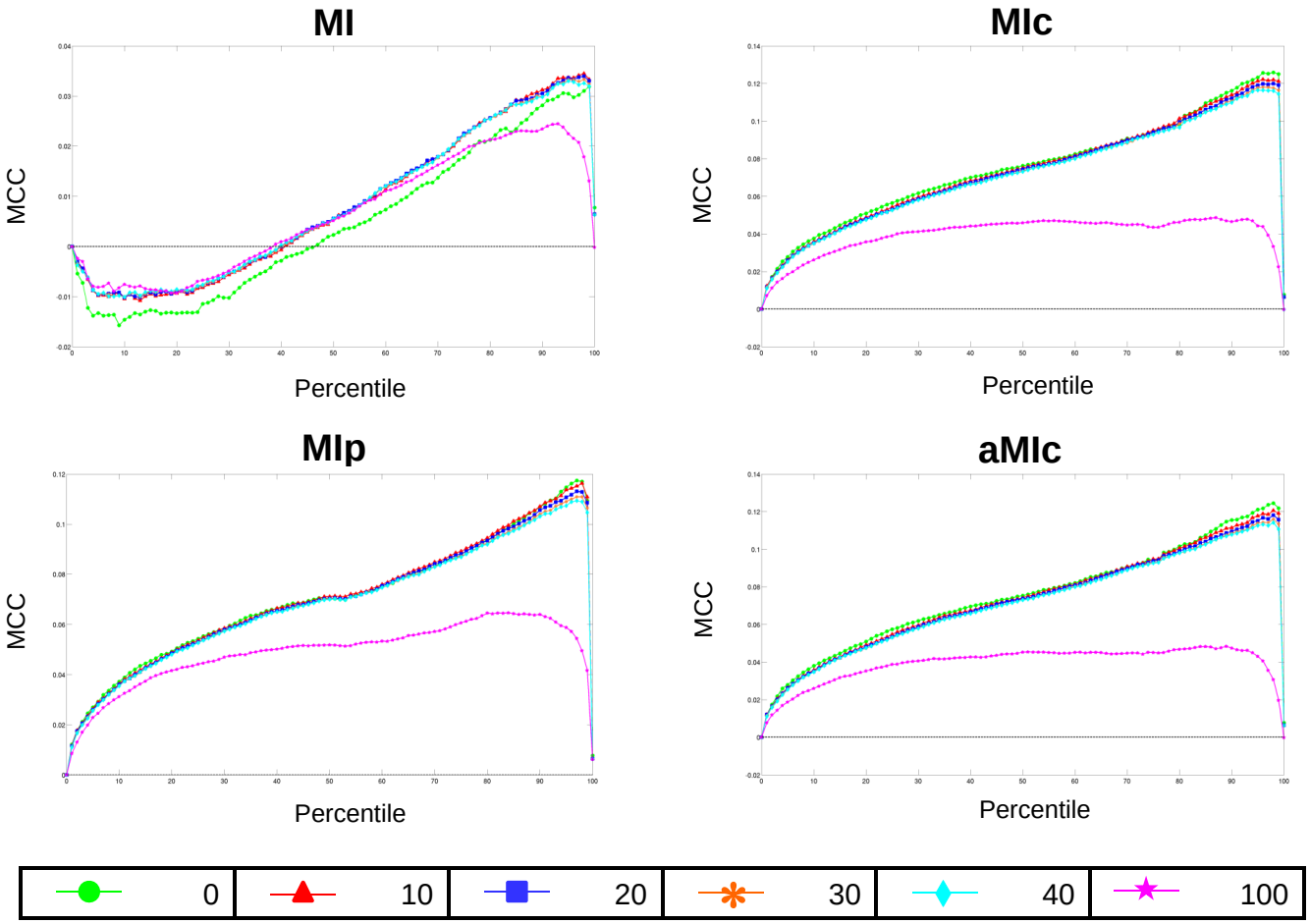

Figure 9: **Contact *versus* non-contact prediction MCC curves at the 0%, 10%, 20%, 30%, 40% and 100% gap cutoff for MI variants on the 80 Hamer test cases.** In each subplot a curve illustrates the performance of the MI variant for classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue.

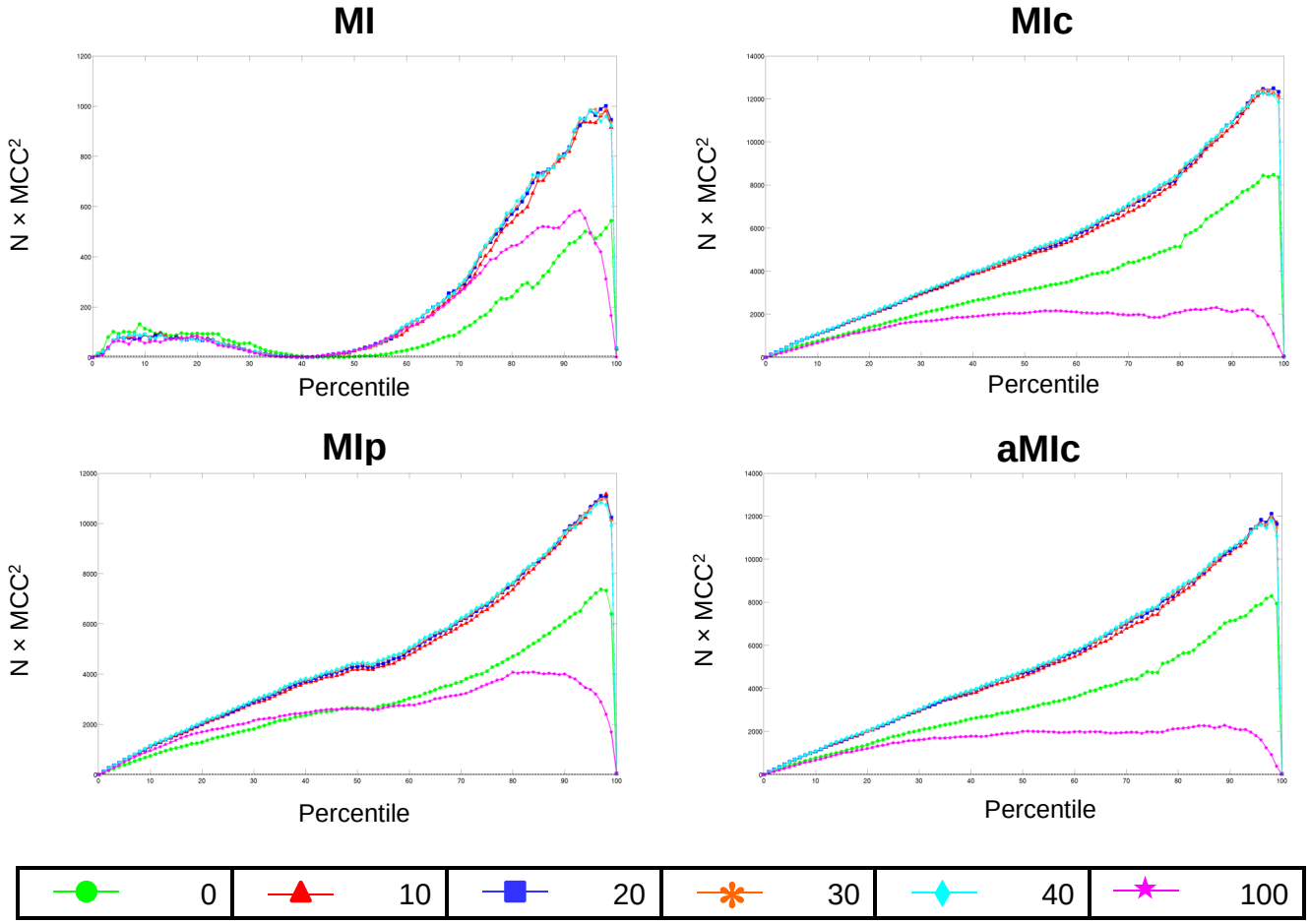

Figure 10: **Contact *versus* non-contact prediction N×MCC^2 curves at the 0%, 10%, 20%, 30%, 40% and 100% gap cutoff for MI variants on the 80 Hamer test cases.** In each subplot a curve illustrates the performance of the MI variant for classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

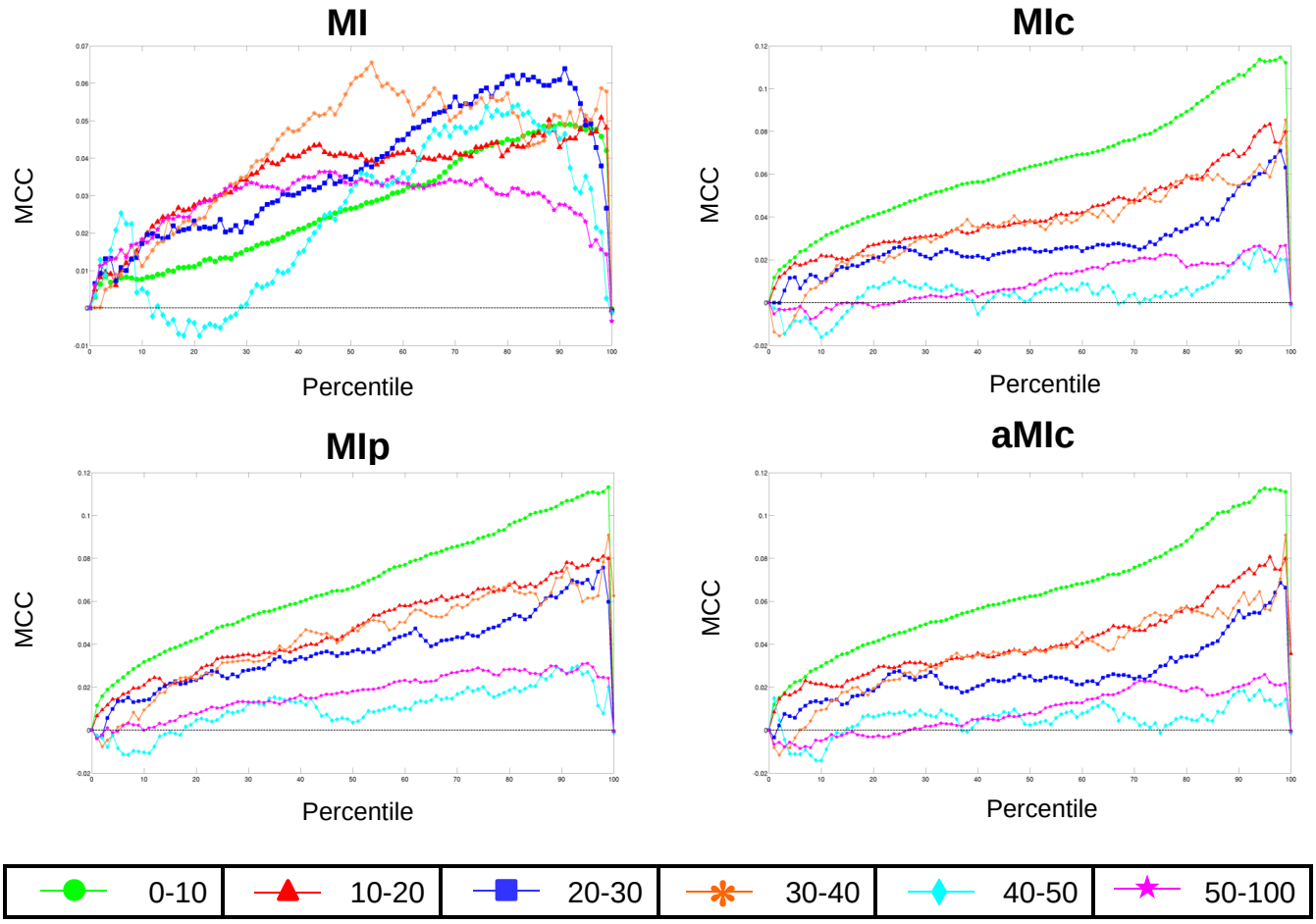

Figure 11: **Contact *versus* non-contact prediction MCC curves considering only the residue pairs that are introduced with each gap cutoff increment for MI variants on the 80 Hamer test cases.** In each subplot a curve illustrates the performance of the MI variant, when distinguishing the newly included contact and non-contact residue pairs for the specified gap cutoff increment. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue.

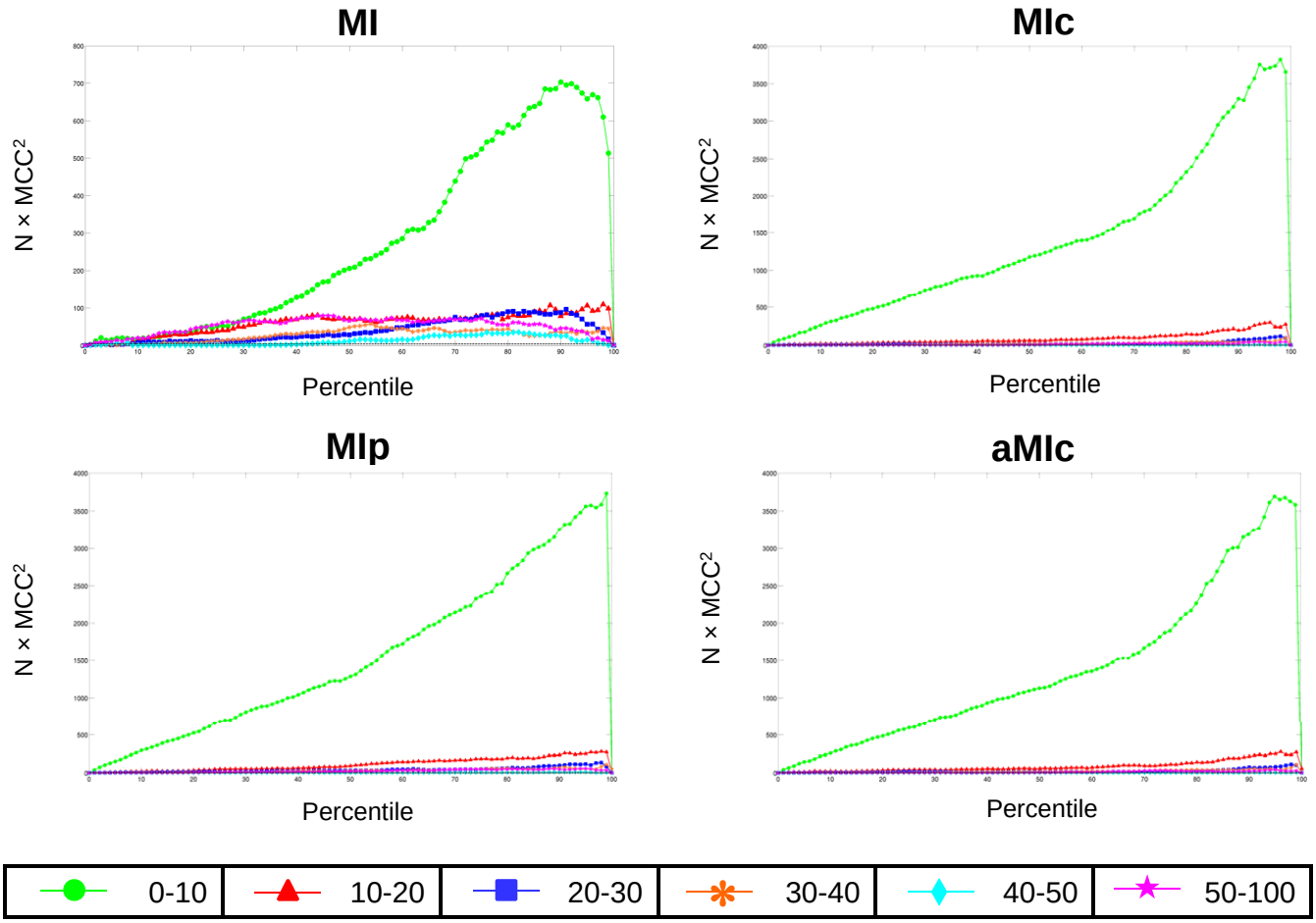

Figure 12: **Contact *versus* non-contact prediction N×MCC^2 curves considering only the residue pairs that are introduced with each gap cutoff increment for MI variants on the 80 Hamer test cases.** In each subplot a curve illustrates the performance of the MI variant, when distinguishing the newly included contact and non-contact residue pairs for the specified gap cutoff increment. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

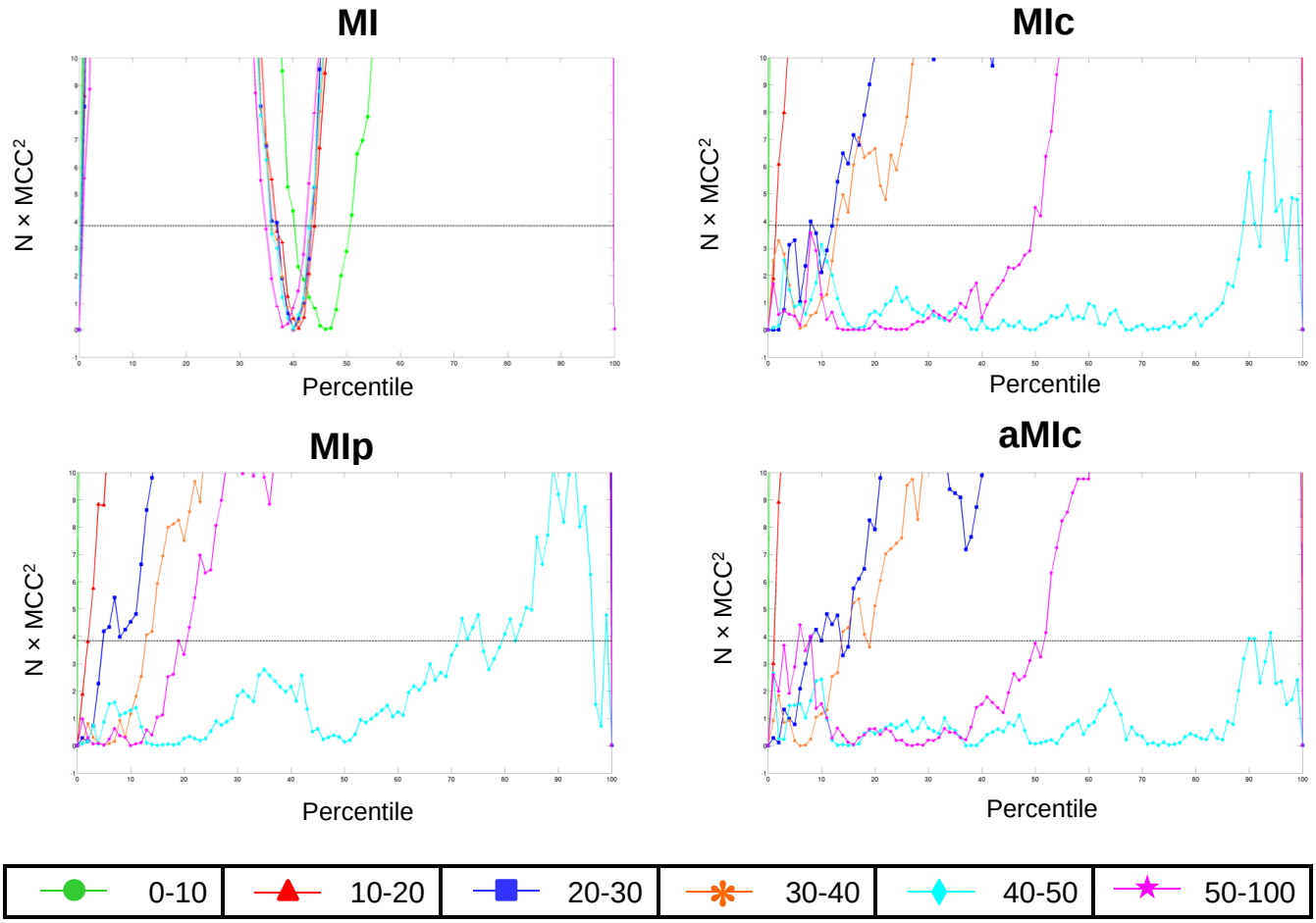

Figure 13: **Contact *versus* non-contact prediction N×MCC^2 curves, around N×MCC^2 = 3.84, considering only the residue pairs that are introduced with each gap cutoff increment for MI variants on the 80 Hamer test cases.** In each subplot a curve illustrates the performance of the MI variant, when distinguishing the newly included contact and non-contact residue pairs for the specified gap cutoff increment. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

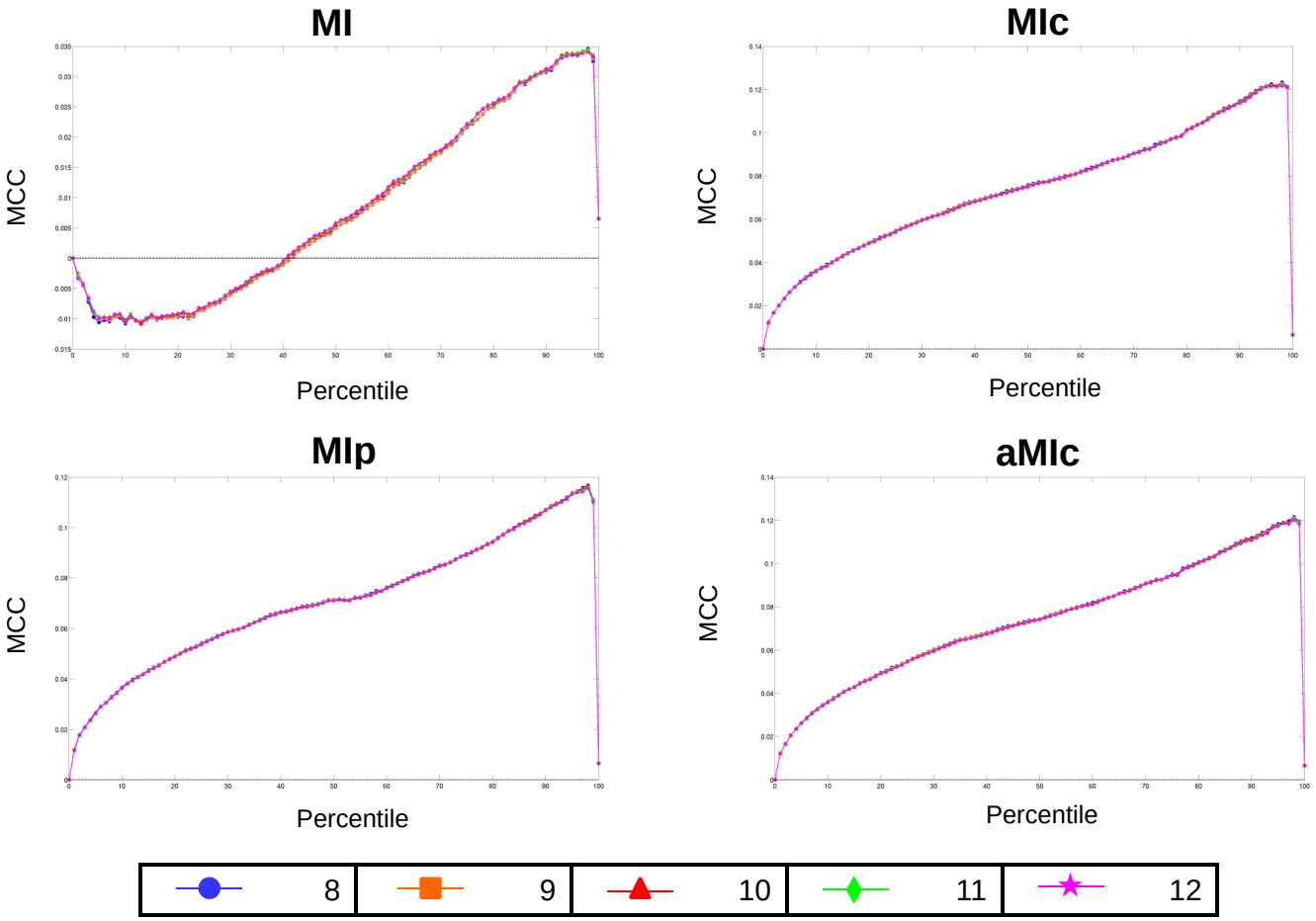

Figure 14: **Contact *versus* non-contact prediction MCC curves when varying the gap cutoff from 8% to 12% for MI variants on the 80 Hamer test cases.** In each subplot a curve illustrates the performance of the MI variant for classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue.

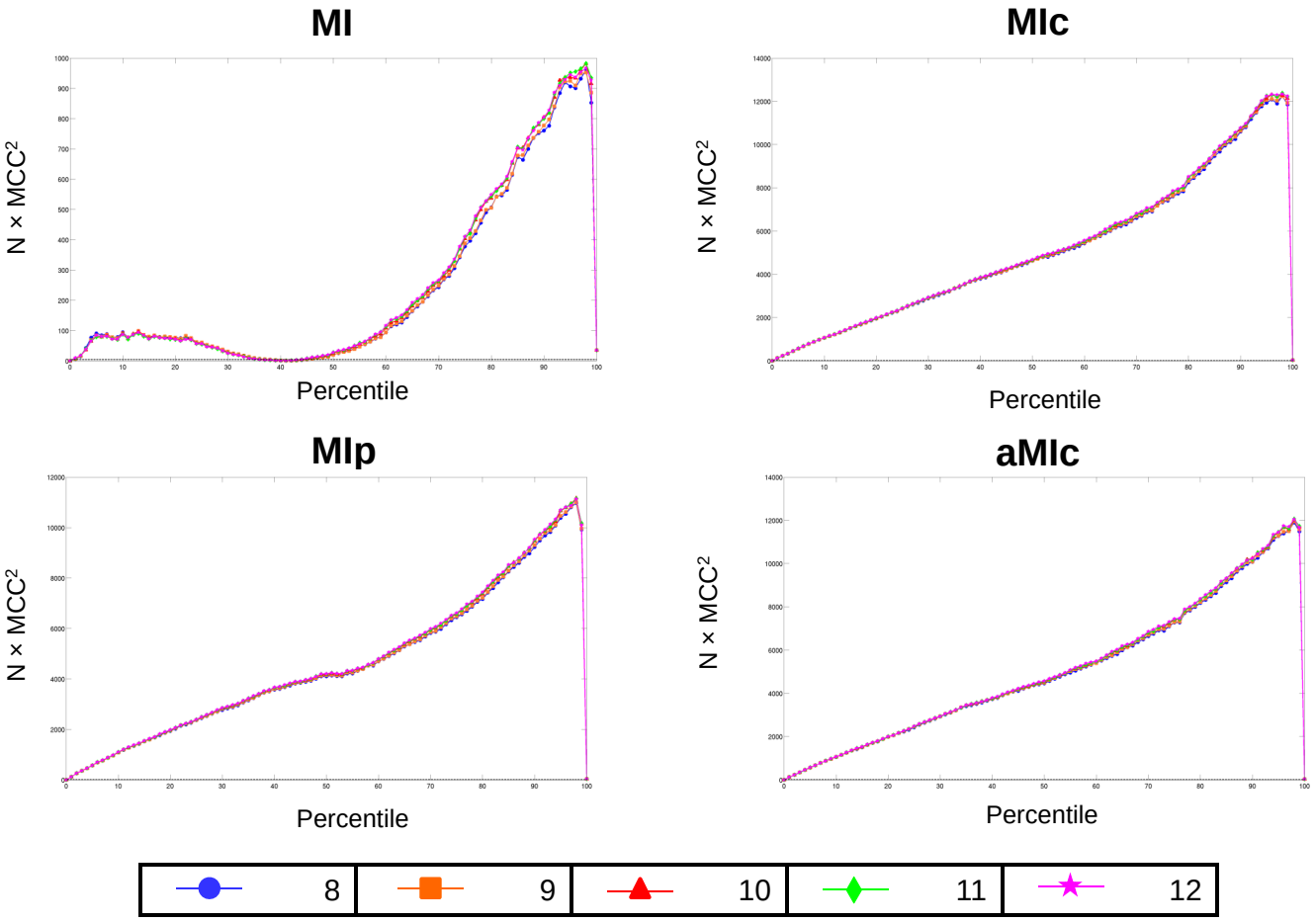

Figure 15: **Contact *versus* non-contact prediction N×MCC^2 curves when varying the gap cutoff from 8% to 12% for MI variants on the 80 Hamer test cases.** In each subplot a curve illustrates the performance of the MI variant for classifying contact *versus* non-contact residue pairs at the indicated gap cutoff. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

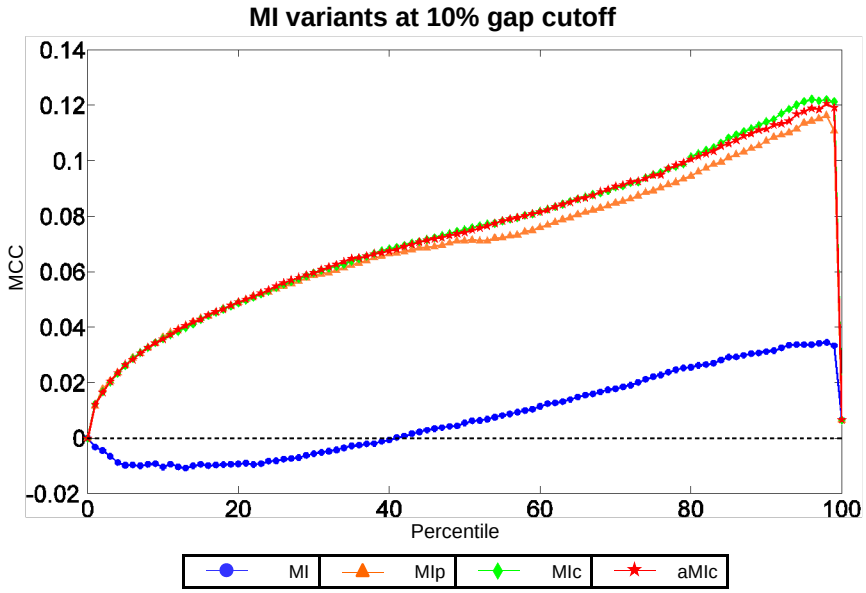

Figure 16: **Contact *versus* non-contact prediction MCC curves at the 10% gap cutoff for MI variants on the 80 Hamer test cases.** Each line on the curve illustrates the performance of MI, MIp, MIc and aMIc respectively when distinguishing contact from non-contact residue pairs at the 10% gap cutoff. The dashed horizontal line at 0 depicts the chance of randomly selecting a contact residue.

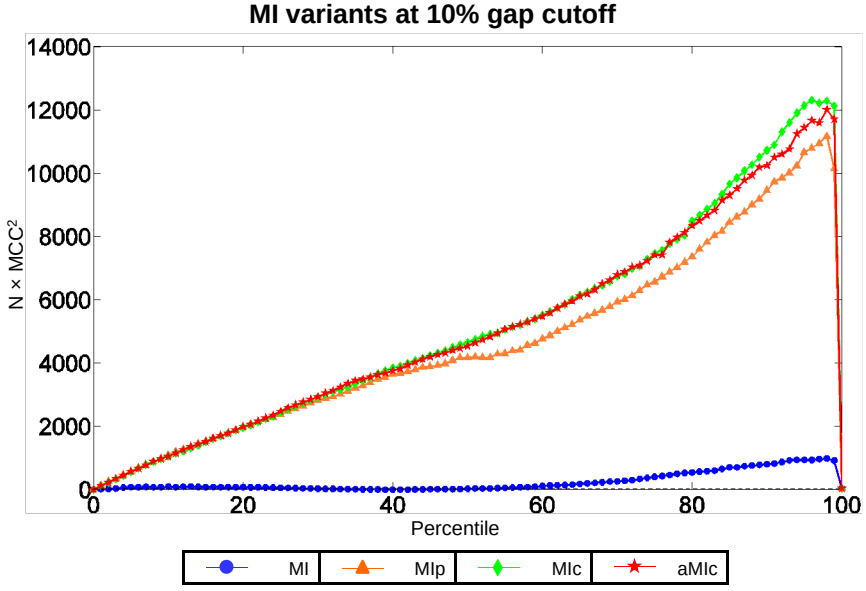

Figure 17: **Contact *versus* non-contact prediction N×MCC^2 curves at the 10% gap cutoff for MI variants on the 80 Hamer test cases.** Each line on the curve illustrates the performance of MI, MIp, MIc and aMIc respectively when distinguishing contact from non-contact residue pairs at the 10% gap cutoff. The dashed horizontal line at 3.84 denotes the chance of randomly selecting a contact residue.

	cutoff	percentile	MCC	cutoff	percentile	N×MCC^2
MIc	0	98	0.1259	20	98	12,400
aMIc	0	98	0.1245	20	98	12,100
MIp	0	97	0.1174	10	98	11,200
MI	10	98	0.0345	20	98	1,000.20

Table 11: **Overall highest MCC and N×MCC^2 achieved, using the Hamer dataset.** MI, MIp, MIc and aMIc were calculated for the 80 test cases for 11 gap cutoffs ranging from 0 to 100%, with 10% cutoff increments. Subsequently MCC and N×MCC^2 curves were plotted. The gap penalty and corresponding percentile at which each of the MI variants achieve the highest MCC and N×MCC^2, respectively, are recorded along with the MCC and N×MCC^2 values.

	percentile	MCC	percentile	N×MCC^2	percentile	F-measure
MIc	96	0.122	96	12,300	64	0.624
aMIc	98	0.120	98	12,006	63	0.624
MIp	98	0.116	98	11,200	63	0.615
MI	98	0.0345	98	982	54	0.510

Table 12: **Highest MCC, N×MCC^2 and F-measure achieved at 10% gap cutoff, using the Hamer dataset.** MI, MIp, MIc and aMIc were calculated for the 80 test cases at a 10% gap cutoff. Subsequently MCC and N×MCC^2 curves were plotted (Figures 5.15 and 5.16), and F-measures calculated. The percentiles at which each of the MI variants achieve the highest MCC, N×MCC^2 and F-measure, respectively, are recorded along with their corresponding values.

References

Alberts, B., Johnson, A., Walter, P., Lewis, J., Raff, M. & Roberts, K. (2007). *Molecular Biology of the Cell*. Garland Science, 5th edn. 3, 4, 6, 8

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410. xiii, 49, 93, 116

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**, 3389–3402. 49, 93, 116

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J.P., Chothia, C. & Murzin, A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, D226–D229. 9

Andreeva, A., Howorth, D., Chandonia, J.M.M., Brenner, S.E., Hubbard, T.J., Chothia, C. & Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, **36**, D419–D425. 15, 16

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230. 8

Aniba, M.R., Poch, O. & Thompson, J.D. (2010). Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Research*, **38**, 7353–7363. 117

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Engineering*, **2**, 101–113. 17

Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W. & Dress, A.W. (2000). Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Molecular Biology and Evolution*, **17**, 164–178. 1, 24, 33

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin,

A. & Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research*, **31**, 400–402. 15

Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, **294**, 93–96. 8

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A.F. & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424. xiv, 45

Barnett, P., Bottger, G., Klein, A.T., Tabak, H.F. & Distel, B. (2000). The peroxisomal membrane protein Pex13p shows a novel mode of SH3 interaction. *The EMBO Journal*, **19**, 6382–6391. 18

Barton, G.J. (1990). Protein multiple sequence alignment and flexible pattern matching. *Methods in Enzymology*, **183**, 403–428. 88, 98, 125

Benner, S.A., Badcoe, I., Cohen, M.A. & Gerloff, D.L. (1994). Bona fide prediction of aspects of protein conformation. *Journal of Molecular Biology*, **235**, 926–958. 88, 98, 125

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The protein data bank. *Nucleic Acids Research*, **28**, 235–242. xv, 4, 14, 19, 21, 51, 90

Bilwes, A.M., Alex, L.A., Crane, B.R. & Simon, M.I. (1999). Structure of CheA, a signal-transducing histidine kinase. *Cell*, **96**, 131–141. 10

Bindewald, E. & Shapiro, B.A. (2006). RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, **12**, 342–352. 33

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370. 13

Bogan, A.A. & Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, **280**, 1–9. 127

Bonneau, R. & Baker, D. (2001). *Ab initio* protein structure prediction: progress and prospects. *Annual Review of Biophysics and Biomolecular Structure*, **30**, 173–189. 5

Bracken, C., Iakoucheva, L.M., Romero, P.R. & Dunker, A.K. (2004). Combining prediction, computation and experiment for the characterization of protein disorder. *Current Opinion in Structural Biology*, **14**, 570–576. 20

Bradford, J.R. & Westhead, D.R. (2005). Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494. 33

Brinda, K.V. & Vishveshwara, S. (2005). Oligomeric protein structure networks: insights into protein–protein interactions. *BMC Bioinformatics*, **6**, 296. 8

Brown, C.A. & Brown, K.S. (2010). Validation of coevolving residue algorithms via pipeline sensitivity analysis: ELSC and OMES and ZNMI, oh my! *PLoS One*, **5**, e10779. xv, 34, 35, 40, 41, 73, 74, 81, 82, 88, 101

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Research*, **33**, D212–D215. 16

Buckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, **45**, 12–19. xv, 43, 44, 49, 64, 76

Burger, L. & van Nimwegen, E. (2010). Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Computational Biology*, **6**, e1000633. 27

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**, 1125–1131. 88

Bustamante, C.D., Townsend, J.P. & Hartl, D.L. (2000). Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Molecular Biology and Evolution*, **17**, 301–308. 66, 116

Camacho, C.J., Weng, Z., Vajda, S. & DeLisi, C. (1999). Free energy landscapes of encounter complexes in protein–protein association. *Biophysical Journal*, **76**, 1166–1178. 18, 21

Carugo, O. & Argos, P. (1997). Protein–protein crystal-packing contacts. *Protein Science*, **6**, 2261–2263. 21, 22

Cavallo, L., Kleinjung, J. & Fraternali, F. (2003). POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Research*, **31**, 3364–3366. 71

Chakrabarti, S. & Panchenko, A.R. (2009). Coevolution in defining the functional specificity. *Proteins*, **75**, 231–240. 24, 88

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

Clarke, N.D. (1995). Covariation of residues in the homeodomain sequence family. *Protein Science*, **4**, 2269–2278. 24, 33, 34

Cover, T.M. & Thomas, J.A. (1991). *Elements of Information Theory*. Wiley-Interscience, 99th edn. 36

Crawford, I.P., Niermann, T. & Kirschner, K. (1987). Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. *Proteins*, **2**, 118–129. 88, 98, 125

Crick, F.H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, **12**, 138–163. 8

Cuff, A.L., Sillitoe, I., Lewis, T., Clegg, A.B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J. & Orengo, C.A. (2011). Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Research*, **39**, D420–D426. 15, 16

Cuff, J.A. & Barton, G.J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511. 71

Davis, F.P. (2011). Proteome-wide prediction of overlapping small molecule and protein binding sites using structure. *Molecular BioSystems*, **7**, 545–557. 33

Deane, C.M., Salwiński, Ł., Xenarios, I. & Eisenberg, D. (2002). Protein interactions. *Molecular and Cellular Proteomics*, **1**, 349–356. 6

Dekker, J.P., Fodor, A., Aldrich, R.W. & Yellen, G. (2004). A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572. 24

DeLano, W.L. (2002). The PyMOL molecular graphics system. 10, 131

Devlin, T.M. (2005). *Textbook of Biochemistry with Clinical Correlations*. Wiley-Liss, 6th edn. 4, 10

Do, C.B., Mahabhashyam, M.S., Brudno, M. & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, **15**, 330–340. 12

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**, 333–340. xi, xiv, 1, 2, 24, 27, 28, 33, 34, 35, 37, 40, 47, 48, 63, 69, 70, 73, 76, 82, 88, 90, 98, 124

Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press. 30, 33, 35, 40

Eddy, S.R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, **7**, e1002195. xiii, 114

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797. xiv, 12, 49, 89, 93, 114

Edgar, R.C. (2010). Quality measures for protein alignment benchmarks. *Nucleic Acids Research*, **38**, 2145–2153. 117

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874. 43, 44

Fernandes, A.D. & Gloor, G.B. (2010). Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics*, **26**, 1135–1139. 33, 36

Fitch, W.M. & Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, **4**, 579–593. 122

Fodor, A.A. & Aldrich, R.W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, **56**, 211–221. 24, 34, 40, 66, 82

Fraczkiewicz, R. & Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry*, **19**, 319–333. 71

Fuller, J.C., Burgoyne, N.J. & Jackson, R.M. (2009). Predicting druggable binding sites at the protein–protein interface. *Drug Discovery Today*, **14**, 155–161. 6

Gajda, M.J., Pawlowski, M. & Bujnicki, J.M. (2011). Protein structure prediction: from recognition of matches with known structures to recombination of fragments multiscale approaches to protein modeling. In A. Kolinski, ed., *Multiscale approaches to protein modeling*, chap. 10, 231–254, Springer New York, New York, New York. 4

Gasteiger, E., Jung, E. & Bairoch, A. (2001). SWISS-PROT: connecting biomolecular knowledge via a protein database. *Current Issues in Molecular Biology*, **3**, 47–55. 13

Geppert, T., Hoy, B., Wessler, S. & Schneider, G. (2011). Context-based identification of protein–protein interfaces and "hot-spot" residues. *Chemistry and Biology*, **18**, 344–353. 126

Göbel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317. 24, 33

Goldman, N., Thorne, J.L. & Jones, D.T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458. 66, 116

Golubchik, T., Wise, M.J., Easteal, S. & Jermiin, L.S. (2007). Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Molecular Biology and Evolution*, **24**, 2433–2442. 117

Gomes, M., Hamer, R., Reinert, G. & Deane, C. (2012). Mutual information and variants for protein domain-domain contact prediction. *BMC Research Notes*, **5**, 472. 35, 62, 73, 124

Gomez, S.M., Noble, W.S. & Rzhetsky, A. (2003). Learning to predict protein–protein interactions from protein sequences. *Bioinformatics*, **19**, 1875–1881. 6

Gouveia-Oliveira, R., Sackett, P.W. & Pedersen, A.G. (2007). MaxAlign: maximizing usable data in an alignment. *BMC Bioinformatics*, **8**, 312. xiv, 13, 49, 89, 93, 114

Guarracino, M.R., Nebbia, A., Manna, V., Chinchuluun, A. & Pardalos, P.M. (2010). Efficient prediction of protein–protein interactions using sequence information. *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*, **0**, 677–682. 6

Haft, D.H., Selengut, J.D. & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371–373. 15

Hakes, L., Lovell, S.C., Oliver, S.G. & Robertson, D.L. (2007). Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proceedings of the National Academy of Sciences*, **104**, 7999–8004. 58

Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, **138**, 774–786. 41, 82

Halperin, I., Wolfson, H. & Nussinov, R. (2006). Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins*, **63**, 832–845. 1, 2, 5, 21, 22, 23, 24, 35, 58, 60

Hamer, R., Luo, Q., Armitage, J.P., Reinert, G. & Deane, C.M. (2010). i-Patch: interprotein contact prediction using local network information. *Proteins*, **78**, 2781–2797. iii, ix, xiii, 2, 25, 29, 30, 31, 32, 33, 41, 42, 48, 49, 51, 54, 60, 63, 67, 72, 74, 76, 77, 78, 83, 84, 85, 86, 89, 93, 114, 116, 123, 125, 126, 140, 141, 142

Harris, B.Z., Hillier, B.J. & Lim, W.A. (2001). Energetic determinants of internal motif recognition by PDZ domains. *Biochemistry*, **40**, 5921–5930. 18

Hodges, P.E., Payne, W.E. & Garrels, J.I. (1998). The yeast protein database (YPD): a curated proteome database for *Saccharomyces cerevisiae*. *Nucleic Acids Research*, **26**, 68–72. 14

Hodis, E., Prilusky, J., Martz, E., Silman, I., Moult, J. & Sussman, J.L. (2008). Proteopedia - a scientific wiki bridging the rift between 3D structure and function of biomacromolecules. *Genome Biology*, **9**, R121. 14

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology*, **233**, 123–138. 10

Holm, L. & Sander, C. (1994). Parser for protein folding units. *Proteins*, **19**, 256–268. 49

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**, 1607–1621. 28, 127, 137

Horner, D.S., Pirovano, W. & Pesole, G. (2008). Correlated substitution analysis and the prediction of amino acid structural contacts. *Briefings in Bioinformatics*, **9**, 46–56. 33

Horton, N. & Lewis, M. (1992). Calculation of the free energy of association for protein complexes. *Protein Science*, **1**, 169–181. 18

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. & Sigrist, C.J. (2006). The PROSITE database. *Nucleic Acids Research*, **34**. 15

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A.F., Rivoire, C., Sangrador-Vegas, A., Selengut, J.D., Sigrist, C.J.A., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P.D., Wu, C.H., Yeats, C. & Yong, S.Y. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, **40**, D306–D312. 16

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, **98**, 4569–4574. 6

Ivanov, Y.D., Kanaeva, I.P., Karuzina, I.I., Archakov, A.I., Hoa, G.H. & Sligar, S.G. (2001). Molecular recognition in the p450cam monooxygenase system: direct monitoring of protein–protein interactions by using optical biosensor. *Archives of Biochemistry and Biophysics*, **391**, 255–264. 18

Janin, J. (1995). Protein–protein recognition. *Progress in Biophysics and Molecular Biology*, **64**, 145–166. 17

Jayasinghe, S., Hristova, K. & White, S.H. (2001). MPtopo: A database of membrane protein topology. *Protein Science*, **10**, 455–458. 14

Jessulat, M., Pitre, S., Gui, Y., Hooshyar, M., Omidi, K., Samanfar, B., Tan, L.H., Alamgir, M., Green, J., Dehne, F. & Golshani, A. (2011). Recent advances in protein–protein interaction prediction: experimental and computational methods. *Expert Opinion on Drug Discovery*, **6**, 921–935. 6, 7, 8

Jones, D.T. (1997). Progress in protein structure prediction. *Current Opinion in Structural Biology*, **7**, 377–387. 5, 127

Jones, D.T., Buchan, D.W., Cozzetto, D. & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190. xv, 5, 23, 27, 28, 56, 61, 127, 137

Jones, S. & Thornton, J.M. (1995). Protein–protein interactions: a review of protein dimer structures. *Progress in Biophysics and Molecular Biology*, **63**, 31–65. 17

Jones, S. & Thornton, J.M. (1996). Principles of protein–protein interactions. *Proceedings of the National Academy of Sciences*, **93**, 13–20. 127

Joo, K., Lee, S.J.J. & Lee, J. (2012). Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins*, **80**, 1791–1797. 71

Kass, I. & Horovitz, A. (2002). Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617. 24, 40, 82

Keskin, O. & Nussinov, R. (2007). Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, **15**, 341–354. 19

Keskin, O., Ma, B. & Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of Molecular Biology*, **345**, 1281–1294. 18

Kim, W.K., Bolser, D.M. & Park, J.H. (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics*, **20**, 1138–1150. 58

Korber, B.T., Farber, R.M., Wolpert, D.H. & Lapedes, A.S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences*, **90**, 7176–7180. 1, 33

Korn, A.P. & Burnett, R.M. (1991). Distribution and complementarity of hydropathy in multisubunit proteins. *Proteins*, **9**, 37–55. 18

Kryshtafovych, A., Fidelis, K. & Moult, J. (2011). CASP9 results compared to those of previous CASP experiments. *Proteins*, **79**, 196–207. 5

Lapedes, A.S., Giraud, B.G., Liu, L. & Stormo, G.D. (1999). Correlated mutations in protein sequences: phylogenetic and structural effects. 236–256. 25

Lee, B.C. & Kim, D. (2009). A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513. xiii, xiv, 22, 29, 34, 35, 37, 38, 39, 47, 48, 63, 73, 83, 88, 90, 95, 96, 97, 98, 123, 124, 125

Letunic, I., Doerks, T. & Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research*, **40**, D302–D305. 15

Lewis, A.C.F., Saeed, R. & Deane, C.M. (2010). Predicting protein–protein interactions in the context of protein evolution. *Molecular BioSystems*, **6**, 55–64. 29

Li, W. & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659. 13, 49, 93

Li, W., Hamill, S.J., Hemmings, A.M., Moore, G.R., James, R. & Kleanthous, C. (1998). Dual recognition and the role of specificity-determining residues in colicin E9 DNase-immunity protein interactions. *Biochemistry*, **37**, 11771–11779. 18

Liang, S., Zhang, C., Liu, S. & Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. *Nucleic Acids Research*, **34**, 3698–3707. 126

Lin, Y.S., Hsu, W.L., Hwang, J.K. & Li, W.H. (2007). Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. *Molecular Biology and Evolution*, **24**, 1005–1011. 66, 116

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**, e4762. 35, 41, 82

Liu, Z.P., Wu, L.Y., Wang, Y., Zhang, X.S. & Chen, L. (2010). Prediction of proteinRNA binding sites by a random forest method with combined features. *Bioinformatics*, **26**, 1616–1622. xiii, 43, 45, 46

Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *Journal of Molecular Biology*, **285**, 2177–2198. 18

Lockless, S.W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299. 24, 122

Lunt, B., Szurmant, H., Procaccini, A., Hoch, J.A., Hwa, T. & Weigt, M. (2010). *Inference of direct residue contacts in two-component signaling*, vol. 471, 17–41. Elsevier. 25

Luque, I. & Freire, E. (2000). Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins*, **Suppl 4**, 63–71. 18

Ma, B., Shatsky, M., Wolfson, H.J. & Nussinov, R. (2002). Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Science*, **11**, 184–197. 18

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**, 7708–7713. 41, 83

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Lu, F., Marchler, G.H., Mullokandov, M., Omelchenko, M.V., Robertson, C.L., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Zheng, C. & Bryant, S.H. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, **39**, D225–D229. 16, 90

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**, e28766. 5, 28, 127, 137

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**, 4116–4124. 1, 2, 24, 29, 33, 34, 35, 66, 69, 78, 81, 88, 90, 132

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, **405**, 442–451. xiv, 43, 76, 141

Meinel, T., Krause, A., Luz, H., Vingron, M. & Staub, E. (2005). The SYSTERS protein family database in 2005. *Nucleic Acids Research*, **33**, D226–D229. 16

Merkl, R. & Zwick, M. (2008). H2r: identification of evolutionary important residues by means of an entropy based analysis of multiple sequence alignments. *BMC Bioinformatics*, **9**. 24, 30

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H. & Thomas, P.D. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, **33**, D284–D288. 15

Mika, S. & Rost, B. (2006). Protein–protein interactions more conserved within species than across species. *PLoS Computational Biology*, **2**, e79. 29

Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Engineering*, **3**, 77–83. 17

Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, **102**, 10930–10935. 18

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. & Overington, J.P. (1998). JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623. xiv, 11, 64, 71

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*, **108**, E1293–E1301. 27, 127, 137

Moreira, I.S., Fernandes, P.A. & Ramos, M.J. (2007). Hot spots–A review of the protein–protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812. 17, 18

Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)–round IX. *Proteins*, **79**, 1–5. 5

Mulder, N.J. (2001). Protein family databases. 15, 16

Neuvirth, H., Raz, R. & Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *Journal of Molecular Biology*, **338**, 181–199. 126

Noble, M.E., Wierenga, R.K., Lambeir, A.M., Opperdoes, F.R., Thunnissen, A.M., Kalk, K.H., Groendijk, H. & Hol, W.G. (1991). The adaptability of the active site of trypanosomal triosephosphate isomerase as observed in the crystal structures of three different complexes. *Proteins*, **10**, 50–69. 70

Notredame, C., Higgins, D.G. & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**, 205–217. 12

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*, **109**, E1540–E1547. 138

Nuin, P., Wang, Z. & Tillier, E. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471. 12

Overington, J., Donnelly, D., Johnson, M.S., Sali, A. & Blundell, T.L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Science*, **1**, 216–226. 66, 116

Pace, C.N., Shirley, B.A., McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB Journal*, **10**, 75–83. 9

Pagel, P., Wong, P. & Frishman, D. (2004). A domain interaction map based on phylogenetic profiling. *Journal of Molecular Biology*, **344**, 1331–1346. 29

Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. (1997). Correlated mutations contain information about protein–protein interaction. *Journal of Molecular Biology*, **271**, 511–523. 24, 29, 33, 47, 49, 123

Pelletier, J.N., Arndt, K.M., Plückthun, A. & Michnick, S.W. (1999). An *in vivo* library-versus-library selection of optimized protein–protein interactions. *Nature Biotechnology*, **17**, 683–690. 6

Petryszak, R., Kretschmann, E., Wieser, D. & Apweiler, R. (2005). The predictive power of the CluSTr database. *Bioinformatics*, **21**, 3604–3609. 16

Pirovano, W., Feenstra, K.A. & Heringa, J. (2006). Sequence comparison by sequence harmony identifies subtype-specific functional sites. *Nucleic Acids Research*, **34**, 6540–6548. 30, 32

Pleiss, J. (2011). Protein design in metabolic engineering and synthetic biology. *Current Opinion in Biotechnology*, **22**, 611–617. 19

Poon, A. & Chao, L. (2005). The rate of compensatory mutation in the DNA bacteriophage phiX174. *Genetics*, **170**, 989–999. 122

Portugaly, E., Linial, N. & Linial, M. (2007). EVEREST: a collection of evolutionary conserved protein domains. *Nucleic Acids Research*, **35**, D241–D246. 15

Pruitt, K.D., Tatusova, T. & Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **35**, D61–D65. 4

Pugalenthi, G., Kumar Kandaswamy, K., Chou, K.C., Vivekanandan, S. & Kolatkar, P. (2012). RSARF: Prediction of residue solvent accessibility from protein sequence using random forest method. *Protein and Peptide Letters*, 50–56. 71

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A. & Finn, R.D. (2012). The Pfam protein families database. *Nucleic Acids Research*, **40**, D290–D301. xv, 15, 28, 88, 90, 92, 125, 131

Read, R.J., Brayer, G.D., Jurasek, L. & James, M.N.G. (1984). Critical evaluation of comparative model building of *Streptomyces griseus* trypsin. *Biochemistry*, **23**, 6570–6575. 4

Rhodes, G. (2006). *Crystallography made Crystal Clear, Third Edition: a Guide for Users of Macromolecular Models*. Academic Press, 3rd edn. 20

Richardson, J.S. (1981). The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, **34**, 167–339. 9, 10

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. & Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, **17**, 1030–1032. 6

Sali, A. & Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, **234**, 779–815. 5

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I.M., Geer, L.Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, J.J., Yaschenko, E. & Ye, J. (2012). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **40**, D13–D25. 13, 49, 93

Schneider, T.D. & Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, **18**, 6097–6100. 30, 31, 32

Sengupta, D. & Kundu, S. (2012). Role of long- and short-range hydrophobic, hydrophilic and charged residues contact network in protein's structural organization. *BMC Bioinformatics*, **13**, 142. 127

Sheinerman, F.B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Current Opinion in Structural Biology*, **10**, 153–159. 18

Shen, W., Yun, S., Tam, B., Dalal, K. & Pio, F. (2005). Target selection of soluble protein complexes for structural proteomics studies. *Proteome Science*, **3**, 3. 20

Shenoy, S.R. & Jayaram, B. (2010). Proteins: sequence to structure and function–current status. *Current Protein and Peptide Science*, **11**, 498–514. 6

Siddiqui, A.S. & Barton, G.J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Science*, **4**, 872–884. 49

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**, 209–225. 5

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M. & Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell*, **133**, 1043–1054. viii, 1, 35, 67, 68, 75, 84, 124

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**, 1315–1317. 6

Sowdhamini, R. & Blundell, T.L. (1995). An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Science*, **4**, 506–520. 49

Stevens, J.M., Armstrong, R.N. & Dirr, H.W. (2000). Electrostatic interactions affecting the active site of class sigma glutathione S-transferase. *The Biochemical Journal*, **347 Pt 1**, 193–197. 18

Stumpf, M.P.H., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M. & Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, **105**, 6959–6964. 6

Suel, G.M., Lockless, S.W., Wall, M.A. & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural and Molecular Biology*, **10**, 59–69. 24

Sułkowska, J.I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J.N. (2012). Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences*, **109**, 10340–10345. 28, 127, 137

Szilágyi, A., Grimm, V., Arakaki, A.K. & Skolnick, J. (2005). Prediction of physical protein–protein interactions. *Physical Biology*, **2**, S1. 6

Tarassov, K., Messier, V., Landry, C.R., Radinovic, S., Serna Molina, M.M., Shames, I., Malitskaya, Y., Vogel, J., Bussey, H. & Michnick, S.W. (2008). An *in vivo* map of the yeast protein interactome. *Science*, **320**, 1465–1470. 6

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680. 12

Thompson, J.D., Linard, B., Lecompte, O. & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS One*, **6**, e18093. 117

Thornton, J.M. (2001). The Hans Neurath award lecture of the protein Society: proteins – a testament to physics, chemistry, and evolution. *Protein Science*, **10**, 3–11. 18

Tillier, E.R.M. & Lui, T.W.H. (2003). Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, **19**, 750–755. 1, 24, 34, 52

Togneri, R. & DeSilva, C.J.S. (2003). *Fundamentals of Information Theory and Coding Design*. CRC Press, Inc., Boca Raton, Florida, USA. 33

Tomovic, A. & Oakeley, E.J. (2007). Position dependencies in transcription factor binding sites. *Bioinformatics*, **23**, 933–941. 33

Tuncbag, N., Kar, G., Keskin, O., Gursoy, A. & Nussinov, R. (2009). A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. *Briefings in Bioinformatics*, **10**, 217–232. 7

Tuncbag, N., Gursoy, A., Nussinov, R. & Keskin, O. (2011). Predicting protein–protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. *Nature Protocols*, **6**, 1341–1354. 19

Tusnády, G.E., Dosztányi, Z. & Simon, I. (2005). PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Research*, **33**, D275–D278. 15

Tyagi, M., Shoemaker, B.A., Bryant, S.H. & Panchenko, A.R. (2009). Exploring functional roles of multibinding protein interfaces. *Protein Science*, **18**, 1674–1683. 18

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. & Rothberg, J.M. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627. 6

UniProt Consortium (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research*, **39**, D214–D219. 14, 16

Uzman, A. (2001). Molecular Cell Biology. *Biochemistry and Molecular Biology Education*, 126–128. 9

Valdar, W.S.J. (2002). Scoring residue conservation. *Proteins*, **48**, 227–241. 30

Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current Opinion in Structural Biology*, **12**, 368–373. 19

Vijayakumar, M., Wong, K.Y., Schreiber, G., Fersht, A.R., Szabo, A. & Zhou, H.X. (1998). Electrostatic enhancement of diffusion-controlled protein–protein association: comparison of theory and experiment on barnase and barstar. *Journal of Molecular Biology*, **278**, 1015–1024. 18

Wang, G. & Dunbrack, R.L. (2004). Scoring profile-to-profile sequence alignments. *Protein Science*, **13**, 1612–1626. 128

Wass, M., Fuentes, G., Pons, C., Pazos, F. & Valencia, A. (2011). Towards the prediction of protein interaction partners using physical docking. *Molecular Systems Biology*, **7**. 6, 7

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M. & Barton, G.J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191. 129, 131

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*, **106**, 67–72. 5, 25, 26, 27, 29, 127, 137, 139

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Helmberg, W., Kapustin, Y., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. & Yaschenko, E. (2006). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **34**. 51

Williams, S.G. & Lovell, S.C. (2009). The effect of sequence evolution on protein structural divergence. *Molecular Biology and Evolution*, **26**, 1055–1065. 8

Wilson, D., Madera, M., Vogel, C., Chothia, C. & Gough, J. (2007). The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Research*, **35**, D308–D313. 15

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*, **97**, 3288–3291. 1, 33, 34

Wu, C.H., Yeh, L.S.S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J. & Barker, W.C. (2003). The protein information resource. *Nucleic Acids Research*, **31**, 345–347. 13

Xia, J.F., Zhao, X.M., Song, J. & Huang, D.S. (2010). APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, **11**, 174. 127, 128

Xu, D., Lin, S.L. & Nussinov, R. (1997a). Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. *Journal of Molecular Biology*, **265**, 68–84. 18

Xu, D., Tsai, C.J. & Nussinov, R. (1997b). Hydrogen bonds and salt bridges across protein–protein interfaces. *Protein Engineering*, **10**, 999–1012. 18

Xu, Y. & Tillier, E.R.M. (2010). Regional covariation and its application for predicting protein contact patches. *Proteins*, **78**, 548–558. 33

Yanofsky, C., Horn, V. & Thorpe, D. (1964). Protein structure relationships revealed by mutational analysis. *Science*, **146**, 1593–1594. 123

Yeats, C., Maibaum, M., Marsden, R., Dibley, M., Lee, D., Addou, S. & Orengo, C.A. (2006). Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Research*, **34**. 15

Yeo, G. & Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, **11**, 377–394. 30

Yona, G., Linial, N. & Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Research*, **28**, 49–55. 16

Young, L., Jernigan, R.L. & Covell, D.G. (1994). A role for surface hydrophobicity in protein–protein recognition. *Protein Science*, **3**, 717–729. 18

Zapf, J., Sen, U., Madhusudan, Hoch, J.A. & Varughese, K.I. (2000). A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction. *Structure*, **8**, 851–862. 11, 19, 22, 25, 68

Zhang, Q.C., Petrey, D., Norel, R. & Honig, B.H. (2010). Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences*, **107**, 10896–10901. 33

Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, **33**, 2302–2309. 128, 129