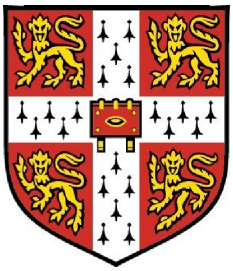
Molecular characterization and evolutionary plasticity of protein-protein interfaces



G R J Bickerton

Emmanuel College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

March 2009

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

For Ruth...

Acknowledgements

First and foremost I would like to thank my supervisor, Professor Sir Tom Blundell FRS, FMedSci. I will always be grateful to him for giving me the opportunity to study in his group as well as for his encouragement, guidance and endless insight he offered throughout.

I would like to take this opportunity to thank the members of the database development team - in no particular order Semin Lee, Sungsam Gong, Adrian Schreyer, Duangrudee Tanramluk, Alicia Higueruelo and Will Pitt. The closely collaborative nature of the development work was one of the most enjoyable aspects of my studies.

My thanks also go to Dr Catherine Worth and Dr Julia Forman for their tremendous help with the work on mutations, particularly in preparation of publications. I am grateful to Dr David Burke for his patient help with BATON, Dr Rinaldo Wander Montalvao for his endless ingenuity and technical acumen and Graham Eliff for his cheerful computing support.

I would also like to extend my thanks to Emmanuel College, most particularly for providing such excellent accommodation for my wife and I during our time in Cambridge. My thanks also go to the BBSRC for their kind financial support.

I am also much indebted to Professor Andrew Hopkins for his kindness and support whilst I completed writing my thesis alongside undertaking a postdoc position in his lab.

The love, support and encouragement that my parents have given me over the years are immeasurable, as is my appreciation for all that they have done. Finally I want to express my gratitude to my wife Ruth. Her unflinching patience, warmth and encouragement have been my inspiration. The sacrifices she has made over

the last three years are greater than my own and are something I shall never forget; it is to her that this thesis is dedicated.

Abstract

The sequencing of the human genome provides the parts list for understanding cellular processes. However, as 70% of eukaryotic genes work through multi-protein systems, it is only through detailed study of the interactions of these components, that a more complete, systems-level understanding can be gained. This thesis is centred on the establishment of PICCOLO - a comprehensive database of structurally characterized protein interactions. In generating the resource, issues of interface definition, quaternary structure, data redundancy, structural environment and interaction type are addressed. The resource enables a variety of analyses to be performed concerning interface properties including residue propensity, hydropathy, polarity, interface size, sequence entropy and residue contact preference.

PICCOLO has been applied to probing the patterns of substitutions that are accepted in protein interfaces across evolution, and whether these patterns are distinguishable from those seen in other structural environments. The derivation of a high-quality set of multiple structural alignments in the form of the database TOCCATA, a prerequisite for such analysis, is described, as well as procedures to derive environment-specific substitution tables.

The Blundell group has contributed a series of methods to predict the likely effect of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability, function and interactions in order to triage the large volumes of data created from high-throughput genetic screening studies, enabling prioritization of those nsSNPs most likely to be phenotypically detrimental. PICCOLO's contribution to these predictions is described.

Historically there has been little focus on protein-protein interactions as drug targets for small-molecule therapeutics. However, alanine-scanning mutagenesis studies have revealed that only a subset of residues contribute the greater part of free energy to binding - so-called "hot-spots". Molecular characterization of hot-spots performed using PICCOLO, probes the molecular basis underlying this important phenomenon leading to the possibility of predictive methods to identify hot-spots *in silico*.

Contents

List of Figures

Nomenclature

Acronyms

$3D$ Three Dimensional

API Application Programming Interface

ASA Accessible Surface Area

ASU Asymmetric Unit

AUC Analytical Ultracentrifugation

$BLAST$ Basic Local Alignment Search Tool

$CAPRI$ Critical Assessment of Protein Interactions

$CASP$ Critical Assessment of Structure Prediction

CPU Central Processing Unit

CSD Cambridge Structural Database

DMC Double Mutant Cycle

DNA Deoxyribonucleic Acid

EBI European Bioinformatics Institute

$EGOR$ Environment-Specific Substitution Table GeneratOR

EM Electron Microscopy

ESST Environment-Specific Substitution Table

ET Evolutionary Trace

FCCS Fluorescence Cross-Correlation Spectroscopy

FGF Fibroblast Growth Factor

FRET Fluorescence Resonance Energy Transfer

GO Gene Ontology

HOMSTRAD Homologous Structure Alignment Database

ITC Isothermal Titration Calorimetry

MAF Minor Allele Frequency

mmCIF macromolecular Crystallographic Information File

MDS Multi Dimensional Scaling

MS Mass Spectrometry

MSA Multiple Sequence Alignment

NCBI National Center for Biotechnology Information

NMR Nuclear Magnetic Resonance

NPC Nuclear Pore Complex

nsSNP non-synonymous Single Nucleotide Polymorphisms

ORF Open Reading Frame

PDB Protein Data Bank

PID Percent Identity

PiQSI Protein Quaternary Structure Investigation

PISA Protein Interfaces, Surfaces and Assemblies

PQS Protein Quaternary Structure

RAID Redundant Array of Inexpensive Disks

RAM Random Access Memory

RCSB Research Collaboratory for Structural Bioinformatics

RDBMS Relational Database Management System

RMSD Root Mean Square Deviation

SCOP Structural Classification of Proteins

SDM Site Directed Mutator

SIFTS Structure integration with function, taxonomy and sequence

SMARTS SMiles ARbitrary Target Specification

SPR Surface Plasmon Resonance

SQL Structured Query Language

SVM Support Vector Machine

TAP Tandem Affinity Purification

TEV Tobacco Etch Virus

TLB Tom Leon Blundell

URL Uniform Resource Locator

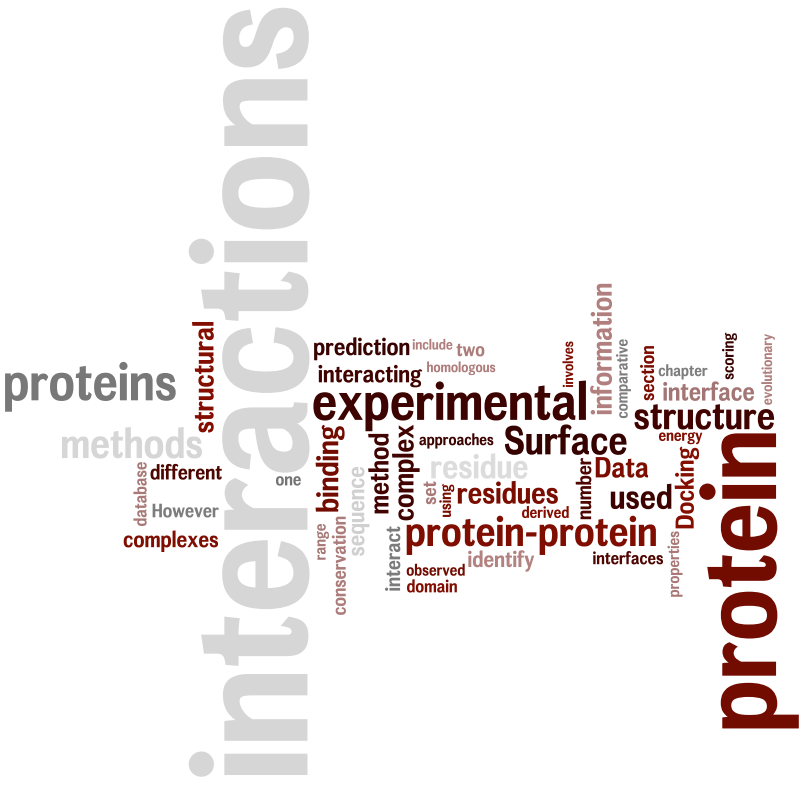
wwPDB Worldwide Protein Data Bank

XML Extensible Markup Language

Y2H Yeast Two-Hybrid System

Chapter 1

Introduction



1.1 Fundamental importance of protein interactions

The sequencing of the human genome provides the parts list for understanding cellular processes (Lander *et al.* (2001); Venter (2001)). However, as 70% of eukaryotic genes work through multi-protein systems, it is only through studying the interactions of these components that a more complete understanding can be gained. The fundamental importance of protein-protein interactions cannot be understated; the cellular processes they mediate include cell communication, proliferation and differentiation, DNA repair and immunity. Importantly, it is often not through pairwise interactions, but rather through weak but synergistic interactions among many components that specific and sensitive regulation is achieved, thus ensuring high fidelity in signal transduction.

As we endeavour to gain a systems level understanding of cellular processes, it is clear that we will require a greater understanding of protein interactions, both at the detailed level of individual interactions as well as broad principles, which might be of general application. A range of experimental and computational techniques has been used to study interactions, each of which provides information of a different nature, resolution and quality.

1.2 Experimental methods for studying interactions

Three-dimensional (3D) structural information of protein complexes deposited in the Protein Data Bank (PDB)(Berman *et al.* (2007)) provides the most complete and highest quality information regarding protein-protein interactions; ultimately it is the fine atomic details of an interaction that determine the affinity and specificity of binding. Despite several technical advances in structural genomics and some excellent recent examples, experimental determination of protein complexes remains difficult. This is reflected in the fact that until recently, less than 1% of all structures solved by the structural genomics consortia

are of protein-protein complexes (Todd *et al.* (2005)). For X-ray crystallography, crystallization remains a bottleneck, particularly for transient complexes. With NMR (Nuclear Magnetic Resonance) techniques there is an upper size limit of around 100kDa (although most are less than 20kDa) for atomic resolution models (although NMR can provide valuable information regarding interacting residues from chemical shift analysis). More encouragingly, cryo-Electron Microscopy (EM) offers much potential in providing lower-resolution structures for large complexes using relatively small amounts of material. Experimental methods to *identify* interactions include co-immunoprecipitation, chemical crosslinking, protein microarrays, synthetic lethality screens, synexpression, phage display, yeast two-hybrid and tandem affinity purification. Methods to *characterize* a known interaction include Surface Plasmon Resonance (SPR), Isothermal Titration Calorimetry (ITC), analytical ultracentrifugation (AUC), fluorescence resonance energy transfer (FRET), Fluorescence Cross-Correlation spectroscopy (FCCS) and alanine scanning mutagenesis. Each of these methods will be briefly reviewed here.

Co-immunoprecipitation is considered to be the gold standard assay for protein-protein interactions. The protein of interest is first isolated using a specific antibody and any binding proteins subsequently identified by western blotting. Although the assay is considered highly reliable, throughput is low so the approach is unsuitable when screening for interaction partners. Pull-down assays are a common variant of co-immunoprecipitation, except that a bait protein is used instead of an antibody, but the higher throughput makes the technique more amenable for screening for interacting partners. **Chemical crosslinking** involves covalently "fixing" interacting proteins in the bound form before trying to isolate and identify the constituents, thereby enabling the determination of near-neighbour relationships and providing some information regarding the distance between interacting molecules. **Protein microarrays** are analogous to DNA microarrays; they use immobilized proteins on a solid glass or membranous surface and use specific antibodies to detect physical binding to provide quantitative information. **Synthetic lethality screens** involve identifying pairs of mutations where each single mutation is individually tolerated but the combination of both mutations renders the cell inviable. Synthetic lethal phenotypes can

be diagnostic of an interaction between the products of the two mutant genes in the cell, but alternatively can be simply an indication of a non-physical, functional interaction. **Synexpression** is the phenomenon by which genes have their expression simultaneously coordinated because their gene products are required in stoichiometric amounts to form subunits of a protein complex. Such patterns can be identified from microarray experiments. However, again, the proteins from genes in a synexpression group are not necessarily physically interacting. **Phage display** uses recombinant methods to create a library of bacteriophages containing peptides embedded in the surface of their protein coats. The target protein of interest is immobilized and phage displaying peptides on their surface that bind to the target can be isolated and easily identified as they carry the gene sequence of the binding peptide. The most widely used methods remain the **yeast two-hybrid system (Y2H)** and **Tandem Affinity Purification (TAP)**. The Y2H method is a rapid, high-throughput *in vivo* screen that identifies the interaction between artificial fusion proteins. The yeast GAL4 transcription factor comprises a DNA binding domain and a transactivation domain. A chimaeric fusion of the "bait" protein with the DNA binding domain is constructed, as is a cDNA library expressing each cloned cDNA in a second chimaeric fusion with the activation domain. If the bait protein interacts with a protein in the library, the interaction of two domains of the GAL4 transcription factor is reconstituted, enabling transcriptional activation of a reporter gene from a GAL4 promoter. However, disadvantages of the method include the fact that it is difficult to apply to extracellular proteins or proteins that initiate transcription, and that construction of the chimaeric fusion protein may alter the protein's structure. Moreover, Y2H has a notorious high false-positive rate which makes confirmatory studies through other methods vital. The accuracy of the high-throughput TAP method, in contrast, approaches that of small-scale experiments (Collins *et al.* (2007)).

The TAP method involves generating a chimaeric fusion of the protein of interest with a TAP tag at the C-terminus. The TAP tag aids protein purification and consists of a recombinant fusion tag formed of two sequentially ordered moieties. The immunoglobulin G (IgG)-binding portion of *Staphylococcus aureus* protein A comprises the distal tag and the calmodulin-binding peptide the proximal tag. The two tags are separated by a tobacco etch virus (TEV) protease

cleavage site. This construct binds to beads coated with IgG, the TAP tag is then broken apart by TEV protease, before the calmodulin-binding peptide part of the TAP tag binds reversibly to calmodulin-coated beads. The protein of interest is then washed through two affinity columns and binding partners identified by either SDS-PAGE or Mass Spectrometry (MS). In MS the proteins are fragmented, typically through electrospray ionization, producing peptide ions in the gas phase, that can then be detected and recorded by means of the mass-to-charge ratios of the peptide ions being assigned to different peaks of the spectrum. The resulting mass fingerprint can be used to search a protein sequence database to identify the protein constituents. However, the TAP tag method requires successive protein purification steps and can therefore be unsuitable for some weak transient interactions. TAP experiments have been performed at genome scale in yeast (Gavin *et al.* (2006); Ho *et al.* (2002); Krogan *et al.* (2006); Uetz *et al.* (2000)). Benchmarking these techniques is difficult as there is no "gold standard" of known interactions and more importantly there is no negative set of proteins known not to interact. Nevertheless, with imperfect benchmark sets, estimates of 30-60% false positives and 40-60% false negatives have been assigned to high-throughput two-hybrid and affinity-purification techniques respectively (von Mering *et al.* (2002)).

Surface Plasmon Resonance (SPR) is a powerful technique that can be used to measure protein-protein interactions in real-time without the use of labels. While one of the interactants is immobilized to the sensor surface, the other is free in solution and passed over the surface. SPR experiments can reveal information regarding the kinetics and dynamics of physical interactions. **Isothermal titration calorimetry (ITC)** is a quantitative biophysical technique that can be used to determine the thermodynamic parameters of protein-protein interactions. It is growing in popularity as it enables the direct measurement of the binding affinity (Ka), changes in enthalpy (ΔH), and binding stoichiometry of the interaction in solution, from which the Gibbs free energy changes (ΔG) and entropy changes (ΔS) can be derived. **Analytical ultracentrifugation (AUC)** is a versatile tool for the study of protein interactions. Monitoring the sedimentation of macromolecular complexes in the centrifugal field allows the characterization of their hydrodynamic and thermodynamic properties in solution. Sedimentation

velocity and sedimentation equilibrium experiments help identify subunit stoichiometry of complexes as well as predicting their equilibrium constants. **Fluorescence resonance energy transfer (FRET)** is a commonly applied *in vivo* assay that exploits the process of energy transfer between two fluorophores, from which the distance between two molecules can be determined. Similarly, **Fluorescence Cross-Correlation spectroscopy (FCCS)** detects the synchronous movement of proteins with two different fluorescent labels. **Alanine scanning** involves systematically mutating each residue in an interaction site and using one of a variety of biophysical methods to establish the contribution of that residue to the energetics of binding. Further details of this method are described in Chapter 6. An extension to this method, **Double Mutant Cycle (DMC)** analysis, where pairs of residues across the interface are systematically mutated, will be described further in Chapter 2.

1.3 Computational methods for studying interactions

An equally broad range of computational methods has been applied to the study of protein-protein interaction (Shoemaker & Panchenko (2007); Valencia & Pazos (2002)). These can be divided into sequence and structure-based methods for identifying interaction partners and those predicting interaction surfaces (starting with the unbound structure).

Phylogenetic profiling uses comparative genomics to identify those pairs of protein domain families whose profile of presence or absence from the genome appears synchronized (Marcotte *et al.* (1999); Pellegrini *et al.* (1999)). The **Gene Neighbourhood** approach uses the fact that interacting prokaryotic proteins are often transcribed from Open Reading Frames (ORFS) on the same operon, such that the likelihood of interaction can be shown to correlate with intergenic distance (Galperin & Koonin (2000)). The **domain fusion** approach, also known as the Rosetta Stone method, relies on finding a precedence for two domain families to be found in a single contiguous gene product, so that in homologous proteins, where the domains are found on separate genes, they can be predicted to interact

(Korbel *et al.* (2004)). Although these approaches benefit from requiring only relatively simple comparative genomic analyses, they predict only a functional association, not necessarily a physical one, such that proteins may simply be part of the same biological process. The **domain interaction propensity** method involves analysis of experimentally determined interaction networks to reveal the likelihood that any two protein domains might interact. The knowledge of these propensities is then used to predict new protein interactions, with the added advantage of directly identifying the putative domain involved in the interaction. Another disadvantage is that these methods are not as effective for eukaryotic species, as fewer genomes are available for comparative studies.

Interacting proteins often co-evolve such that substitutions in one protein may lead to compensatory substitutions in a binding partner. Pairs of residues exhibiting such co-variation can be identified through systematic analysis of multiple sequence alignments (MSAs) of putatively interacting pairs of protein families, identifying pairs of **correlated mutations** (Pazos & Valencia (2002)). This method has the advantage of identifying the precise region involved in interactions. Halperin *et al.* (Halperin *et al.* (2006)) benchmarked the performance of various co-evolution measures in prediction of contacts and found most to hold weak predictive power. Co-evolution can also be reflected through similarity between the phylogenetic trees derived from multiple sequence alignments of non-homologous protein families known to interact. So called "**mirror tree**" methods use this principle by quantifying this similarity by calculating the correlation coefficient between the two distance matrices underpinning the phylogenetic trees, thereby estimating the degree of co-evolution (Goh & Cohen (2002); Pazos & Valencia (2001)). Both the correlated mutations and mirror tree methods require large, high quality alignments of equivalent orthologues, and both assume that the evolution will primarily be driven by correlated changes in the binding epitope, which in general is not the case (Drummond (2005)). Other interaction partner prediction methods include **domain signature analysis**, identifying statistical over-representation of certain domains in interacting proteins (Sprinzak & Margalit (2001)) and **linear motif detection** (Neduva *et al.* (2005)).

1.4 Interaction site prediction

In the absence of a structure of the bound complex, a broad range of *in silico* methods has been developed in order to identify patches on the surface of the protein that may mediate interactions.

Jones and Thornton (Jones & Thornton (1997)) used contributions of six descriptive parameters characterizing surface patches (hydrophobicity, protrusion, planarity, residue interface propensity, solvation potential and solvation accessible surface area) to generate a single combined score which was used to identify putative interaction surface patches. The algorithm has recently been deployed as an interactive web server called **SHARP2** (Murakami & Jones (2006)). **Optimal Docking Area (ODA)** identifies continuous surface patches with desolvation energy based on atomic solvation parameters (Fernandez-Recio *et al.* (2005)). **PPI-PRED** (Bradford & Westhead (2005)) distinguishes interacting from non-interacting surface patches by using interface properties (surface shape, hydrophobicity, conservation, electrostatic potential, residue interface propensity and solvent accessible surface area) as input to a support vector machine (SVM). Hoskins (Hoskins *et al.* (2006)) used terms reflecting solvent accessibility, residue propensity, hydrophobicity and secondary structure data as prediction parameters identifying abnormally exposed amino acid residues in protein interaction sites.

Protein IntErface Recognition (**PIER**) predicts interfaces from a single protein structure using local statistical properties of the molecular surface at the level of atomic groups (Kufareva *et al.* (2007)). **InterProSurf** predicts interacting residues in proteins that are most likely to interact with other proteins based on solvent accessible surface area, a propensity scale for interface residues and a clustering algorithm to identify regions with residues of high interface propensities (Negi *et al.* (2007)). **cons-PPISP** is a consensus neural network method that uses position-specific sequence profiles and solvent accessibility information for each residue and its adjacent neighbours (Tjong *et al.* (2007)). **SPPIDER** is another neural-network prediction method that uses enhanced relative solvent accessibility prediction terms as its input (Porollo & Meller (2007)) **Promate** locates protein-protein binding sites, by using a composite probability derived

from 13 different interface properties (Neuvirth *et al.* (2004)). **PINUP** uses an empirical scoring function combining terms for side-chain energy (a composite function that includes terms to describe atom-contact surface area, overlap volume, hydrogen bonding energy, electrostatic interaction energy, buried hydrophobic solvent accessible surface, buried hydrophilic solvent accessible surface between side-chain rotamers and the fraction of the buried surface of non-hydrogen-bonded hydrophilic atoms), interface propensity and residue conservation (Liang *et al.* (2006)). In a progression analogous to that of CASP (Critical Assessment of Structure Prediction (Kryshtafovych *et al.* (2005))), the diverse range of approaches has inspired the development of various meta-servers pooling the results of other servers e.g. **meta-PPISP** (Qin & Zhou (2007)) (combines cons-PPISP, PINUP, and Promate) and **metappi** (Huang & Schroeder (2008)) (combines PPI-Pred, cons-PPISP, PINUP, Promate and SPPIDER). Zhou and Qin (Zhou & Qin (2007)) attempted to benchmark several of these methods against a set of 35 enzyme structures. Overall, at a coverage of 50%, the ranking and accuracies of six web servers was PPI-Pred (27%) <SPPIDER (33%) <cons-PPISP (36%) <Promate (38%) <PINUP (48%) <meta-PPISP (50%).

The rationale of using evolutionary information is that structural and functional constraints impose selective pressures, such that those regions of the protein surface that are engaged in interactions often evolve at a slower pace than elsewhere. Prediction methods based on evolutionary conservation carry the advantage that they are of completely general application. The **Rate4Site** algorithm estimates the rate of evolution of amino acid sites through maximum likelihood (Pupko *et al.* (2002)). Phylogenetic trees are derived from multiple sequence alignments to reflect evolutionary relationships amongst homologous proteins to identify functional interfaces. The **ConSurf** web server projects the conservation scores from Rate4Site onto the molecular surface to reveal patches of highly conserved residues (Landau *et al.* (2005)). Similarly the **Evolutionary trace** (ET) method involves the phylogenetic partitioning of specificity-determining residues across an MSA and clustering the results onto the surface of the solved structure (Yao *et al.* (2003)). ET differs from Rate4Site in that the focus is on the identification of class-specific residues. **PatchFinder** works by assigning conservation scores to each residue position on the protein surface and then generating a

score for each putative non-overlapping patch (Nimrod *et al.* (2005)). **WHISCY** uses an multiple sequence alignment to establish the sequence distance between the structure and its homologues to give the expected degree of mutation. Any residues whose observed mutation is less than the expected is given a positive WHISCY score. These scores are then combined with residue propensities and any resulting neighbouring residue predictions are smoothed to give the final surface patch predictions (de Vries *et al.* (2006)).

The TLB group has developed the application of environment-specific substitution tables (ESSTs) (Overington *et al.* (1992)) in the prediction of interaction sites. The ESSTs are based on the rates of observed substitutions in a library of 64 different structural environments and are described in greater detail in Chapter 4. The structural environments are defined on the basis of secondary structure, solvent accessibility and hydrogen bonding and are derived from a set of high quality structural alignments (Stebbings (2004)). The program *Crescendo* uses this information to identify evolutionary restraints on protein sequence and structure (Chelliah *et al.* (2004)). This is achieved by comparing, for each amino acid position, the sequence conservation *observed* across a multiple sequence alignment of the homologous family of proteins with the degree of conservation *expected* on the basis of amino acid type and local structural environment. This identifies those residues that have a higher degree of conservation than expected *for a given environment* and are therefore likely to be involved in interactions. The resulting residue scores are mapped onto the surface of the protein structure and contoured to identify clusters of residues contributing to a functional site. Importantly this method differs from other published techniques by successfully distinguishing those restraints that arise from maintaining structure from those that mediate intermolecular interactions.

1.5 Prediction of protein complexes

The obvious importance of protein-protein interactions coupled with the shortfall of available structural data means that fast, reliable, *in silico* methods for prediction of the structure of protein complexes are highly desirable. Such methods begin with some structural representation of each of the constituent proteins

(either experimentally solved structures or comparative models) and attempt to produce an accurate 3D model of the complete complex. Comparative modelling of individual proteins is used to extend the coverage of structural representation of protein sequences. Analogously, comparative modelling of protein interactions can be used to extend coverage of structural representation of interactions. Indeed where available such information can be invaluable in predicting the structure of the complex (Korkin *et al.* (2006)). However, care must be taken in such an approach. Only a subset of homologues of an interacting pair will themselves interact i.e. form "interologs" (Yu *et al.* (2004)), for reasons of opportunity (differing sub-cellular localization, expression profile) and capacity (interface regions have incorporated such residue substitutions as to render them incompatible). In practical terms, there are therefore two key stages to this problem. Firstly, identification of a suitable homologous complex and secondly, given such a homologous complex, assessment of whether the mode of interaction is preserved. If the mode of interaction is preserved, the problem then becomes one of comparative modeling with suitable restraints (Sali & Blundell (1993)). Equally, if there is no homologous complex, or there is one but the mode of interaction appears not to be preserved, and nonetheless the proteins are known to interact, then protein-protein docking approaches, described below, are appropriate. Such a scheme is described in Figure 1.1.

Russell and Aloy (Aloy *et al.* (2003)) have suggested that when sequence similarity is above 25-30% proteins are highly likely to interact in the same way. However, Park *et al.* (Park *et al.* (2004)) highlighted the interesting exception of the interaction between the P2 domain from histidine kinase CheA and its phosphorylation target CheY from *Thermotoga maritima* and the equivalent pair of proteins from *Escherichia coli*, where, despite the binding sites comprizing the same residues, the orientation of the CheA P2 domains differs by a $\sim 90^{\circ}$ rotation (Figure 1.2). All of these approaches assumes *a priori* that the components are known to interact. Such homology-based approaches can be extended to identify novel interactions or those proteins in a candidate set that are most likely to interact compatibly.

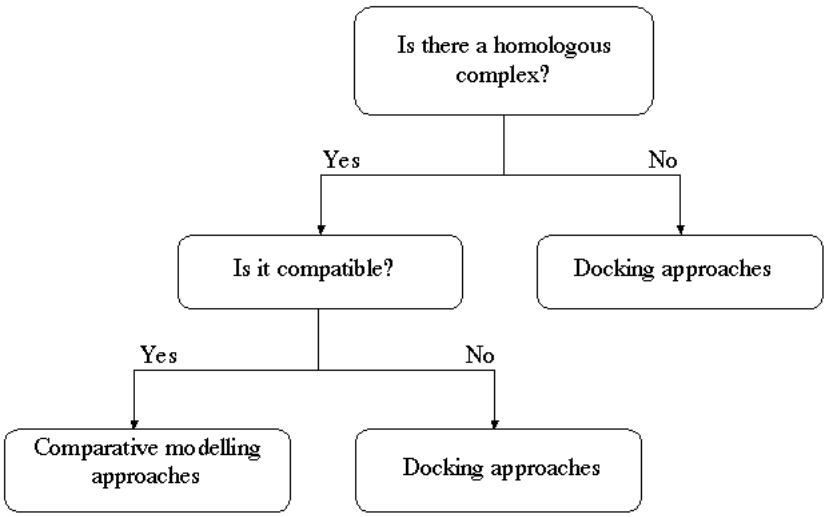


Figure 1.1: Alternative approaches to predicting protein complexes.

1.6 Comparative modelling of interfaces

The same comparative modeling approaches used to model monomeric proteins (Sali & Blundell (1993)) can be applied to interactions (Fukuhara & Kawabata (2008)). However, in a landmark paper Alber *et al.* (Alber *et al.* (2007)) extended the approach to produce a detailed architectural map of a large multiprotein assembly, the nuclear pore complex (NPC). They translated the results of a combination of experimental methods, including ultracentrifugation, affinity purification and electron microscopy, into spatial restraints which were then optimized to generate an ensemble of structures consistent with the data. A number of approaches related to comparative modelling have also been developed. **INTERPRETS** (INTeraction PREdiction by Tertiary Structures)(Aloy & Russell (2003)) uses empirical pair potentials derived from a molar-fraction random state model based on the observed tendency of residues to interact across interfaces, to assess how well a homologous pair of sequences fit into a complex structure. The method was validated on the fibroblast growth factor receptor (FGFR) system and on yeast two-hybrid data. **MULTIPROSPECTOR** (Lu *et al.* (2002)) involves a multimeric threading approach. The algorithm first performs traditional

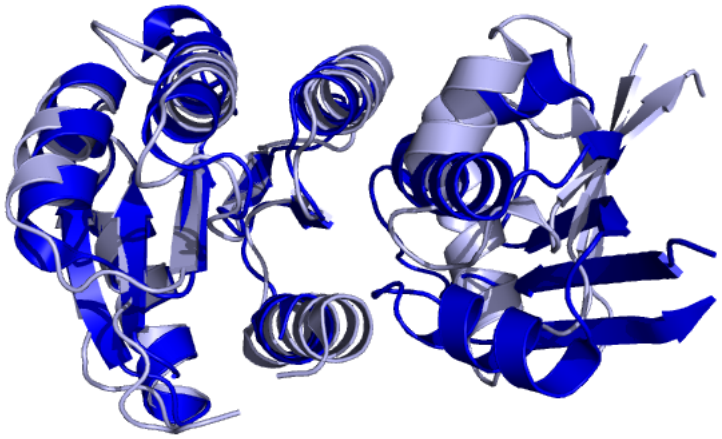


Figure 1.2: Interaction between the P2 domain from histidine kinase CheA and its phosphorylation target CheY from *Thermotoga maritima* in dark blue (PDB entry 1u0s) and the equivalent pair of proteins from *Escherichia coli* in light blue (PDB entry 1ffg).

monomeric threading to generate a set of potential structures for each partner query sequence. An empirical protein-protein interfacial energy score, derived from a non-redundant, high-quality dimer database (Lu *et al.* (2003a)), is used in combination with the threading Z-score to identify true multimers. This has also been applied at large scale to the genome of *Saccharomyces cerevisiae* (Lu *et al.* (2003b)). The **PRISM** algorithm (Protein Interactions by Structural Matching) predicts novel protein-protein interactions by combining structure similarity and evolutionary conservation in interfaces. Surface regions of target proteins are structurally aligned to a set of template interfaces and any matches of sufficient similarity are predicted to interact through these equivalent regions.

1.7 Protein-protein docking

Protein-protein docking involves the prediction of the 3D structure of a protein complex from the individual structures of its components, which are known *a priori* to interact. In essence this is much like solving a 3D jigsaw puzzle, though

the problem is made more difficult as each piece of the jigsaw can often rearrange itself on binding. These rearrangements can be small changes in residue side-chain conformations, local backbone movements or large conformational changes involving entire secondary structure units or even complete domains. Such mobility represents one of the largest hurdles in protein-protein docking.

The field is now well established with more than 20 algorithms published (Andrusier *et al.* (2008); Halperin *et al.* (2002); Vajda & Camacho (2004); Vakser & Kundrotas (2008)). Despite the large number of published algorithms, many of them follow the same overall two-stage scheme. A brute force rigid-body search of a mobile ligand around a fixed receptor generates a large set of configurations, which, ideally, includes at least one nearly correct pose. This is followed by a scoring phase, where each of the vast number of generated poses are scored and ranked, preferably distinguishing the correct configurations from the others. The brute force search involves sampling six dimensions (three rotational and three translational), the combinatorics of which can potentially generate billions of poses, the actual number depending on the granularity of sampling. This vast search space is computationally expensive, so many of the algorithms first reduce the receptor and ligand molecules from all-atom models to pseudo-atoms or more commonly a simplified cubic lattice model. A major breakthrough came with the use of Fast Fourier Transform techniques (Eisenstein & Katchalski-Katzir (2004); Katchalski-Katzir *et al.* (1992)), which have provided a 10^7 improvement in performance. Acting on the cubic lattice model, for the subset of scoring functions that are discrete convolutions, configurations related to each other by translation of one protein by an exact lattice vector can all be scored almost simultaneously by applying convolution theorem.

An ideal scoring scheme would reliably distinguish nearly native configurations from the rest. Most scoring schemes are based on molecular mechanics type functions, variously including terms for van der Waals contacts (often with explicit complementarity terms), electrostatics, hydrogen-bonding and desolvation (Camacho *et al.* (1999); Gabb *et al.* (1997); Halperin *et al.* (2002)). The importance of, and weighting given to, each of these components varies significantly depending on the system under investigation. Other groups have attempted this using more esoteric parameters (Gottschalk *et al.* (2004); Murphy *et al.* (2003);

Tress *et al.* (2005)). Typically the poses generated from complex generation do include the native conformation, though current scoring functions cannot necessarily distinguish it from the others. As such, scoring functions remain an area of intense study.

When substantial conformational change occurs at the time of complex formation, rigid-body docking is inadequate. A variety of approaches has been used to try to deal with this flexibility. In so-called "soft" docking the van der Waals term is modified to permit some local plasticity (Fernandez-Recio *et al.* (2002); Palma *et al.* (2000)). In flexible docking, bond angles, bond lengths and torsion angles of the components are modified during complex generation. However, scoring all possible conformational changes is prohibitively computationally expensive. Flexible docking procedures must therefore intelligently select a small subset of possible conformational changes for consideration. This can be achieved either by using precedence-based rotameric libraries or molecular dynamics approaches, either at the stage of complex generation or scoring (Fernandez-Recio *et al.* (2003); Schueler-Furman *et al.* (2005); Smith *et al.* (2005)).

1.7.1 Guided docking

Recently some notable success has been achieved by including information from external sources in so-called data-driven or guided docking (de Vries (2006); van Dijk *et al.* (2005a,b)). Examples of the kinds of information that can be used are shown in Table 1.1. They vary in both quality and the level of resolution they provide. The information they provide can be used either in the complex generation phase or the scoring phase, or indeed both, either in the form of constraints or restraints.

In the TLB group the results of running *Crescendo* to predict interaction sites are used as a restraint in the guided docking program PyDock (Chelliah *et al.* (2006)). This is achieved through the PyDockRST module of PyDock. Essentially an interaction restraint by a given residue is satisfied if any of its atoms is less than 6Åfrom any atom of the partner molecule. For each docking solution, the method computes the percentage of satisfied restraints with respect

Technique	Resolution	Issues
Mutagenesis with binding assay	Residue	Loss of structure
Cross-linking with Mass Spectrometry (MS)	Distance information	Attachment & detection
NMR H/D exchange	Residue	Indirect effects
NMR chemical shift perturbation	Atomic	Indirect effects
Cryo EM	Orientation	Low resolution
Small angle X-ray scattering	Shape	Low resolution
In silico conservation methods	Residue	Conservation from other determinants

Table 1.1: Experimental sources of restraints used in data-driven protein-protein docking.

to the total number of possible restraints, and this number is converted to energy by the equation:

$$restraint\ energy = -1.0\ kcal/mol \times \%\ satisfied\ restraints \tag{1.1}$$

1.7.2 The CAPRI experiment

Progress in protein-protein docking is assessed objectively in the form of the CAPRI experiment (Janin (2005)). CAPRI (Critical Assessment of Protein Interactions) is similar in spirit to the CASP (Critical Assessment of Structure Prediction (Kryshtafovych *et al.* (2005))) in that it is a blind assessment; the coordinates of solved complexes are held privately by the assessors, with the cooperation of the structural biologists who determined them. As such, it provides a uniform benchmark set for objective comparison of the performance of the different prediction algorithms. Three quality measures are used to evaluate success. The first is the fraction of native contacts, defined as the number of correct residue-residue contacts in the predicted complex divided by the number of con-

tacts in the target complex. A pair of residues on different sides of the interface is considered to be in contact if any of their atoms are within 10Å. The second measure is the ligand backbone root-mean-square deviation (RMSD) of atomic positions after superimposition of the receptor. The final measure is the binding site RMSD, defined as the ligand RMSD calculated only for those ligand residues in contact with the receptor (Vajda & Camacho (2004)). The CAPRI experiment attracts a high level of participation (37 groups participated worldwide in round seven) and provides a valuable way of monitoring progress and stimulating discourse. However, the results are of little statistical significance as the number of targets in each round is small. Furthermore, the range of targets has historically been somewhat non-representative, with antibody-antigen and enzyme-inhibitor systems dominating. It can be argued these examples are somewhat atypical from an evolutionary point of view: antibodies have evolved to recognized antigens; enzymes have evolved to recognize or bind substrates, albeit in a transition state.

1.8 Different properties of protein interactions

The composition and anatomy of protein-protein interfaces has been intensively studied through several statistical analyses of the properties of protein-protein interfaces (Ansari & Helms (2005); Caffrey *et al.* (2004); Gruber *et al.* (2006); Janin & Chothia (1990); Jones & Thornton (1996); Lo Conte *et al.* (1999); Ofran & Rost (2003); Reš & Lichtarge (2005); Yan *et al.* (2008)). Some of the properties investigated include interface size, shape, planarity, complementarity, residue propensity, segmentation, evolutionary conservation, residue pairing preferences, polarity, hydrophobicity, gap volume, secondary structure preferences and many different bonding types. The conclusions from these studies can be contradictory, typically due to the use of different data sets and assumptions, but some common themes have emerged. The area buried per subunit generally ranges from $\sim$350Å^2 to $\sim$4500Å^2. Interactions typically occur between segmented, discontinuous patches either between the same or different secondary structure types. The interface region is generally more hydrophobic than the remainder of the protein surface but less hydrophobic than the protein core. This is related to the

observed enrichment of hydrophobic residues in the interface, though arginine, histidine and tyrosine have also been observed to be over-represented.

1.9 Different types of protein interactions

Protein interactions are highly heterogeneous and their properties reflect this. Many studies have been performed partitioning structurally observed interfaces by a variety of criteria (Ansari & Helms (2005); De *et al.* (2005); Jones & Thornton (1996); Mintseris & Weng (2005); Nooren & Thornton (2003); Reš & Lichtarge (2005)). Protein-protein interactions between identical chains are termed homo-oligomers; those between non-identical chains hetero-oligomers. A further distinction can be made between homologous hetero-oligomers, where the constituent chains share a common evolutionary origin (e.g. α_2 β_2 haemoglobin), and unrelated hetero-oligomers (e.g. fibroblast growth factor and its receptor). Each monomer constituting a homo-oligomer (or indeed a homologous hetero-oligomers) related by a two-fold structural symmetry axis contributing equivalent surfaces are described as interacting in an isologous manner (e.g. nerve growth factor and its receptor). In heterologous interactions the constituents contribute distinct interaction surfaces, such that, in the absence of cyclic symmetry, they can aggregate indefinitely (e.g. Tobacco Mosaic Virus coat protein).

A further significant partitioning of interfaces is between obligate interactions, where the constituent proteins are unstable in the unbound form and are not observed independently *in vivo*, and non-obligate interactions. Similarly interactions can be further partitioned with respect to the lifetime of the complex. While obligate interactions are almost invariably permanent, non-obligate interactions exhibit a wide range of longevities: from weak transient interactions equilibrating dynamically in solution with dissociation constants in the mM to μM range (e.g electron transport); through interactions in the intermediate range with dissociation constants in the μM to nM range (e.g signal transduction); to strong or even permanent interactions with dissociation constants in the nM or even fM range (e.g protease-inhibitor systems) which may require an external effector to trigger dissociation (affinity ranges taken from Nooren and Thornton (Nooren & Thornton (2003))). In reality a continuum exists between all of these differing

interaction types and the stability of all complexes very much depends on the subcellular location, concentration and local physico-chemical environment.

1.10 (Non-structural) protein interaction databases

There have been a number of efforts to organize the results of published high-throughput protein interaction studies (Rohl *et al.* (2006)). These will be briefly reviewed here (data resources with a focus on 3D structure of interactions are discussed in Chapter 2). The **Database of Interacting Proteins (DIP)** contains high quality, manually-verified data regarding experimentally determined protein-protein interactions. Interaction data are initially obtained from a number of sources; PDB complexes, literature; high-throughput methods Y2H protein microarrays screens; TAP-MS analysis of protein complexes; and organism-specific and pathway databases. As of August 2008, the relational database holds 19,935 proteins from 204 organisms totalling 56,638 interactions (Salwinski (2004)). The popular **Biomolecular Interaction Network Database (BIND)** is no longer publicly available due to lack of funding (Bader *et al.* (2003)). **IntAct** is a freely-available open source molecular interaction database and software suite from the EBI (http://www.ebi.ac.uk/intact). The data come entirely from published literature and are manually annotated by expert biologists to include details of experimental methods, conditions and interacting domains. As of August 2008, IntAct contains 63,824 proteins and 111,834 interactions and complexes. The **Molecular Interaction Database (MINT)** focuses primarily on protein-protein interactions from mammalian genomes that have been experimentally verified. Interactions are initially derived from the scientific literature using text mining software and then reviewed by expert curators and includes both direct physical interactions and indirect functional relationships. MINT contains 28,817 proteins and 105,899 interactions (Chatr-Aryamontri *et al.* (2007)). **STRING** is a database of known and predicted direct physical and indirect functional protein-protein interactions derived from literature, genomic context information, high-throughput experiments and conserved co-expression for a large number of organisms, and transfers information between these organisms wherever applicable. The database currently contains 1,513,782 proteins from 373 species (von

Mering *et al.* (2007)). **MPI-LIT** is a literature-curated dataset focusing on microbial binary protein-protein interactions with associated experimental evidence manually curated from 813 papers, comprizing 746 non-redundant interactions of which 88% are not reported in public databases (Rajagopala *et al.* (2008)). DIP, BIND, IntAct, MINT and MPI-LIT are all members of the IMEx consortium (Orchard *et al.* (2007)), an international group of data resources to facilitate exchange of data and avoidance of duplication.

1.11 Aims of this thesis

Chapter 2 of this thesis deals with the process of establishing PICCOLO - a comprehensive database of structurally characterized protein interactions. The name PICCOLO, while following the TLB group's tradition of musical names for software and tools, is also an approximate acronym of Protein Interaction Collection. Issues of interface definition, quaternary structure, data redundancy, local structural environment and different interaction types are addressed. Chapter 3 goes on to describe a variety of pursuant analyses of the properties of protein-protein interfaces enabled by the data stored in PICCOLO, including residue propensity, hydropathy, polarity, interface size, sequence entropy and residue contact preference.

Chapter 4 addresses the question of what patterns of substitutions are accepted in protein interfaces across evolution, and whether these patterns are distinguishable from those seen in other structural environments. A pre-requisite for answering such questions is a high-quality set of multiple structural alignments. The derivation of such a set, in the form of the database TOCCATA is discussed, along with ancillary applications of this information, before procedures to derive environment-specific substitution tables are described.

The TLB group has been working for some years on methods to predict the likely effect of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability, function and interactions. The rationale behind this is to triage the large volumes of data created from high-throughput genetic screening studies, enabling clinicians to prioritize those nsSNPs that are most likely deleterious to

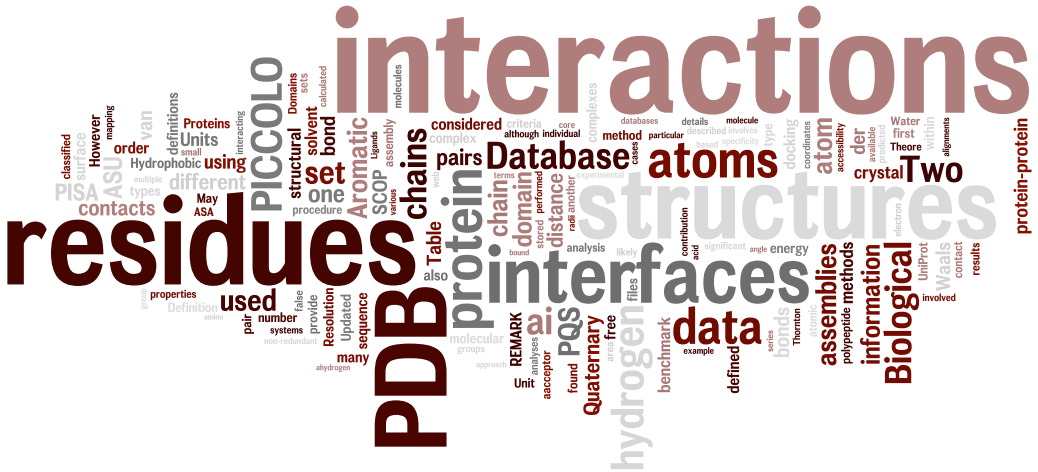
a protein and thereby be phenotypically detrimental. One aspect of this work involves running the program *Crescendo* (Chelliah *et al.* (2004)) in order to predict protein interaction sites. However, the method offers the advantage of being generally applicable and so provides high coverage. PICCOLO offers an opportunity to supplement these predictions with the comprehensive set of observed protein interaction sites. This has the advantage of providing a higher quality data set, though coverage will be lower as the data only reflects that which is available in the PDB. These ideas will be explored further in Chapter 5 which includes the results of a recently published benchmark study where PICCOLO has been used to assess the impact of nsSNPs on protein interactions.

From a therapeutic viewpoint, historically there has been little focus on protein-protein interactions as small-molecule drug targets; they are perceived as being planar and hydrophobic in nature (Wells & Mcclendon (2007)), properties that would tend to make them somewhat intractable. Traditionally, the pharmaceutical industry has been led by functional assays, and target-based approaches have been focused on receptors, channels and enzymes. However, the discovery from alanine-scanning mutagenesis studies of so-called "hot-spots" - small subset of residues that contribute the greater part of the free energy of binding - offers more opportunity for small molecule therapeutics (Bogan & Thorn (1998); DeLano (2002); Keskin *et al.* (2005a)). There is a small but growing number of protein interaction-based drug discovery programmes in progress. Examples include: interleukin-2 (IL-2) and the α-chain of its receptor (IL-2Rα); B7 and CD28; B-cell lymphoma-2 (BCL-2) and BAK (Bcl-2-antagonist/killer); Lymphocyte function-associated antigen-1 (LFA-1) and intercellular adhesion molecule-1 (ICAM-1); inducible Nitric Oxide Synthase (iNOS) dimerization and Nerve Growth Factor (NGF) and its receptor; cytokine tumour-necrosis factor (TNF) its receptors, TNFR1 and TNFR2; FtsZ and ZipA; Human papilloma virus (HPV) E1 and E2; and human protein double minute 2 (HDM2) and p53; (Arkin & Wells (2004); Wells & Mcclendon (2007)). Molecular characterization of the properties of hot-spots can be derived from the data in PICCOLO. This can not only lead to deeper understanding of the molecular mechanisms underlying this important phenomenon but could also lead to novel predictive methods to identify hot-spots

in silico. Success in this area could assist in druggability assessment for target prioritization as well as contributing to lead compound identification and selection. These ideas are discussed in Chapter 6.

Chapter 2

Building PICCOLO



2.1 Introduction

In order to gain a full understanding of the principles that underlie protein-protein interactions it is vital to have access to high quality data on the molecular details of these interactions. Access to comprehensive data enables both detailed analyses of individual systems and broad systematic comparative analyses. Such data enable the pursuit of a number of fundamental questions surrounding the nature of molecular interactions: What properties distinguish interfaces from the rest of the protein surface? What properties differ between the various types of interfaces? What properties vary between the various sub-regions comprising the anatomy of an interface? What evolutionary constraints apply to interfaces? Can we harness these properties in a predictive manner - to identify interaction sites, to predict novel interactions or to predict the structure of a complex? What are the features of interfaces that can be useful in identifying the cases that are most likely to be amenable to modulation by drug-like small molecule ligands? I show that PICCOLO can contribute towards answering each of these questions.

Recent years have seen a surge of interest in protein-protein interactions and a number of databases characterizing the 3D structures of interactions have been published. These are reviewed in Table 2.1.

Database	Input Source	Interface Definition	Comments	Resolution	Scale	Availability
3DID 3D Interaction Domains (Stein (2004))	PDB ASU, Pfam	hydrogen bond: $d(a_i, a_j)$ ≤3.5Å where $E(a_i) = N,\ E(a_j) = O$ salt bridge: $d(a_i, a_j)$ ≤5.5Å where $E(a_i) = N,\ E(a_j) = O$ van der Waals: $d(a_i, a_j)$ ≤5 Å where $E(a_i) = C,\ E(a_j) = C$ ≥5 interactions	Intra- and inter-molecular interactions between Pfam domains from high-resolution crystal structures. Includes GO-based functional annotations and predicted similarity of interactions across families. Basis for InterPrets(Aloy & Russell (2003)).	domain	5,000 unique Pfam domain pairs. 119,042 chain interfaces	Updated weekly. Web interface and database download.
DAPID Domain Annotated Protein-protein Interaction Database (Chen *et al.* (2006))	PDB ASU, Pfam	$d(a_i, a_j)$ ≤8Å where a_i and a_j are Cα (or GLY Cβ) ≥5 residues	Domain-annotated protein interactions from PDB. Basis for prediction of novel interactions through 3D-domain interologs.	domain & residue	1,008 observed interacting domain pairs 101,511 predicted interactions	Last update 2006. Web interface.
DOCKGROUND (Douguet *et al.* (2006); Gao *et al.* (2007))	PDB Biological Units	Mean ASA ≥ 250Å^2	Focuses on dynamic generation of non-redundant bound-bound datasets for docking. Structure and sequence-based redundancy filtering.	chain	102,527 complexes 24,596 PDB entries 86,856 Biounit chains	Updated monthly. Web interface for generating user-specified non-redundant sets.
ICBS (Dou *et al.* (2004))	PDB, PQS	Continuous interchain β-sheet	Specialist database of interactions mediated by interchain β-sheet formation.	residue	6,599 matches from 31,483 scanned structures	Last update 2006. Web interface.

InterPare (Gong *et al.* (2005a))	PDB ASU, SCOP	3 methods : Euclidean distance (PsiBase) Accessible surface area Voronoi-based definition	Derived from PsiBase. Focuses on interface details.	domain	31,620 inter-chain interfaces. 12,758 intra-chain interfaces	Last update 2004. Web interface and database download.
PIBASE (Davis & Sali (2005))	PDB, PQS, CATH, SCOP	$d(a_i, a_j)$ ≤6.05Å Interface surface area ≥300Å^2	Polar and non-polar surface area calculated.	domain	105,061 SCOP domains 158,915 SCOP domain pairs	Last update 2005. Web interface and database download.
Protein3D & PRINT (Keskin *et al.* (2004); Ogmen *et al.* (2005))	PDB ASU	$d(a_i, a_j) \leq vdw(a_i) + vdw(a_i)$+0.5Å ≥10 contacting residues. "Nearby" residues defined as Cα within 6Å of Cα of interacting residue	Interfaces clustered using sequence-order independent method. Used to predict novel interactions through surface structural similarity (Ogmen *et al.* (2005)).	chain	21,704 interfaces	Last update 2002. Server currently offline for update.
PROTCOM Database of PROTein COMplexes (Kundrotas & Alexov (2007))	PDB ASU	$d(a_i, a_j) \leq vdw(a_i) + vdw(a_i)$+2.8Å	Intermolecular interactions from PDB chains as well as intra-chain domain-domain interfaces.	chain & residue	1,350 heterodimers 7,773 homodimers 1,589 intra-chain	Last update 2007. Web interface and data download.
ProtBud Protein Biological Unit Database (Xu *et al.* (2006))	PDB ASU, Biological Units, PQS	$d(a_i, a_j)$ ≤6Å	Focuses on comparison of PDB ASU, Biological Units & PQS complexes in framework of SCOP families.	chain & residue	64,732 PDB entries	Updated weekly. Web interface and database download.
PsiBase (Gong *et al.* (2005b))	PDB ASU, SCOP	$d(a_i, a_j)$ ≤5Å≥5 residue pairs (known as 5-5 rule)	Focuses on network of family and superfamily relationships	domain	60,532 domain pairs	Last update 2005.

SCOPPI Structural Classification of Protein-Protein Interfaces (Winter (2006))	PQS, SCOP	$d(a_i, a_j)$ ≤5ÅInteracting residues have ≥5 such contacts. Interacting domains have ≥5 residue pairs	Interfaces classified by geometry of domain pairs. Includes multiple sequence alignments and Gene Ontology (GO) terms.	domain & residue	102,084 domain pairs 3,358 family pairs	Last update 2005. Web interface.
SCOWLP Structural Characterization Of Water, Ligands and Proteins (Teyra *et al.* (2006, 2008))	PDB ASU, SCOP	$d(a_i, a_j)$ ≤9Åallowing for up to two bridging water molecules	Explicitly deals with small protein ligands & solvent	domain & residue	60,664 structural units 4,907 interfaces 2,093,976 residue pairs	Last update October 2006. Web interface and database download.
SNAPPI-DB Database of Structures, iNterfaces & Alignments of Protein-Protein Interactions (Jefferson *et al.* (2007))	PQS, SCOP, CATH, Pfam	$d(a_i, a_j) \leq vdw(a_i) + vdw(a_i)$+0.5Å ≥ 10 interacting residue pairs between domains	Includes Pfam, SWISSPROT, InterPro, GO terms, secondary structures and multiple alignments. Domain-domain pairs clustered by family and superfamily pairs.	domain		Updated every 6 months. Database download and query API.

Table 2.1: $d(a_i, a_j)$ is the distance between atoms a_i and a_j on opposing interface surfaces. $vdw(a_i)$ is the van der Waals radius and $E(a_i)$ is the element type of atom a_i.

At the time of writing several previously-published resources appeared to be no longer available - iPFAM(Finn *et al.* (2005)), BID (Fischer *et al.* (2003)), Binding Motif Pairs (Li & Li (2005)), PINT (Kumar (2006)) and SPIN-PP .

2.1.1 Docking benchmarks

In order to benchmark docking algorithms objectively, it is necessary to have structures of the constituent proteins in the bound and the unbound forms, so that the result of the docking algorithm can be compared to that of the known complex. Beyond the resources listed above, other resources have been developed whose focus is primarily on generating test sets for protein docking. The popular Protein Docking Benchmark (Chen *et al.* (2003); Hwang *et al.* (2008); Mintseris *et al.* (2005)), now on version 3.0, began as a hand-picked set of complexes, but now is generated through a semi-automated process. Docking test cases are grouped as Enzyme/Inhibitor, Antibody/Antigen and Others and are further classified as rigid-body (88 cases), medium (19 cases) or difficult (17 cases). Automatic Generated Test-Sets Database for Protein-Protein Docking (AGT-SDP) (Zollner *et al.* (2005)) and Unbound-Unbound Protein-Protein Docking Dataset (UUPPDD) are two fully automated methods for identifying docking test sets.

2.1.2 Web servers

Aside from these databases, several interactive web servers are available to calculate properties of protein-protein interfaces (Laskowski (2009); Saha *et al.* (2006); Tina *et al.* (2007)). However, there is a fundamental and important distinction to be made between tools that provide the capacity to analyse a particular interface and those that systematically perform such analyses on the complete set of available structures and make the results available. The Protein-Protein Interaction Server (Jones & Thornton (1996)) has now been updated as Protorp which characterizes interfaces in terms of size, shape, secondary structure, hydrogen bonds, salt bridges and gap volume (Reynolds *et al.* (2008)).

2.1.3 PICCOLO criteria

The first stage in the development of PICCOLO was to identify a set of criteria that any new resource should fulfill in order to maximize functionality. Previously published databases were then reviewed in light of these criteria. Four criteria were identified. First, despite the rapid growth of the PDB there is a paucity of coverage of protein-protein complexes (Russell *et al.* (2004)). Therefore to maximize the potential of what information is available any such resource should be fully comprehensive, and have the capacity to be updated periodically to reflect future PDB depositions. Second, the interface definition would have to be robust and accurate. In order to investigate the evolutionary plasticity of interfaces, quantitative information on accepted residue substitutions would be required. As such the level of resolution required would be at least that of individual residues. However, work in the TLB group has established that residue substitutions depend on the details of molecular interactions in which the residue partakes (Overington *et al.* (1992)). Therefore it would be necessary to capture data at the highest possible resolution for interactions - therefore the third criterion would be to have interaction data stored at atomic resolution. The PDB is inherently redundant, with the same protein sequence often solved multiple times under different experimental conditions, with different ligands, in different conformations and so forth, and the same is true of protein-protein complexes. Over-representation of particular subsets can skew subsequent analyses. Therefore the fourth criterion would be for the resource to have the capacity to appropriately define a non-redundant set.

The various resources differ significantly with respect to their definitions, scope, coverage, resolution, quality, clarity, availability and frequency of updates. No published resource fitted the requirements exactly. For this reason, as well as for more pragmatic considerations such as control of data quality and updates, the decision was taken to develop PICCOLO locally - as a comprehensive database of the molecular details of structurally characterized protein-protein interactions.

Subsequent to the initial development of PICCOLO, parallel sister databases dealing with protein interactions with other classes of molecule were devised.

Semin Lee has developed BIPA, concerning the interactions of proteins with nucleic acids and Adrian Schreyer has developed CREDO, concerning the interactions of proteins with small-molecule heteratomic ligands. TIMBAL, developed by Alicia Higueruelo, is a hand-curated database comprising small molecule ligands known to disrupt protein-protein interactions published in the literature. TIMBAL comprises 105 small molecules, from 21 protein-protein interaction systems, 13 of which have some structural representation and can be cross-referenced to PICCOLO enabling insights into the type of molecular interactions favoured by inhibitors of protein-protein interactions. Close collaboration has ensured that these databases are compatible with one another: they were designed using largely the same interaction definitions and they share the same PDB residue identifiers, thereby enabling useful comparative cross-queries.

2.1.4 A menagerie of interactions

Protein folding, assembly and interactions are largely governed by the non-covalent interactions between residue side chains. In order to capture computationally the molecular details of such interactions, for the purposes of precise interface definition, it is first necessary to understand their physico-chemical origins. These will be reviewed briefly here for a series of such non-covalent interactions, broadly in order of their significance to protein-protein interactions.

2.1.4.1 van der Waals

Van der Waals interactions are a consequence of quantum dynamics inducing fluctuating polarizations in the electron cloud of nearby particles. They are composed of a short-range repulsive component, due to steric hindrance when neighbouring atoms have overlapping electron clouds, and a longer-range attractive term, due to the coupling of dipoles in the electron cloud of neighbouring atoms, known as London dispersion forces. The two components are usually combined and, although many different functional forms have been suggested, they are most commonly described by the so-called Lennard-Jones potential shown in Equation 2.1 and Figure 2.1:

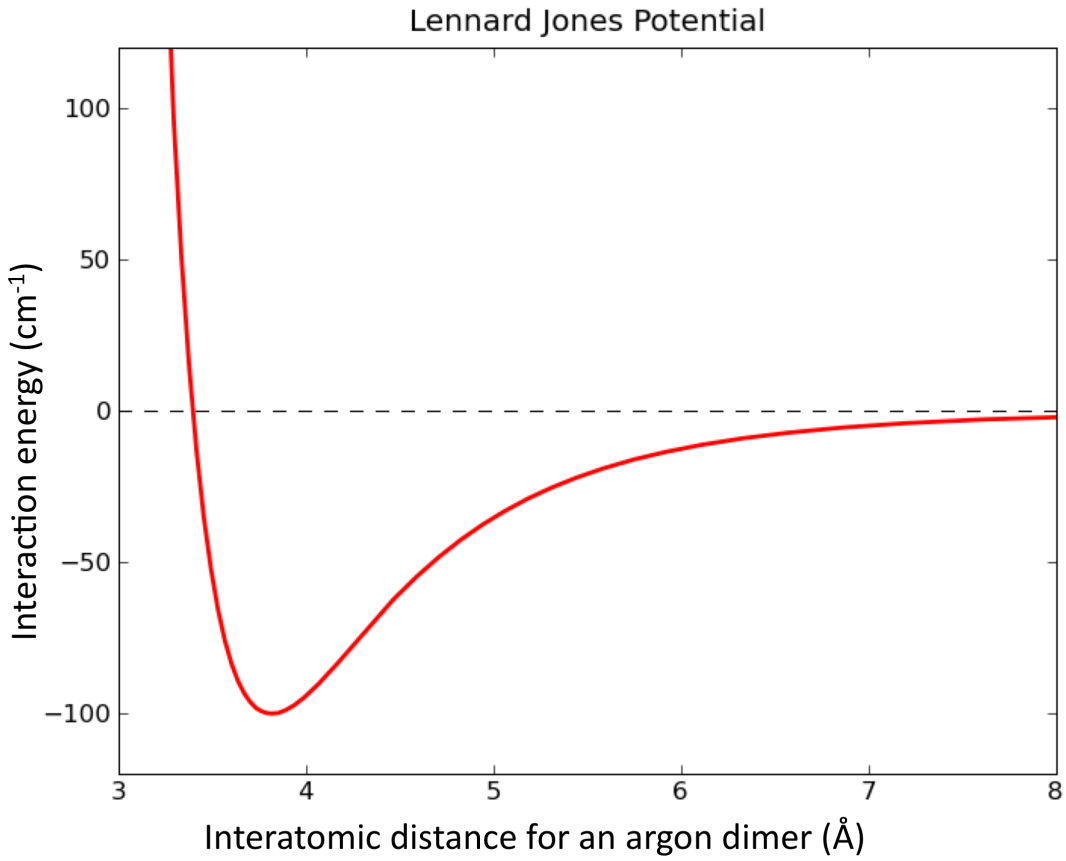


Figure 2.1: Lennard Jones potential for an Argon dimer. Short range repulsions are mediated by the r^{-12} component whereas London dispersion-attraction forces are accounted for by the r^{-6} term.

$$V(r) = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^{6}] \tag{2.1}$$

where r is the interatomic distance, ϵ is the depth of the potential well and σ is the distance at which the potential between the atoms is zero.

Figure 2.1 indicates a weak attraction at large distances and strong repulsion at very close distance. Van der Waals interactions are typically around 2.8Å to 4.0Å in length. The difference between sum of the van der Waals radii of the two atoms and the point of lowest energy is of the order of 0.3Å to 0.5Å . Relative to other forces governing conformation, the binding energies of van der

Waals interactions are individually very low (typically < 1 kcal/mol), but the large number of such interactions makes them significant for protein folding and binding.

2.1.4.2 Hydrogen bonds

A hydrogen bond is an attractive interaction between two electronegative atoms competing for the same hydrogen atom. A hydrogen atom is formally covalently bound to the donor and aligned between the donor and acceptor atoms. The strength of the hydrogen bond varies broadly depending on various factors including the linearity of the interaction, but typically ranges from 2-10 kcal/mol.

2.1.4.3 Hydrophobic interactions

The increase in entropy gained by removing surfaces of hydrophobic side chains from ordered solvent is amongst the most significant factors for protein folding. A convenient handle for characterizing this phenomenon is to consider the effect in terms of favourable interactions between hydrophobic side chains. However, it should be borne in mind that it is not the interactions between these side chains that is the source of the favourable interactions, rather the entropic gain from exclusion of solvent. As such, hydrophobic residues tend to cluster in the protein core and hydrophilic residues on the surface (Tsai *et al.* (1997)). Based on experimental data Kyte and Doolittle (Kyte & Doolittle (1982)) derived a hydropathy scale to describe the differing hydrophobic capacity of each residue type.

2.1.4.4 Ionic interactions

In electrostatic interactions charges on nuclei and electrons interact according to the Coulomb equation:

$$V = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{2.2}$$

where q_i and q_j are the magnitude of the charges, r_{ij} is their separation, ϵ_0 the permittivity of free space and ϵ_r the relative dielectric constant of the medium.

Estimates of the free energy of formation of a solvent exposed salt bridge on the protein's surface vary but values of the order of -1.0 kcal/mol have been suggested (Schulz & Schirmer (1996)).

2.1.4.5 Aromatic interactions

In the side chains of aromatic amino acids π-electron orbital systems are delocalized on both sides of a planar ring, generating a small partial negative charge on each face, and a small partial positive charge on the peripheral hydrogens. 60% of aromatic side chains in protein domains are involved in aromatic pairings (Burley & Petsko (1985)). It is commonly perceived that aromatic groups stack on top of one another in a face-to-face manner. In fact, detailed analysis of protein structures (Hunter *et al.* (1991)) backed up the results of previous molecular modelling studies (Hunter & Sanders (1990)) suggesting that such a stacking arrangement is disfavoured due to π-electron repulsion. Instead, offset stacked interactions and edge-to-face interactions are marginally favoured. Analysis of thermodynamic cycles (Horovitz (1996)) indicates that the contribution to protein stability by the interaction energy between two aromatic groups is 1.3 kcal/mol, only marginally higher than the stabilization expected from the hydrophobic contribution from burying the surface area between them.

2.1.4.6 π-cation

The π-cation interaction is increasingly recognized as an important non-covalent binding interaction. The effect arises from the electrostatic interaction of a cation with the negative face of an aromatic π-system. A survey of high resolution structures (Gallivan & Dougherty (1999)) indicated that one out of every 77 residues is involved in an energetically meaningful π-cation interaction. Of the cationic residues, arginine participates in almost twice as many π-cation interactions as lysine; of the aromatics, tryptophan participates more commonly than phenylalanine or tyrosine. Interestingly 26% of tryptophans were involved in π-cation interactions, with the preferred geometry being the cation positioned over the 6-atom ring. The free energy contribution of π-cation interactions across protein-

protein interfaces has been estimated as around 3 kcal/mol on average (Crowley & Golovin (2005)).

2.1.4.7 Disulphide

Disulphide bonds are formed by the oxidation of two cysteine residues to form a covalent sulphur-sulphur bond coupling the two thiol groups. Calculations suggest that a disulphide bond should give rise to 2.5 - 3.5 kcal/mol of stabilization but experimental values vary greatly (Thornton (1981)).

2.1.4.8 Aromatic sulphur interactions

Interactions between the non-polar aromatic and sulphur-containing amino acids occur most frequently in the interior of proteins. About half of all side-chain sulphur atoms are in contact with aromatic groups (Zauhar *et al.* (2000)). The geometry of such interactions suggests that sulphur atoms, unlike carbon and nitrogen, predominantly approach the edge of aromatic rings rather than interacting in a planar stacking fashion (Pal & Chakrabarti (2001); Reid *et al.* (1985)). Aromatic-sulphur interactions have been predicted to provide between -0.7 and -2.6 kcal/mol of free energy, depending on local geometry (Ringer *et al.* (2007)).

2.1.5 Methods for identifying contacts

The three most commonly used methods for defining a protein interface in the structure are (i) changes in the solvent accessible surface area upon complex formation, (ii) radial cutoff and (iii) Voronoi polyhedra.

The solvent accessible surface area (ASA) of a protein molecule, measured in Å^2, can be calculated from the atomic coordinates by the program NACCESS (Hubbard (1993)) implementing the method first described by Lee and Richards (Lee & Richards (1971)). This algorithm involves rolling a probe sphere around the van der Waals surface of the molecule. In this study the default probe radius of 1.4Å was used, approximating the radius of a water molecule. The ASA is generated by combining the individual atomic surfaces to give the non-overlapping consensus surface, effectively defining the distance of closest approach for a water molecule to the protein.

To calculate the surface area that becomes buried when two molecules associate three separate calculations are performed. First the ASA of chain A and chain B are calculated separately, followed by the ASA of the A-B complex. The size of the protein-protein interface (ΔASA) is then given by:

$$\Delta ASA = ASA_A + ASA_B - ASA_{AB} \tag{2.3}$$

Relative accessibilities can also be calculated by expressing the accessible surface of each residue X relative to that observed in an Alanine-X-Alanine tripeptide.

ASA can easily be used to define the interface, in that those residues that exhibit a change in ASA between the bound and unbound forms are considered to be involved in the interaction. The ASA method has the drawback that the surface area of an interface can be overestimated, and the standard implementation identifies atoms or residues involved in the interface - not pairwise contacts.

Another straightforward method to identify pairwise interactions is to apply a simple radial cutoff, the premise being that a radius can approximate an atom's sphere of influence. Implementations of this method vary in the radius used and their choice of radii centres (heavy atoms (Ofran & Rost (2003)), Cα atoms, Cβ atoms (Glaser *et al.* (2001)), residue side-chain centroids (Caffrey *et al.* (2004)). Choice of radial cutoff threshold involves a trade-off between sensitivity and specificity. Various thresholds have been applied (see Table 2.1) however, no single value can adequately account for variations in atom sizes and irregular packing (Tsai & Gerstein (2002); Tsai *et al.* (1999)). Furthermore the radial cutoff is a binary description of interactions; two atoms equidistant from a third atom may have different contact areas or interaction free energies. For purposes of comparison, an example of the difference between defining the interface of the same system using a radial cutoff versus solvent accessibility can be seen in Figure 2.2.

Voronoi tessellation involves decomposition of Euclidean space, described as a set of points (typically describing atoms or pseudo-atoms), by assigning each point to a convex, polyhedral region (Gore *et al.* (2005); Richards (1974)). Voronoi methods have been applied to identifying protein interfaces (Bernauer *et al.* (2008); Cazals *et al.* (2006)) and assessing atomic packing (Lo Conte *et al.* (1999)).

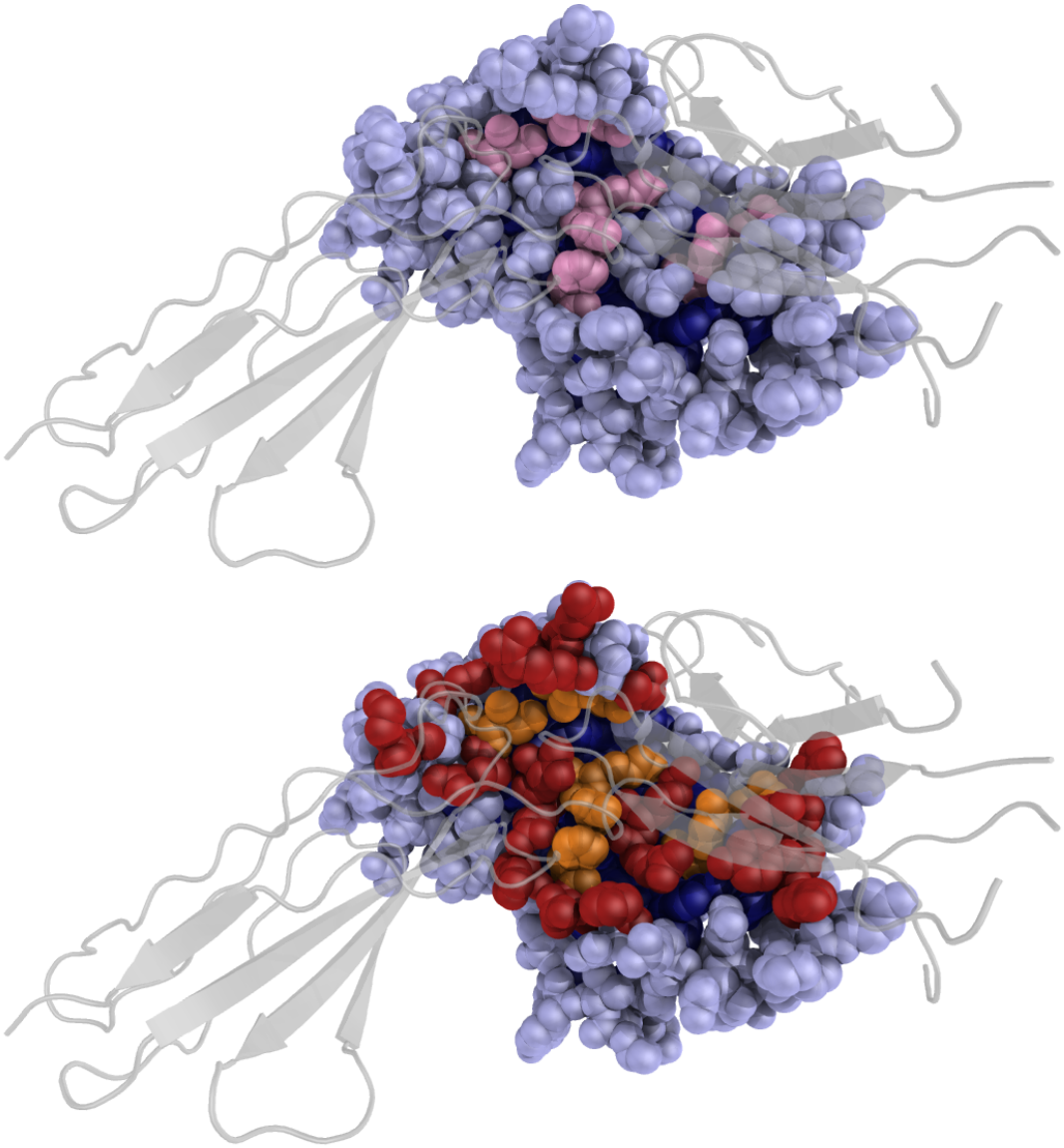


Figure 2.2: Different approaches to defining the protein interface using the example of the human growth hormone-prolactin receptor complex (PDB entry 1bp3). In the upper panel the interface was defined using solvent accessibility (Hubbard (1993)). In the lower panel the interface residues were identified using the augmented radial-cutoff approach implemented by PICCOLO.

Voronoi tessellation does give robust neighbourhood definition between interaction sites, however calculating polyhedra is computationally intensive, and special criteria need to be applied at the molecular surface.

2.1.5.1 Benchmark of methods for prediction of interactions

Benchmarking these methods, in order to establish which is superior, is troublesome, as computationally it is difficult to assign residues that are definitively interacting with one another and those that definitively are not. Fischer *et al.*(Fischer *et al.* (2006)) turned to experimental methods, namely site-directed mutagenesis, to resolve this issue. Site-directed mutagenesis is a widely-used approach to assess the contribution of a particular residue to binding (see Chapter 6 for more details). However, many residues typically interact with more than one residue across the interface, and single-residue mutagenesis reveals little about the pairwise contribution of an interaction between two residues across an interface. The Double-Mutant Cycle (DMC) (Schreiber & Fersht (1995)) is a valuable thermodynamic tool in the study of protein interactions that can circumvent this issue and isolate individual weak non-covalent interactions from the noisy background. The procedure involves mutating, both singly and doubly, pairs of residues (X and Y). This gives a coupling energy $\Delta\Delta G_{int}$, defined as:

$$\Delta\Delta G_{int} = \Delta\Delta G_{X-A,Y-B} - \Delta\Delta G_{X-A} - \Delta\Delta G_{Y-B} \tag{2.4}$$

where $\Delta\Delta G_{X-A}$ is the change in free energy of mutating a single residue X to A, and $\Delta\Delta G_{Y-B}$ a second residue Y to B and $\Delta\Delta G_{X-A,Y-B}$ the change in free energy on the simultaneous mutation of X to A and Y to B. $\Delta\Delta G_{int}$ measures the degree co-operativity of interaction between the two mutated residues - if the effects of the mutations are non-co-operative the difference in free energy for the double mutant is the sum of those for the two single mutations. If the mutated residues are coupled, then the change in free energy for the double mutant will be different from the sum of the two single mutants. A significant pitfall of the DMC approach is that strong coupling energies may reflect indirect interactions, whereas weak coupling energies may arise from local structural rearrangements.

The study by Fischer *et al.* used the results of five DMC experiments where the crystal structure of the complex was also available. The complexes are 1dqj (HyHEL-63 antibody complexed with hen egg white lysozyme (Li *et al.* (2003))), 1brs (complex of barnase and barstar (Schreiber & Fersht (1995))), 1a4y (complex of ribonuclease inhibitor and angiogenin (Chen & Shapiro (1999))), 1vfb (IgG1-kappa D1.3 fv (Dall'Acqua *et al.* (1998))) and 3hfm (HyHEL-10 antibody complexed with hen egg white lysozyme (Pons *et al.* (1999))). For this study the data were augmented by including data from 1lfd (GTPase HRas and Ral guanine nucleotide dissociation stimulator) and 1gua (Ras-related protein Rap-1A and RAF proto-oncogene serine/threonine-protein kinase (Kiel *et al.* (2004))), increasing the size of the data set by ∼80%. These data are shown in Appendix B.1.

2.1.6 Quaternary structures

Quaternary structure concerns the interactions of distinct polypeptide chains to form a protein oligomer. Its consideration is vital to a proper understanding of a protein's biological function.

The atomic coordinates deposited in PDB files reflect the contents of the asymmetric unit (ASU). The ASU is a set of atoms which, when operated on by the crystallographic symmetry operations defined by the spacegroup, generates the complete crystal. The space group symmetry operations are restricted to rotations and translations in biological systems. As such, although the ASU can represent the biologically functional assembly of the protein, often it comprises multiple biological molecules or even a portion of a biological molecule. Proteins crystallize in a highly non-physiological environment, at low temperatures, artificially high protein concentrations and in the presence of organic solvents and crystallization buffers, which can lead to the formation of extensive non-specific crystal packing interfaces. This has important implications for PICCOLO generated using ASU data. The presence of non-specific crystal contacts introduces false positive data to PICCOLO. Conversely, where the ASU comprises a subset of the biologically functional oligomer, the absence of genuine interactions

implicitly introduces false negatives. Figure 2.3 has examples of each of these cases.

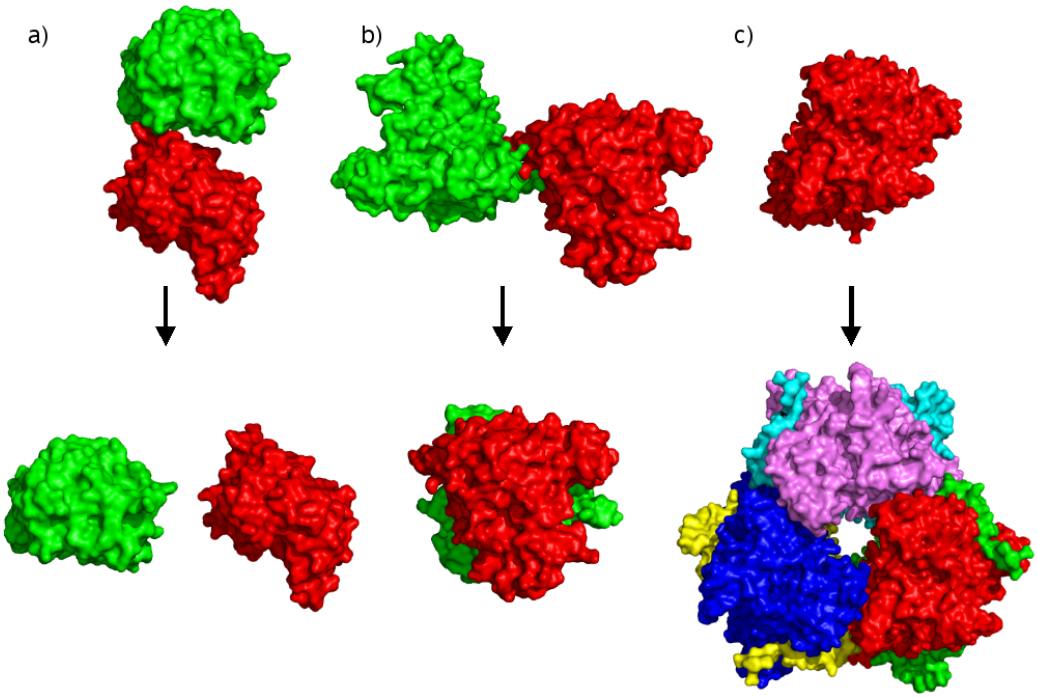


Figure 2.3: Three examples of the difference between the PDB ASU and PISA-predicted assemblies. In panel **a)**, the dimer of murine mitochondrial carbonic anhydrase V observed in PDB entry 1dmx is predicted by PISA to be dissociated. In panel **b)**, the dimer of rat NAD(P)H:quinone reductase observed in the ASU of PDB entry 1qrd, is predicted by PISA to adopt an alternative dimeric conformation. In panel **c)**, the monomeric form of Gal6/bleomycin hydrolase oberserved in the ASU of PDB entry 3gcb is predicted by PISA to form a homohexamer.

2.1.6.1 Biological units

The Worldwide Protein Data Bank (wwPDB) provide information on the biologically functional assembly for each deposition. Since 1999 this information has

been provided by depositors. Prior to that, in cases where no information was provided by depositors, supporting information from UniProt (Uniprot-Consortium (2009)) or PQS (Henrick & Thornton (1998)) (described below) was used. The problem of reliably distinguishing biologically significant assemblies from crystal contacts is non-trivial and several approaches have been proposed (Bernauer *et al.* (2008); Carugo & Argos (1997); Dasgupta *et al.* (1997); Janin & Rodier (1995); Mintseris & Weng (2003); Rodier *et al.* (2005)). Ponstingl *et al.* derived a pair-frequency scoring function to discriminate between homodimeric and monomeric proteins in the crystalline state (Ponstingl *et al.* (2000)). Valdar and Thornton (Valdar & Thornton (2001)) assessed the utility of combining interface size and sequence conservation to discriminate between biological and non-biological contacts, concluding that interface size alone is such a powerful discriminant that the additional predictive power contributed by conservation information is marginal at best. Bahadur *et al.* (Bahadur *et al.* (2004)) used combined terms describing amino acid propensity, hydrophobic interaction and atomic packing to distinguish the different types of interfaces. By combining the non-polar interface area with the fraction of buried interface atoms and residue propensity score, they were able to assign the quaternary structure correctly for 93-95% of their data set.

2.1.6.2 PInS

The Protein Interface Server (PIns) was introduced during the course of this work (Bordner & Gorin (2008)). The PInS procedure involves generating all neighbouring symmetry-related molecules and then calculating the probability that each interface is a specific biological interface using a Random Forest machine learning algorithm trained using information on residue pair counts, residue propensities, evolutionary conservation, interface area, number of intermolecular bonds, packing density, interface type and symmetry information.

2.1.6.3 PQS

Protein Quaternary Structure (PQS) server (http://www.ebi.ac.uk/msd-srv/pqs/) is a resource from the European Bioinformatics Institute (EBI) that provides coordinates for likely quaternary states for crystallographic structures in

the PDB (Henrick & Thornton (1998)). The procedure involves generating putative quaternary assemblies by recursively adding monomeric chains and then discriminating biological interfaces from crystal contacts using a weighted score based on interface properties including the solvent accessible area, number of interacting residues, solvation energy, salt bridges and disulphide bonds. The PQS complexes are then screened by expert annotators helping to reduce errors and inconsistencies that may result from a fully automated procedure. Even with this manual screening step their own internal benchmark on 218 structures suggests a 78% accuracy for PQS (Ponstingl *et al.* (2003)).

2.1.6.4 PISA

More recently the EBI has released PISA (Protein Interactions, Surfaces and Assemblies) (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html) (Krissinel & Henrick (2007)). Jones and Thornton (Jones & Thornton (1996)) have pointed out that no single parameter has been identified for robustly discriminating biologically relevant protein interfaces. Assembly of an oligomeric complex is a cooperative process, typically involving multiple interfaces. Whereas PQS evaluates the properties of individual interfaces, PISA includes contributions from all interfaces constituting the complete oligomeric assembly. The PISA procedure involves representing the crystal as a periodic graph, with monomeric chains as vertices and interfaces as edges, and then enumerating all possible assemblies. Each possible assembly is then evaluated for stability using an empirical estimate of the contribution of enthalpy and entropy to the free energy of binding, leaving only sets of potentially stable assemblies ranked by their predicted stability. The assembly is stable if the predicted free energy of dissociation ΔG_{diss} is positive:

$$\Delta G_{diss} = -\Delta G_{int} - T\Delta S \tag{2.5}$$

where ΔG_{int} is an enthalpic term describing protein affinity including terms for solvation energy, hydrogen bonds and salt bridges; ΔS is an empirical estimate of entropy and T is the temperature. Small molecule ligands and nucleic acids are included in the free energy calculation. Using the same benchmark set of 218 structures used to evaluate the performance of PQS (Ponstingl *et al.*

(2003)), the performance of PISA was estimated at over 90%, a considerable improvement on PQS. However, anecdotal observations suggest that in particular, binary complexes with small interaction surfaces (e.g. enzyme-inhibitor systems) are often predicted to be uncomplexed.

However, the problem is fundamentally problematic as, even in cases where experimental evidence is available, it can be contradictory, possibly because a protein's oligomeric state itself is not fixed and can vary depending on various factors including subcellular location (Brewer *et al.* (1994)) and ligand binding (Latif *et al.* (2002)). Jefferson *et al.* quantified the increased coverage of interactions by including PQS definitions of quaternary structures (Jefferson *et al.* (2007)) (see also SNAPPI-DB Table 2.1). They found that by considering PQS assemblies, the number of unique SCOP (Structural Classification of Proteins) (Hubbard *et al.* (1999)) family pair interactions was increased by 13.3%. This proportion increases to 34.5% when the relative orientation of the domains was also considered. The vast majority of these additional interactions are homo-oligomeric.

2.1.6.5 ProtBud

The ProtBud database (Xu *et al.* (2006))(also see Table 2.1) aims to facilitate comparison of the ASU, PDB Biological unit and PQS contents of related structures. Their analysis reveals that the ASU differs from PDB Biological Unit or PQS complexes for 52% of crystal structures, and that PQS and PDB Biological Units disagree on 18% of entries.

2.1.6.6 3DCOMPLEX

3DCOMPLEX (`http://www.3dcomplex.org/`) (Levy *et al.* (2006)) is a structural classification of protein complexes, essentially an attempt to provide a classification of quaternary structures in an analogous manner to that of the SCOP classification of tertiary structures (Hubbard *et al.* (1999)). 3DCOMPLEX is likely to be a valuable contribution to studies of the evolution of quaternary structures. A hierarchical classification of complexes is constructed by first representing each PDB Biological Unit complex as a graph, where each chain is

a node and each interface an edge. An interface is here defined as ten or more residues having contacts (any pair of atoms within the sum of their van der Waals radii plus 0.5Å). Graphs are then compared to generate the hierarchy. The levels of the hierarchy are defined using graph topology, SCOP domain architecture, sequence similarity and finally symmetry of the complex.

2.1.6.7 PiQSi

The PiQsi resource (Levy (2007)) for Protein Quaternary Structure Investigation assists the investigation of the quaternary structure of protein complexes in the PDB. More than 10,000 PDB Biological Units have been manually annotated using experimental information from the literature and, crucially, quaternary structure information of protein's evolutionary relatives. This allows an assessment of any errors in the Biological Units in the form of false positive or false negative interface predictions, deriving either from the automated PQS predictions or author misannotations. According to this analysis approximately 14% of biological units are thought to be in error. Corrected atomic coordinates are not supplied. The data from PiQsi make an ideal benchmark for methods that predict protein quaternary structure.

2.2 Materials and Methods

2.2.1 Relational databases

Relational databases consist of a collection of interconnected sets of data stored in efficiently-indexed tables. The fundamental concepts that underly them derive from relational algebra, as first described in 1970 by E. F. Codd (Codd (1983)). Interaction with the database is mediated by a relational database management systems (RDBMS) through SQL (Structured Query Language) - a standardized language for the addition, modification and retrieval of data, the creation and alteration of schema, and the management of database access. MySQL (Widenius *et al.* (2002)) (http://www.mysql.com) was chosen as our RDBMS because it is open source (and therefore free of charge); it is broadly considered to provide good performance and reliability; and its wide popularity means that it is still

under active development. The software runs as a server providing multi-user access to several databases. To support this work a dedicated MySQL server was established with 8 Intel® Xeon® CPU E5320 processors giving 16Gb of RAM and for storage a 6 disk RAID (Redundant Array of Inexpensive Disks) array totalling 3Tb.

2.2.2 Upstream preparation

PICCOLO, CREDO and BIPA all require comprehensive, up-to-date reference information regarding all structures, their chains and residues currently available in the PDB. This information is housed centrally in a shared hierarchical PDB schema (Figure 2.4) with automatic updates synchronized with updates to the PDB mirror.

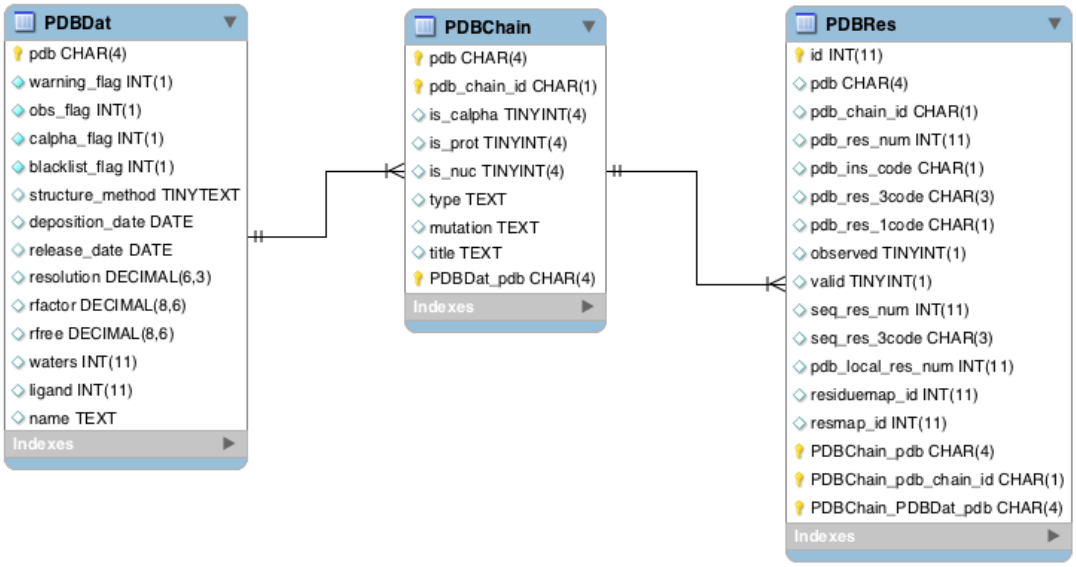


Figure 2.4: Database schema for shared PDB database.

The PDB uses macromolecular Crystallographic Information File (mmCIF) data dictionaries to describe the information content of each structure deposition. The RCSB (Research Collaboratory for Structural Bioinformatics) provides a helper tool called Db Loader to convert mirrored mmCIF data into a form that

is directly loadable into a MySQL relational database. A scheduled process ensures that the MMCIF tables are emptied and repopulated biweekly. From this primary data a series of tables are subsequently derived, containing core data regarding information on PDB Structures (PDBDat), chains (PDBChain) and residues (PDBRes), the details of which are outlined below. Semin Lee, in the TLB group, manages the initial mmCIF loading process.

2.2.3 Core PDB data

The **PDBDat** table contains information regarding the PDB deposition including: experimental method (X-ray crystallography, NMR, fibre diffraction etc); deposition and release dates; crystallographic information (resolution, R-factor and free R-factor) for X-ray structures; the number of waters found in the structure; a flag indicating the presence of any ligand groups; and a flag indicating that the structure contains Cα only chains (see below).

Information on polymeric chains is stored in the **PDBChain** table. As well as the chain identifier this table stores flags indicating whether the chain is polypeptide or polynucleotide and whether the chain is Cα only. In poorly diffracting crystal structures there may be only sufficient electron density to resolve the Cα atoms (or occasionally the backbone atoms) and side-chains may not be reliably resolved. Cα only chains are defined here as those chains where the ratio of Cβ atoms to non-glycine Cα atoms is $<$0.8. Without sidechains, information regarding secondary structure, solvent accessibility and hydrogen bonding cannot be determined. Such structures provide very limited information and are therefore not considered further. This is the case for more than 300 structures (or 0.6% of the PDB).

Great lengths have been taken to provide a consistent inventory of each individual residue in the PDB, providing globally unique identifiers which act as a common currency which is shared between the various databases, greatly facilitating comparative analyses. This information is housed in the **PDBRes** table. Importantly, each individual residue in the PDB repository can be uniquely identified through its parent PDB code, chain identifier, PDB residue number and insertion code. The PDB residue number is the numeric label of the residue

taken from the PDB file (N.B. the first residue in each chain does not necessarily have PDB residue number 1). The insertion code is sometimes used to preserve a certain desirable residue numbering scheme. For example, PDB entry 4est, a serine protease, has residue identifiers as follows: PHE 65, ARG 65A, VAL 66. In this way the residue numbering scheme stays in tune with that of homologous structures to keep the numbering of the catalytic triad consistent. Currently 2,267 structures (or 4.3% of the PDB) employ insertion codes. Further fields hold: the amino-acid residue type; a flag indicating whether the residue is observed in the electron density map; for those residues that are observed, a validity flag indicating whether they are one of the 20 canonical residue types; finally for those residues that are both observed and valid, a serial counter whereby the first residue in each chain is always numbered 1. This last field is invaluable when mapping the output of external structural analysis programs into the database, obviating the need for elaborate alignment procedures.

2.2.4 Mapping structure to sequence

UniProt (Uniprot-Consortium (2009)) is a comprehensive central repository of protein sequence data with manual annotations. The capacity to map accurately, at the resolution of individual residues, between a protein's sequence - as observed in a PDB structure - to its cognate UniProt record, is both important and, somewhat unexpectedly, non-trivial. There are many benefits of such a mapping. Firstly, UniProt residue annotations can trivially be transferred to PDB structures. Further, multiple subtly different structures of the same protein can be grouped and aligned, providing a simple but robust method to cluster similar proteins to help remove redundancy. Such a scheme has been used to provide a non-redundant set of interfaces in PICCOLO. Most significant though is the application of this mapping to the TLB group's efforts on predicting the effects of mutations on protein structure, function and interactions. As described in Chapter 5, genome-level annotations from Ensembl (Birney (2006)), in particular information on non-synonymous Single Nucleotide Polymorphisms (nsSNPs) has been mapped to cognate UniProt sequences. The sequence-structure mapping

therefore facilitates simple navigation from mutations to sequences to structures and from there details of protein function.

One of the main difficulties in determining this sequence-structure residue mapping is that many structures in the PDB have regions of unobserved residues within polypeptide chains due to poorly defined regions of structure, such as flexible loops. Such gaps in the sequence are not taken into account by traditional sequence alignment algorithms, leading to incorrect alignments for regions flanking the unobserved regions. Other features that can introduce differences between the sequences of subsequent structures of the "same" protein include the presence of artefactual constructs to aid cloning (e.g. starting methionine), protein purification (e.g. hexa-HIS tag), or crystallization (e.g. trimmed domain boundaries), as well as isoforms, natural and engineered mutants and modified residues (e.g. selenomethionine).

The Structure Integration with Function, Taxonomy and Sequence (SIFTS) resource (Velankar *et al.* (2005)) from the EBI accurately maps the sequences from PDB entries on to corresponding UniProt entries. To circumvent this problem they modified the standard alignment protocol by using sequences of the observed regions of the protein structure and producing separate alignments for these segments with the complete sequence of the protein that was used in the experiment (from the SEQRES record in the HEADER of the PDB entry). These separate alignments were then merged together to assemble a complete consensus sequence that does not have gaps reflecting unobserved residues. A similar procedure was carried out to obtain alignments between this consensus sequence and the corresponding UniProt entry. These two composite alignments are then merged to give the complete residue-level mapping between the sequence of the complete polypeptide from the experiment and its UniProt counterpart. This procedure copes with the more complex situation in chimaeric structures, where sequences from two or more UniProt entries are involved, in which case the correct boundaries are confirmed manually.

Although the SIFTs records were of high-quality a frustrating lack of coverage and the sporadic nature of data updates meant that, somewhat reluctantly, a similar procedure had to be devised and implemented locally to provide com-

prehensive coverage of sequence-structure mapping. Sungsam Gong in the TLB group developed and manages such a procedure.

2.2.5 Structure quality

The PDB is highly redundant at various levels. The structures of many proteins have often been solved several times under different experimental conditions, in different conformations, with different ligands, and with different modifications. To prevent this redundancy biasing any analysis it is necessary to perform some form of clustering to provide a non-redundant set. The question then arises, which of a set of "equivalent" structures within a cluster should be selected as a representative? Rather than selecting an arbitrary cluster member, each structure is assigned a quality score, or QScore, based on the structure's resolution, R-factor and the number of absent internal residues, with the resolution dominating:

$$QScore = ((\frac{1}{resolution}) + (0.1 - Rfactor)) \times (1 - proportion\ missing\ residues) \tag{2.6}$$

This is analogous to the SPACI score (Summary PDB ASTRAL Check Index) used by Brenner *et al.* (Chandonia *et al.* (2004)) in deriving the ASTRAL compendium. This has the effect that non-X-ray structures are deprioritized. Of all members of a cluster, the member with the best QScore is chosen. Fig 2.5 shows the frequency distribution of Resolution, R-factors, proportion of missing residues and QScore across the whole of the PDB in PDBDat. Fig 2.5 illustrates the contribution of the R-factor and missing residue terms to modulating the resolution to generate the QScore. The QScore data have been applied both in clustering protein interfaces in PICCOLO and SCOP domains in TOCCATA.

2.2.6 Inconsistencies in the PDB

Although the data found in PDB files are clearly invaluable, high-throughput processing of every structure in the repository can be hampered by the inherently heterogeneous and inconsistent nature of certain aspects of the data. A small minority of troublesome structures often require an incommensurate amount of

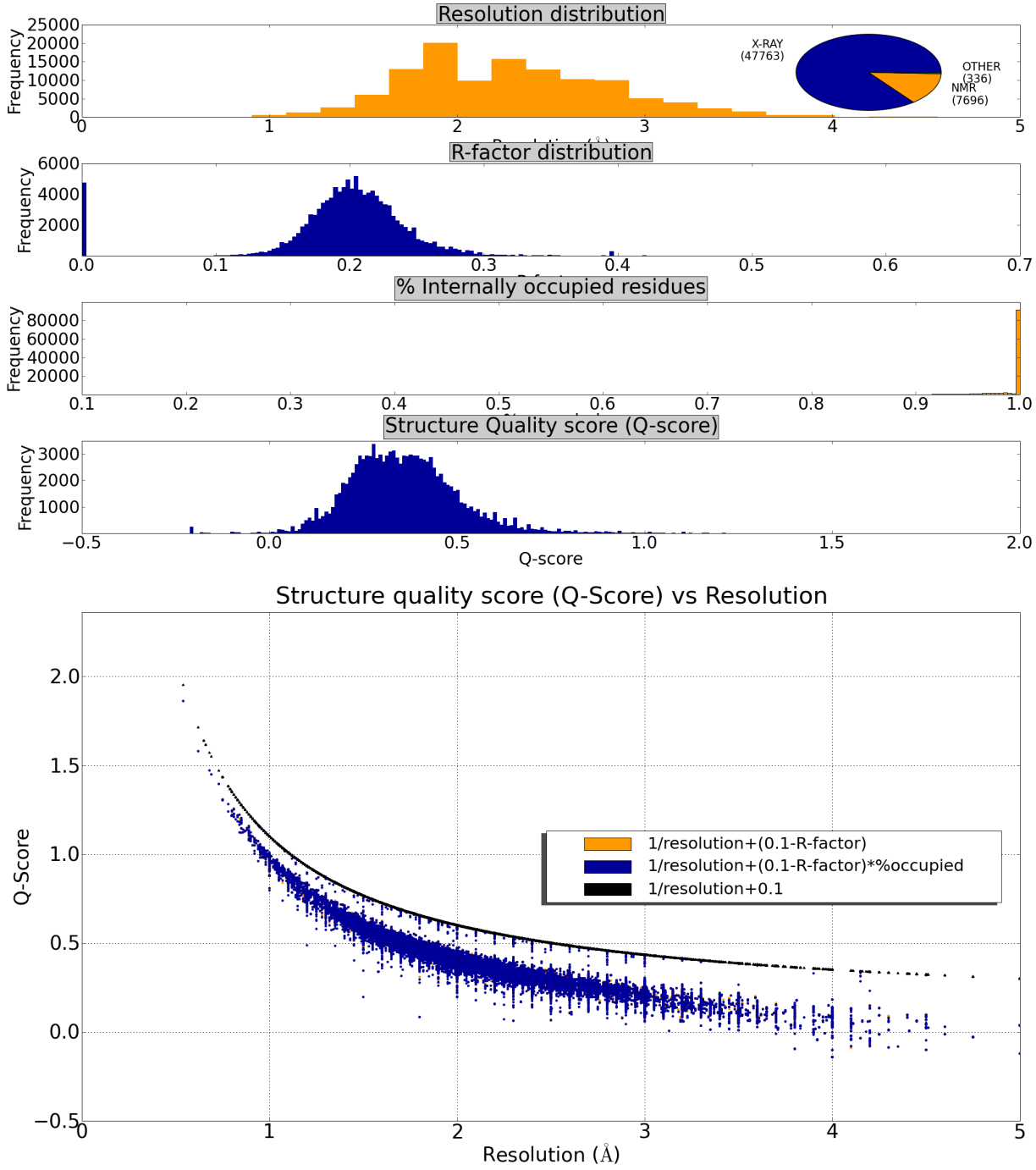


Figure 2.5: Components of structural quality score (QScore). Upper panel shows distributions of Resolution, R-factor, % missing residues and Qscore. The scatter plot in the lower panel shows the relationship of QScore to Resolution and the impact of including terms to describe R-factor and % missing residues.

attention to negotiate them successfully. Many of the problems stem from the fact that the PDB is not a relational database (Schierz *et al.* (2007)). The situation has been improved somewhat by the efforts of the recent PDB remediation project (Henrick *et al.* (2007)), however many problems remain. Many of the issues can be attributed to the limitations of the particular experimental methods used to solve the structure whereas some are due to differing assumptions made by the many thousands of different depositors over the years. Some of the issues a robust software system must handle include: crystal structures with multiple-occupancy atoms; multiple models from NMR ensembles; residue numbers with alphabetic insertion codes; inconsistent presence of water molecules; inconsistent presence of hydrogen atoms; absent residues in crystal structures owing to missing electron density; low-resolution structures consisting solely of Cα backbone atoms; structures containing no peptide or nucleic acid polymer residues; >300 different non-standard amino acid residue types (naturally-occurring or engineered modifications forming part of the polypeptide backbone); low resolution structures with unassigned residue types; lower-case and numeric chain identifiers. Unfortunately many of the standard software tools in common use today do not handle these relatively common circumstances consistently.

2.2.7 Sanitizing PDB data

In order to isolate and avert such issues, an automated system to "sanitize" all structures on the data mirror has been devised. Such pre-processing of the raw PDB data addresses the inconsistencies upstream of other processes, thereby greatly simplifying all downstream procedures and reducing the requirement for each component to perform elaborate error checking.

This sanitizing process involves using the PDB module from BioPython (Hamelryck & Manderick (2003)) to read each structure in turn, optionally perform a series of cleaning steps before re-writing a consistently formatted PDB file. This process ensures that only those residues that are already captured in the database are included in the outputted PDB files, ensuring that every residue is validated and uniquely identifiable as part of a *bona fide* polypeptide or nucleic acid chain, thereby guaranteeing self-consistency between the PDB flat files and

the database. The optional cleaning processes that can be performed include: selection of highest-occupancy atoms only; stripping of hydrogen atoms; removal of all but the first model in multi-model structures; removal of ligands; stripping of waters; and repair of the most common modified residues to their "parent" residues. Even though more than 300 different non-standard polypeptide residues can be found in the PDB, more than 90% of the total are selenomethionine (MSE), methyllysine (MLY) or hydroxyproline (HYP). Heavy selenomethionine residues are routinely synthetically engineered into proteins to help crystallographers solve the phase problem, whereas the others are more likely to be naturally occurring. The modification to their parent amino acid residue means that any such affected structures can now be appropriately handled by downstream legacy software that would otherwise fail, but it does carry a small risk of incurring artefactual results (most likely false negative contact identification).

2.2.8 PDB flavours

This cleaning procedure also has the benefit of permitting the creation of a series of different "flavours" of the PDB data mirror - each with slightly different contents reflecting different states of the protein. For example, to generate the TOCCATA database of family structural alignments, PDB formatted files for each SCOP domain are generated. In order to create the protein-ligand database CREDO, a flavour of the PDB containing only those structures containing heteroatom ligands is generated. Similarly for BIPA, a flavour containing only those structures that contain both polypeptide and nucleic acid chains is created. Furthermore, this creates the opportunity to centralize the execution of a series of structure analysis programs on each of the flavours. In an automated, weekly process managed by Semin Lee in TLB group, NACCESS (Hubbard (1993)) is performed to calculate solvent accessibility, HBPLUS (McDonald & Thornton (1994)) to identify hydrogen bonds and JOY (Mizuguchi *et al.* (1998a)) to provide secondary structure and other structural annotation. An example of the benefits of this system is illustrated in PICCOLO, wherein in order to calculate the solvent accessible surface area involved in macromolecular interactions, it is necessary to assess the solvent accessibility of all unbound protein chains as well

as their bound counterparts for each pair of interacting PICCOLO chains. By having the appropriate forms of both PDB files ready-generated, automatically updated each week, with pre-calculated NACCESS accessibility data, subsequent analyses are greatly eased.

2.2.9 Generating contact data

All chains in the PDBChain table are classified as either valid or invalid. Invalid chains are those that are non-polypeptide (i.e. nucleic acid or carbohydrate), Cα only, obsolete, blacklisted (a small number of manually-identified troublesome PDB entries), 100% unobserved or comprising only non-standard residues. To generate PICCOLO all PDB entries containing more than one valid chain are first identified. For each of these entries every unique pair of non-identical chains was examined. Therefore for n chains $n(n-1)/2$ comparisons were performed e.g. for a PDB entry with four chains A,B,C and D, six comparisons are performed: AB, AC, AD, BC, BD and CD. Note that the chain pairs are always ordered alphanumerically. This scheme does prevent needless duplication of the measurement and storage of pairwise contacts, as distance calculations are commutative, but this also has important implications for the database schema and subsequent queries.

When examining each pair, for each atom in the first chain all atoms within a fixed search radius are identified. If any of these atoms belong to the second chain the pair is flagged as a potential inter-chain contact, the details of the two atoms are logged and the inter-atomic Euclidean distance is measured in Ångstroms (Figure 2.6). Euclidean distance between two atoms i and j with three-dimensional co-ordinates x, y, z is defined in Equation 2.7:

$$d = \sqrt{(i_x - j_x)^2 + (i_y - j_y)^2 + (i_z - j_z)^2} \tag{2.7}$$

By default a fixed search radius of 6.05Å is used. This value was chosen as the maximum length of a water-mediated hydrogen bond (Robert & Janin (1998)). Neighbour search algorithms such as this can be computationally expensive. However, the PDB module of BioPython implements a NeighbourSearch method using the *kd*-tree algorithm (de Berg *et al.* (1997)). The *kd*-tree family of algorithms use

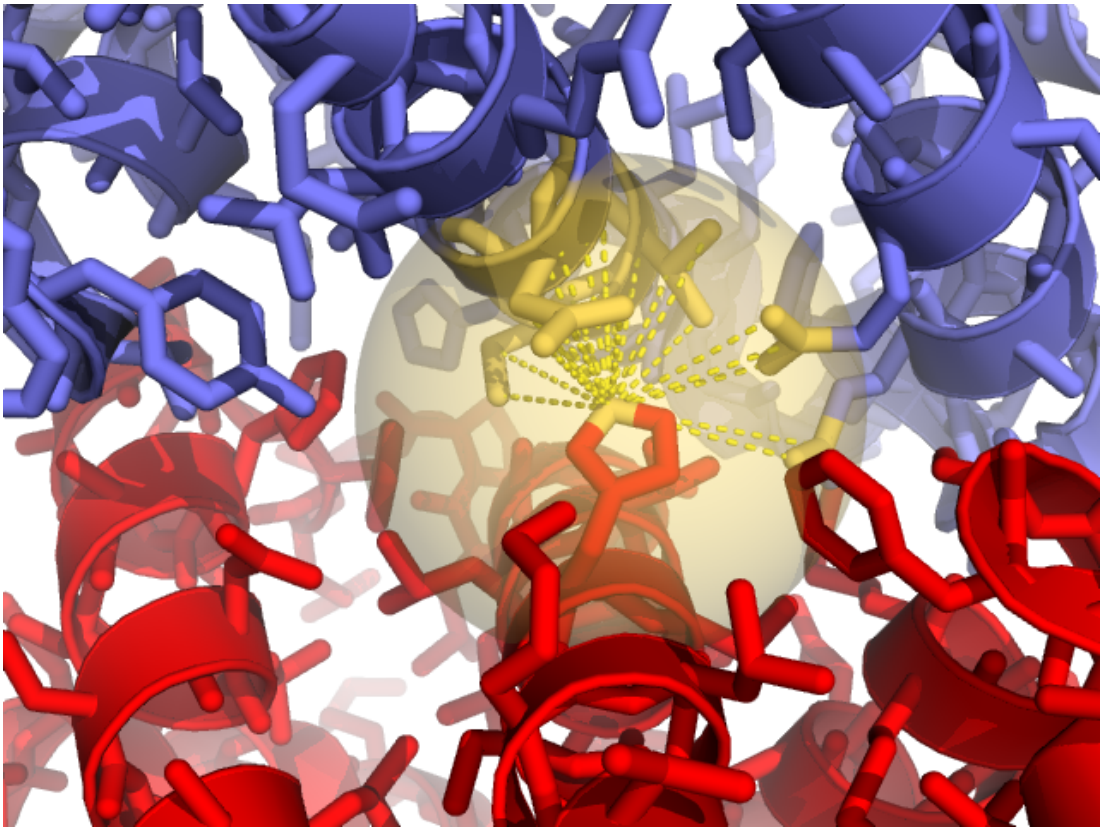


Figure 2.6: Illustration of the radial-cutoff method. The interface of human α and β haemoglobin is used as an example (PDB entry 1y4v). All atoms on the β chain of haemoglobin within 6.05Å of the $NE2$ atom of the side-chain of histidine 103 on the α chain are considered *proximal*, highlighted in yellow and are considered for further annotation.

efficient hierarchical space-partitioning data structures for recursively organizing points in a k-dimensional space. This gain in efficiency means that PICCOLO can be run over the entire PDB overnight on a Linux workstation with Intel® Core 2 CPU 6600 with 2Gb of RAM.

Defining contacts based on inter-atomic distance is a commonly used approach (see Table 2.1). However, it is clear that although it is possible for atoms 6.05Å from one another to be engaged in an interaction, the vast majority of atoms this far from one another are likely to be false positives. In other words, this method is sensitive but not specific. Although reducing the default threshold is

likely to increase specificity it would come at the cost of sacrificing sensitivity by increasing false negatives. To resolve this issue, based upon the chemical nature of the two atoms involved in the putative interaction and the distance between them, each of the potential inter-atomic contacts are classified into a series of specific interaction types. These interaction types can be seen in Table 2.2.

In order to achieve this, each atom of the 20 canonical residues is assigned van der Waals (non-covalent) and atomic (covalent) radii as well as being assigned a series of property flags indicating the types of interactions in which they have the capacity to participate. These are described below and summarized in Table A.1 and figures in Appendix I. The values for the van der Waals and atomic radii come from intermolecular distance calculations on >30,000 high-resolution crystal structures of small organic compounds from the Cambridge Structural Database (CSD) (Allen (2002)) that contain the same atomic groups as those found in proteins, so that the radius for an atom of a given element is residue-specific (Tsai *et al.* (1999)) (http://bioinfo.mbb.yale.edu/geometry/geom-mbg/data/README.htm). This set of radii has previously been used to calculate protein volumes (Tsai & Gerstein (2002)). Flags indicating those atoms that are considered hydrophobic, aromatic, cationic or anionic are set by applying SMARTs queries (SMiles ARbitrary Target Specification) (http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html) to structures of the 20 canonical residues, followed by manual inspection. Each of the 20 canonical residue types also has an extra negatively ionizable OXT atom defined, to include the acidic carboxyl group when the residue is chain-terminating.

Van der Waals contacts, the most common type of interaction, are assigned as those pairs of atoms whose interatomic distance is less than the sum of the van der Waals radii plus 0.5Å (Jefferson *et al.* (2007); Keskin *et al.* (2004)). No restriction is placed on atom type. This contact definition alone is more sophisticated than many of the fixed cutoff values described in Table 2.1. **Van der Waals clashes** are those contacts where the interatomic distance is less than the sum of the van der Waals radii. Similarly **covalent contacts** are those where the interatomic distance is less than the sum of the atomic radii. The vast majority of covalent contacts are disulphides. By definition, covalent interactions

Interaction type	**Type atom i**	**Type atom j**	**Distance Criteria**	**Angle criteria**
van der Waals	any	any	$d(a_i, a_j) \leq vdw(a_i)+vdw(a_j)+0.5$Å	-
van der Waals clash	any	any	$d(a_i, a_j) \leq vdw(a_i) + vdw(a_j)$	-
hydrogen bond*	hydrogen bond donor	hydrogen bond acceptor	$d(a_i, a_j) \leq 3.9$Å $d(a_h, a_{acc}) \leq 2.5$Å	$\theta(a_{don}, a_h, a_{acc}) \geq 90°$ $\theta(a_{don}, a_{acc}, a_{acc-antecedent}) \geq 90°$ $\theta(a_h, a_{acc}, a_{acc-antecedent}) \geq 90°$
water-mediated hydrogen bond*	hydrogen bond donor or acceptor	hydrogen bond donor or acceptor	$d(a_i, a_j) \leq 3.9$Å $d(a_h, a_{acc}) \leq 2.5$Å	$\theta(a_{don}, a_h, a_{acc}) \geq 90°$ $\theta(a_{don}, a_{acc}, a_{acc-antecedent}) \geq 90°$ $\theta(a_h, a_{acc}, a_{acc-antecedent}) \geq 90°$
amino-aromatic hydrogen bond*	hydrogen bond donor	amino-aromatic hydrogen bond acceptor	$d(a_i, a_j) \leq 3.9$Å $d(a_h, a_{acc}) \leq 2.5$Å	$\theta(a_{don}, a_{acc}, N_{aromatic-plane}) \leq 20°$ $\theta(a_{don}, a_h, N_{aromatic-plane}) \leq 20°$
hydrophobic contact	hydrophobic	hydrophobic	$d(a_i, a_j) \leq 5$Å	-
ionic	cation	anionic	$d(a_i, a_j) \leq 6$Å	-
aromatic	aromatic	aromatic	$d(a_i, a_j) \leq 6$Å	-
π-cation	cationic	aromatic	$d(a_i, a_j) \leq 6$Å	-
disulphide	sulphur residue: cys	sulphur residue: cys	$d(a_i, a_j) \leq 2.08$Å	-
aromatic-sulphur	sulphur	aromatic	$d(a_i, a_j) \leq 5.3$Å	-
covalent	any	any	$d(a_i, a_j) \leq cov(a_i) + cov(a_j)$	-
proximal	any	any	$d(a_i, a_j) \leq 6.05$Å	-

Table 2.2: Interaction classification scheme. Interactions between atom i and atom j are classified based on the atom types of i and j, the distance d between them and angle criteria. In all cases i and j are exchangeable.
$*$ Interaction defined by HBPLUS.

are a subset of van der Waals clashes, which themselves are a subset of van der Waals contacts.

Unlike all other interaction types **hydrogen bonds** and **water-mediated hydrogen bonds** are located by running an external program, HBPLUS (McDonald & Thornton (1994)). The algorithm, developed by McDonald and Thornton, involves first positioning the hydrogen atoms, followed by calculation of the hydrogen bonds. An interaction is considered to be a hydrogen bond if one atom of the pair is listed as a donor and the other as an acceptor (Table A.1 in Appendix I), and the angles and distances formed by the relevant atoms meet the appropriate criteria (Table 2.2). Studies have suggested that the π-electron shells of aromatic rings may also act as weak hydrogen bond acceptors (Mitchell (1994)). In order to implement this the -R option on HBPLUS has been set to allow atoms in the aromatic rings of tyrosine, tryptophan and phenylalanine to accept these amino-aromatic hydrogen bonds.

Only in a minority of very high resolution (<1.0Å) crystal structures can hydrogen atoms be accurately resolved and little or no difference can typically be determined between carbon, nitrogen and oxygen atoms. For structures solved at resolutions above 1.0Å, atoms in the majority of side-chains can be uniquely identified from the electron density map, but for asparagine, glutamine and histidine, whose side-chains appear symmetrical in the electron density, certain atoms can only be identified on the basis of their local structural context and in particular their hydrogen bonds. To resolve this issue HBPLUS implements an option (-x) to explore potential hydrogen bonds that would be formed if the CD2 of histidine was actually ND1, CE1 was NE2 and the nitrogens and oxygens of the asparagine and glutamine amide groups were exchanged. Note that some atoms are capable of acting as either hydrogen-bond donors or acceptors depending on the details of their local structural context (SER OG, THR OG1, HIS ND1, CYS SG1, TRP NE1, TYR OH).

Water molecules are present in 80% of PDB structures (data from PDBDat). Although relatively rare in intra-molecular interactions, water-mediated hydrogen bonds make a significant contribution to inter-molecular interactions. Water can mediate between two hydrogen bond donors, two acceptors or from a donor to an acceptor. One difficulty in identifying water-mediated contacts is that in

many lower-resolution crystal structures water molecules may be inappropriately modelled into patches of electron density or added during refinement to improve the calculated structure factors. In this work "genuine" structured waters were distinguished by considering only those water molecules that engage in more than one hydrogen bond. Conveniently, any water molecule suggested by HBPLUS to be hydrogen bonded to two residues on different chains, by definition already meets this definition. Hydrogen bonds and water-mediated hydrogen bonds are further sub-classified as being between either two main-chain atoms, two side-chain atoms or between main-chain and side-chain.

Hydrophobic interactions are those where both atoms are labelled as hydrophobic and the inter-atomic distance is less than 5Å (Tina *et al.* (2007)). In defining **ionic interactions**, the strictly correct way to calculate the electrostatic interaction for two point charges would be to use quantum chemical methods to solve the Coulomb equation separately for each nucleus, but this is somewhat impractical for large biological systems. A simpler approach is to consider only the formal charges on the protein (whether an electron has been lost or gained). Carboxyl groups are deprotonated and carry a negative charge delocalized over the two oxygen atoms, while amino groups are protonated and carry a positive charge delocalized over the three hydrogen atoms. The protonation state of amino acid residues in free solution at pH7 can be determined from model pKa values defined for each residue. However, the protonation state of ionizable residues in the folded protein depends also on the local structure environment, including exposure to solvent, proximity to other titratable groups or permanent charges in the protein. Methods that take these factors into account (Davies *et al.* (2006); Dolinsky *et al.* (2007); Li *et al.* (2005)) are again not practical to run at large scale, so the solution pKa values are used for ionizable residues and pH7 is assumed. The distance threshold was taken from Barlow and Thornton (Barlow & Thornton (1983)).

Aromatic interactions are defined when two criteria are met. When a pair of aromatic atoms is within the appropriate distance threshold then the centroids of the two parent planar ring systems are calculated. If the centroids are also within the distance threshold, then the contact is considered aromatic. As part of his work on CREDO, the protein-ligand interaction database, Adrian Schreyer in

the TLB group developed a procedure to sub-classify aromatic contacts as being either "face-to-face", "edge-to-face" or "displaced edge to face". To achieve this, for each pair of atoms involved in an aromatic contact, the normals of the two parent planar ring systems are calculated using Newell's method (Kirk (1994)). The dihedral angle between the two planes is defined as the angle between the normals. The displacement angle is defined as the angle between the normal of the first ring and the vector between the two ring centroids. The aromatic interaction is classified as "edge to face" where the dihedral angle is greater than 30°. Dihedral angles less than or equal to 30° are classified as "face to face" where the displacement angle is less than or equal to 20° and "displaced face to face" otherwise.

π-cation interactions are defined when a cationic atom and an aromatic atom approach within 6Å threshold of one another (Gallivan & Dougherty (1999)). **Disulphide bonds** are those where two sulphur atoms from cysteine residues approach within 2.08Å (Thornton (1981)). **Aromatic-sulphur interactions** are those where an aromatic atom approaches within 5.3Å of a sulphur atom (Zauhar *et al.* (2000)).

Even though these interaction definitions are not rigorous, they are each precedented, robust and rapid to calculate. Note that an exclusive classification of inter-atomic interactions would require artificial prioritization of one interaction type above another. In this work, interactions are classified equivocally so each atom pair can simultaneously exhibit the character of more than one interaction type. Each atom-pair can therefore be thought of as being represented as a binary bit-string vector - an ordered sequence of qualitative indicators that provides co-ordinates in an arbitrary space that can be used to locate interactions relative to one another. This results in overlaps between for example, van der Waals contacts and shorter hydrogen bonds, hydrogen bonds and shorter ionic interactions and hydrophobic and aromatic interactions. This deliberate ambiguity arguably reflects the somewhat amorphous nature of molecular interactions.

All atom pairs within the original 6.05Å distance threshold of one another are only considered as being in contact with one another if one of the above criteria are met (i.e. the logical "OR" of all interaction types). Atom pairs not

meeting any of these criteria are still stored in PICCOLO as being proximal to one another, but are in general not considered in any further analyses.

2.2.10 Benchmark of methods to identify interactions

A short benchmark study was performed to quantify the benefit, if any, of adapting the radial cutoff method with the various detailed interaction type definitions to specify a more refined interaction definition. The Double Mutant Cycle data used in this study are taken from the literature and are shown in Appendix B.1. Figure 2.7 shows a histogram of the free energy differences for the systems used in this benchmark in kcal/mol. Only mutations to alanine were considered.

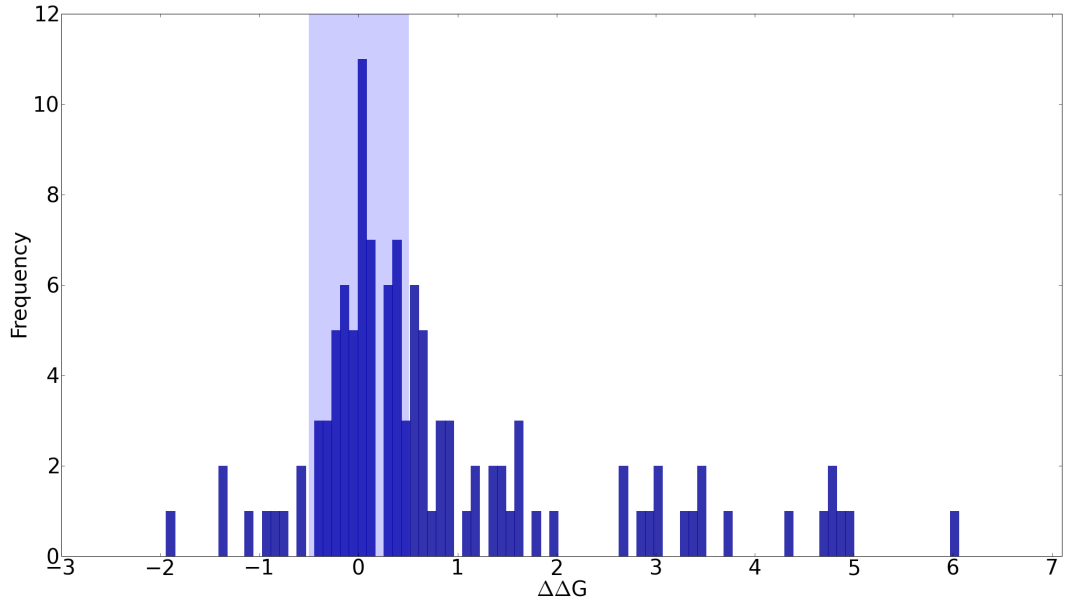


Figure 2.7: Histogram of $\Delta\Delta G$ values in kcal/mol for the Double Mutant Cycle data in Table B.1 on page 222. The region marked in pale blue (between -0.5 and 0.5 kcal/mol) constitute the true-negative set used in the benchmark study, and the remainder the true-positive set.

The following performance measures were used:-

$$Sensitivity = 100 \times \frac{TP}{(TP + FN)} \tag{2.8}$$

$$Specificity = 100 \times \frac{TN}{(FP + TN)} \tag{2.9}$$

$$Accuracy = 100 \times \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{2.10}$$

where TP, FP, TN and FN are the numbers of true positives, false positives, true negatives and false negatives respectively. When evaluating the performance of any predictive method it is important to focus on both the sensitivity and specificity. In practical terms high sensitivity means little if specificity is poor. However, evaluation of specificity requires data regarding true negatives which is often unpublished. Here $\Delta\Delta G$ values less than -0.5 kcal/mol or greater than 0.5 kcal/mol are considered significant.

2.2.11 Solvent accessibility

Absolute solvent accessibility data as calculated by NACCESS are stored in PICCOLO for the *apo* and bound forms, as well as the difference between the two, to give the area contributed to the interface at the level of individual amino acid residues and complete polypeptide chains. Relative accessibilities are stored at the level of individual residues. Accessibility data are stored for all residues from structures engaged in interactions - not just those residues mediating interactions.

2.2.12 PICCOLO schema

The result of running PICCOLO is, for each PDB file and for each pair of chains, a series of six tab-delimited output files, corresponding to atom pairs, residue pairs, chain pairs, atom-level summaries, residue-level summaries and chain-level summaries. Although the output files are amenable to direct analysis themselves, each file corresponds to a particular table in the PICCOLO MySQL database schema into which they can be loaded without further processing. The schema for PICCOLO is shown in Figure 2.8.

2.2.13 Structural environments

Several key properties of protein residues depend acutely on the local structural environment in which they are found. All residues in PICCOLO are classified

Figure 2.8: Database schema for PICCOLO.

depending on whether or not they are engaged in interactions and their exposure to solvent. Residues in the interface engaged in interactions are classified as "Interface Core" if they are buried in the bound form and "Interface Periphery" otherwise. Residues not engaged in interactions are similarly classified as "Core" if they are buried and "Exposed" otherwise. The definition used to assign a residue as buried is that it must have a relative accessibility of less than 7% (Mizuguchi *et al.* (1998a)). This is analogous to the definitions used by Guharoy and Chakrabarti (Guharoy & Chakrabarti (2005)) although they use 100% burial as a criterion for core. Solvent accessibility is always calculated with respect to a single pair of chains. This means for that for complexes of more than two chains, a particular residue may be found in different environments, depending on which chains are being examined. To circumvent such troubling ambiguous classifications, the classes are prioritized such that residues are preferentially classified as Interface Core > Interface Periphery > Core > Exposed.

2.2.14 PyMOL integration

With large and complex data sets of this nature visualization tools to aid analysis are paramount. PyMOL (Delano (2002)) is an open-source molecular visualization system that extends, and is extensible by the Python programming language (Van Rossum (2003)). This enables Python functions to be written that connect to the MySQL database and extract annotations from PICCOLO describing which atoms and residues are involved in interactions, and for these annotations to be displayed in the PyMOL window. Using this approach useful visualizations of the four residue classifications can be generated for any complex. Examples of this are shown in Figure 2.9.

This methodology can be extended to visualize atomic interactions by using different colour and dash parameters to indicate different interaction types. Figures 2.10, 2.11 and 2.12 (on pages 63 and 64) show examples of three different protein interfaces indicating the jungle of molecular interactions.

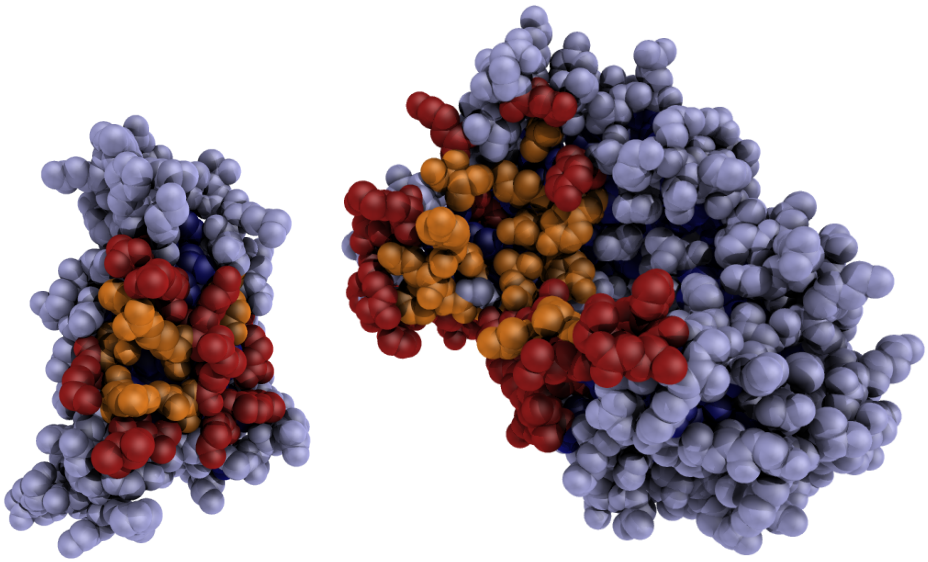


Figure 2.9: Two examples of automatically generated images of the four interface residue environment classifications. Residues in the interface core are shown in orange, interface periphery in dark red, non-interface exposed surface in light blue and buried protein core in dark blue. Human interleukin-4 is shown on the left (PDB entry 1iar) and D-alanine aminotransferase on the right (PDB entry 1daa).

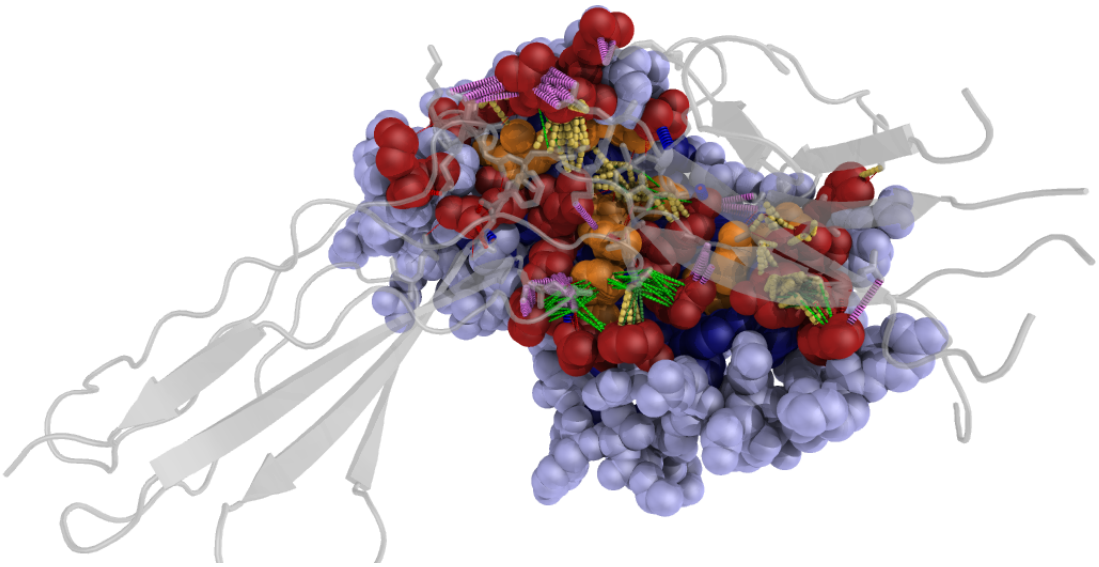


Figure 2.10: Complex of human somatotropin and the prolactin receptor (PDB entry 1bp3). Interaction types are coloured as follows: hydrogen bonds in dark blue; water mediated hydrogen bonds in light blue; π-cation interactions in green; ionic interactions in pink; hydrophobic contacts in yellow; and van der Waals in red. The same colouring scheme is used in Figures 2.11 and 2.12.

Figure 2.11: Complex of human plasmin with Streptococcal Streptokinase C (PDB entry 1bml).

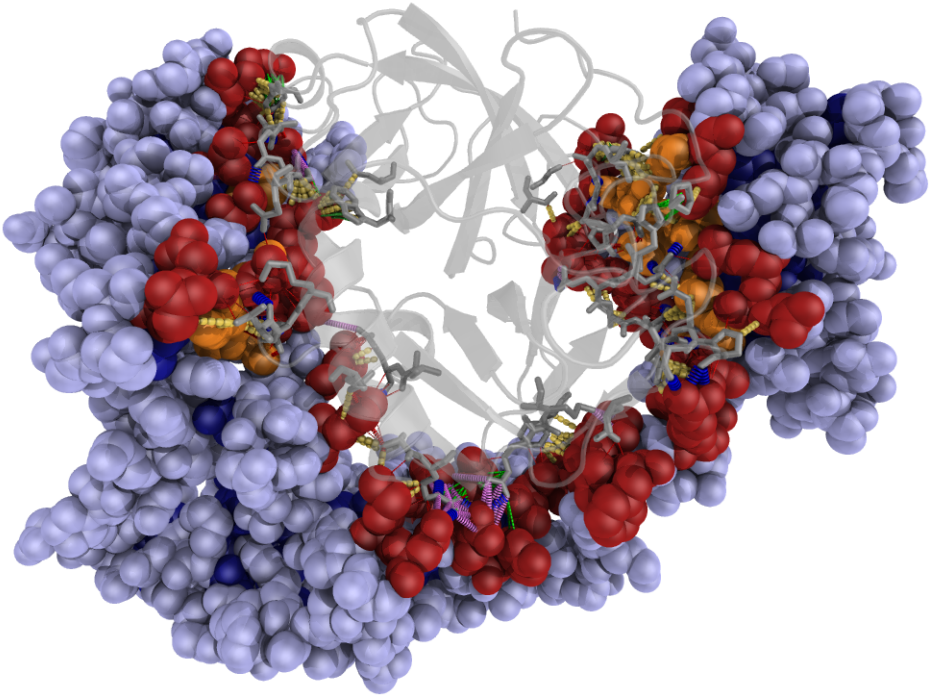
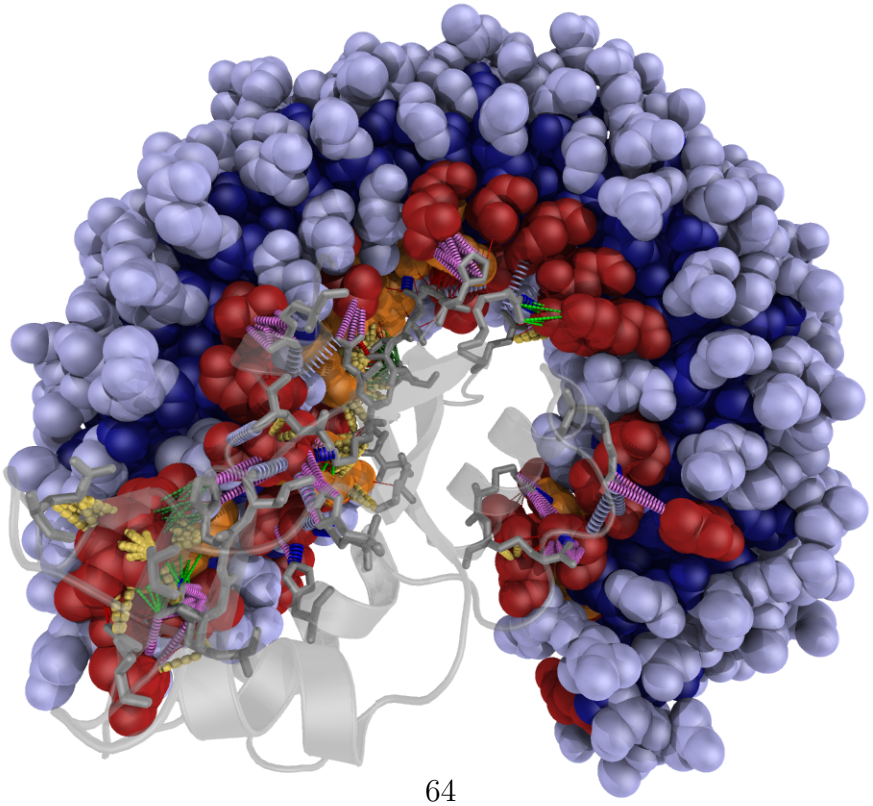


Figure 2.12: Complex of human ribonuclease inhibitor with angiogenin (PDB entry 1a4y).



2.2.15 Generating assemblies

To generate assemblies prior to building PICCOLO, three sources of quaternary structures were used: PDB Biological Units, PQS and PISA.

2.2.15.1 Biological units

Although wwPDB provides Cartesian coordinates for PDB Biological Unit assemblies in PDB format, water molecules, which are required in order to identify water-mediated hydrogen bonds, unfortunately are not transformed. Therefore the biological units must be regenerated locally. To achieve this, biological unit information is parsed from the HEADER records of the PDB format ASU files and the resulting data loaded into a database table. REMARK 300 provides a description of the biological unit in free text and REMARK 350 presents the crystallographic and non-crystallographic transformations as rotation-transformation matrices to operate on the atomic coordinates to generate the biological unit. In the example shown in Figure 2.13, four distinct transformations are applied to transform a monomeric chain in the ASU into a homotetramer. Each transformation is described by a 3×3 rotation matrix and a 3×1 transformation matrix in Å.

```
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: P
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
REMARK 350   BIOMT3   1  0.000000  0.000000  1.000000        0.00000
REMARK 350   BIOMT1   2  0.000000 -1.000000  0.000000       46.85000
REMARK 350   BIOMT2   2  1.000000  0.000000  0.000000       46.85000
REMARK 350   BIOMT3   2  0.000000  0.000000  1.000000        0.00000
REMARK 350   BIOMT1   3  0.000000  1.000000  0.000000      -46.85000
REMARK 350   BIOMT2   3 -1.000000  0.000000  0.000000       46.85000
REMARK 350   BIOMT3   3  0.000000  0.000000  1.000000        0.00000
REMARK 350   BIOMT1   4 -1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   4  0.000000 -1.000000  0.000000       93.70000
REMARK 350   BIOMT3   4  0.000000  0.000000  1.000000        0.00000
```

2.2.15.2 PQS

Transformation matrices required to generate PQS assemblies are parsed from the HEADER records of PQS records in a similar manner to the Biological Unit records above. The only difference being that in PQS records the translation matrix is stored in fractional coordinates. Before they can be used the fractional coordinates need to be scaled by the dimensions of the unit cell, which are captured from the CRYST record in the PDB header.

2.2.15.3 PISA

Unfortunately PISA does not provide Cartesian coordinates of the predicted assemblies. However, Eugene Krissinel, the PISA developer, kindly provided a URL from where XML files containing all data pertinent to the predicted assemblies could be downloaded. These XML files were parsed and loaded into the database. NMR structures naturally have no information on crystal symmetry so they are implicitly absent from PISA. As of May 2008 (using PISA software version 1.14) PISA comprised 103,779 assemblies in 92,953 assembly sets. Assembly sets may include more than one assembly in cases where multiple biological units are found in the ASU e.g. PDB 1c3h consists of two distinct homotrimers. The PISA procedure can identify multiple assembly sets for each PDB entry. Only the top-ranked assembly set i.e. the most stable, for each PDB entry was considered further, leaving 31,697 assemblies in 27,281 assembly sets (30.5% of the original assemblies). Often assemblies in the top-ranked set are not confidently predicted to be stable. 26,923 assemblies are labelled as "stable in solution" and it is this set that will be considered further - the remaining assemblies of lower levels of predicted stability are discarded.

Given that rotation-translation matrices for Biological Units, PQS and PISA assemblies are all stored in identical relational format, the same method can be used for generating assemblies for all three resources. A Python script uses the BioPython library to read in the coordinates of the ASU of each structure and apply the relevant transformations stored in the database. Prior to transformation any water molecules within 5Å of each polypeptide chain have their chain identifier set to that chain. An example is shown in Figure 2.13. Importantly

a mapping is maintained between the PDB chain identifiers in the original ASU PDB files and these newly-generated assemblies. There can be more than one biomolecule for each PDB entry.

2.2.15.4 Overlaps of Biological units vs PQS vs PISA

A requirement for a further analysis of the properties of protein interfaces is a non-redundant set of reliable quaternary structure predictions for high-quality crystal structures. For the reasons explained previously quaternary structure data are considered superior to those found in the PDB ASUs. Originally it was intended to build a non-redundant set from the union of the assemblies captured from the PDB Biological Units, PQS and PISA. The manually curated data from PiQSi allowed a brief benchmark study to be performed. PiQSi assesses the correctness of PDB Biological Units only. Therefore a quality assessment can be performed on the set of Biological Units common to either PISA or PQS or both. Commonality here was defined as those assemblies comprising precisely the same set of transformations.

This analysis revealed that the frequency of the Biological Unit being annotated as incorrect was 7.01% (568 out of 8101 assemblies) for those Biological Units that are found within the PISA set and 46.7% (858 out of 1838 assemblies) for those Biological Units that are not found within the PISA set. This significant discrepancy between the PISA and non-PISA derived assemblies clarified the selection of only PISA-derived assemblies for further analysis and derivation of a non-redundant set. The PQS and PDB Biological Unit data were not considered further.

2.2.16 PICCOLO flavours

The name PICCOLO refers to both the software program, written in Python, and the MySQL database where the results of running the program systematically are stored. The database itself comes in three flavours, with essentially identical schemas. The first is based on PDB files provided by the wwPDB i.e. representing the ASU of the crystal structure. The second is based on predicted

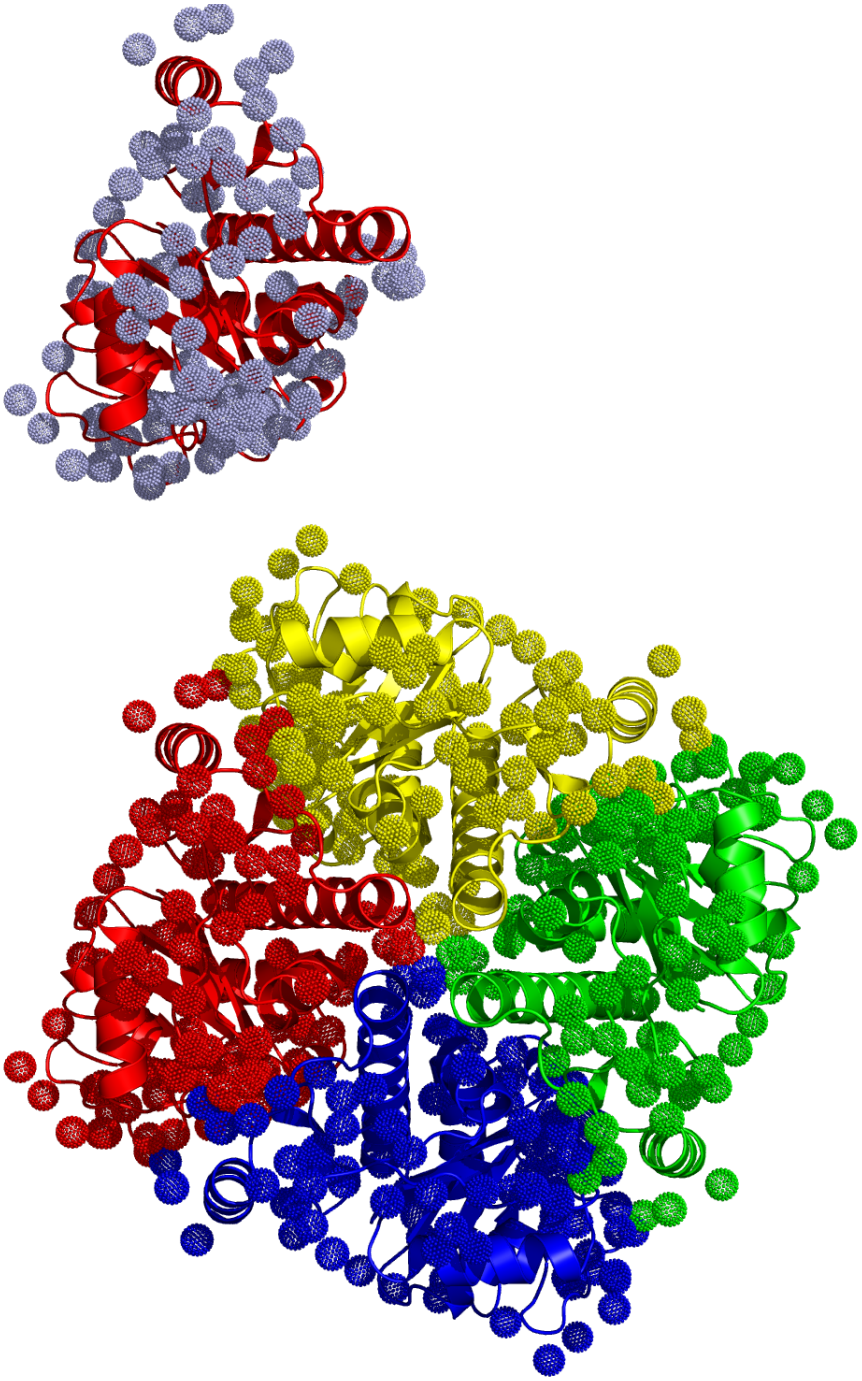


Figure 2.13: Generation of a PISA-predicted assembly. PDB entry 1dzx contains one monomer of L-fuculose-1-phosphate aldolase from *Escherichia coli* in the ASU shown in the upper panel. Application of the 4 transformations shown in the PDB excerpt on page 65 generates the homotetramer in the lower panel. Water molecules are shown coloured by their "adoptive" parent PDB polypeptide chain (see text).

	PDB ASU	PISA Quaternary Structures	Multidomain Chains
PDBs	24,944	23,718	5,809
(Assemblies)		(28,691)	
Chains	88,098	99,440	21,463
Chain pairs	99,588	130,337	12,736
Residues	5,627,364	8,012,212	783,869
Residue pairs	9,165,464	13,168,061	1,332,076
Atoms	30,682,029	44,033,147	4,400,865
Atom pairs	116,563,917	167,835,728	17,538,172

Table 2.3: Overall PICCOLO summary statistics.

quaternary structures from PISA, which are believed to be more likely to represent the physiological oligomeric state of the protein. Finally PICCOLO has also been calculated on *intra*-chain domain-domain interactions from 5,809 polypeptide chains found in SCOP that comprise more than one structural domain - the corresponding *inter*-chain domain-domain interactions can be found in the existing ASU and PISA sets. However, much of the focus of the research into protein interactions is around the physiological interaction of non-covalently bound proteins (e.g. protein docking, comparative modelling and drug discovery). As such, it is the results of running PICCOLO on the PISA-predicted quaternary structures that is of greatest interest and that will be the main focus of the subsequent analyses.

2.3 PICCOLO results

2.3.1 Database summary statistics

A summary of the number of data points in each of the three flavours of PICCOLO is shown in Table 2.3.

Energetically Significant	PICCOLO predicted	Result class	PICCOLO radial cutoff	PICCOLO molecular interactions
YES	YES	True Positives	51	46
YES	NO	False Negatives	10	15
NO	YES	False Positives	29	35
NO	NO	True Negatives	24	18
		Specificity	54.7	66.0
		Sensitivity	83.6	75.4
		Accuracy	70.2	71.1

Table 2.4: Results of benchmark of interaction detection methods.

2.3.2 Benchmark of prediction methods

The results of the benchmark of interface definition methods are shown in Table 2.4. The results suggest that the enhanced molecular interaction definitions provide a significant increase in specificity, albeit at the expense of sensitivity, which together gives a marginal increase in overall accuracy. It can be argued that data quality is more important for many applications than data coverage. Inspection of the 15 False Negatives in the set of specified molecular interactions indicates that these residues are often >10Å from one another. Therefore they must presumably either be engaged in extended indirect interactions or else be the result of some synergistic rearrangements. Similarly many of the False Positive set appear to engage in valid short-range hydrogen or ionic bonds, together suggesting that the results of this benchmark data may represent the lower bound of genuine interaction prediction performance. The overall effect of these definitions across the entire database is that 21.3% atom pairs exhibit one or more of these interaction types. This greater specificity means that only this set was considered for subsequent analyses.

2.3.3 Contribution of input sources

A simple quantitative assessment of the contribution of quaternary structures to annotations of interacting residues can be performed using the mapping of PDB residues to UniProt residues in ResMap. Figure 2.14 shows the comparison of the relative contributions of PDB ASU and PISA quaternary structure data with respect to *unique* UniProt residues. The 102,245 residues that are unique to PDB ASU contacts are likely to be dominated by non-specific crystal contacts. Conversely the 125,834 residues that are unique to the PISA-predicted quaternary structures represent a 33% increase in the number of residue annotations that would not otherwise be considered.

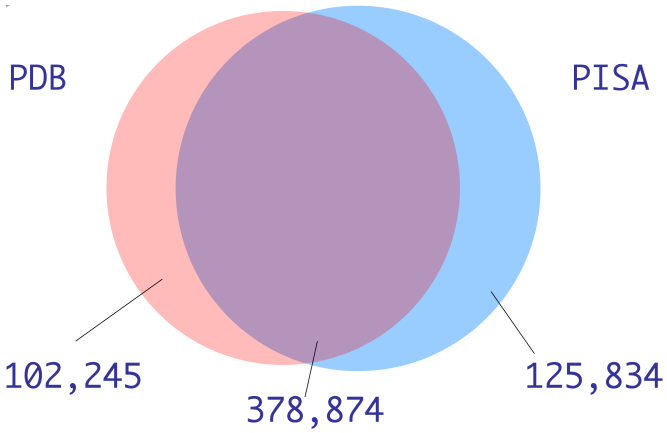


Figure 2.14: Overlap of unique UniProt residues from PICCOLO built from PDB ASU data *versus* PISA generated assemblies.

2.3.4 PICCOLO applications

The establishment of PICCOLO enables a series of analyses to be performed with the overarching goal of increasing understanding of the fundamentals of protein-protein interactions, the results of which will be reported in Chapter 3. Chapter 4 goes on to describe the properties of interfaces from an evolutionary perspective through assessment of observed substitution patterns. Aside from these systematic analyses, PICCOLO also has a number of more immediate practical applications in several areas of interest to researchers in the TLB group. In

Chapter 5 I show that PICCOLO can be used to assist in the prediction of the effects nsSNPs. PICCOLO can also be applied to unravel the principles behind "hot-spots" in protein interfaces, as described in Chapter 6.

Further applications of PICCOLO, not addressed at length in this thesis, have arisen in the areas of structural modelling of interfaces, interface site prediction and protein-protein docking. The fact that PICCOLO is non-redundant and comprehensive means that it would provide the ideal source database for identifying template structures for comparative modelling of protein-protein interactions. A wide range of predictive methods for the identification of interaction sites on the surface of proteins were highlighted in Chapter 1. PICCOLO provides an ideal benchmark set for the objective assessment of the performance of these methods.

In the field of protein-protein docking, principles derived from PICCOLO have been used to derive novel docking scoring functions. Jawon Song in the TLB group is assessing the performance of PICCOLO-derived terms describing residue contact preferences and interaction type density profiles in discriminating genuine interaction poses from docking decoys. Potential also exists in describing residues interacting across an interface as a graph, with residues as nodes and interactions as edges (see Figure 2.15 for an example) and using graph properties to derive a new scoring function.

Furthermore, in order to assess any progress in docking studies, availability of reliable benchmark sets is critical. Any such benchmark should consist of good quality structures of the complex under scrutiny and importantly both proteins solved independently in their uncomplexed form. PICCOLO is ideally suited to identify such a comprehensive benchmark set.

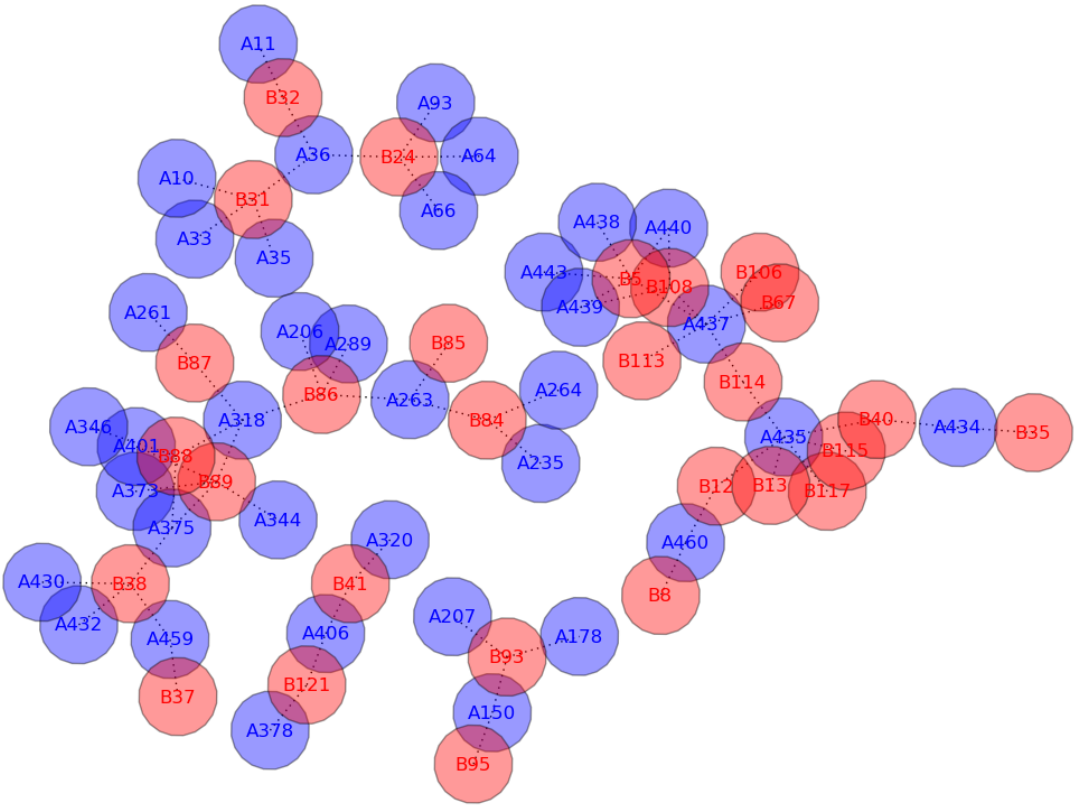


Figure 2.15: Automatically generated graph of the interface of ribonuclease inhibitor (blue) and angiogenin (red) (PDB entry 1a4y). Nodes correspond to residues, edges to contacts identified by PICCOLO. Nodes are labelled with chain identifier and residue number. The location of residue nodes does not reflect their physical location in the protein structure, rather it is determined automatically by the layout algorithm in GraphViz (Jünger (2003)) that attempts to minimize edge crossings. Representation of protein-protein interfaces as graphs makes them amenable to assessment using graph theory methods and enables such visualizations.

Chapter 3

PICCOLO analysis

3.1 Introduction

The establishment of PICCOLO as a comprehensive resource of structurally characterized protein-protein interfaces was described in Chapter 2. This Chapter describes a series of studies enabled by PICCOLO, investigating interface properties including residue propensity, hydropathy, polarity, interface size, sequence entropy and residue contact preference. The raw data from PICCOLO derive from the PDB and as such are inherently highly redundant. Before any analysis is performed it is vital to derive a non-redundant set of interfaces to avoid any

potential skewing of subsequent analyses. Protein-protein interfaces are highly heterogeneous. Chapter 1 described previous work that had discriminated different interface types and found significant distinctions in the properties of the various classes. Amongst the most important distinctions to be made are those that distinguish homo-oligomers from hetero-oligomers and obligate from transient systems. Deriving overall interface properties from PICCOLO without paying attention to the different sub-classes of interfaces risks obscuring significant underlying trends. Procedures to deal with these issues of data redundancy and heterogeneity are described. Subsequent derivation of the various interface properties are then discussed.

3.2 Methods

3.2.1 Filtering and clustering

Some of the inherent redundancies in the PDB were introduced in Chapter 2. Any analysis of interface properties not taking such biases into account is likely to be skewed by over-represented systems. Therefore before interface properties are analysed the data sets were filtered and clustered to provide a reliable non-redundant set.

The procedure of applying PISA-derived rotation-translation matrices to generate biological assemblies removes artefactual non-specific crystal packing interfaces. Despite this a small number of insignificant interfaces remain in the PISA-derived assemblies. These typically comprise only a handful of residues, and manual examination reveals they are almost exclusively due to peripheral contacts of non-neighbouring chains in high order multiprotein systems.

No chain length filtering criteria were applied prior to generation of PICCOLO. This was a deliberate choice; interactions of proteins with small peptides are of interest when considering the effects of mutations on protein function. However for the purposes of systematically deriving properties of protein interfaces, it is the interaction surfaces of globular proteins that are of most interest. Interactions of small peptidic polypeptide chains of less than 15 valid amino acid residues were therefore removed.

Figure 3.1 shows the number of residues contributed by each side of the interface (R_i and R_j). For clarity the pairs of interfaces have been ordered by size, with the chain contributing the most residues shown on the x-axis. The inset indicates a close-up of the smallest interfaces. Although a threshold of a minimum of 5 contact residues per protein ($R_i \geq 5$ and $R_j \geq 5$) was initially considered (red dashed line) this would exclude a small number of genuine interfaces. Instead the criterion that the *product* of the number of residues from each interface is greater than or equal to 25 was used ($R_i \times R_j \geq 25$) (solid red line). Collectively these filters remove 28,152 interfaces (21.6% of the original 130,336).

It would be anticipated that each side of a protein-protein interface would contribute approximately the same number of residues and this is borne out by Figure 3.1. The largest single interface, in terms of the number of residues involved, is that of homodimeric pyruvate-ferredoxin oxidoreductase (PDB entry 1kek) with more than 300 residues contributed from each partner in an extended, interdigitated surface as shown in Figure 3.2.

Typical procedures to deal with redundant data involve performing cluster analysis whereby the objects are partitioned into subsets such that the data in each agglomerated subset are co-proximal, as defined by a particular distance measure. Selection of one representative from each subset provides a non-redundant set. An example of such a procedure for clustering homologous protein sequences is described in Chapter 4. However, identifying a non-redundant set from a *pairwise* set of proteins, such as that in PICCOLO is not so straightforward. Any upstream sequence-based clustering of PDB polypeptides cannot be performed, as two protein structures with identical sequences may exist in different states: one may be complexed and the other bound; and even if both are bound, they may be bound to different partners; and even if both bind the same partner there is no guarantee the interaction surface or mode of interaction will be maintained.

For this reason the following clustering procedure was devised. All pairwise interfaces were first grouped by the unique ordered combination of UniProt identifiers of *both* component proteins. Then *within* these UniProt pair clusters, each cluster member pair was compared to all other cluster member pairs and the overlap of unique UniProt residue numberings (pre-calculated and stored in the

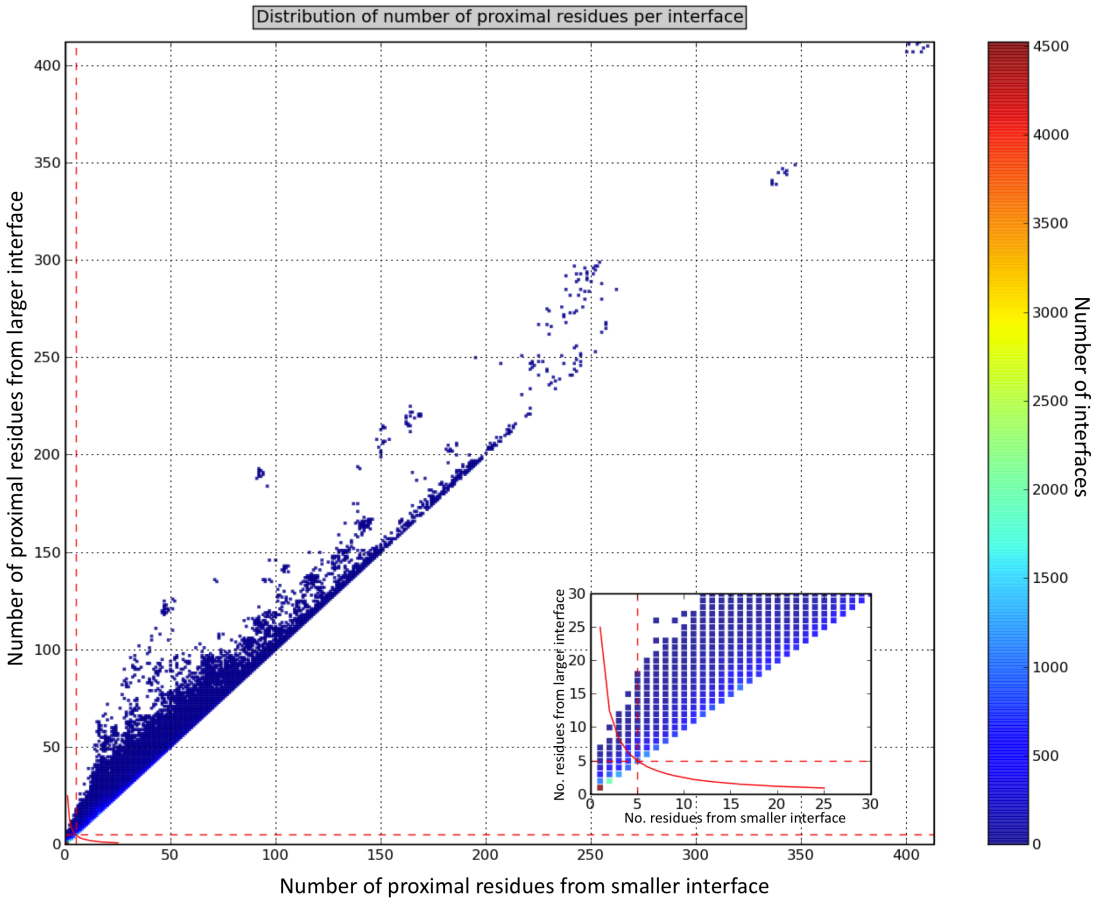


Figure 3.1: Scatter plot of the number of residues contributed by the larger side of each PICCOLO interface (R_i on y-axis) against the number of residues contributed by the smaller side (R_j on x-axis). Colour indicates the total number of interfaces at each point, reflecting the fact that many interfaces share the same number of contributing residues. The red dashed line indicates a threshold of a minimum of 5 contact residues per interface ($R_i \geq 5$ and $R_j \geq 5$) that was initially considered. The solid red line indicates a threshold where the product of the number of residues from each interface is greater than or equal to 25 ($R_i \times R_i \geq 25$) that was used. The inset shows a close-up of the lower left corner of the larger plot, highlighting the smallest interfaces.

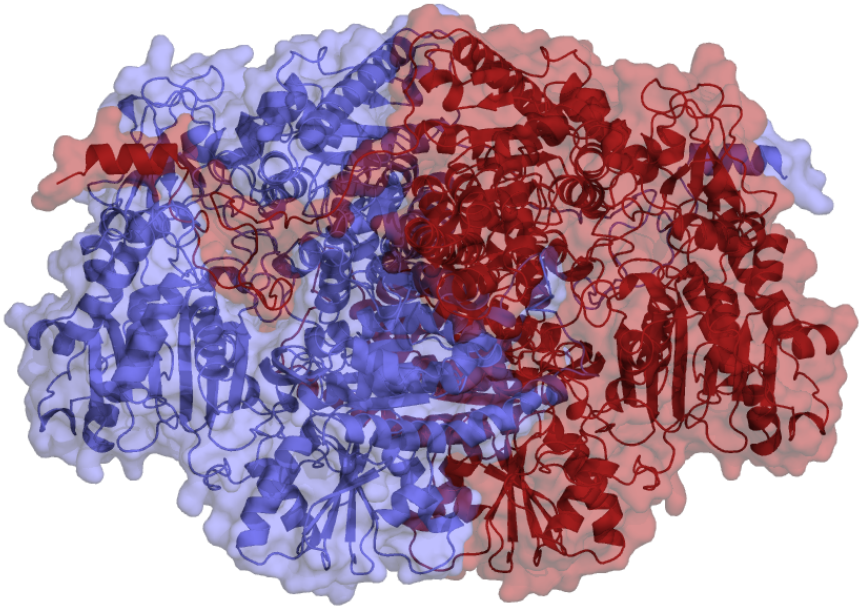


Figure 3.2: Cartoon representation of the homodimeric interface of pyruvate-ferredoxin oxidoreductase (PDB entry 1kek) with more than 300 residues contributed by each surface is the single largest interface in PICCOLO. Chain A is shown in blue and chain B in red. Figure generated using PyMOL (Delano (2002)).

ResMap table) for both constituents was assessed reciprocally. If *both* sides of the interface share more than 75% of unique residue positions in common with another pairwise interaction then the interfaces were co-clustered. 75% was chosen as a sparsely populated region that gave good separation of some manually selected test cases.

In order to choose representatives to form the non-redundant set, rather than simply choose an arbitrary member of each cluster, the representative complex for each cluster was chosen as that complex whose mean QScore of the two constituent chains was highest (QScore is a property of each polypeptide chain as it depends partly on the number of missing residues). Note that this process results in a non-redundant set of interfaces, not oligomeric assemblies.

3.2.2 Partitioning interface types

Classification of interactions as being either homo- or hetero-oligomers was relatively straightforward, given the information stored in the PDBRes table concerning UniProt sequence identifiers with respective assignment of relevant sequence boundaries. Partitioning interfaces as being either transient or permanent was less straightforward. Manually curated sets are available from the literature but they suffer from the drawback that they are inherently small in size and may not be representative.

An alternative approach was to use NOXclass, a published, automated protein-protein interaction classification algorithm (Zhu *et al.* (2006)). The algorithm involves a support vector machine (SVM) algorithm to partition interfaces as either biological obligate, biological non-obligate or non-specific crystal packing. The discrimination is based on the following interface properties: interface area (calculated by NACCESS (Hubbard (1993))), interface area ratio, area-based residue composition, correlation between area-based residue compositions of interface and non-interface surface, gap volume index and conservation score. The developers assessed various approaches and parameter sets, concluding the most accurate method was a multi-stage classifier that separates data progressively (firstly biological interaction versus crystal packing contact, followed by obligate versus non-obligate) using three parameters describing interface area, interface

area ratio, and area-based residue composition interfaces. The authors claim an accuracy of 91% for the classification of three types of interactions using a leave-one-out cross-validation procedure. The NOXClass program was run on the set of non-redundant representative PICCOLO interfaces that resulted from the clustering procedure (14,658 interfaces). The results suggested that 36% of the interfaces were non-specific crystal contacts. Given that most non-specific crystal contacts were already removed through use of PISA biological assembly data and interface filtering, coupled with some manual inspection of the predictions, it was concluded that this result was unlikely to be genuine. Furthermore, when applied to the published data used to train the NOXClass algorithm, 87% of obligate interfaces and 68% of transient interfaces were correctly predicted. As this was the training set, these values are likely to represent the upper bound of performance. Unfortunately therefore, the results of NOXClass do not appear to be sufficiently reliable to be considered further.

The alternative option of using literature-derived data sets was therefore used. Two published sets were downloaded and a consensus set of the two produced. Zhu *et al.* (Zhu *et al.* (2006)) derived their training data of 75 obligate and 62 non-obligate interfaces used in generating NOXClass (http://noxclass.bioinf.mpi-sb.mpg.de/trainingdata_bncpcs.htm) from Bradford and Westhead (Bradford & Westhead (2005)) and Neuvirth (Neuvirth *et al.* (2004)). Mintseris and Weng (Mintseris & Weng (2005)) provide a set of 115 obligate (http://zlab.bu.edu/julianm/obligate.txt) and 212 transient (http://zlab.bu.edu/julianm/transient.txt). Consolidating these sets, and incorporating UniProt information from PDBRes to flag them as homo- or hetero-oligomers, results in the data shown in Table 3.1. The transient Homo-oligomer set was too small to be considered further.

3.2.3 Physico-chemical properties

Interfaces were characterized by various physico-chemical properties including hydropathy, polarity and number of interactions. Values for interface hydropathy were estimated by applying the experimentally observed values from the Kyte and Doolittle hydropathy index (Kyte & Doolittle (1982)). Interface polarity

Hetero or Homo	Obligate or Transient	Number of Interfaces
Hetero-oligomer	Obligate	122
Hetero-oligomer	Transient	171
Homo-oligomer	Obligate	60
Homo-oligomer	Transient	2

Table 3.1: Number of each interface type in the literature-derived consolidated set.

was estimated by considering the number of nitrogen, oxygen and sulphur atoms that engage in molecular interactions. Frequency counts of interaction types can also provide information that can be used for discriminating different interaction types. As polarity and interaction type counts are to some extent dependent on the size of the interface concerned, they were each normalized by the size of the interface, as calculated by NACCESS, in order to enable valid comparison of interfaces of different sizes.

3.2.4 Residue propensity

Previous studies on residue propensities in protein-protein interfaces have drawn somewhat contradictory conclusions (Ansari & Helms (2005); Jones & Thornton (1996); Ofran & Rost (2003); Ponstingl *et al.* (2005); Yan *et al.* (2008)). However much of these differences can be attributed to differences in data sets, interface definition, source of background frequency data and differing approaches to partitioning of interaction types. Importantly many studies do not distinguish between different anatomical regions of the interface region. In this study the interface core and periphery are distinguished based on solvent accessibility, as described in Chapter 2. Here, background residue frequency (B_i) is defined, independently of structural environment, as follows for each residue type i:

$$B_i = \frac{F_i}{\sum F_i} \tag{3.1}$$

where F_i is the count of each residues type calculated using all residues found in PICCOLO structures, not just interface residues. The environment-dependent

residue frequency (E_i) is defined for each combination of structural environments e and residue types i:

$$E_i = \frac{F_{ie}}{\sum F_{ie}} \tag{3.2}$$

The normalized environment-dependent propensity (R_{ie}) is then the ratio of the environment-dependent frequency (E_i) to the background frequency (B_i):

$$R_{ie} = \frac{E_{ie}}{B_i} \tag{3.3}$$

3.2.5 Sequence entropy

Interfaces can also be characterized from an evolutionary point of view. Identification of residues that are conserved across a multiple alignment of evolutionarily related protein sequences is a widely used approach to probe protein function, as highly conserved residues tend to correlate with those that carry structural or functional importance. Chapter 4 describes procedures to generate a library of structural alignments of homologous protein domains, in the form of the relational database TOCCATA. Combination of the structural data on protein interfaces from PICCOLO with evolutionary data from TOCCATA allows the evolutionary properties of interfaces to be probed. TOCCATA alignment sets are generated at a range of redundancy levels (fully redundant, 95%, 90%, 70%, 50% and 30% - see Chapter 4 for more details). Choice of which alignment set to use involves a trade-off between maximizing data coverage and minimizing data redundancy. The 95% redundancy set was chosen, in which no two sequences have greater than 95% identity. Note that the clustering procedure used to generate the different redundancy levels is entirely sequence-based with the result that cluster representatives may be unbound structures absent from PICCOLO. Therefore in order to link the two databases, all residues from structures found in PICCOLO are first mapped to their identical counterparts in the fully redundant alignments. These are then mapped to their 95% representatives through the redundant alignments which are then themselves extracted from the 95% alignment set. The upshot of this procedure is that the TOCCATA structural domains used for the evolutionary analysis may not themselves be engaged in interactions, but they are identical

(or almost identical) to another structural domain that is. Alignments comprising fewer than five structures are not considered and neither are alignment column positions where more than 50% of elements are gaps.

Although many different measures of residue conservation have been proposed (Valdar (2002)), Shannon's entropy is amongst the most widely-used measures. The entropy score for each aligned column in a multiple sequence alignments can be expressed as:

$$S_{entropy} = -\sum p_i \log_2 p_i \tag{3.4}$$

where p_i represents the observed frequency of residue type i in the aligned column.

However, although widely used, Wang and Samudrala (Wang & Samudrala (2006)) pointed out that Shannon entropy does not incorporate background amino acid frequencies, and as such is sub-optimal for assessing conservation. They proposed a modified entropy term incorporating background frequencies of each residue type:

$$S_{relative-entropy} = -\sum p_i \log_2 (\frac{p_i}{p_{ib}}) \tag{3.5}$$

where p_{ib} represents the background amino acid frequency found in naturally occurring proteins.

Wang and Samudrala found that including such background frequency information significantly improved the performance of functional site prediction. Intuitively, an invariant tryptophan (∼1.1% natural abundance) is more likely to be significant than an invariant leucine (∼9.6% natural abundance) (http://expasy.org/tools/pscale/A.A.Swiss-Prot.html). The Shannon entropy score would be the same for these two circumstances, whereas the relative entropy measure assigns a higher score to the invariant tryptophan.

Both entropy measures will be used in this study. The background frequencies chosen for the $S_{relative-entropy}$ measure were derived from the entire 95% redundancy set. Frequencies derived from solved structures are likely to differ from overall proteomic background frequencies as trans-membrane segments,

low-complexity sequences and natively disordered peptides are likely to be under-represented in crystal structures.

3.2.6 Contact pairing preferences

The frequency of pairwise residue interactions (P_{ij}) can be derived for the PICCOLO derived non-redundant set:

$$P_{ij} = \frac{C_{ij}}{\sum C_{ij}} \tag{3.6}$$

where C_{ij} represents the number of times residue type i is observed engaging in contacts across the interface with residue type j.

The individual frequencies (W_i) reflecting the amino acid composition of each residue type i can be defined as:

$$W_i = \frac{F_i}{\sum F_i} \tag{3.7}$$

where F_i represents the number of residues engaged in contacts.

If interfacial amino acid residues exhibit no preference as to which residues they contact across the interface, the expected frequency of any particular residue-pair interaction would be simply the product of the two individual residue frequencies ($W_i \times W_j$).

Any such interaction preference can be quantified by calculating the log odds ratio of the *observed* interaction frequency to the *expected* interaction frequency:

$$L(i,j) = \log_2(\frac{P_{ij}}{W_i W_j}) \tag{3.8}$$

This measure is commonly used (Moont *et al.* (1999)) but it does not take into account differing residue sizes (intuitively larger residues have greater surface area and therefore greater opportunity to interact with one another). Glaser *et al.* (Glaser *et al.* (2001)) used residue volume data to normalize the expected frequencies only. In this study we normalize both the expected and observed frequencies using ASA data for each residue from NACCESS (Hubbard (1993)). Thus, the propensity of residue-residue contacts, $L(i,j)$, is defined as in Equation 3.8, but with P_{ij} and W_i replaced as follows:

$$P_{ij} = \frac{C_{ij} \times ASA_i \times ASA_j}{\sum C_{ij} \times ASA_i \times ASA_j} \tag{3.9}$$

$$W_i = \frac{F_i \times ASA_i}{\sum F_i \times ASA_i} \tag{3.10}$$

$$L(i,j) = \log_2(\frac{P_{ij}}{W_i W_j}) \tag{3.11}$$

3.3 Results

3.3.1 Number of subunits per assembly

Figure 3.3 shows the number of subunits per assembly for both the PDB ASU data in red, and the PISA predicted quaternary structures in blue. The PISA predicted assemblies are marginally larger (mean 3.61 subunits) than PDB ASU assemblies (mean 3.53 subunits) but they do have a greater density of interactions: on average each PISA subunit interacts with 3.05 other subunits, whereas PDB ASU subunits interact with 2.26 others.

3.3.2 Interface clustering

Figure 3.4 shows the data used to perform the interface clustering. Interfaces sharing the same ordered pair of UniProt identifiers are grouped, and within each group interfaces are compared all against all with respect to their set of UniProt residue mappings. The axes in Figure 3.4 represent the percentage overlap of each side of the interface. Data in the upper right corner, where both sides of the interfaces overlap by more than 75% are co-clustered. If either side of the interface overlaps by less than this threshold the interfaces are not co-clustered. Through this procedure each unique pair of UniProt proteins can have multiple interface-regional clusters that are identified by their interface cluster serial identifier. This is best illustrated by the example of $\alpha_2\beta_2$ haemoglobin in Figure 3.5. The clustering procedure successfully discriminates between the two regional interfaces between α and β haemoglobin (shown in orange and purple) while at

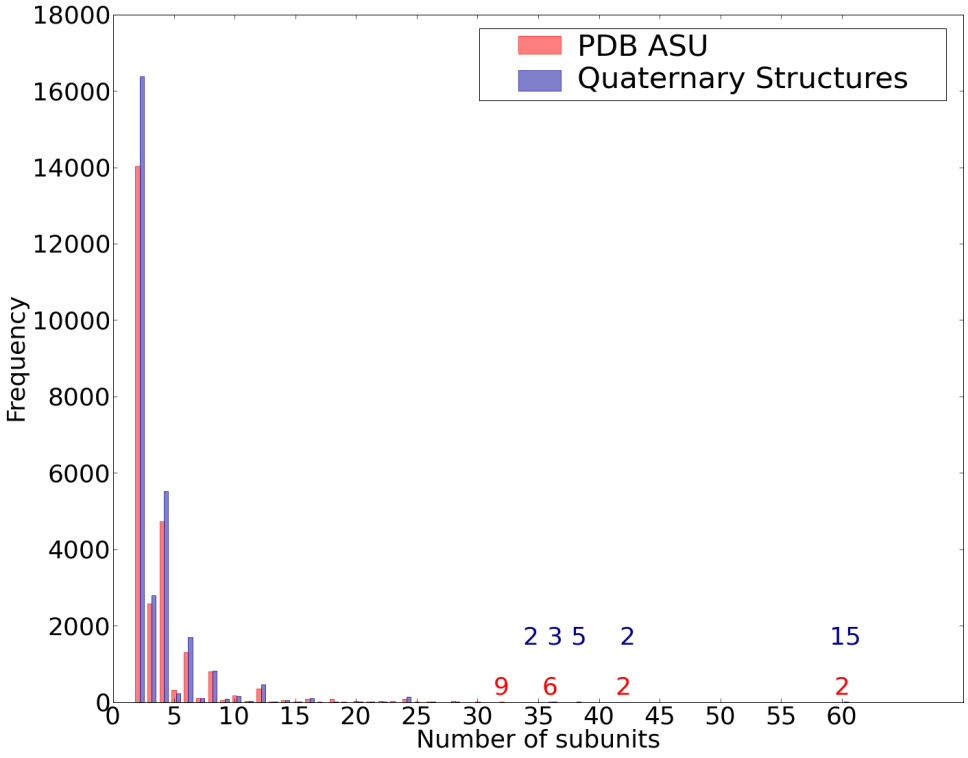


Figure 3.3: Histogram of the number of subunits in assemblies from PDB ASU (red) versus PISA predicted quaternary structures (blue). Inset numbers reflect the frequencies of those assemblies comprising more than 30 subunits.

the same time successfully groups the two symmetry-related versions of the two regional interfaces.

The clustering procedure began with 91,651 interfaces - the result of the initial interface size filtering. If interfaces were clustered only by the criterion of unique UniProt pair there would be 9,060 clusters. However the combined UniProt pair and interface regional clustering results in a non-redundant set of 14,658 cluster representatives, of which 12,227 are homodimers and 2,381 are heterodimers. Figure 3.6 shows the distribution of cluster sizes. The 20 largest clusters are shown in Table 3.2. The set of non-redundant representatives is used for all subsequent analyses.

3.3.3 Contribution of each structural environment

The pie chart in Figure 3.7 shows the numbers of residues found in each of the structural environments in the non-redundant set of 14,658 interfaces from 9,142 structures.

3.3.4 Interface solvent accessibility

Figure 3.8 shows the distribution of interface sizes. The top panel shows the distribution of the complete non-redundant set of interfaces. The remaining panels show the distributions for obligate homodimers, obligate heterodimers and transient heterodimers respectively. Obligate homodimers have the largest interfaces followed by obligate heterodimers and then transient heterodimers. These results suggest somewhat larger average interface sizes than previously published results (Janin & Chothia (1990); Jones & Thornton (1996)). Chothia and Janin suggest a standard size of 1600 $\pm$ 400Å^2. The likely explanation is that as protein biochemistry and crystallographic techniques improve, the structures of larger complexes become increasingly solvable. Note that in some studies the interface area is defined as *half* of the difference between the sum of the accessible surface area of the complex and the unbound constituents (Jones & Thornton (1996)).

Residue level information on solvent accessibility is also stored in PICCOLO. The question of whether different amino acid residues exhibit distinguishable patterns of exposure can be probed by comparing, for each residue type, the ratio of

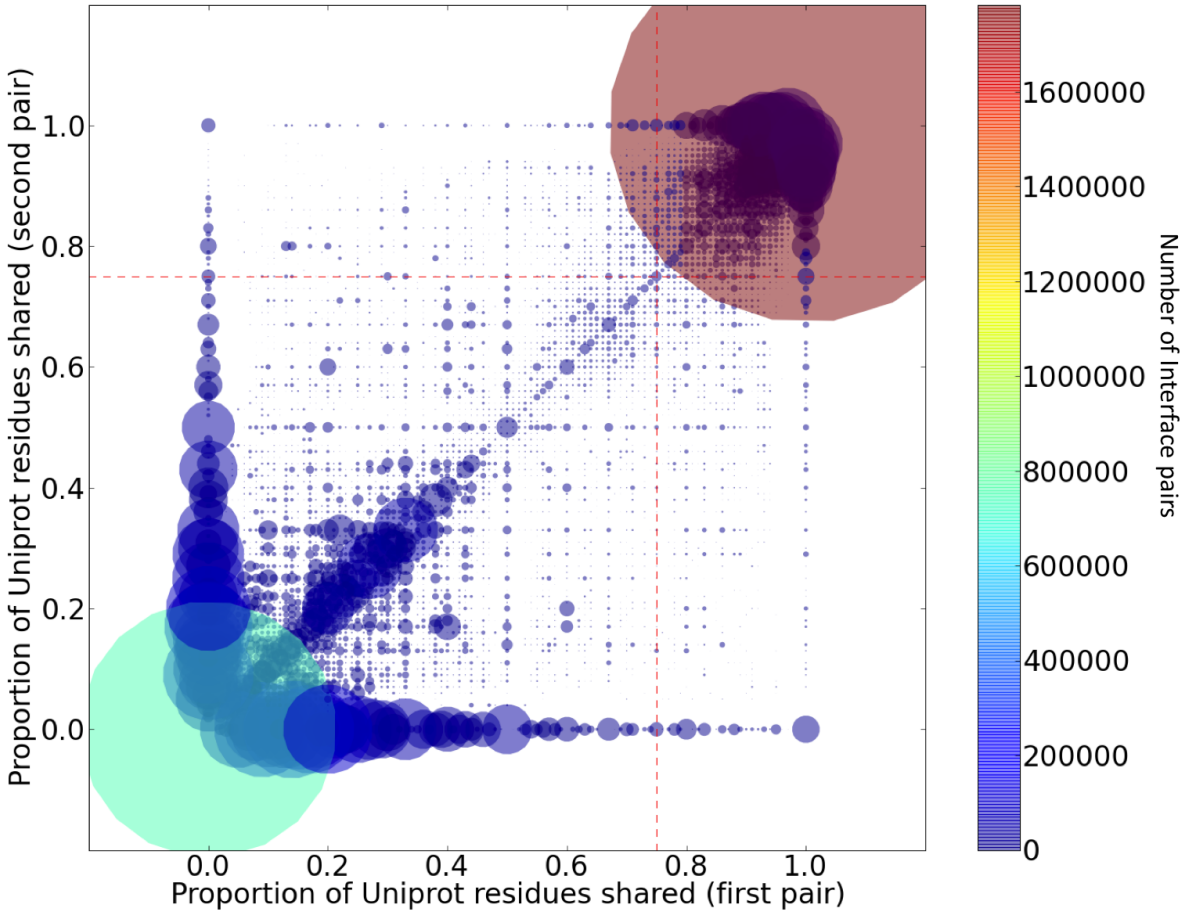


Figure 3.4: Scatter plot describing the interface clustering procedure. Every pair of PDB chains that share the same pair of parent UniProt identifiers is compared. Each point reflects the percentage overlap of each side of the interface with respect to common UniProt residues. The size and colour of each point reflects the number of interface pairs sharing that location. The red dashed line indicates the 75% interface overlap threshold - interface pairs passing this threshold on both sides of the interface are clustered. The vast majority of interface comparisons result in either zero overlap on either side (lower left hand corner) or complete overlap on both sides (upper right hand corner).

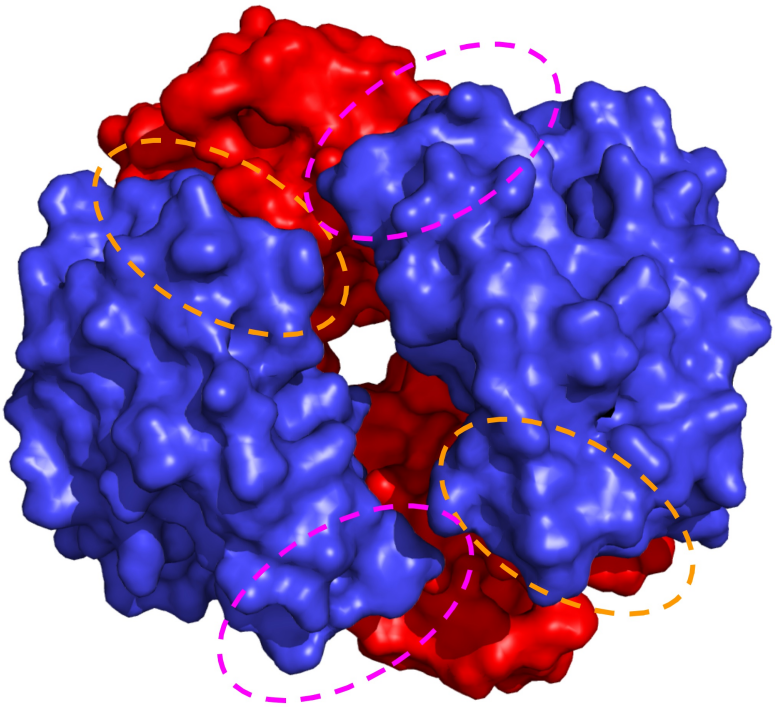


Figure 3.5: An example of how a single pair of UniProt proteins can have multiple interface-regional clusters identified by their interface cluster serial identifier. The clustering procedure successfully discriminates the two regional interfaces between α and β haemoglobin, shown in orange and purple, while at the same time successfully groups the two symmetry-related versions of the two regional interfaces (PDB entry 1y4v).

Complex	Cluster Index	Cluster Size
Complex of haemoglobin α and β subunits (P69905 and P68871)	4	329
6,7-dimethyl-8-ribityllumazine synthase homodimer (O66529)	3	280
Complex of haemoglobin α and β subunits (P69905 and P68871)	2	275
Insulin homodimer (P01308)	1	264
Streptavidin homodimer (P22629)	6	263
Aspartate carbamoyltransferase catalytic chain homodimer (P0A786)	3	255
Complex of protocatechuate 3,4-dioxygenase α and β chains (P00436 and P00437)	3	252
3-dehydroquinate dehydratase homodimer (P15474)	1	252
Streptavidin homodimer (P22629)	3	251
6,7-dimethyl-8-ribityllumazine synthase homodimer (O66529)	1	250
Complex of aspartate carbamoyltransferase catalytic and regulatory chains (P0A786 and P0A7F3)	1	242
Ferritin light chain homodimer (P02791)	1	240
Ferritin light chain homodimer (P02791)	2	240
Coat protein homodimer (Q9EB06)	1	240
Prothrombin homodimer (P00734)	12	228
Dihydrolipoyllysine-residue acetyltransferase homodimer (P10802)	2	216
Coat protein homodimer (Q9EB06)	3	193
Ferritin heavy chain homodimer (P02794)	2	192

Table 3.2: 20 largest interface clusters. Heterodimers are shown in pale grey. Cluster index reflects cases where interacting partners share multiple interfaces (index numbering is arbitrary).

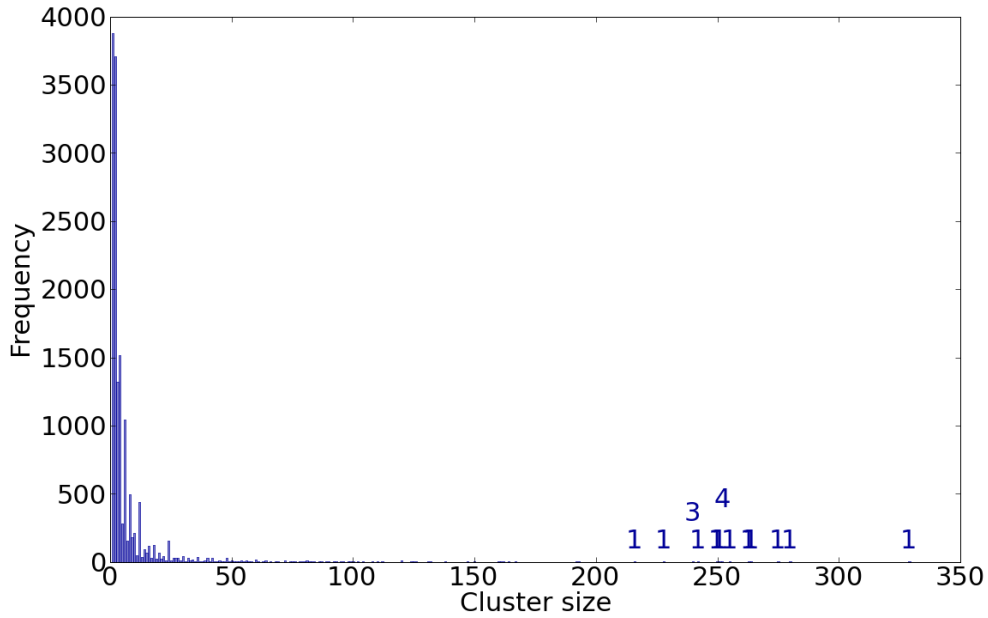


Figure 3.6: Distribution of cluster sizes resulting from the pairwise redundancy filtering.

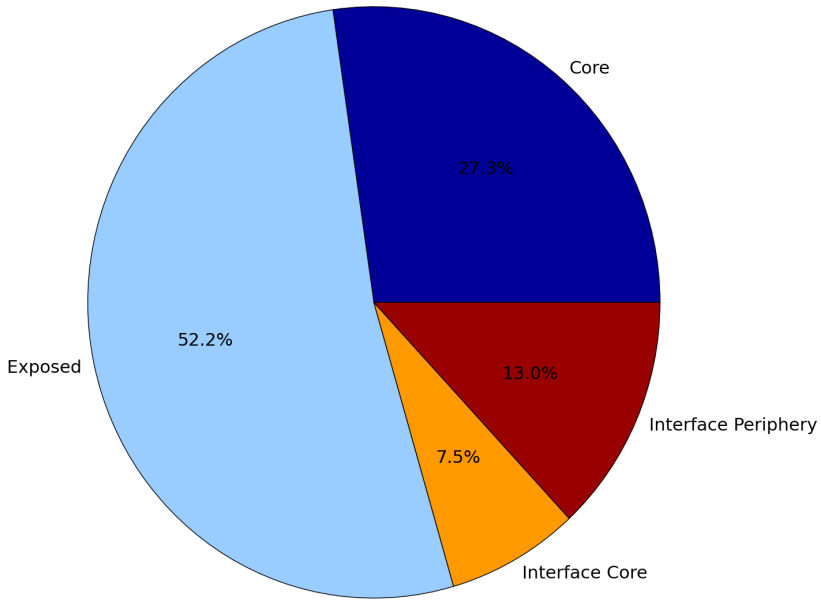


Figure 3.7: Relative contribution of the four structural environment classifications for 5,786,031 residues in the non-redundant set.

area buried on complex formation to area exposed in the unbound form. Figure 3.9 shows the average values for each residue type in the interface core and interface periphery. In the interface core there is a strong dependence of average area buried on the residue size, with the smaller residues having a greater proportion of their exposed surface buried in the interface. In the interface periphery this is not evident, instead the charged residues seem to bury the least proportion of their superficial surface, suggesting that they may be found at the extreme periphery of the interface region where they can interact with solvent.

3.3.5 Interface hydropathy

Table 3.3 shows the hydropathy of each structural environment for each of the different interface classes as calculated by summing values from the Kyte and Doolittle hydropathy index.

These data suggest that across all interface types the protein core is consistently the most hydrophobic environment, followed by the interface core. The in-

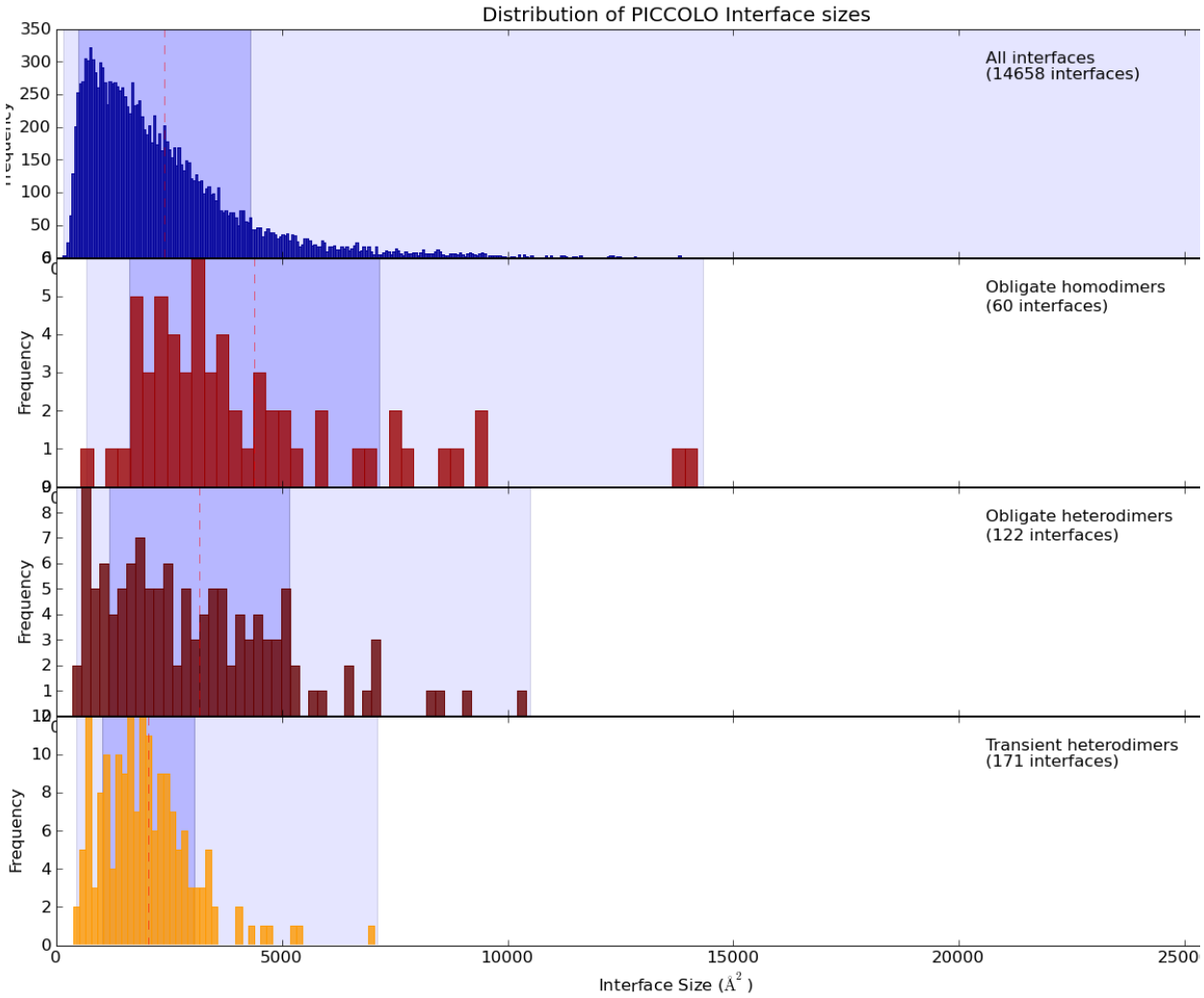


Figure 3.8: Interface size distribution for different interface types. The red dashed line indicates the mean of each distribution, the blue region one standard deviation either side of the mean and the pale blue the range of the distribution.

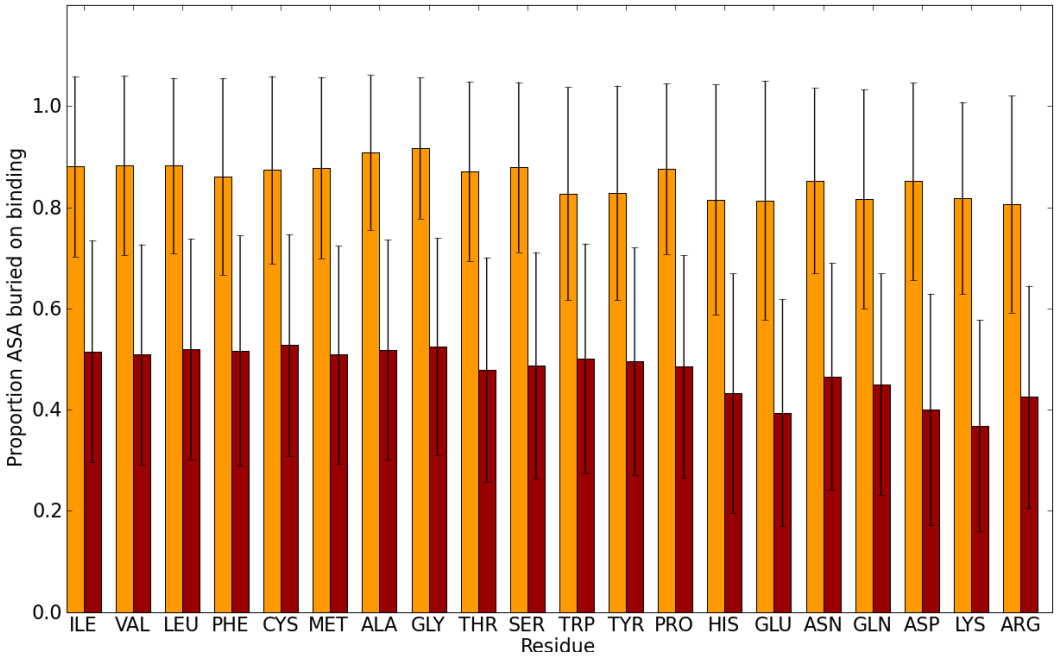


Figure 3.9: Mean proportion of the solvent exposed surface of each residue that is buried on binding, for each residue in the Interface Core (orange) versus the periphery (dark red).

		Residues	Hydropathy	
			Mean	S.D.
Overall	Core	1,578,895	1.62	2.64
	Exposed	3,020,585	-1.08	2.86
	Interface Periphery	754,667	-1.10	2.98
	Interface Core	431,884	0.83	2.93
Obligate Homodimers	Core	10,589	1.49	2.70
	Exposed	18,110	-1.13	2.83
	Interface Periphery	3,487	-1.10	2.92
	Interface Core	2,812	0.72	2.95
Obligate Heterodimers	Core	18,377	1.35	2.72
	Exposed	31,385	-0.92	2.89
	Interface Periphery	10,272	-1.03	2.95
	Interface Core	6,474	0.33	2.94
Transient Heterodimers	Core	17,667	1.69	2.61
	Exposed	37,078	-1.08	2.80
	Interface Periphery	8,031	-1.28	2.89
	Interface Core	4,431	0.52	2.93

Table 3.3: Hydropathy of anatomical regions of different interface types.

terface peripheries are the least hydrophobic regions in the heterodimers whereas in obligate homodimers the hydrophobicity of the interface periphery is comparable to that of the remainder of the surface. Of the interface cores those in obligate homodimers are more hydrophobic than transient heterodimers and obligate heterodimers. However the fact that the mean hydrophobicity of all of the hand-selected sets is less than the overall average suggests that these sets may not be fully representative of the larger non-redundant sets. The large standard deviations mean that all of these results should be treated with caution.

3.3.6 Interface polarity

Interface polarity can be measured in terms of the number of polar atoms engaging in interactions for each 100Å^2 of interface surface. However, when comparing the polarity of different structural environments it is important to take into account the differing densities of all interacting atoms in the different structural environments. Table 3.4 shows the number of polar atoms per 100Å^2 and for comparison the total number of interacting atoms per 100Å^2.

The numbers of polar atoms per unit area are higher in the interface core than the periphery for all interface types. However, this is a result of the differing atomic densities in the structural environments. For all interface types there is significantly higher interacting atom density in the interface core than the periphery - an inevitable result given the definitions used and the fact that atoms in the interface core have more opportunities to engage in interactions. The ratio of polar atoms per unit area to total atoms per unit area is perhaps more revealing. For each interface type the interface core is less polar than the interface periphery. Obligate homodimers have the highest atomic density and lowest polarity, followed by the hetero-transient interfaces then homo-transient interfaces. However, given the large standard deviations, the results must be treated cautiously.

The results of the interface polarity analysis are congruent with the parallel hydropathy analysis; the two phenomena are really two sides of the same coin. In general much of the free energy of binding comes from non-polar, hydrophobic

		Mean number polar atoms/ 100Å^2 (S.D.)	**Mean number all atoms/ 100Å^2 (S.D.)**	**Mean proportion polar atoms**
Overall	Interface Periphery	0.72	2.15	0.34
		(0.45)	(1.27)	(0.09)
	Interface Core	1.43	5.24	0.29
		(3.03)	(5.79)	(0.16)
Obligate Homodimers	Interface Periphery	1.22	3.56	0.34
		(0.31)	(0.69)	(0.06)
	Interface Core	2.01	7.85	0.25
		(0.79)	(1.78)	(0.08)
Obligate Heterodimers	Interface Periphery	0.71	2.11	0.34
		(0.44)	(1.25)	(0.08)
	Interface Core	1.42	5.00	0.29
		(1.09)	(3.25)	(0.11)
Transient Heterodimers	Interface Periphery	0.88	2.53	0.36
		(0.49)	(1.31)	(0.08)
	Interface Core	1.93	6.74	0.29
		(1.48)	(4.29)	(0.13)

Table 3.4: Polarity of anatomical regions of different interface types.

interactions. The gain in free energy from polar hydrogen and ionic interactions across an interface typically only compensates for the free energy lost from favourable interactions from solvent. Instead, such interactions often provide the required specificity through complementary polar contacts. This subjects transient protein assemblies to opposing evolutionary selective pressures with respect to their hydrophobicity. Whilst they require sufficient hydrophobicity to generate binding affinity, an excess of solvent exposed hydrophobic residues risks non-specific protein association and its potentially fatal consequences (Calloni *et al.* (2005)). Obligate assemblies are not typically exposed to solvent during their lifetime and are therefore not subject to such constraints to the same degree.

3.3.7 Interactions per unit area

The numbers of each of the major interaction types, normalized by area, are shown in Figures 3.10, 3.11, 3.12 and 3.13. The data are derived by dividing the sum of interaction counts for each chain pair by the interface size as calculated by NACCESS. Figure 3.10 shows the overall data for the non-redundant set of PISA-predicted interfaces. Figures 3.11, 3.12 and 3.13 show the same information for obligate homodimers, obligate heterodimers and transient heterodimers respectively. Red dashed lines indicate the mean of the distribution. The central shaded region indicates one standard deviation either side of the mean. The pale shaded region is the range of the distribution. The numbers in brackets are the percentage of interfaces having non-zero values for this interaction type. Only 79% of PDB structures contain co-ordinates of water molecules. When only structures that contain water are considered, the number of water-mediated hydrogen bonds rises to 0.305 (81.6% of interfaces have 1 or more water-mediated hydrogen bond). These data must be interpreted carefully, as a pair of residues engaging, for example, in an ionic or aromatic interaction, may have several atoms multiply interacting in a combinatorial fashion which can result in a significant proportional increase in the interaction counts. Conversely, hydrogen-bonds and water-mediated hydrogen bonds, for example, are calculated as an interaction between specific donor and acceptor atoms so do not contribute combinatorially in this manner.

Figure 3.10: Figures 3.10, 3.11,3.12 and 3.13 show frequency distributions of the numbers of each of the major interaction types per unit area. The red dashed lines indicate the mean of the distribution. The central shaded region indicates one standard deviation either side of the mean. The pale shaded region is the range of the distribution. The numbers in brackets are the percentage of interfaces having non-zero values for this interaction type. Interaction types per unit area for the overall non-redundant set.

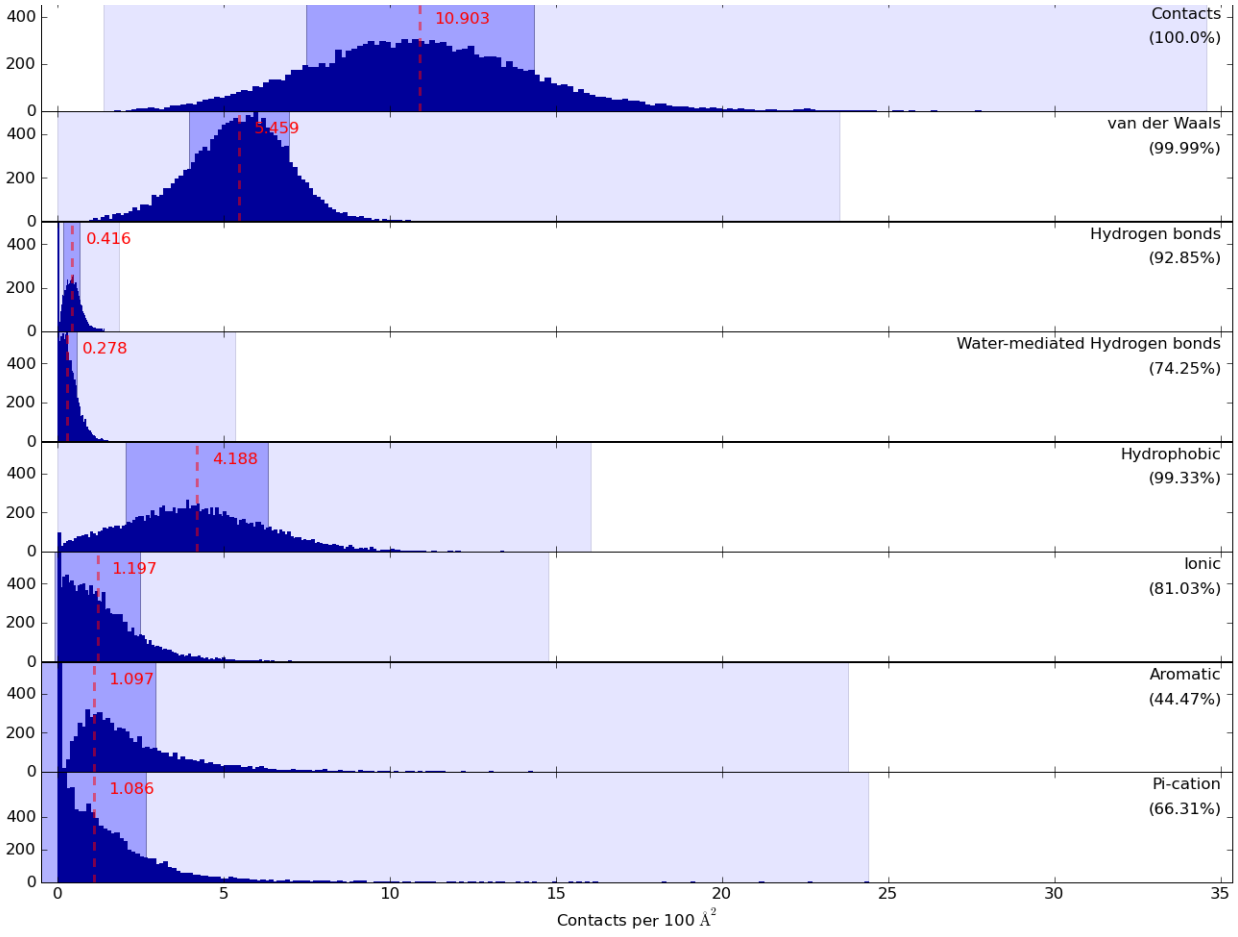


Figure 3.11: Interaction types per unit area for obligate homodimers.

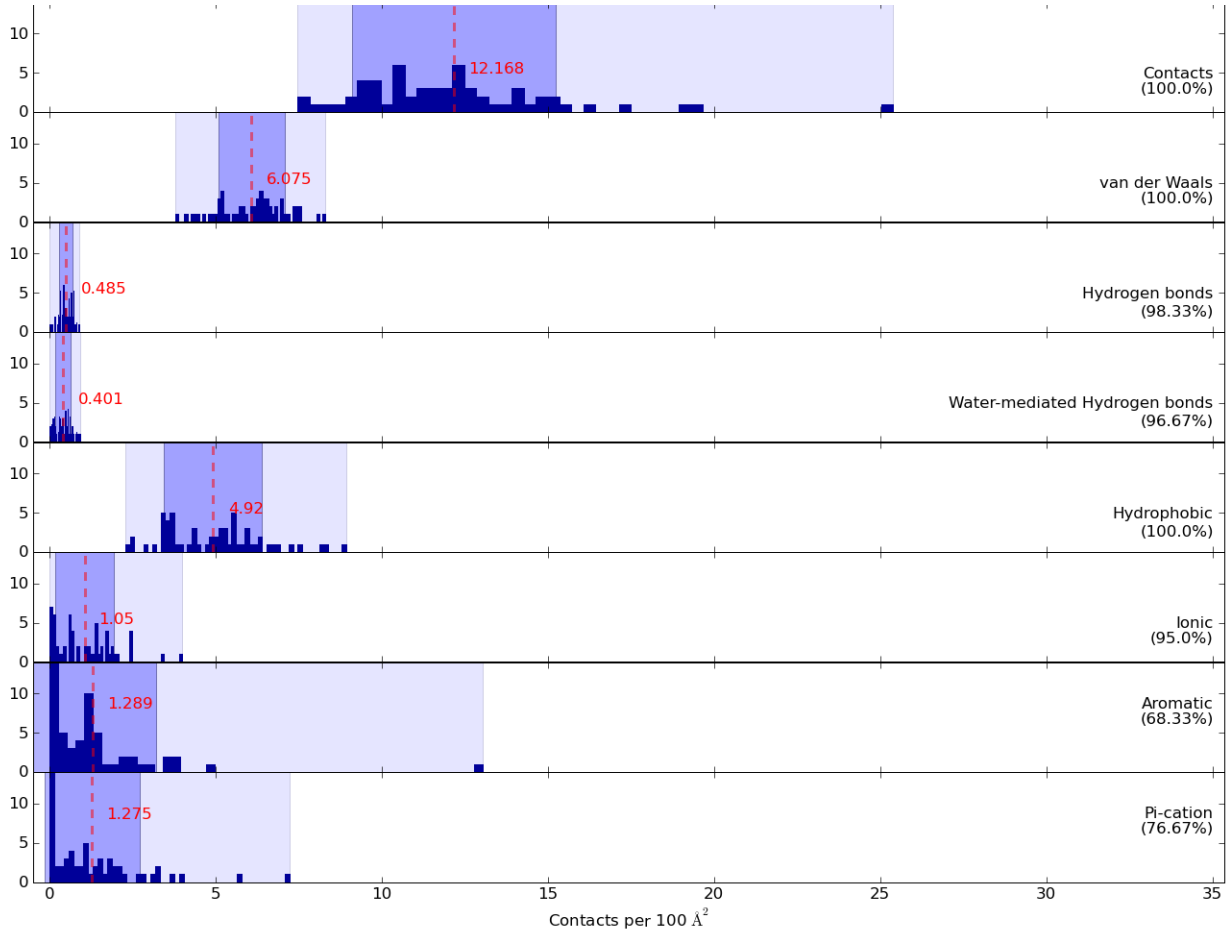


Figure 3.12: Interaction types per unit area for obligate heterodimers.

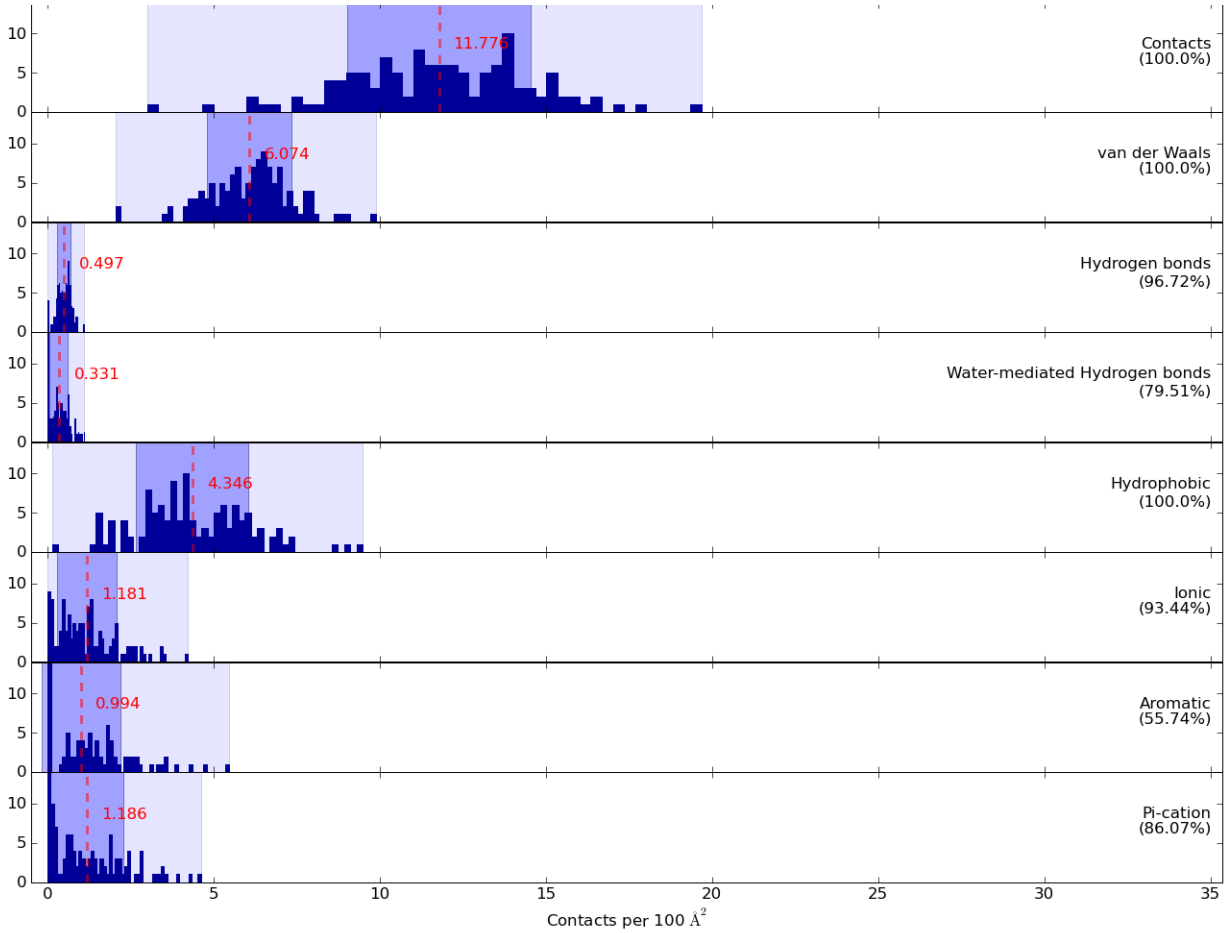
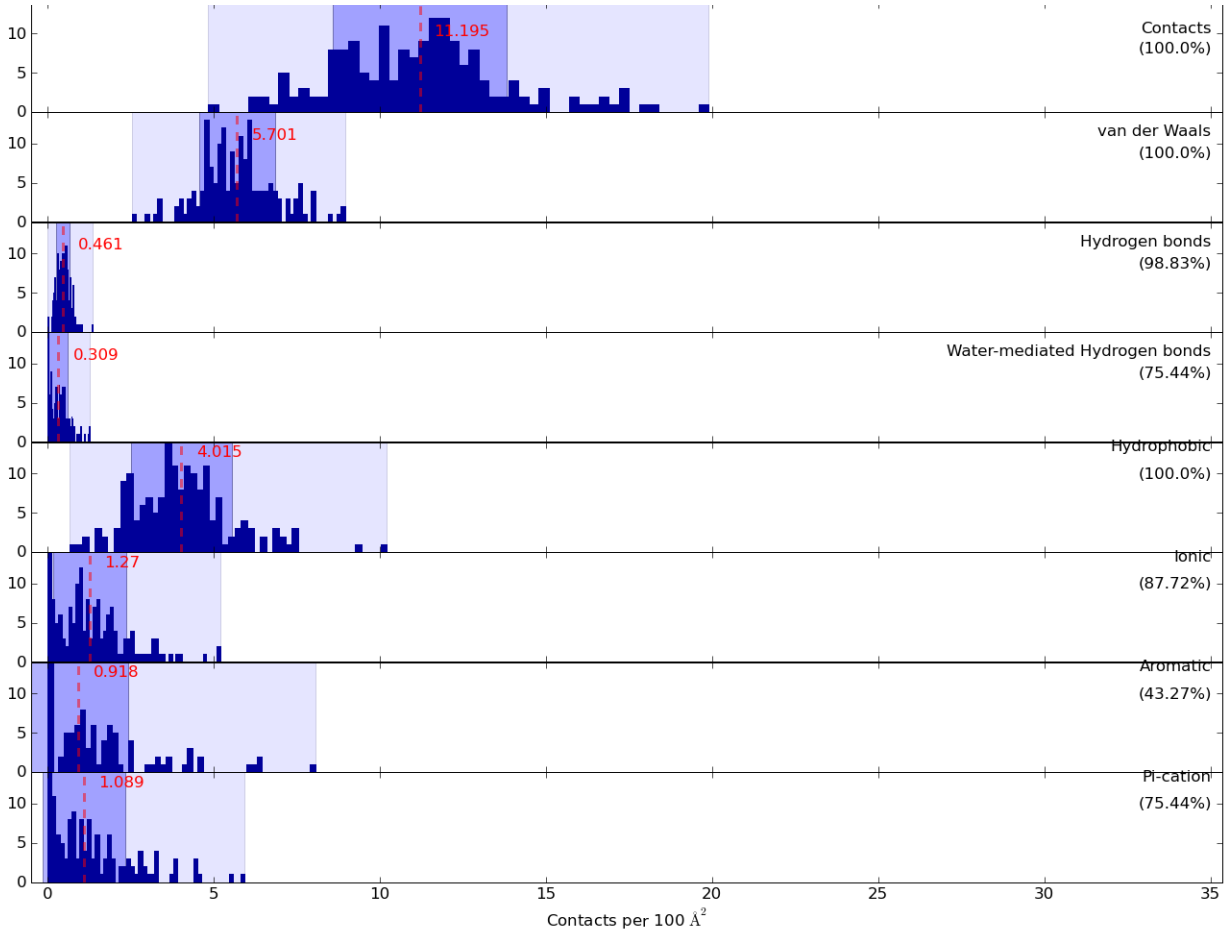


Figure 3.13: Interaction types per unit area for transient heterodimers.



A potential application of these terms would be to establish whether collectively their profile can act as a fingerprint to distinguish non-specific crystal contacts from genuine biological contacts, or randomly oriented docking decoys from poses close to the native complex.

3.3.8 Residue propensity

Figure 3.14 shows the residue propensities of each of the twenty standard residues for each of the four structural environments. One overall trend is that the hydrophobic residues (Ile, Val, Leu, Phe, Met and Ala) are enriched in the protein core and interface core and conversely are depleted in the exposed surface and the interface periphery. While most of these residues are relatively more enriched in the protein core, phenylalanine is as prevalent in the interface core and not significantly depleted in the interface periphery. The polar and ionizable residues (Asp, Gln, Asn, Glu, Lys and Arg) exhibit reciprocal behaviour: they are significantly enriched on the surface and the interface periphery. Lysine is highly disfavoured in the protein core and interface core.

For the majority of residues the interface core and periphery are intermediate between the protein core and exposed surface, with the interface periphery most similar to the exposed protein surface and the interface core most similar to protein core. The exceptions to this scheme are methionine, glycine, alanine, histidine, tryptophan, tyrosine and arginine. Of these, alanine and glycine, the two smallest residues, are disfavoured at the interface periphery. Histidine and arginine, two positively charged residues, are favoured at the periphery - in fact this is the structural environment in which these residues are most enriched. Arginine is capable of multiple types of favorable interactions: it can simultaneously form up to five hydrogen bonds and an ionic salt-bridge with the positive charge carried on its guanidinium motif. Tryptophan, tyrosine and methionine, three large, hydrophobic residues that can engage in a range of interactions, are all favoured at the interface core, corresponding with the observations of Ofran and Rost (Ofran & Rost (2003)). The enrichment of aromatic tyrosine may be explained by its contribution to the hydrophobic effect without a large entropic penalty due the side chain having few rotatable bonds as well the hydrogen bonding capacity of

Figure 3.14: Residue propensity for the overall non-redundant set.

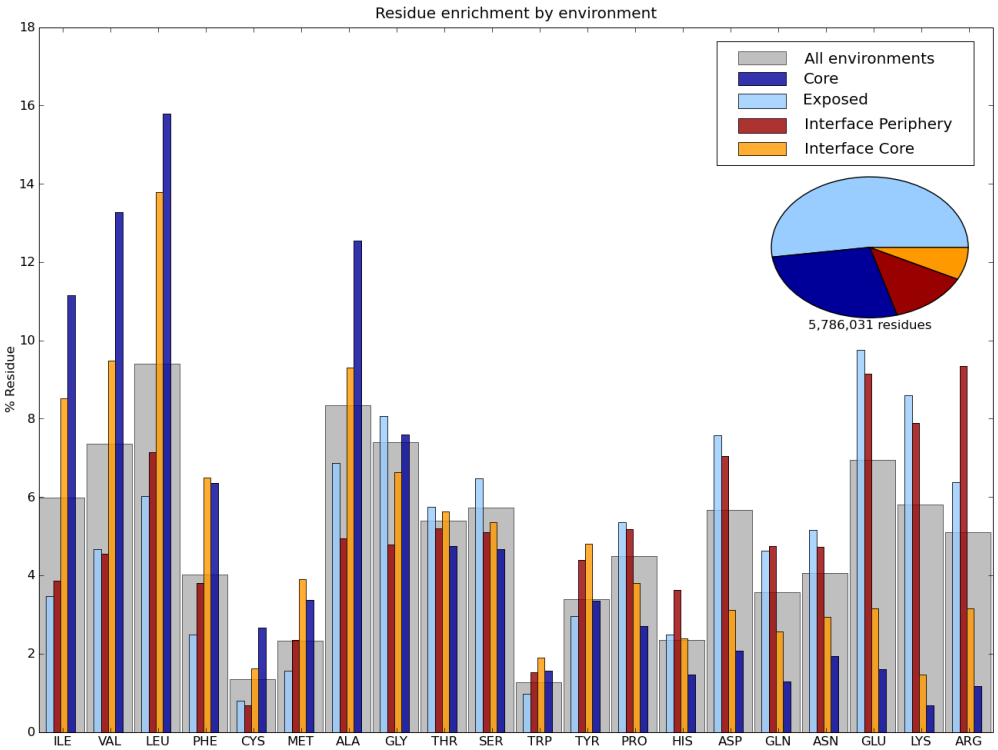


Figure 3.15: Residue propensity for the obligate homodimers.

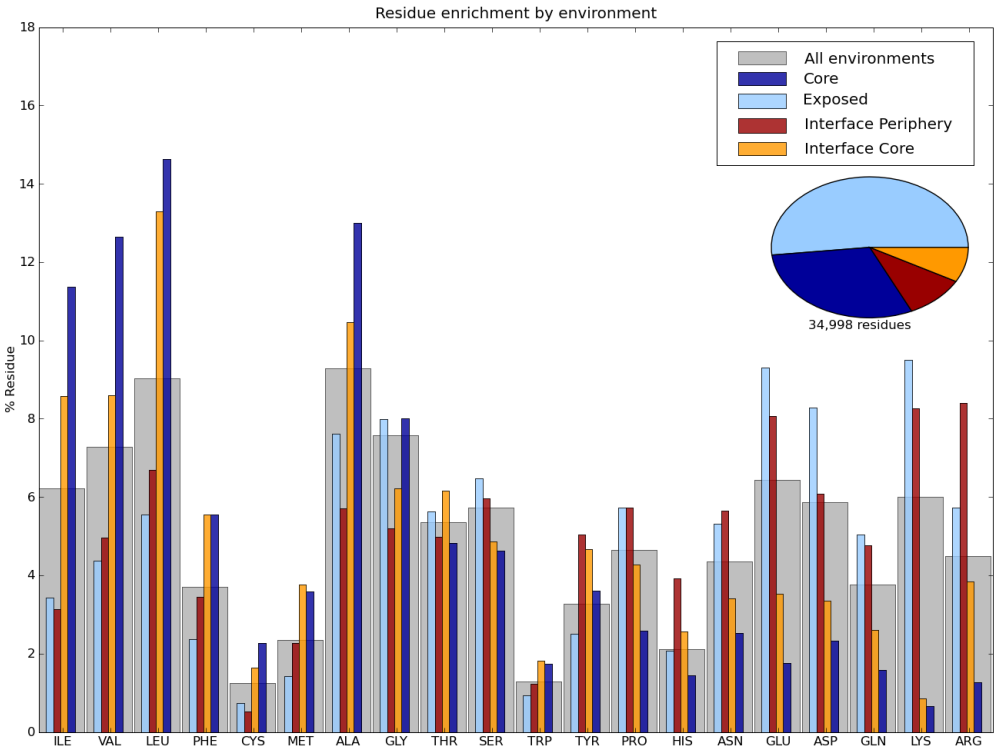


Figure 3.16: Residue propensity for the obligate heterodimers.

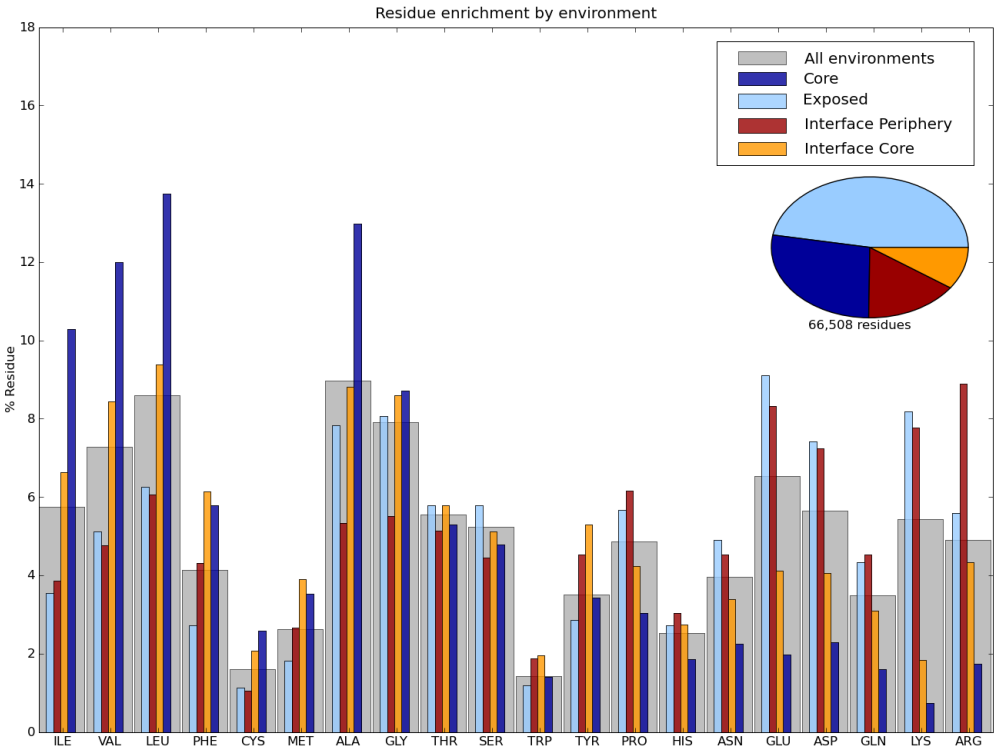
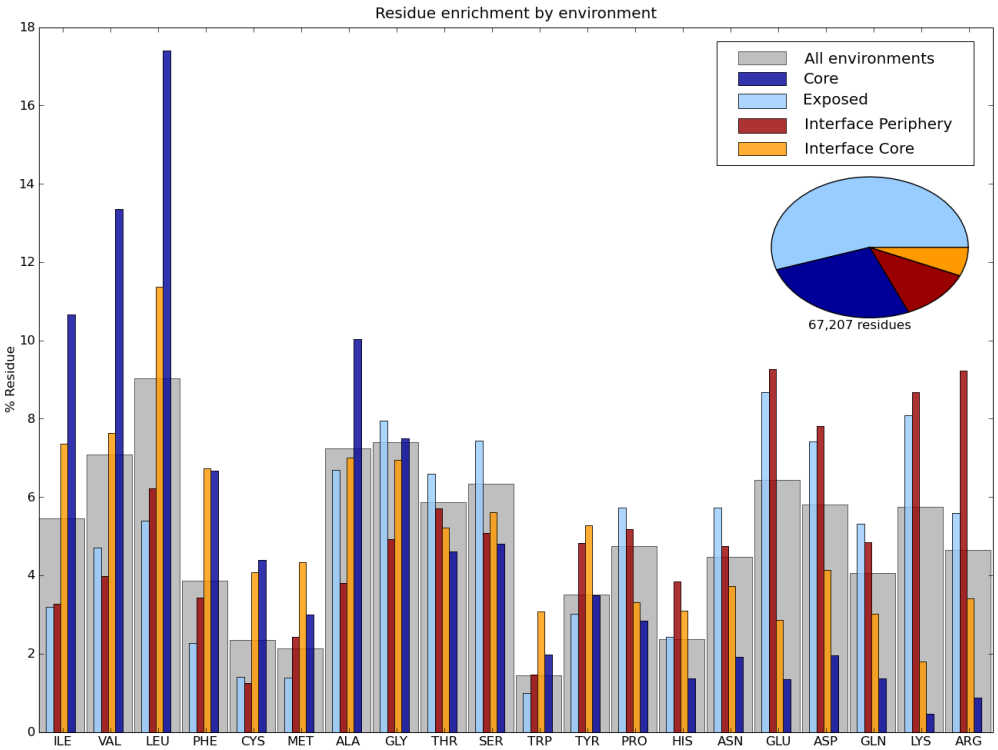


Figure 3.17: Residue propensity for the transient heterodimers.



its 4-hydroxyl group. Tryptophan has a very large aromatic side chain that can mediate aromatic π-interactions, act as a hydrogen bond donor, as well as form extensive hydrophobic contacts.

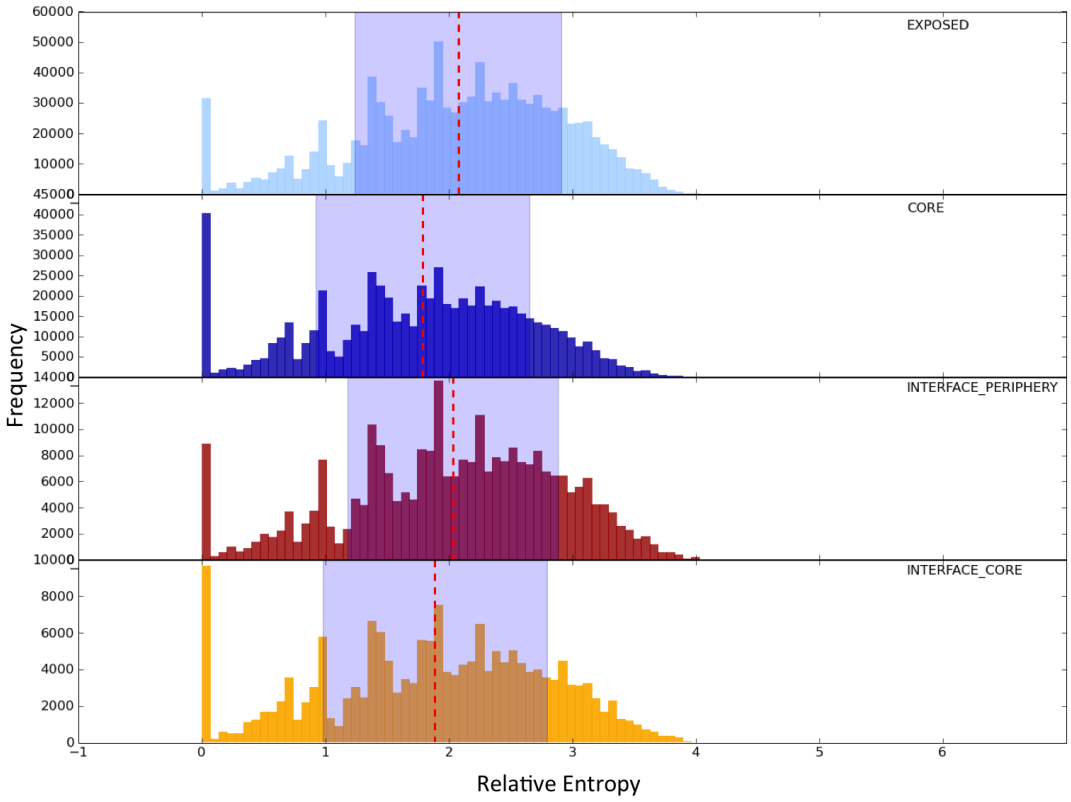
Figures 3.15, 3.16 and 3.17 show residue propensity data for obligate homodimers, obligate heterodimers and transient heterodimers respectively. Comparison of the interface regions of obligate homodimers and transient heterodimers suggests that the obligate homodimers have a greater proportion of hydrophobic residues and a lesser proportion of polar residues. This corresponds with the observations of Jones and Thornton (Jones & Thornton (1996)). There is likely to be strong selection against large hydrophobic patches in transiently-interacting proteins in order to avoid potentially harmful aberrant aggregation. This difference is less obvious when comparing the homo-obligate and hetero-obligate set but in most other respects these two sets appear to exhibit similar propensities. Some other differences of interest are apparent between the obligate sets and the transient heterodimers. Alanine residues would appear to be less favoured in the hetero-transient set whereas cysteines are more favoured. Tryptophan and methionine residues both exhibit comparable overall frequencies with respect to their transient and obligate sets but both are enriched in the interface core region of the transient set. The charged and polar residues would appear to be more abundant in their hetero-transient set than they are in the obligate systems (though as a whole they are still disfavoured here). Overall, while there are unambiguous differences in the residue propensities of the different anatomical regions of a protein structure, the results presented here suggest that differences between different interface types are less clear cut. Larger sample sizes are required to assess the significance of the observations.

3.3.9 Sequence entropy

Sequence variability for the different structural environments for the non-redundant set of interfaces is shown in Figure 3.18 and 3.19, plotting Shannon entropy and Relative entropy respectively. The red dashed lines indicate the mean of the distribution and the central shaded region indicates one standard deviation either side of the mean. Note that for Shannon entropy lower values indicate greater

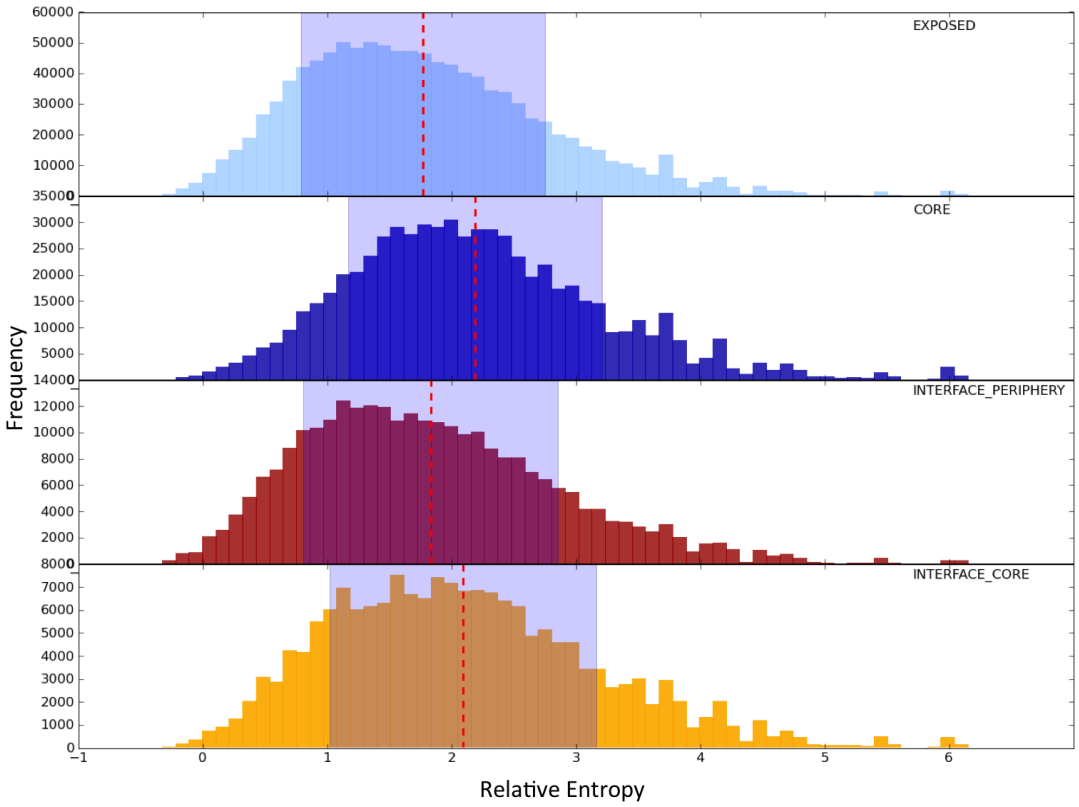
conservation, whereas for Relative entropy the opposite is true (Relative entropy actually reflects the difference between two distributions). By both measures the protein interior is the least variant, followed by the interface core. The mean values for the protein exterior and interface periphery are almost indistinguishable.

Figure 3.18: Frequency distribution of Shannon entropy values for each of the four structural environments. The red dashed lines indicate the mean of the distribution and the central shaded region indicates one standard deviation either side of the mean. For Shannon entropy lower values indicate greater conservation. The peak at zero Shannon entropy corresponds to invariant alignment columns.



To test the significance of the difference between the entropy scores for the protein core and interface core Welch's t-test was used (Welch (1947)). The results of this test suggest that at 99% confidence the two distributions are independent

Figure 3.19: Relative entropy values for each of the four structural environments. For Relative entropy higher values indicate greater conservation.



(t-statistic of 47.1).

An alternative way to probe such trends is to assess the number of interfaces where the average entropy of the core is greater than that of the periphery. Using the Shannon entropy measure, in 69.6% of interfaces (7,123 of 10,227) the interface core is more conserved than the periphery. The equivalent figure for Relative entropy is 80.1% (8,188 of 10,227).

These results confirm that the interface tends to be more conserved than the rest of the exposed surface, validating the results of previously published studies, but using a data set two orders of magnitude larger (Caffrey *et al.* (2004); Mintseris & Weng (2005); Valdar & Thornton (2001)). Guharoy and Chakrabarti (Guharoy & Chakrabarti (2005)) found, using a smaller data set and the Shannon entropy measure, that 73.6% of homodimers (89/121) and 68.1% of heterocomplexes (47/69) had an interface core more conserved than a solvent-exposed rim. Nooren and Thornton (Nooren & Thornton (2003)) analysed a small set of 39 transient heterodimers and also found (using a different measure of conservation and interface definition) that the interface core was more conserved than periphery.

Figures 3.20 and 3.21 are box plots showing Shannon and Relative Entropy results as calculated previously but partitioned first by structural environment and then by amino acid residue. Inspection of these figures indicates which residues, if any, exhibit greater or lesser variability in each structural environment. By either measure the results suggest that the protein core and interface core are more conserved than the exposed surface and the interface periphery. The difference is most pronounced in the polar and charged residues, and least in the hydrophobic and aromatic residues. Cysteine does not obey this trend and appears to be more conserved in all environments, presumably due to its important role in disulphide bridges.

Figures 3.22 and 3.23 are derived from precisely the same data but this time arranged first by amino acid residue and then by structural environment. This arrangement indicates in which structural environment each residue exhibits greater or lesser variability. By Shannon entropy cysteine, glycine and proline are the most conserved in all environments. Each of these can perform unique structural

Figure 3.20: Shannon entropy values for each residue in each of the four structural environments.

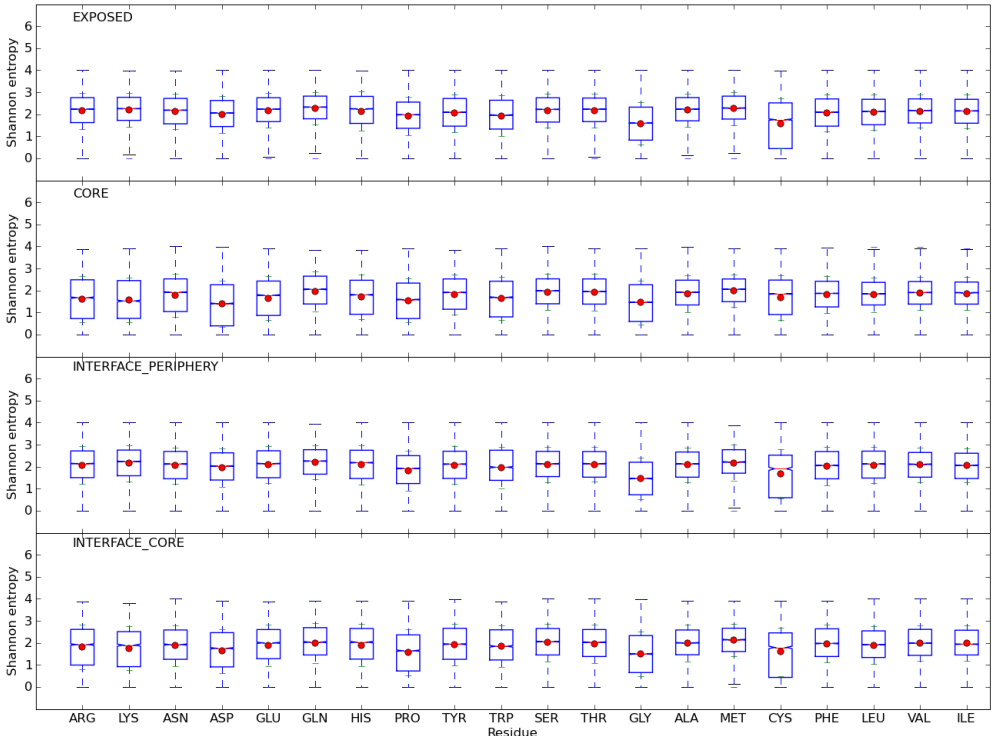
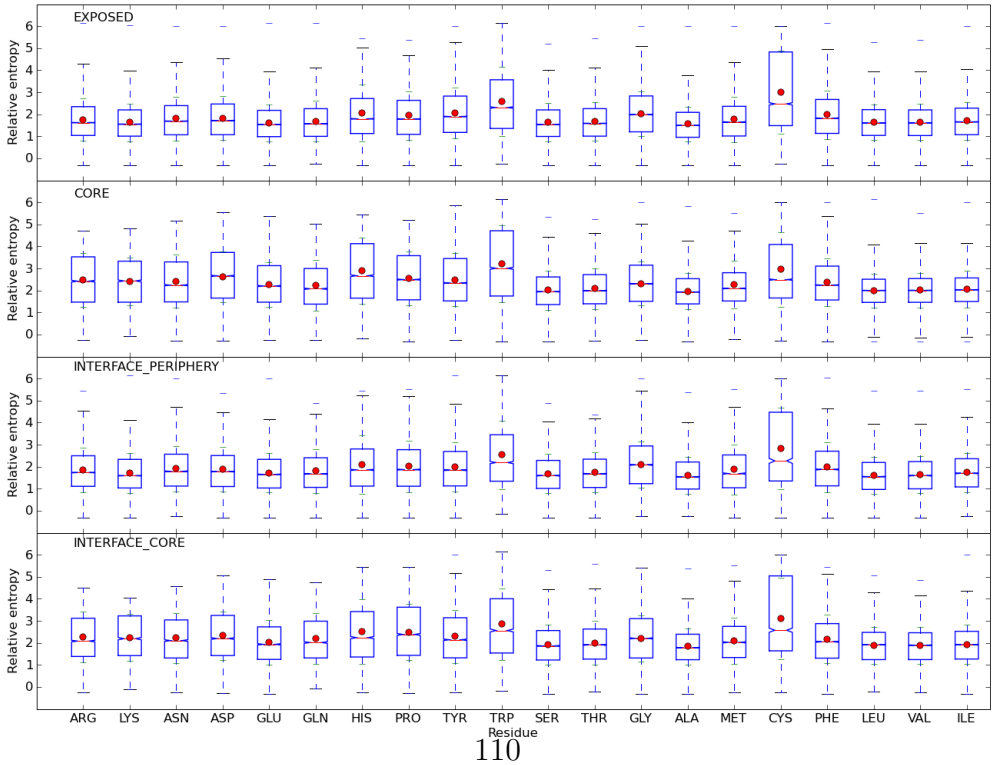


Figure 3.21: Relative entropy values for each residue in each of the four structural environments.



roles - cysteine in disulphide bonds, glycine in positive ϕ torsion angle environments and proline as a helix kink and in solvent exposed turns.

By Relative entropy cysteine and tryptophan appear to be the most conserved in all environments (although this may be an artefact of their extremely low abundance in the Relative entropy calculation). Relative entropy generally appears to provide the greater discrimination. In each of Figures 3.20, 3.21, 3.22, 3.23 the standard deviations are large so results must be interpreted cautiously. One general trend is that in the protein core and interface core, the polar and charged residues are consistently more conserved than the hydrophobic residues, although the effect is less pronounced with the Shannon entropy measure. This result corresponds with the finding of Worth and Blundell (Worth & Blundell (2009)) that buried polar side chains are amongst the most conserved residues of all.

3.3.10 Contact preferences

Figures 3.24, 3.25, 3.26 and 3.27 (pages 113 and 116) show a series of matrices used in the derivation of a contact preference matrix. The progression is shown to enable assessment of the contribution of the different terms to the final contact preference matrix. Figure 3.24 shows the raw observed contact matrix. Here leucine-leucine contacts dominate, however this is largely due to the high occurrence of leucine in interfaces. To assess the impact of residue abundance, comparison should be made with the expected contact matrix in 3.25. As described previously, both the observed and expected contact matrix are normalized by the ASA of each residue. The pairwise ASA data (independent of interface contacts and residue frequencies) are shown in Figure 3.26. Finally, Figure 3.27 shows the final contact preference matrix - essentially the log ratio of the ASA-normalized observed to ASA-normalized expected contacts.

The final preference matrix reveals some interesting patterns consistent with previously published studies (Ansari & Helms (2005); Moont *et al.* (1999); Ofran & Rost (2003); Yan *et al.* (2008)), summarizing much of what is already established regarding macromolecular interactions - hydrophobic interactions, salt bridges and disulphide bonds are all important in protein-protein interactions.

Figure 3.22: Shannon entropy values for each structural environment for each residue type.

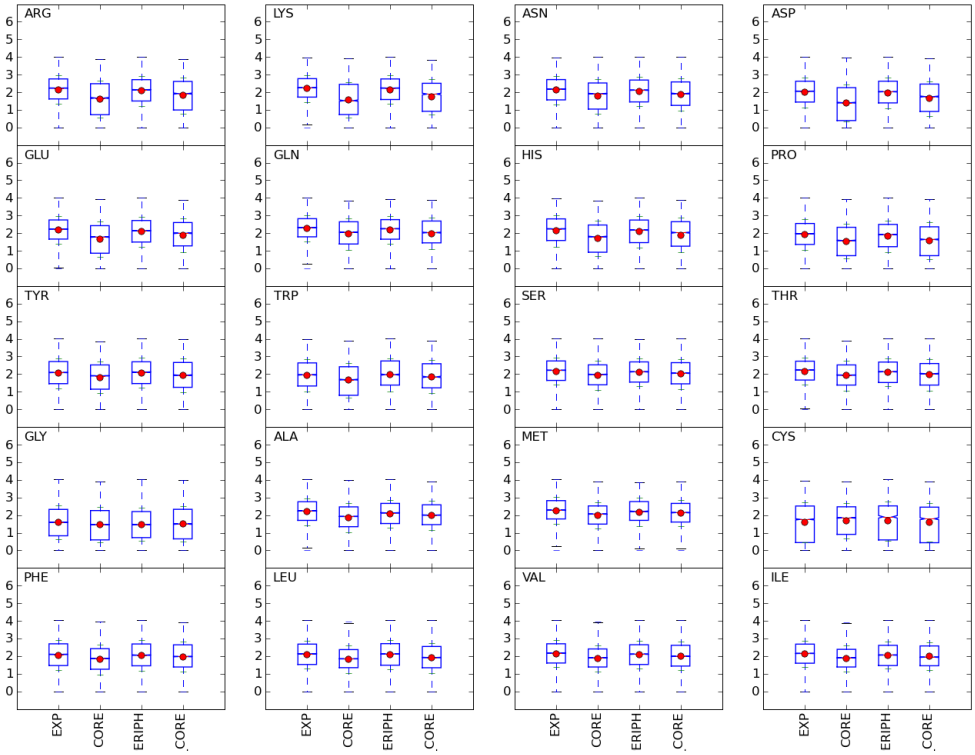


Figure 3.23: Relative entropy values for each structural environment for each residue type.

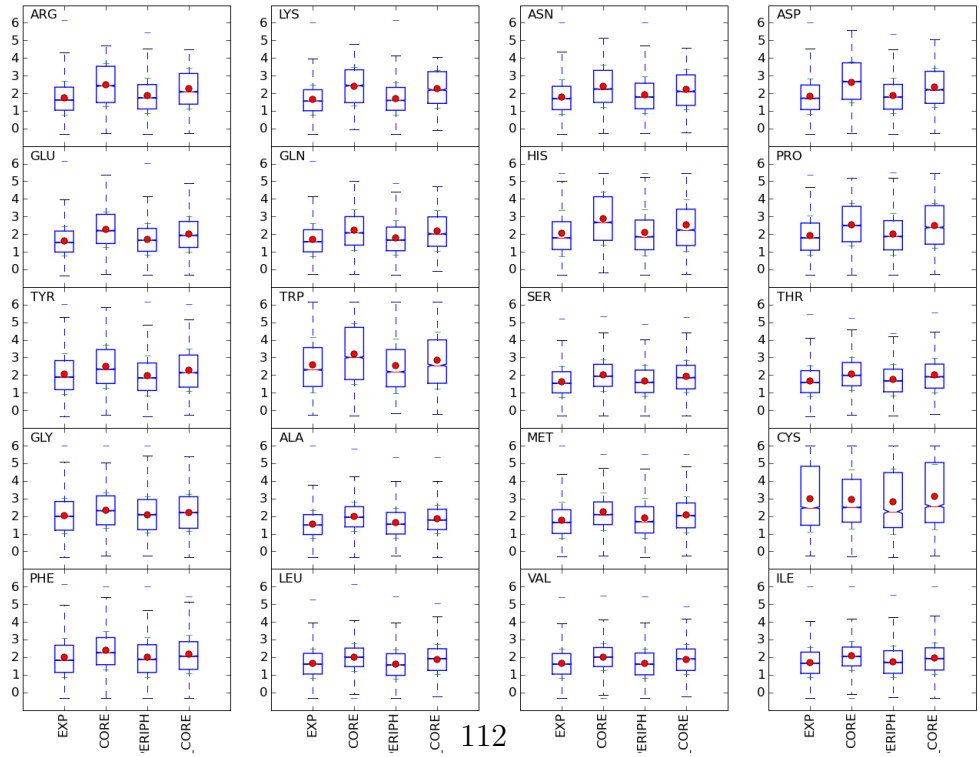


Figure 3.24: Observed contact matrix. Colours correspond to the proportion of each residue pair observed in the non-redundant set in PICCOLO (P_{ij} as described in Equation 3.6)

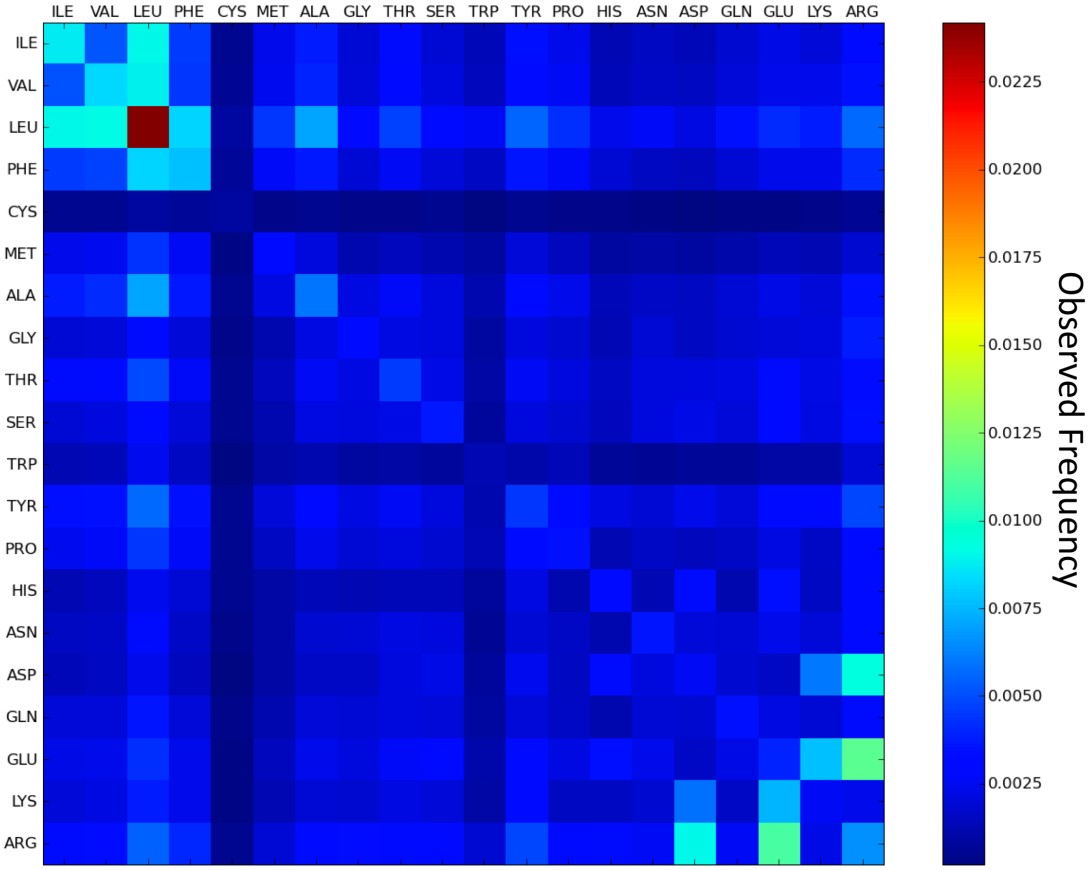


Figure 3.25: Expected contact matrix. Colours represent the expected frequency of residue pairs based solely on the occurrence of each residue in interface regions, independent of contacts actually observed in PICCOLO ($W_i \times W_j$ as described in Equation 3.7).

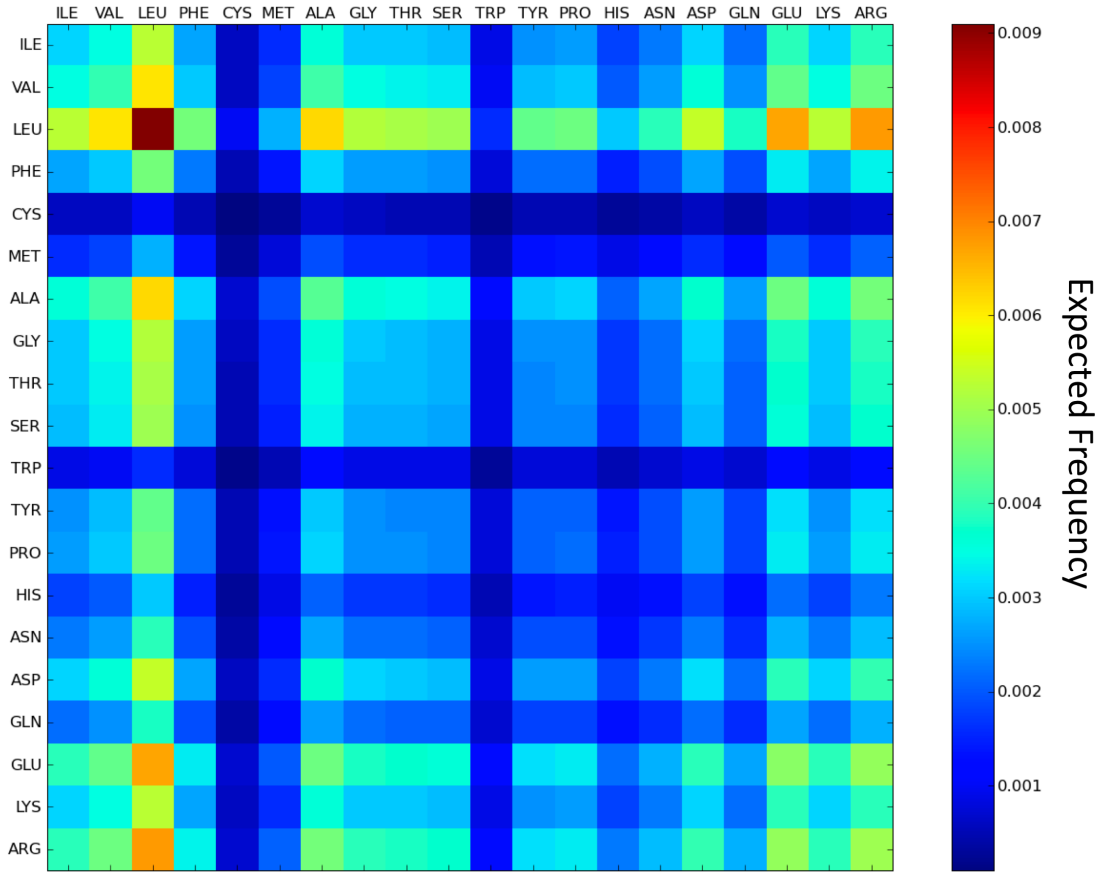


Figure 3.26: Pairwise ASA matrix. Colours represent the proportion of combined ASA of the pair of residues, independent of contacts actually observed in PICCOLO.

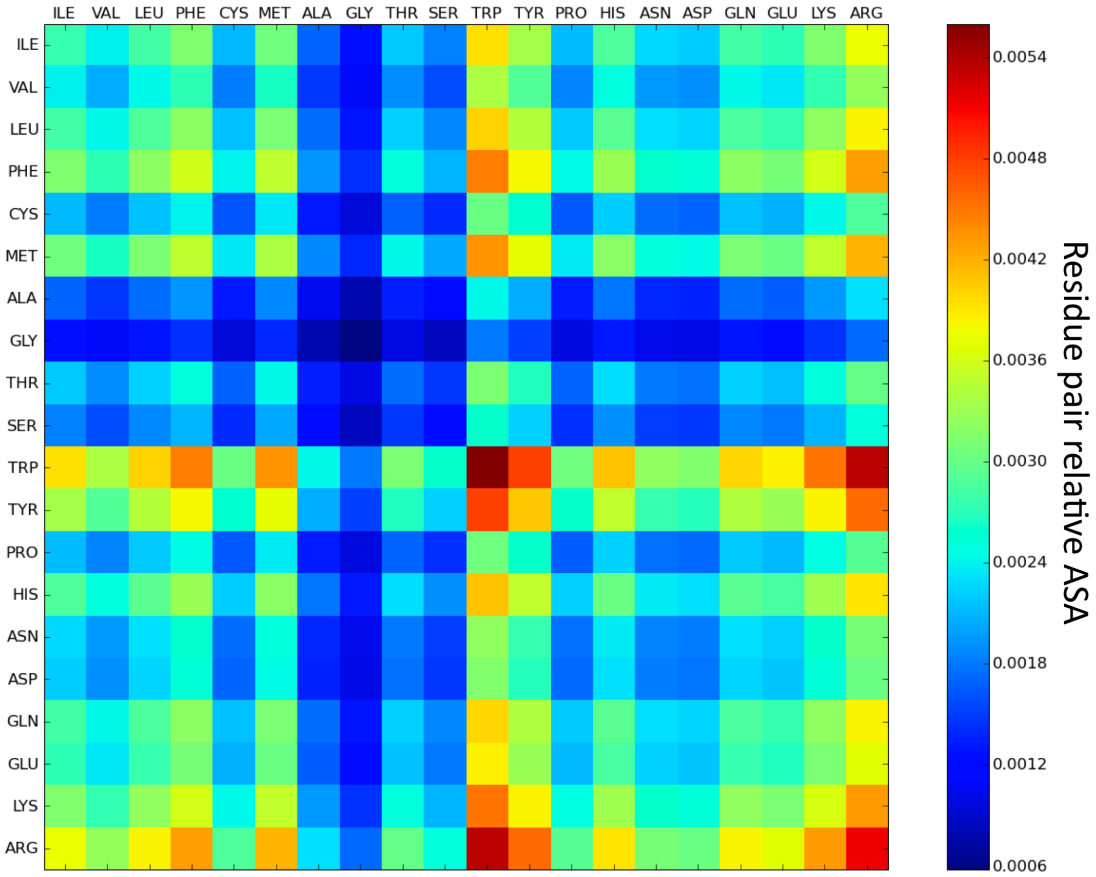
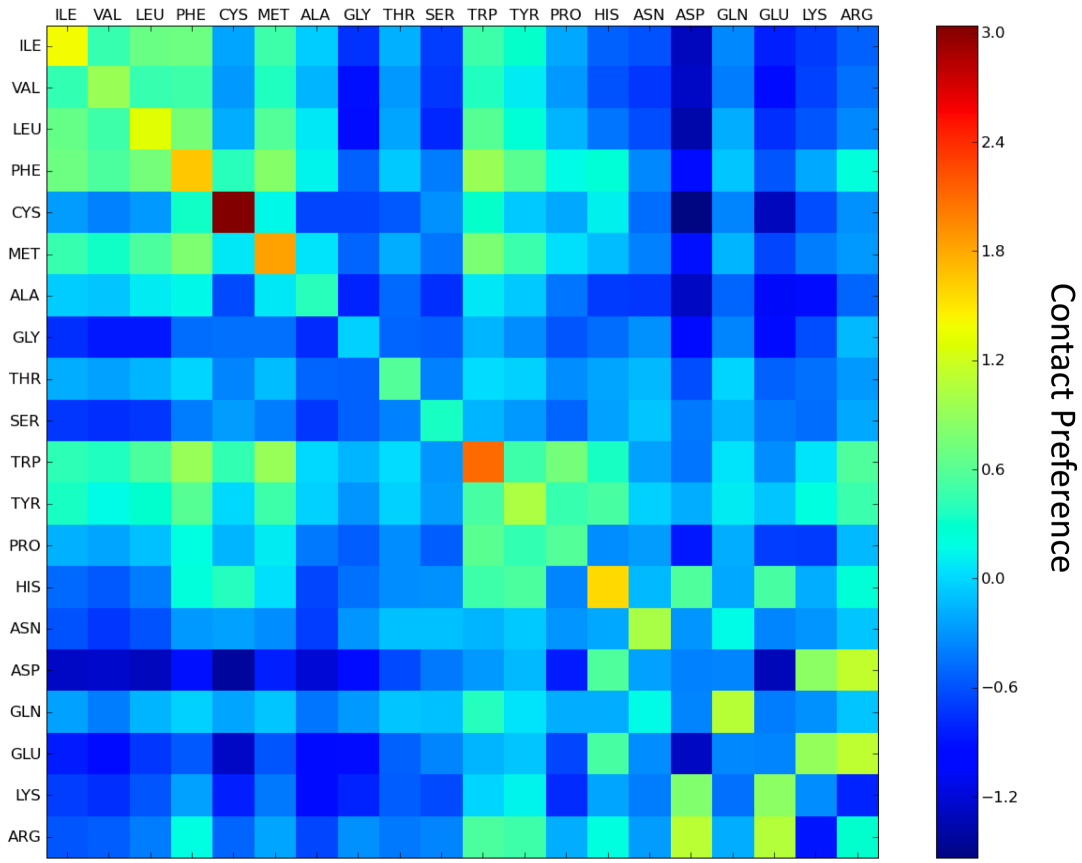


Figure 3.27: Contact preference matrix. Colours represent the log ratio of the ASA-normalized observed and expected residue frequencies $L(i, j)$ as described in Equation 3.8.



Hydrophobic residues favour other hydrophobic residues and disfavour the charged and polar residues, as would be expected from desolvation behind the hydrophobic effect. Aromatic residues favour one another, as do hydrophobics and aromatics. Proline shows a preference for aromatic residues - indeed it has been suggested that the interaction between a proline ring and an aromatic ring resembles the interaction between two aromatic rings (Glaser *et al.* (2001); Yan *et al.* (2008)). Residues of opposing charge favour one another (arginine, lysine and histidine versus glutamate and aspartate) enabling electrostatic complementarity to be established. Like charge interactions are predictably disfavoured for the glutamate and aspartate residues carrying a negative charge. The pattern for residues carrying positive charge is less clear cut. Lysine-lysine interactions are disfavoured, whereas histidine-histidine, arginine-arginine and arginine-histidine are favoured.

Examination of the atomic interaction details stored in the atom_pairs table revealed that a range of interactions types contribute to the histidine-histidine result, including aromatic, van der Waals, π-cation and hydrogen bonding interactions. The arginine-arginine preference is due in part to some hydrophobic interactions between the Cβ and Cγ atoms as well as some hydrogen bonding between main chain and side chain atoms. The arginine-histidine result can be largely attributed to π-cation interactions with some side chain to side chain hydrogen bonding.

The diagonal of the matrix is generally favoured (except for lysine pairs and aspartate and glutamate pairs), likely due to the preponderance of self-interacting residues from homodimers with a 2-fold symmetry axis. The most preferred contact pairs are cysteine-cysteine followed by tryptophan-tryptophan and methionine-methionine - the three least abundant residues (see Figure 3.14). The disulphide capacity unique to cysteines plays a critical role in stabilization of small secreted proteins. Methionine-methionine pairs are dominated by hydrophobic interactions. The tryptophan-tryptophan pairwise interactions have contributions from van der Waals and hydrophobic contacts but are dominated by edge to face type aromatic interactions. Figure 3.28 shows as a typical example the homodimeric complex of isopentenyl-diphosphate delta-isomerase (PDB entry 1ow2) with two pairs of symmetry related tryptophans.

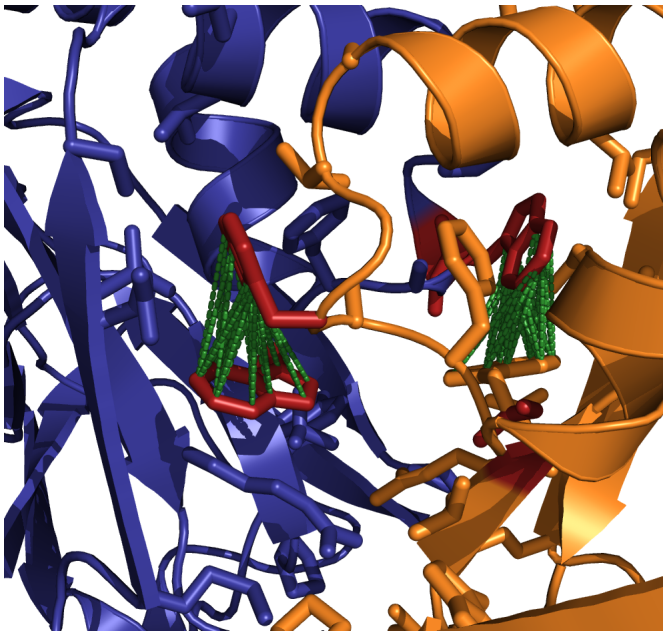


Figure 3.28: Cartoon representation of a typical example of tryptophan-tryptophan contacts from the homodimeric complex of isopentenyl-diphosphate delta-isomerase (PDB entry 1ow2). Chain A is shown in blue and Chain B in orange. Two pairs of symmetry-related tryptophan residues are represented as sticks in red, with aromatic-aromatic contacts stored in PICCOLO as green dashes.

These results indicate some differences to those presented by Glaser *et al.* (Glaser *et al.* (2001)), who found arginine-tryptophan to be the most favoured pair, which is only marginally favoured here. However that study used a different interface definition and, as indicated in the Methods, they use residue volumes to normalize the expected interactions only.

3.3.11 PICCOLO analysis conclusions

Although it is difficult to draw broad conclusions across the different properties investigated (residue propensity, hydropathy, residue contact preference, sequence entropy) an underlying theme does emerge. This is that the different anatomical regions of protein interfaces exhibit different qualities, and these differences can be more striking than the differences observed between different interface types. Overall the core of the interface is most similar to the core of the protein domain, whereas the periphery of the interface is, in most respects, similar to remaining solvent exposed surfaces of the protein. The next chapter investigates whether these patterns hold true when examining the nature of substitutions that are accepted in protein interfaces across evolutionary time.

Chapter 4

TOCCATA

4.1 Introduction

Structural domains are independently folding units that typically form the fundamental modular units of protein evolution. Protein domains with similar sequences have similar folds and always share a common evolutionary origin. However, this property is non-commutative; domains with similar folds do not necessarily have detectably similar sequences. Protein domains with similar folds *generally* share a common evolutionary origin (genuine convergent evolution is

rare) (Lesk *et al.* (1989)). An understanding of this relationship between protein sequence and structure is crucial to understanding evolution and aiding in the analysis of the results of the various genome projects. This is manifested in the principle that sequence alignment can unambiguously identify proteins who share similar structures; for proteins of reasonable length, when the pairwise sequence identity is above 30% the proteins can confidently be assigned as sharing a common evolutionary origin. Between 20-30% sequence identity this relationship becomes ambiguous, hence this window is described as the "twilight zone" (Rost (1999)). Pairwise identities below 15-20% mean that although the proteins may still share a common evolutionary origin, other evidence would be required to gain confidence. This principle, that structure is more conserved than sequence, can be harnessed to improve the quality of an alignment of sets of homologous proteins. Wherever available, structural information should be used: features including secondary structure, solvent accessibility, hydrogen bonding, disulphide bonding and positive ϕ torsion angles can assist accurate alignment of residues where the sequences have diverged such that no obvious similarity remains.

Sequence-structure homology recognition is a crucial step in genome annotation and is also the first step in comparative modelling as it is the means by which putative template structures are identified. The TLB group has been contributing to the area of comparative modelling for several years and is currently working towards applying these approaches at genome-scale (Burke *et al.* (2007); Worth *et al.* (2007a)), as discussed in Chapter 5. A pre-requisite for such work is well-organized structural information in the form of a comprehensive set of multiple structural alignments. Multiple alignments offer improved performance in homology recognition searches over pairwise alignments, as they permit profile methods to be used. Such methods have been shown to offer greater sensitivity than pairwise search methods (Altschul *et al.* (1997); Durbin *et al.* (1999)). Ideally such an alignment set should cover all available structures from the Protein Data Bank (PDB) in order to maximize the likelihood of finding a structural match. Maintaining such coverage with incremental PDB updates represents a considerable challenge.

Aside from the purpose of comparative modelling, such alignment data are also required when exploring how the properties of protein interfaces extend to

homologues. A further application of such a resource is the potential for the derivation of new amino acid residue substitution tables.

4.1.1 Structural alignments

The TLB group has previously developed HOMSTRAD (HOMologous STRucture Alignment Database) (Johnson *et al.* (1993); Mizuguchi *et al.* (1998b)) a web-accessible resource of structural alignments of homologous protein families. The family definitions in HOMSTRAD came from various databases including SCOP (Hubbard *et al.* (1999)), Pfam (Finn *et al.* (2008)), PROSITE (Bairoch (1991)) and SMART (Schultz *et al.* (1998)) as well as the results of sequence similarity searches using PSI-BLAST (Position Specific Iterative Basic Local Alignment Search Tool) (Altschul *et al.* (1997)) and *Fugue* (Shi *et al.* (2001)) and a small number of manually defined families. Most of the HOMSTRAD alignments have been generated using the programs *Comparer* (Sali & Blundell (1990)) and MNYFIT, although a small number were generated by hand. A key feature of HOMSTRAD is that the alignments are provided annotated using the program JoY (Mizuguchi *et al.* (1998a)) to highlight a number of important structural features including secondary structure, relative sidechain accessibility, hydrogen bonding to the main chain amide, main chain carbonyl or other sidechains, disulphide bonds and positive ϕ torsion angle. This annotation is critical to understanding how these various structural features contribute to the determinants of conservation across the family. HOMSTRAD is commonly used as a search database for the sequence-structure homology recognition program *Fugue* to identify homologous proteins of known structure that potentially act as templates for comparative modelling.

HOMSTRAD requires substantial manual curation, both to seed new families and to classify weekly PDB updates, and this is simultaneously one of its strengths and one of its main drawbacks. As dozens of new structures are published each week, unless continual effort is made to curate and update classifications (a task for which resources are unavailable) HOMSTRAD's coverage undergoes attrition. A further drawback of HOMSTRAD is that it has no consistent underlying

domain-family definition either in terms of regional delineation of domains or evolutionary breadth (for those cases where several related families share common structural fold). For these reasons the decision was taken to derive an alternative library of structural alignments in the form of the relational database TOCCATA, with a focus on trying to achieve and maintain greater structural coverage by using automated procedures for generating multiple alignments. Although fully automated methods are unlikely to achieve the same alignment quality as manual approaches, the increasing volumes of structural data published in light of structural genomics projects (Chandonia & Brenner (2006)) means that methods that require manual curation are increasingly unrealistic.

4.1.2 Environment Specific Substitution Tables

Intuitively it would be predicted that the likelihood that an amino acid substitution will be accepted through evolution depends strongly on the local environment of the amino acid sidechain. This is illustrated by the fact that residues buried within the core of a protein tend to be more conserved than those on the surface on account of buried residues having a role in maintaining the structure of a protein, with buried polar residues being the most conserved of all (Worth & Blundell (2009)). This notion of context-dependent substitution likelihood has been established qualitatively and quantitatively through careful observation of amino acid substitutions in divergent evolution compiled from HOMSTRAD. Overington *et al.* (Overington *et al.* (1992)) used HOMSTRAD to derive a library of environment-specific substitution tables (ESSTs).

ESSTs have been successfully applied to a series of key problems in structural bioinformatics: secondary structure prediction (Wako & Blundell (1994a)); sequence-structure homology recognition as *Fugue* (Shi *et al.* (2001)); structural model validation as *Harmony* (Pugalenthi *et al.* (2006)); prediction of stability changes upon mutation as SDM (Site Directed Mutator) (Topham *et al.* (1997)); and functional site prediction as *Crescendo* (Chelliah *et al.* (2004)). *Crescendo* works by comparing the observed substitution patterns from an alignment of the protein of interest with its orthologues, which are under both functional and structural constraints, with those that are expected on the basis of structure taken

from the library of pre-calculated ESSTs. Recently Gong and Blundell (Gong & Blundell (2008)) found that the performance of *Crescendo* could be markedly improved by more extensive masking of functionally annotated residues during ESST generation.

In much of the preceding work on ESSTs, the local structural environment has typically been defined based on (1) main-chain conformation and secondary structure, (2) solvent accessibility, and (3) hydrogen bonding. By combining four classes of secondary structure (α-helix, β-strand, coil and residue with positive ϕ main-chain torsion angle), two classes of solvent accessibility (accessible and inaccessible), two classes of main-chain carbonyl hydrogen bonding (bonded and unbonded), two classes of main-chain amide hydrogen bonding (bonded and unbonded), and finally two classes of sidechain hydrogen bonding (bonded and unbonded), a total of 64 ESSTs can be combinatorially derived ($4\times2\times2\times2\times2$=64). Thus these combined structural features define the local structural context and permit quantitative exploration of the different restraints placed by the environment on the probability of substitutions being tolerated.

The overall procedure for generating a substitution table involves first using the raw counts of observed substitutions across an alignment or series of alignments to give an Accepted Replacement Matrix (ARM). The ARM can be readily converted to an Observed Frequency Matrix (OFM) by converting the raw replacement counts to respective frequencies.

$$OFM : P(b|a,E) = \frac{A^E_{ab}}{\sum_c A^E_{ac}} \tag{4.1}$$

where $P(b|a,E)$ is the probability that amino acid a in environment E is substituted by amino acid b.

An Expected Frequency Matrix (EFM) can be derived based on the observed background occurrence of each target residue, independent of substitutions.

$$EFM : q_b = \frac{\sum_{a,E} A^E_{ab}}{\sum_{a,b,E} A^E_{ab}} \tag{4.2}$$

where q_b is the background probability of observing amino acid b. Subsequently a Substitution Frequency Matrix (SFM) can be derived as the simple division product of the OFM over the EFM. The SFM is then converted to a log-odds matrix (LOM) by taking the logarithm of the SFM probabilities and applying a scaling factor:

$$LOM : s(a, E \rightarrow b) = \frac{3}{\log 2} \times \log(\frac{P(b|a, E)}{q_b}) \tag{4.3}$$

One of the key challenges in the derivation of ESSTs is that of data sparsity. The availability of observed substitution data is limited and when partitioned over many structural environments data coverage can be meagre, particularly for those less abundant environments with severe physico-chemical constraints. This generates a tension between the desire to reflect accurately each residue's local environment with a large number of structural descriptors and the necessity for each environment to comprise sufficient observations to attain statistical significance. One obvious way to overcome this challenge is to increase the number of observations by increasing the number and size of the multiple alignments. TOCCATA attempts to maximize the available structural information by including all SCOP families, although coverage of the PDB is not complete owing to the infrequency of SCOP updates.

An important distinction lies between ESSTs that are conformationally restrained and those that are unrestrained. Conformationally constrained substitutions are those where, despite the residue substitution, the local structural environment remains unaltered. Such a constraint can be crucial for example for the purpose of using ESSTs to predict the effect of mutations on protein stability (Topham *et al.* (1997)). In contrast ESSTs used for the purpose of assessing sequence-structure alignments are unrestrained and include all substitutions regardless of any change in local environment (Shi *et al.* (2001)). However the use of conformational restraints has the added affect of reducing the number of substitution counts that can be included, thereby exacerbating sparsity issues.

In generating ESSTs for use in prediction of the effects of mutations on protein stability, Topham *et al.* originally defined 216 environments (9 secondary structure terms, 3 solvent accessibility terms and 8 hydrogen bonding terms)

(Topham *et al.* (1993)). However data sparsity issues led them to reduce this to 54 environments by collapsing the 8 hydrogen bonding terms to 2.

4.1.2.1 Smoothing

Entropy-based smoothing procedures, originally devised by Sippl (Sippl (1990)), were implemented by Topham *et al.* (Topham *et al.* (1993)) (in the form of the Fortran program *maksub* for use in predicting the effects of mutations on stability) and Sali and Blundell (Sali & Blundell (1993)) to help resolve data sparsity issues. Shi *et al.* (Shi *et al.* (2001)) updated the ESST generation procedure in the form of the Fortran program *subst* (written by Kenji Mizuguchi) for use in the sequence-structure homology recognition program *Fugue*. As well as updating the smoothing procedure to compensate for data sparsity they also included a clustering scheme to correct for sampling bias and filtering procedures to reduce interference from non-structural restraints by masking functional residues (defined as those interacting with small-molecule ligands or other structural domains). The smoothing procedures attempt to obtain better estimates of substitution probabilities by replacing Equation 4.1 for the OFM with:

$$P(b|a,E) = \omega_1^{a,E} A(b|a,E) + \omega_2^{a,E} W(b|a,E) \tag{4.4}$$

where $W(b|a,E)$ is the OFM as defined in Equation 4.1 (with $P(b|a,E)$ replaced with $W(b|a,E)$) and $A(b|a,E)$ is an *a priori* probability distribution (Sali & Blundell (1993)) and the weights $\omega_1^{a,E}, \omega_2^{a,E}$ are given by:

$$\omega_1^{a,E} = \frac{1}{(1 + \frac{N^{a,E}}{\sigma n})} \tag{4.5}$$

$$\omega_2^{a,E} = 1 - \omega_1^{a,E} \tag{4.6}$$

where $N^{a,E}$ is the total number of observed substitutions of amino acid a in environment E, n is the number of bins for the probability distribution (20 in this case) and the constant parameter σ determines the relative contributions of the *a priori* distribution and the observed distribution. When the average number of counts per bin ($N^{a,E}/n$) is less than σ, the *a priori* distribution dominates and when greater than σ the observed probability distribution dominates. In this

work, the recommended empirically-derived σ value of 5 has been used throughout. The *a priori* distribution $A(b|a, E)$ for each environment is determined by iteratively generating subset combinations of all features contributing to the environment (Sali & Blundell (1993)).

4.2 Methods

4.2.1 SCOP

A key decision in developing TOCCATA was to identify a suitable fundamental structural unit for alignment. Harnessing an external domain definition resource would provide a consistent domain family definition for each structural alignment while removing the requirement for continual manual curation. The two most popular resources for classifying structural domains are CATH and SCOP. Whilst they share some similarities there are important distinctions. CATH is an acronym for Class, Architecture, Topology and Homology (http://www.cathdb.inf/) (Greene *et al.* (2007)) - the four main levels in the classification: Class represents the overall secondary-structure content of the domain; Architecture is a broad association of similar topologies which share particular structural features; Topology captures significant similarity in structure but no clear evidence of sequence homology; and Homologous superfamily indicates a clear evolutionary relationship.

The Structural Classification of Proteins (SCOP) resource (http://scop.mrc-lmb.cam.ac.uk/scop/) (Andreeva *et al.* (2008)) from the MRC, aims to provide a full description of the structure-evolution relationships between proteins of known structure from the PDB. Chains from the PDB structures are first delineated into structural domains, and these domains are classified into a multi-level hierarchy. The principal levels of the hierarchy are family, superfamily and fold. Proteins in the same SCOP family show clear evolutionarily relatedness. In practice this typically means that pairwise identity between members of the family is above the twilight zone. Proteins in the same SCOP superfamily are believed to share a common evolutionary origin despite having low or undetectable sequence similarity. The evidence suggesting a common evolutionary

origin would typically include particular structural motifs or functional residues. Proteins in the same SCOP fold share the same relative arrangement of major secondary structures with the same topology. They may differ in the length and conformation of loops and the presence of peripheral secondary structure embellishments. Although proteins in the same fold from different superfamilies may share a common evolutionary origin, such relationships typically cannot be distinguished from rare cases of convergent evolution. Importantly for the quality of classifications, SCOP is manually generated by visual inspection and structure comparison and it is for this reason that SCOP is chosen as the basis for this work.

The family level was selected as the most appropriate level in the SCOP hierarchy to generate alignments. Although the higher superfamily level would provide greater evolutionary diversity, this would also make the alignment generation more challenging, possibly reducing alignment quality. The SCOP classes "low-resolution protein structures" and "Designed proteins" containing non-natural sequences were excluded, as were Cα only PDB structures, leaving 94,387 SCOP domains classified into one of 3,967 SCOP families. 753 of these families comprised a single domain; these were included in TOCCATA but naturally no alignment stage was required. Figure 4.1 shows the distribution of the number of domains comprising each SCOP family. PDB format files corresponding to each SCOP domain definition were pre-generated as part of the sanitization process described in Chapter 2.

SCOP data is provided in the form of a series of formatted text files, describing domain delineation (in terms of PDB chain identifiers and residue numbers forming the domain boundaries) and their classification in the SCOP hierarchy. SCOP data is typically released approximately every 1-2 years. The current release, version 1.73 released November 2007, contains 97,178 domains from 34,495 PDB entries. One drawback of SCOP is this slow update cycle; the increasing rate of publication of new structures in the PDB means that the gap between published structures and classified structures is ever widening. As of August 2008 there are 50,754 PDB entries containing protein, meaning SCOP coverage is currently around 68%. More recently the SCOP curators have provided pre-SCOP

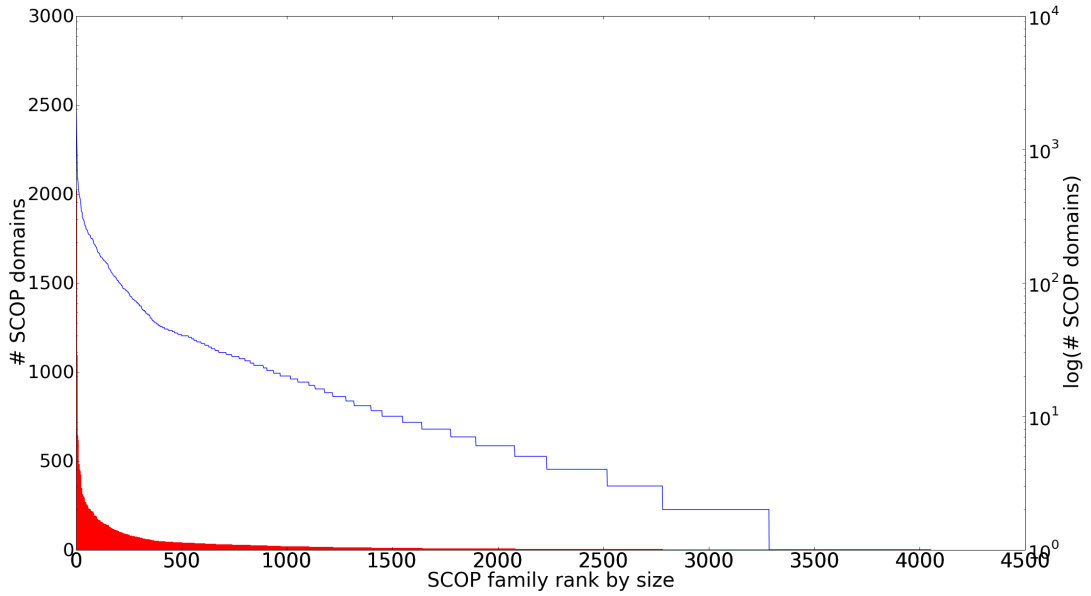


Figure 4.1: Distribution of the number of domains per family in SCOP depicted in red in the form of a bar chart (left hand y-axis) and the same data as a log-linear plot in blue (right-hand y-axis).

as an incremental update of provisional classifications. These data have been incorporated but currently only contributes an additional 554 structures, or about 1% additional coverage.

The SCOP text data are parsed and loaded into a locally-generated relational database version of SCOP. The schema comprises three core tables shown in Figure 4.2, which have been designed to be tolerant of some of the atypical situations found in SCOP. The *domains* and *fragments* tables hold data corresponding to each structural domain. Although 96% of domains comprise a single fragment, domains can comprise multiple fragments, either as discontinuous fragments within the same polypeptide chain (e.g. catalytic domain of matrix metalloprotease gelatinase A in PDB entry 1ck7) or as a combination of distinct polypeptide chains (e.g. human thrombin in PDB entry 1h8d). The chain identifier is stored as a property of the fragment. The *hierarchies* table stores data concerning the multi-level hierarchical classification of each domain (class, fold, superfamily, family, protein, species and domain). A PDB polypeptide chain may comprise a single domain. In such cases SCOP does not provide domain boundaries. However these values are required for generating TOCCATA alignments, and so are retrieved using the information stored in PDBRes. Finally a linking table to join the SCOP domain definitions to the residue-level data in PDBRes (as described in Chapter 2) is generated in order to provide residue level annotations of all residues falling within SCOP classified domains.

4.2.2 Sequence clustering

Each SCOP family contains many redundant structural domains. For many purposes, not least for the generation of TOCCATA, it is useful to derive non-redundant sets at higher resolution than the SCOP family. In order to achieve this, the BlastClust sequence clustering program was used, which uses scores from the BLAST (Basic Local Alignment Search Tool) sequence similarity search tool (Altschul *et al.* (1997)) to perform single-linkage clustering. For each SCOP family, unaligned protein sequences are first generated, delineated to correspond to the domain boundaries specified by SCOP. This ensures that each domain sequence in the family should be of similar length. The input parameters were

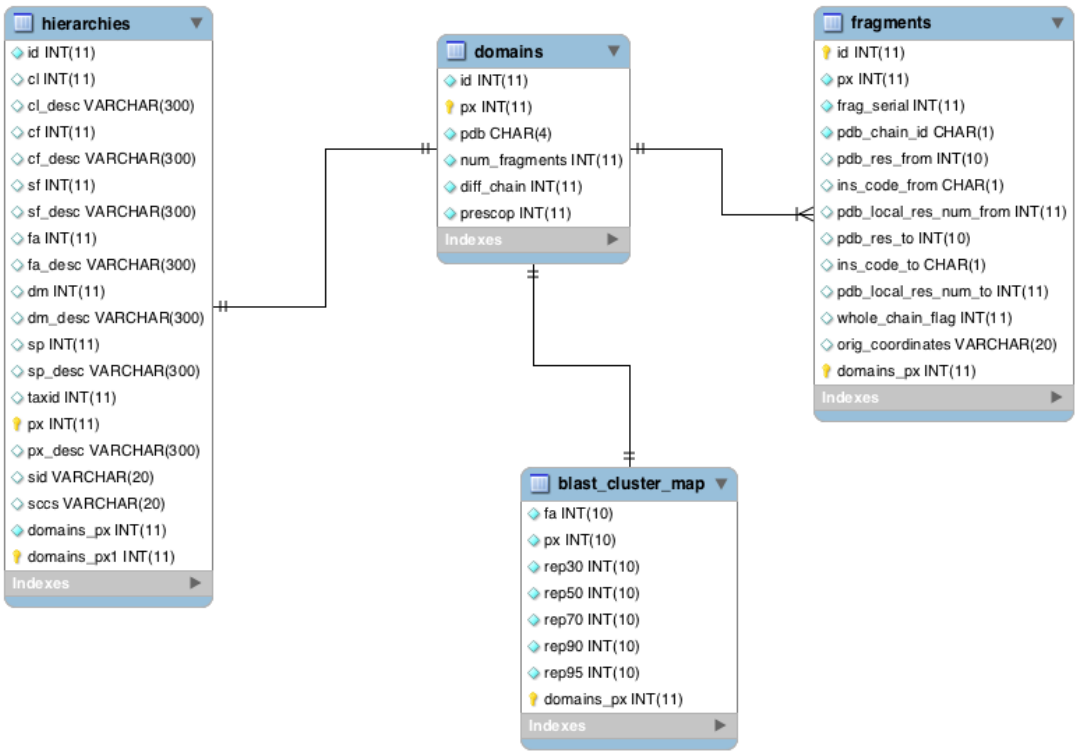


Figure 4.2: Schema of the SCOP database.

selected such that, if the coverage of the aligned region, as a proportion of the length of the sequence, was above 90% (in both directions) and the percent identity (PID) was above the specified threshold, the two sequences are considered to be neighbouring (neighbour relationships are therefore symmetrical). Single-linkage clustering adds a sequence to a cluster if the sequence is a neighbor to at least one sequence in the cluster. Once the clusters have been generated a representative structural domain is chosen as that member having the best QScore. The clustering was repeated at five PID thresholds (95%, 90%, 70%, 50% and 30%) to provide some flexibility in choice of resolution, and the results loaded into the *blast_cluster_map* table in the SCOP relational database.

4.2.3 BATON

The in-house structural alignment program BATON was used to generate the TOCCATA alignments. BATON has been developed over several years in the TLB group, initially in the form of COMPARER (Sali & Blundell (1990)) and more recently rewritten and updated by David Burke. As well as sequence similarity terms BATON uses a number of structural descriptors including secondary structure, hydrogen-bonding, solvent accessibility, disulphides, dihedral angles and sidechain orientation. When aligning two proteins of length i and j, BATON first generates an $i \times j$ matrix with each element populated with the sum of the difference scores of the different structural descriptors. The contribution S of each of structural descriptor D for each matrix element would be:

$$S = D_w \times \frac{(D_i - D_j)}{100} \tag{4.7}$$

where D_w is the relative weight assigned to the feature. As such, where structural features for the two residues are the same, the contribution to the total for that matrix element would be zero. Once all of the matrix elements have been populated, the standard Smith-Waterman dynamic programming algorithm is applied (Smith & Waterman (1981)) to identify the optimal lowest scoring diagonally-traversing path (Zhu *et al.* (1992)). Subsequent proteins can be iteratively aligned in a similar fashion by aligning the new sequence to existing alignment.

BATON uses as input the results of running the structural annotation program JoY (Mizuguchi *et al.* (1998a)) on each of the domains to be aligned. These residue-level annotations themselves comprise a valuable resource and as such are captured, reformatted and loaded into TOCCATA in relational form, providing pre-calculated structural attributes for every residue in every SCOP domain for use in later analyses.

The main output of BATON is the multiple alignment stored as text in "ali" format. These are reformatted and loaded into the TOCCATA database, the schema of which is shown in Figure 4.3. Importantly, the precise details of residue equivalences are stored in relational form. This is achieved by storing all alignment information with respect to the alignment column position, such that, for each alignment column, for each sequence, either an amino acid residue (with appropriate residue identifier) or a gap is stored. This scheme of storing the precise details of residue equivalences in high granularity offers the benefit of allowing several valuable analyses to be performed through simple SQL (Structured Query Language) database queries. For example, sequence entropies, relative sequence entropies (as described in Chapter 2), pairwise percent identity matrices and amino acid substitution tables can all be generated using only SQL. Further, the original alignments can be reconstituted into their original form but can also easily be sorted and filtered with respect to constituent proteins and residue range.

4.2.4 ESST generation

PICCOLO and TOCCATA were developed with distinct applications in mind. However, the fact that they share the same amino acid residue identifiers enables the resources to be combined to pursue several interesting avenues of enquiry - in particular investigation of the evolutionary properties of protein-protein interfaces. The first example of this is that of calculation of sequence entropies for the four structural environments described in Chapter 3 (interface core, interface periphery, core and exposed).

Combination of these structural environment residue annotations with the TOCCATA alignments permits exploration of the evolutionary plasticity of each

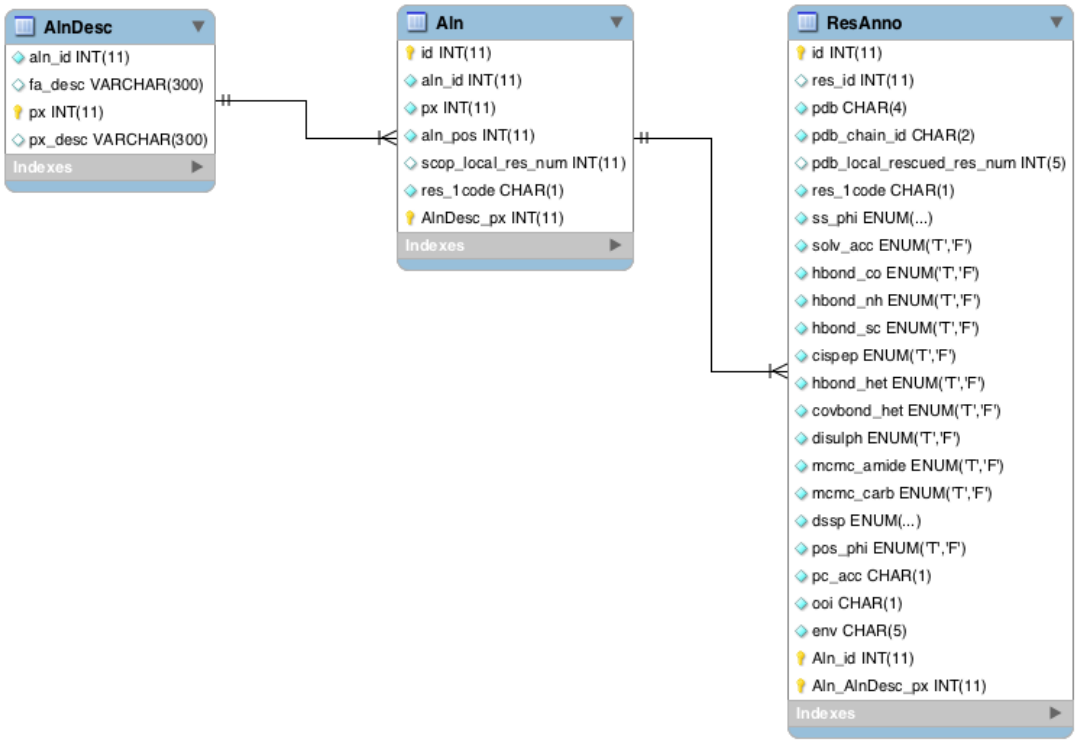


Figure 4.3: Schema of the TOCCATA database.

environment in the form of ESSTs. This is achieved by identifying those alignments from TOCCATA that include domains corresponding to structures from the PICCOLO PISA-derived quaternary structure complexes (see Chapter 2). Note that the fundamental unit of PICCOLO is that of the polypeptide chain, whereas that of TOCCATA is the structural domain, and as such the mapping between the two resources was performed at residue level. The resulting TOCCATA-PICCOLO residue mappings yielded 1,249 alignments of two or more domains.

In this work two series of ESSTs were generated. The first series is independent of protein interface definitions and was derived from a previously published series of 177 HOMSTRAD alignments (http://www-cryst.bioc.cam.ac.uk/~kenji/subst/) and was generated to validate the approach and to provide an overview of the prevailing structural determinants of substitutions observed in proteins, without reference to their oligomeric state. The 64 structural environments used correspond to the canonical environments used in earlier studies (i.e. combining terms describing secondary structure, solvent accessibility and hydrogen-bonding).

The second series extends this notion to include the impact of protein-protein interaction sites on the observed substitution patterns. The obvious route of simply extending the first series of 64 environments with additional interface terms was not viable, as the large number of environments results in data sparsity issues. Instead a series of 48 ESSTs were generated using a combination of:

- 4 categories of interface accessibility environments (interface core(i), interface periphery(I), core(a) and exposed(A))
- 4 categories of mainchain conformation and secondary structure (helix (H), strand(E), coil(C), and positive ϕ torsion angle(P))
- 2 categories of PICCOLO-derived intermolecular hydrogen-bonding (bonded(B) and unbonded (b))
- 2 categories of intramolecular hydrogen-bonding (bonded(W) and unbonded (w)) taken from the JoY annotations.

The set of combinations of these structural descriptors is enumerated in Figure 4.4. This series derives from the comprehensive set of TOCCATA-PICCOLO mapped alignments.

Unfortunately the data sets reflecting the different interface classes used in Chapter 3 (obligate homodimers, obligate heterodimers and transient heterodimers) were too small to permit derivation of class-specific ESSTs. Prototypic ESSTs were originally generated using native SQL by directly querying the relational database forms of TOCCATA and PICCOLO. Although this proved a rapid and effective way of generating conformationally-constrained ESSTs, the more sophisticated features such as smoothing and clustering are more straightforward to implement in a procedural context. Fortunately during the course of this work, Semin Lee, as part of his parallel work on protein-nucleic acid interactions, developed *Ulla* (`http://github.com/semin/ulla`). *Ulla* is a Ruby implementation of the legacy Fortran *subst* program and importantly includes functionality to use BLOSUM-like weighting procedure as well as the entropy-based smoothing procedures. Version 0.05 of *Ulla* was used in this work with BLOSUM-like weighting used at 60% identity (the default). For the interface-independent 64 ESST series partial smoothing was enabled and no conformational constraints were used. For the interface-dependent 48 ESSTs series substitution counts were conformationally constrained such that only those substitutions where the interface accessibility environment and secondary structure remained unaltered, were included. Note that the two series are not directly comparable as they derive from different alignment series with different structural descriptors and different input parameters.

4.2.5 Multidimensional scaling

Multidimensional scaling (MDS) is a useful tool for visualization of high dimensional data. The procedure aims to detect meaningful underlying dimensions that allow the observer to explain observed similarities or dissimilarities between the objects under investigation. For a matrix of n objects, theoretically $n-1$ dimensions would be required to visualize the data accurately. However, in cases where data inherently clusters (which is often the case with biological objects due to

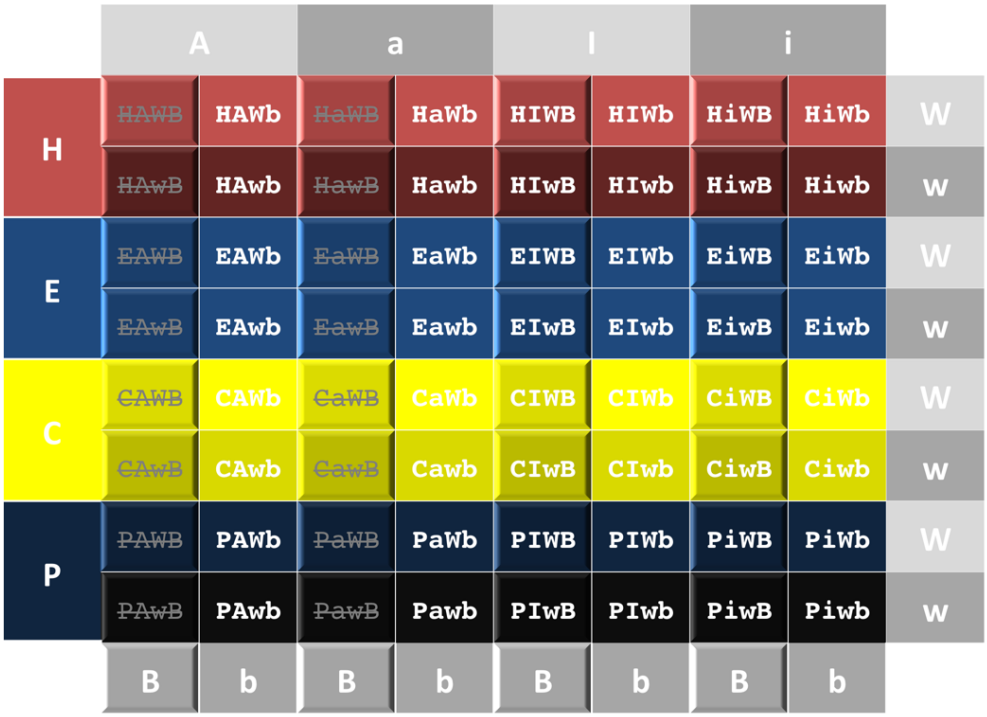


Figure 4.4: The set of combinations of structural descriptors used to generate the series of 48 interface-dependent ESSTs. Terms on the left hand side denote secondary structure (H=helix, E=extended, C=Coil, P=Positive phi). Terms across the top denote interface-dependent solvent accessibility environment (A=non-interface accessible, a=non-interface buried, I=Interface accessible, i=Interface buried). Terms on the right hand side denote intra-molecular hydrogen bonding (W=engaged in one or more intramolecular hydrogen bonds, w=engaged in no intramolecular hydrogen bonds). Terms along the bottom denote inter-molecular hydrogen bonding (B=engaged one or more intermolecular hydrogen bonds, b=engaged in no intermolecular hydrogen bonds). Note that by definition non-interface environments (a and A) cannot have inter-molecular hydrogen-bonds(B)(scratched environments).

the constraints of evolution), MDS can enable meaningful representation of the contents of the matrix in a lower number of visualizable dimensions (i.e. 2 or 3). Starting with a symmetrical $n \times n$ distance matrix of n objects, MDS attempts to position points in space so as to reproduce the observed distances optimally. To achieve this points are initially assigned to arbitrary spatial coordinates. In classical MDS the Euclidean distance amongst these points is calculated, to generate a new pairwise matrix. This matrix is compared with the input matrix by evaluating a stress function, such that the smaller the value, the greater the correspondence between the two. The coordinates of each point are then adjusted in the direction to optimally minimize the stress function and the procedure iterated to convergence. Two outputs are generated - the 3D co-ordinates and the Eigen values for each dimension. The absolute co-ordinates for each data point are oriented arbitrarily, but the relative proximity of the points to one another indicates their similarity. The Eigen values represent the information content of each dimension. The dimensions are ordered by their information content such that a scree plot of the Eigen values indicates what proportion of the total information in the matrix is being represented in a two- or three-dimensional visualization, allowing an objective assessment of the validity, or otherwise, of the analysis. Johnson *et al.* used a similar approach in the analysis of the determinants of various substitution tables (Johnson *et al.* (1993)).

4.3 Results and Discussion

The key features that distinguish TOCCATA from HOMSTRAD are the fact that the data are stored in relational database form, the consistent domain-family definitions (courtesy of SCOP) and the automated nature of the data generation. The current release of TOCCATA, based on SCOP version 1.73 plus pre-scop update from February 2008 comprises 89,964 SCOP domains in 3,965 alignments.

BATON, described above, was developed to align up to a few dozen structures at a time. Manual inspection of the results of the first attempts at broad scale alignment of several thousand SCOP families (some of which each contain several hundred constituent domains) revealed that although the vast majority of inspected alignments met with expectations, a significant minority exhibited

some problems. The main problem was that for some families, output alignments included unaligned (i.e. non-overlapping) sequences, even where the correct alignment appeared obvious, such that the subsequent sequences were simply concatenated at the C-terminal end of the alignment.

A separate issue was that some of the larger families failed to generate any alignment output. Through iterative interaction with David Burke, the BATON developer, these issues were largely resolved through a combination of bug-fixes and re-parameterization through exploration of new weighting schemes for the structural descriptors. A crude but useful metric for quantifying the effect of the "concatenation" problem was to compare, for each alignment, the ratio of alignment length to mean constituent sequence length. Figure 4.5 indicates the scope of the improvement achieved through this iterative parameter optimization process. The pronounced shift to the right-hand end indicated that the current alignments are more compact with significantly fewer gaps.

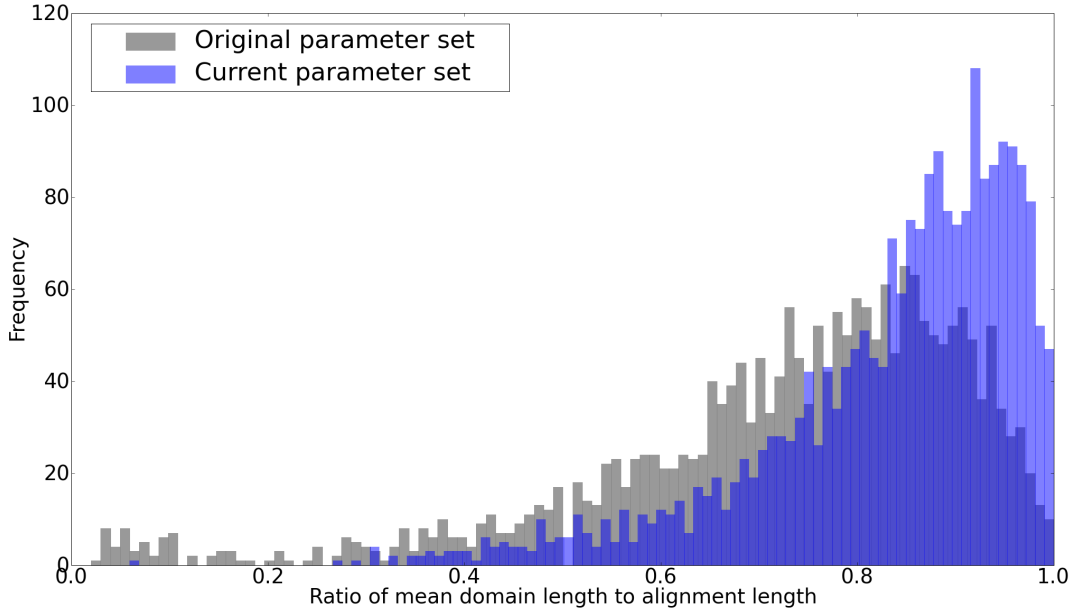


Figure 4.5: Distribution of ratios of alignment length to mean constituent sequence length for the original BATON parameter sets (grey) and the latest parameter set (blue).

The alignments generated by BATON are generally of high quality but despite these improvements they remain imperfect. Anecdotal observation suggests that the alignment of core secondary structure units appears robust, whereas long flexible loops and the N- and C-terminal region in particular, have more alignment gaps inserted than would a comparable hand-curated alignment. However, this could be attributed to BATON using particularly strict definition of structural equivalence. It is hoped that with continued development BATON can be further improved to provide routinely robust and accurate alignments of even the largest and most diverse families.

4.3.1 TOCCATA web interface

A simple web interface has been established to permit visualization of the TOCCATA alignments, primarily for the purposes of performing some quality control. JoY residue annotations are valuable for interpreting the determinants for evolutionary conservation across a family. It is also useful to have the capacity to examine subsets of the alignment at different levels of redundancy. This presents a problem in that under normal circumstances if any alterations or selections are made to the underlying alignment, JoY has to be re-run to re-annotate the alignment, requiring the presence of the appropriate domain-delineated PDB files, resulting in some difficult technical issues in dynamically presenting the formatted alignment through the web. A specially designed Cascading Style Sheet (CSS) was devised that included 64 distinct styles - one for each of the observed structural environments. Thereby, instead of running JoY "on the fly", the precalculated residue annotations stored in TOCCATA are combined to generate dynamically a style for each residue, resulting in a JoY-style rendering of the alignment annotations. This decoupling of the data from the software has the additional advantage that the database can be distributed with its interface without the dependency of having to install JoY and its ancillary programs.

`http://www-cryst.bioc.cam.ac.uk/toccata/toccata.php`

An example page is shown in Figure 4.6, a screenshot of a typical TOCCATA alignment, that of the interleukin 8-like chemokine family.

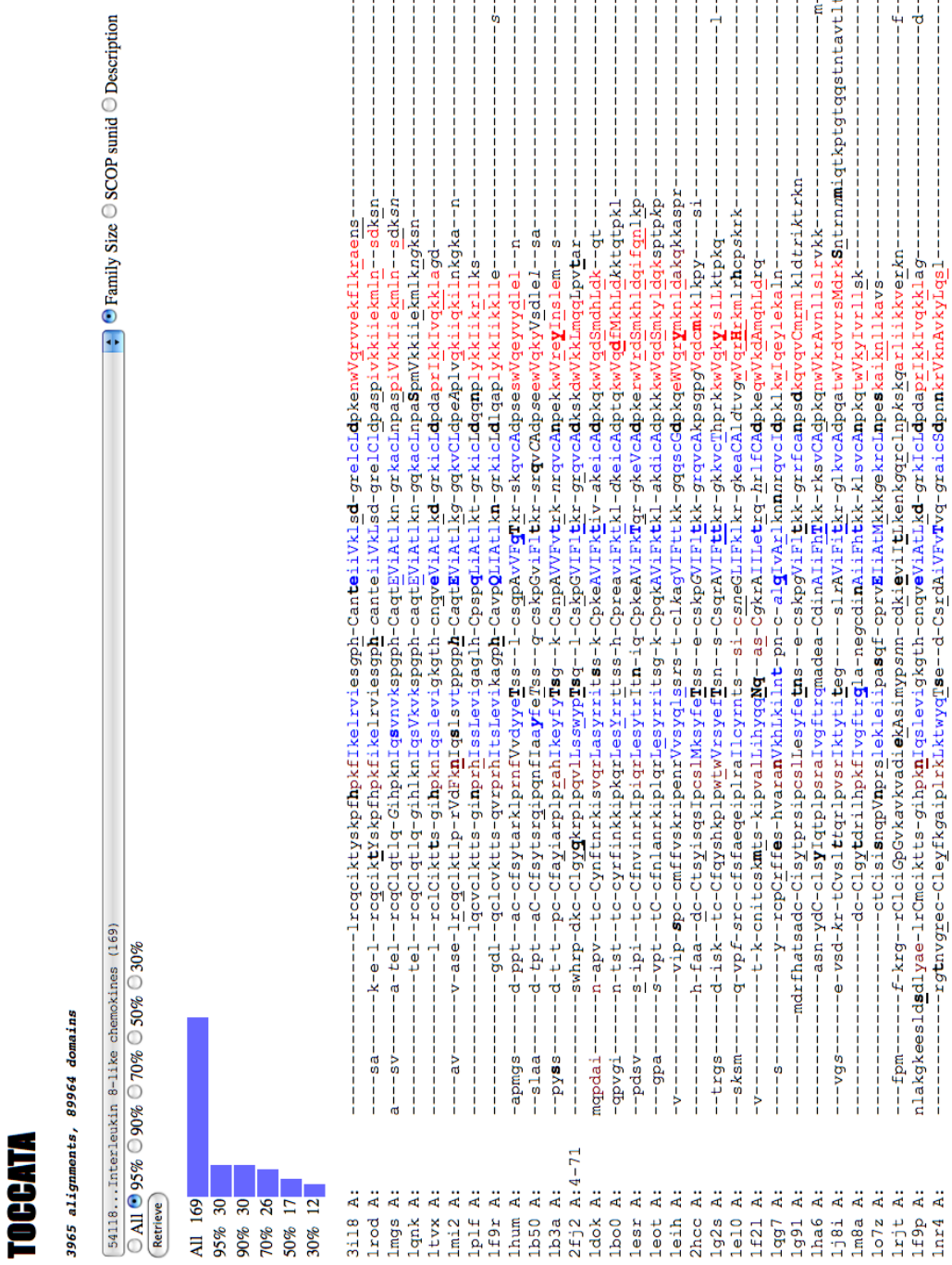


Figure 4.6: Screenshot of a typcial example of a TOCCATA alignment, that of the interleukin 8-like chemokine family. Residues are highlighted with JoY-style annotations. Solvent accessiblie residues are shown in lower case, buried residues in upper case. Residues with positive ϕ torsion angles are shown in italics. α-helices are shown in red, β-strands in blue, 3_{10} helices in maroon and coil in grey. Residues with hydrogen bonds to maichain amide are shown in bold, to mainchain carbonyl underlined.

The web page is written in PHP, a widely-used scripting language for generating dynamic web pages. The layout includes: a drop-down menu listing all available SCOP families (including the family identifier and the number of constituent domains in brackets), sortable by family size (the default), SCOP family identifier or family description; an option box where the level of redundancy to be retrieved can be selected; a graphical view of the size of the family at the various levels of redundancy; and finally the alignment itself displayed with JoY-style annotations. The sequences are ordered by the complete set of each their respective sequence cluster representatives, themselves ordered by increasing specificity (that is 30% identity cluster representative followed by 50%, 70%, 90% and then 95%) to give an approximation of a tree-based proximity ordering.

During the development process the issue arose as to whether residue equivalences in a structure-based alignment depend on their context; that is, if a subset of proteins is extracted from a multiple alignment, is the resulting alignment meaningfully distinguishable from the *a priori* alignment of the subset? In order to establish the extent of this issue a second flavour of TOCCATA was generated. In the original version of TOCCATA a single redundant alignment was generated for all domains in each SCOP family. In order to view each alignment, rows corresponding to each sequence cluster representative are extracted from the parent redundant alignment with no further re-alignment taking place (although alignment columns comprising only gaps are excluded). In the second flavour of TOCCATA, entitled TOCCATANR, the non-redundant sets are used as input to BATON, that is 6 alignments are performed for each family, corresponding to the 5 levels of sequence clustering plus the full redundant set. TOCCATANR results can be viewed through the following URL:

`http://www-cryst.bioc.cam.ac.uk/toccata/toccatanr.php`

The two URLs permit comparative visualization of the two data sets. Although the majority of alignments are almost indistinguishable, anecdotal evidence suggests that the TOCCATANR alignments are marginally more compact.

4.3.2 ESST distance matrix

The result of running *Ulla* is a library of ESSTs. Figure 4.7 displays the 64 ESSTs comprising the interface-independent series of substitution tables. Figure 4.8 displays the equivalent visualization of the 48 ESSTs comprising the interface-dependent substitution tables.

Whereas each of these substitution tables can be examined independently, comparative analyses can be difficult to interpret. A crude but informative overview of the whole library can be produced by calculating the Euclidean distance between each pair of matrices to generate a summary 64x64 distance matrix. This distance matrix is then amenable to analysis through MDS to provide a two dimensional projection of the data. Figure 4.9 shows the results of performing MDS on the distance matrix derived from 64 ESSTs comprising the interface-independent series.

Although the results initially appear unclear there are some strong underlying patterns. The projection suggests that the single strongest determinant appears to be made by the solvent accessibility terms - the buried and exposed environments are distinct from one another. With respect to the hydrogen-bonding terms, at this low resolution, no significant discrimination is exhibited between environments corresponding to mainchain amide, mainchain carbonyl and sidechain to sidechain hydrogen bonding. However there is a clear axis (from top to bottom) with respect to the number of occupied hydrogen bonding terms. Those towards the top of the projection have no hydrogen bonds (and exhibit the greatest diversity from one another). Below this a band can be found corresponding to environments having one occupied hydrogen bond term (either mainchain carbonyl, mainchain amide or sidechain), followed by a band where two hydrogen bonding environments are occupied. Finally towards the bottom of the projection environments corresponding to residues that have at least three hydrogen bonds (one to mainchain carbonyl, one to mainchain amide and one to other sidechain) can be found.

However, it should be borne in mind that these methods will always reflect the relative importance of the environmental descriptors chosen in the analysis. Many

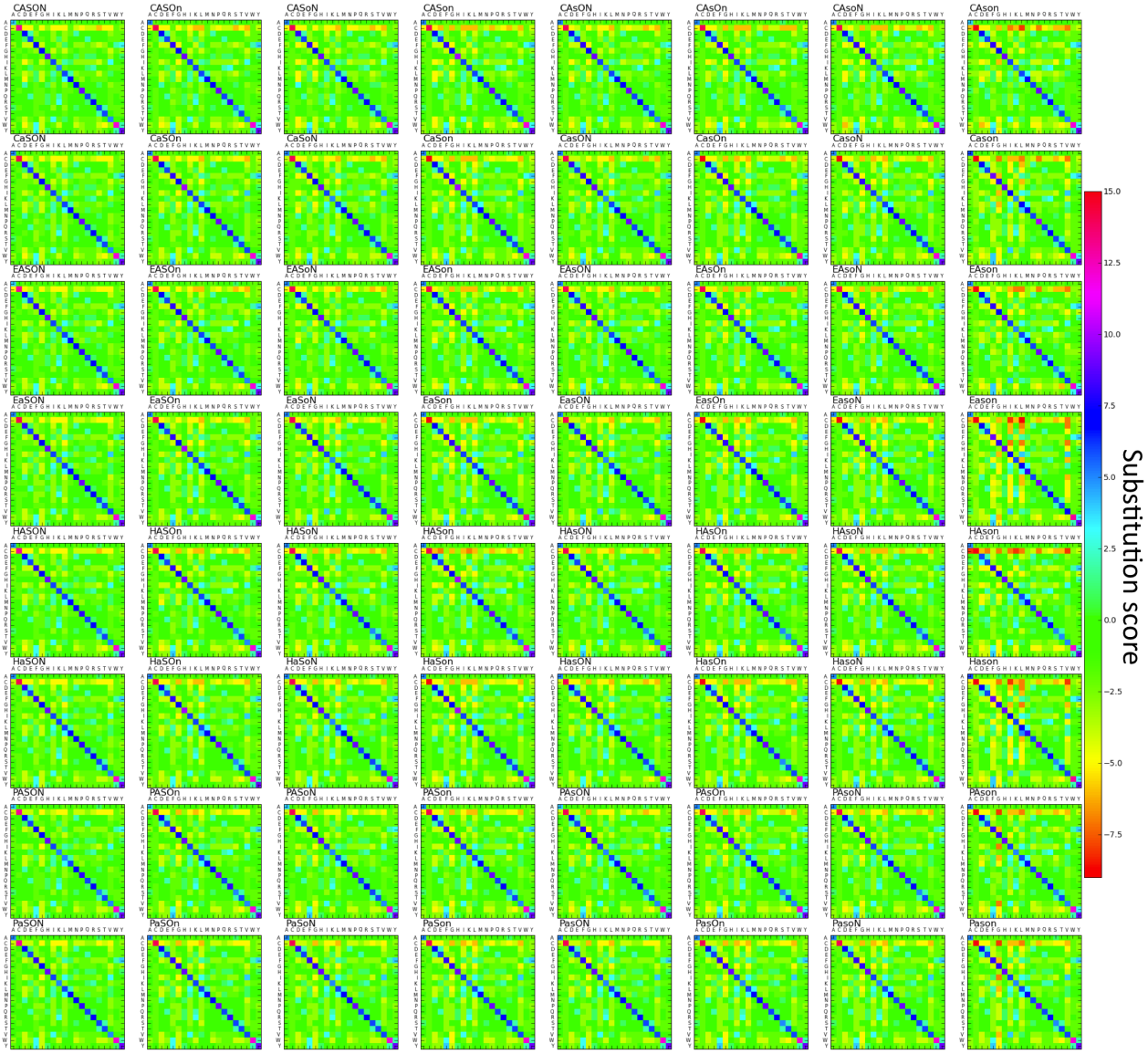


Figure 4.7: Library of ESSTs for each of the 64 interface-independent structural environments. Each 20 × 20 matrix reflects the likelihood of substituting one residue with another in a particular structural environment. Colours represent residue substitution scores calculated as the log-odds of the Substitution Frequency Matrix as described in Equation 4.3. Environment labels describe secondary structure, solvent accessibility and hydrogen bonding status as described in Figure 4.9. A larger version of this image can be found at http://www-cryst.bioc.cam.ac.uk/~richard/64ESSTs.png.

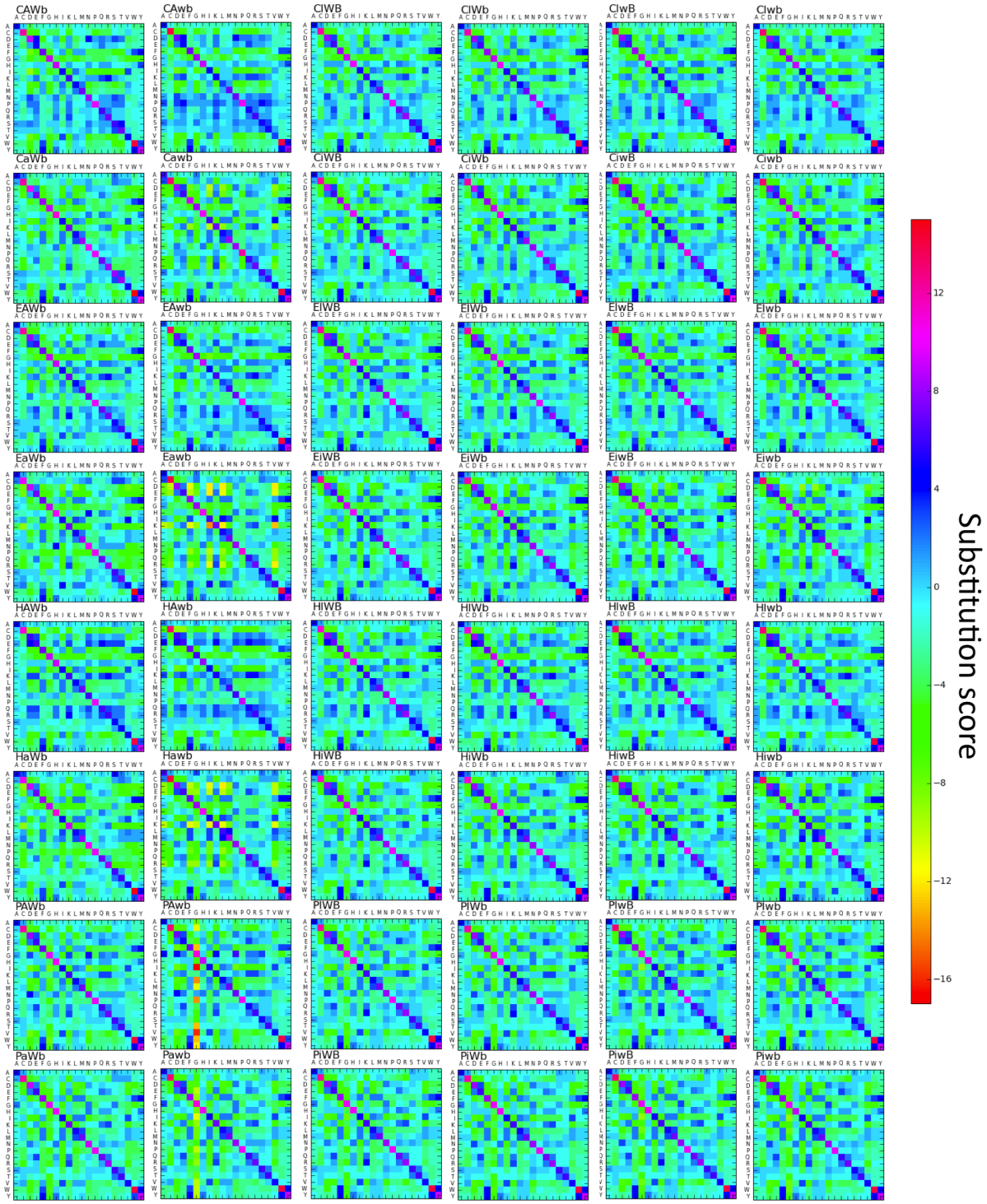


Figure 4.8: Library of ESSTs for each of the 48 interface-dependent structural environments. Colours represent substitution scores as described in Equation 4.3. Environment labels describe secondary structure, interface-dependent solvent accessibility and hydrogen bonding status as described in Figure 4.4. A larger version of this image can be found at http://www-cryst.bioc.cam.ac.uk/~richard/48ESSTs.png.

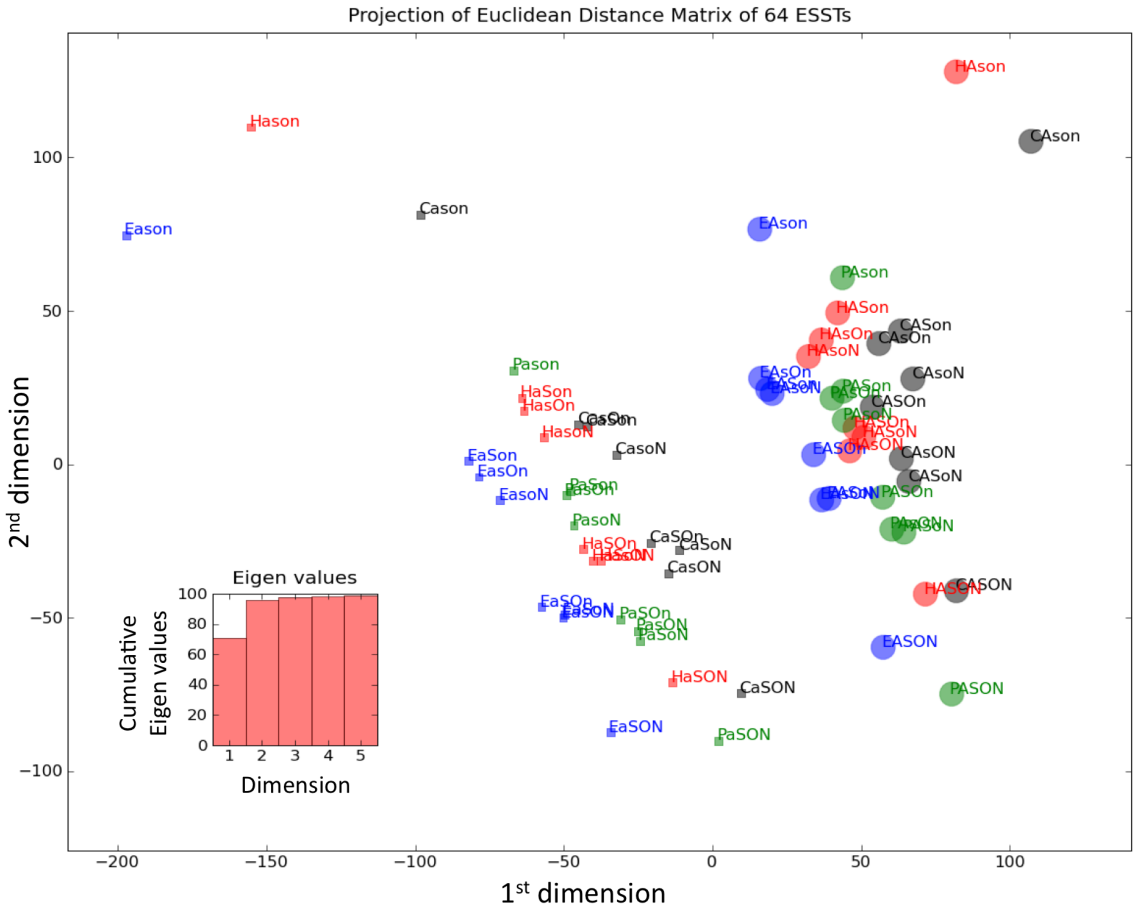


Figure 4.9: MDS projection of the 64 ESSTs of the interface-independent series. In such projections the absolute coordinates are meaningless; relative proximity of points indicates similarity. Small squares represent buried environments and large circles exposed. Colour indicates secondary structure (helices in red, strands in blue, coil in grey and positive ϕ torsion angle residues in green). Point labels use a 5 character shorthand for each structural environment where the first character describes one of the 4 secondary structure environments (H=helix, E=extended, C=Coil, P=Positive ϕ), the second character represents the solvent accessibility (A=accessible, a=buried), the third character represents hydrogen bonding status to a sidechain or heterogen (S=True, s=False), the fourth character represents hydrogen bonding status to a mainchain carbonyl (O=True, o=False), and the fifth character the hydrogen bonding status to a mainchain amide (N=True, F=False). The inset shows a histogram of the cumulative Eigen values for the first 5 dimensions, suggesting that 95.7% of the total information in the matrix can be visualized in the first two dimensions.

other determinants, structural and otherwise, are likely to be making contributions. The recent work of Worth and Blundell (Worth & Blundell (2009)) suggests that saturation of the hydrogen-bonding capacity of buried, polar residues can be a key indicator of a residue's importance in stabilizing structural domains, as reflected by their observed high levels of conservation. Such hydrogen-bonding saturation was not explicitly captured in this study (although it may be reflected in the highly-bonded terms) but will be captured unambiguously in future work. Within each of these groups the subsequent determinant would appear to be the secondary structure. Environments corresponding to helices, strands and coils can be discriminated within each accessibility/hydrogen-bonding occupancy combination.

The smoothing procedures are valuable for correcting for partially missing data. However, ESSTs derived from environments that are particularly sparse will be composed largely of information derived from adjacent environments. As such, the smoothing procedures will have the effect of dampening the observed difference between environments. However, the fact that coherent discrimination is evident, corresponding plausibly with prior knowledge of structural determinants, suggests that sufficient signal remains. One way to appraise such issues quantitatively is to assess the occupancy of each environment prior to smoothing, that is, examine the proportion of each of the 400 possible residue combinations that is populated by one or more observed substitutions. These data are shown for the 64 environments from the interface-independent set in Figure 4.10. The data suggest that in this series only a small proportion of environments approach full occupancy. Environments corresponding to positive ϕ torsion angles have particularly low occupancy; these environments are dominated by glycine (the absence of a sidechain gives it particular flexibility conformational). Furthermore, occupancy is a qualitative measure, in that an element is either occupied or unoccupied - many environment specific residue-substitutions may be occupied but have only a small number of observations giving marginal statistical significance and as such occupancy represents the upper bound for estimating ESST information quality. Together these aspects indicate that the results ought to be interpreted cautiously.

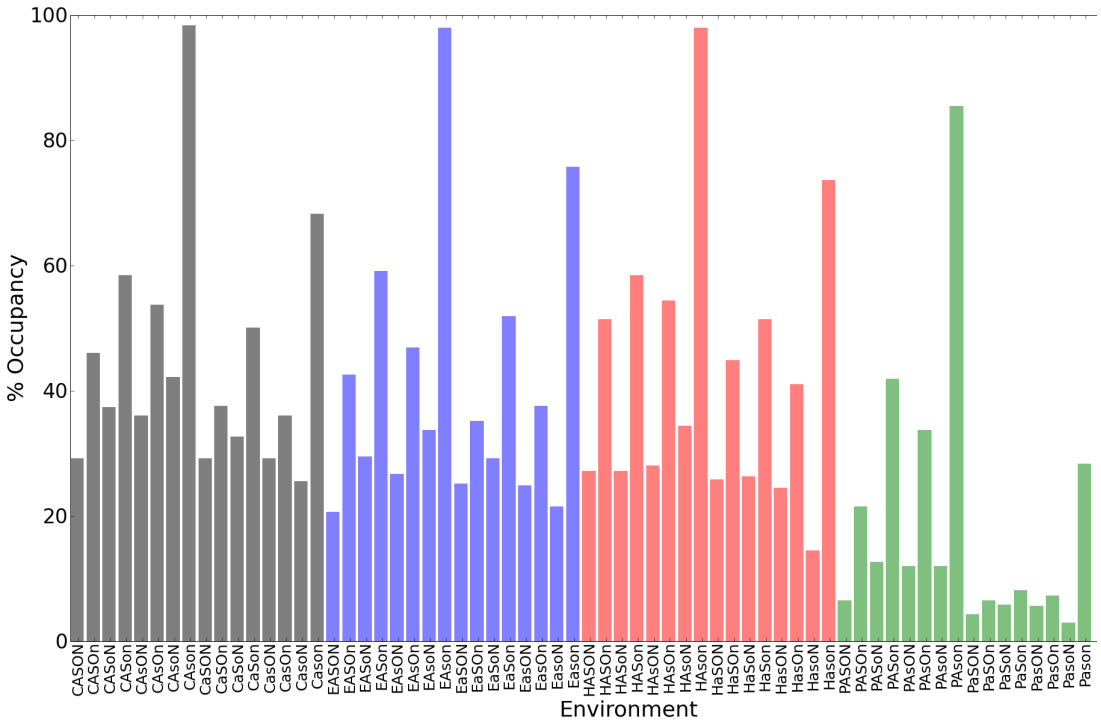


Figure 4.10: % occupancy of the 64 ESSTs from the interface-independent series.

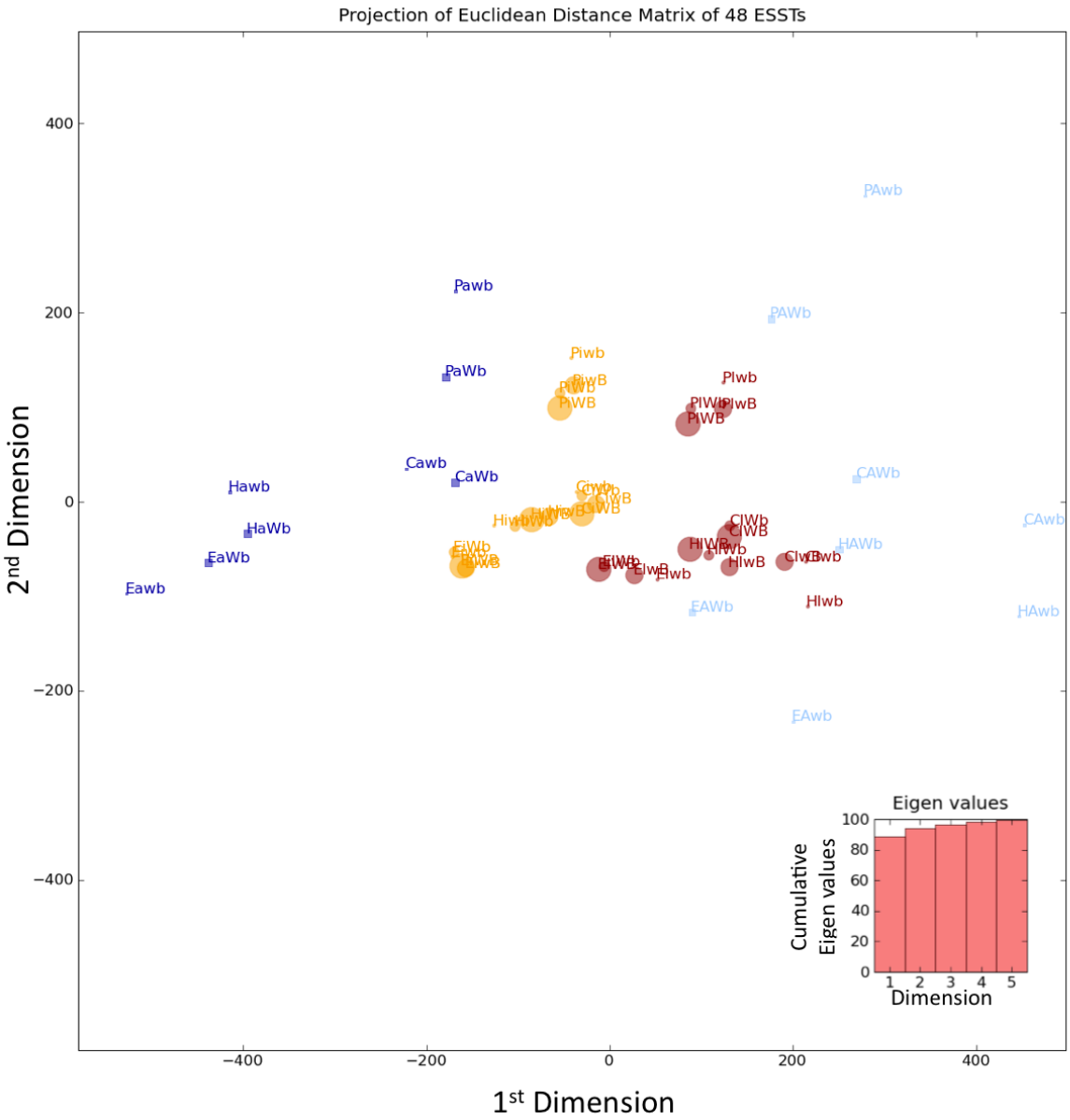


Figure 4.11: MDS projection of the 48 ESSTs of the interface-dependent series. Interface core environments (i) are shown as orange circles, interface periphery(I) as red circles, non-interface exposed environments(A) as light blue squares and non-interface buried environments(a) as dark blue squares. Increasing size corresponds to increasing number of hydrogen bonds: No hydrogen bonds (wb) < Intra-molecular hydrogen bonds only (Wb) < Inter-molecular hydrogen bonds only (wB) < Inter- and Intra- molecular hydrogen bonds (WB). The Eigen value analysis (inset) suggest that 94.8% of the total information in the matrix can be visualized in the first two dimensions.

Figure 4.11 shows the results of performing MDS on the distance matrix derived from 48 ESSTs comprising the interface-dependent series. Here the strongest determinant is the interfacial accessibility environment (I/i/A/a). The interface environments are intermediate between the exposed surface and buried core. Further the interface core is more similar to the buried protein core and the interface periphery is more similar to the exposed surface. The next strongest determinant would appear to be secondary structure (H/E/P/C), with environments corresponding to positive ϕ torsion angles being the most differentiated. Of the non-positive ϕ torsion angle terms, the non-interface environments are less congregated than those participating in interfaces. At this resolution hydrogen bonding (either inter- or intra- molecular) would appear to be a weak determinant, particularly for the interfacial environments, although non-hydrogen bonded conditions (wb) tend to be outliers of their respective groups.

Figure 4.12 shows the % occupancy of the 48 ESSTs from interface-dependent series. This series has significantly higher occupancy than the interface-independent series (Figure 4.10) which adds to the confidence in the analysis. Once more the environments corresponding to the positive ϕ torsion angle have the lowest occupancy. The higher occupancy in this series is due largely to the increased number of alignments used in this series although the conformational constraints reduce the number of possible observations.

Overall this evolutionary analysis largely corroborates the results of the physicochemical analysis of interface properties in Chapter 3 in that the properties of the core and periphery of the protein interface are distinguishable from one another with the core most resembling the buried protein core and the periphery the remaining exposed surface.

4.3.3 Future developments

Using family definitions from SCOP carries the advantage of providing a consistent underlying framework for alignment delineation and evolutionary breadth, and obviates the bulk of the manual curation. However, the drawback of this dependency is that the alignments are only as up to date as the latest SCOP release. Ideally SCOP would provide more frequent updates but the workaround

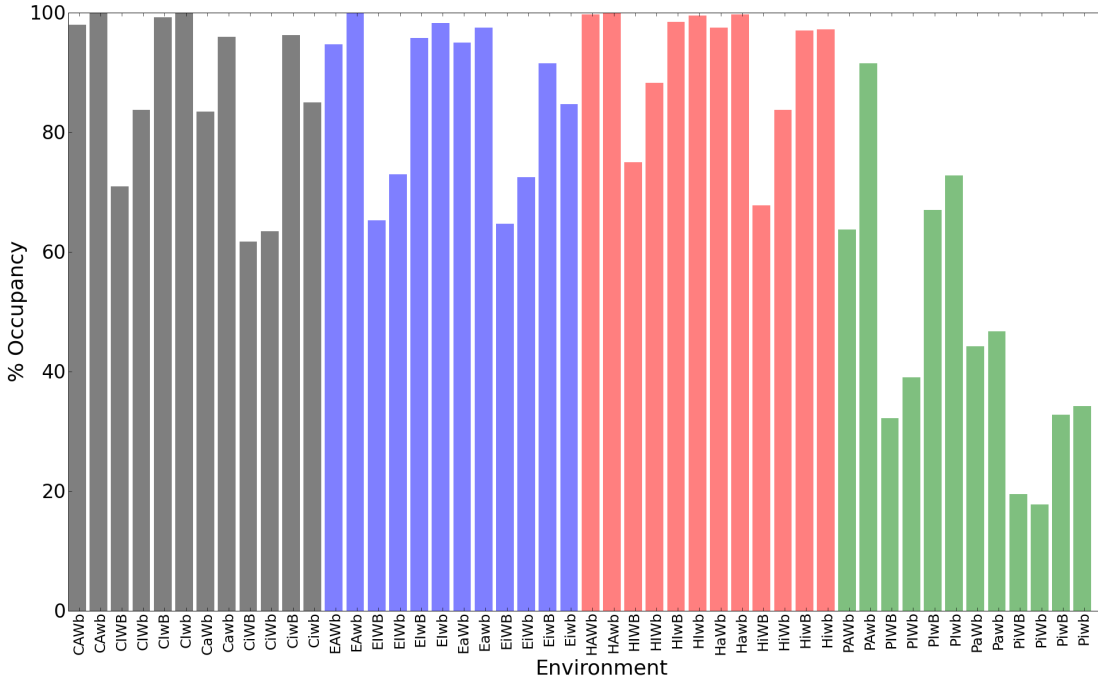


Figure 4.12: % occupancy of the 48 ESSTs from the interface-dependent series.

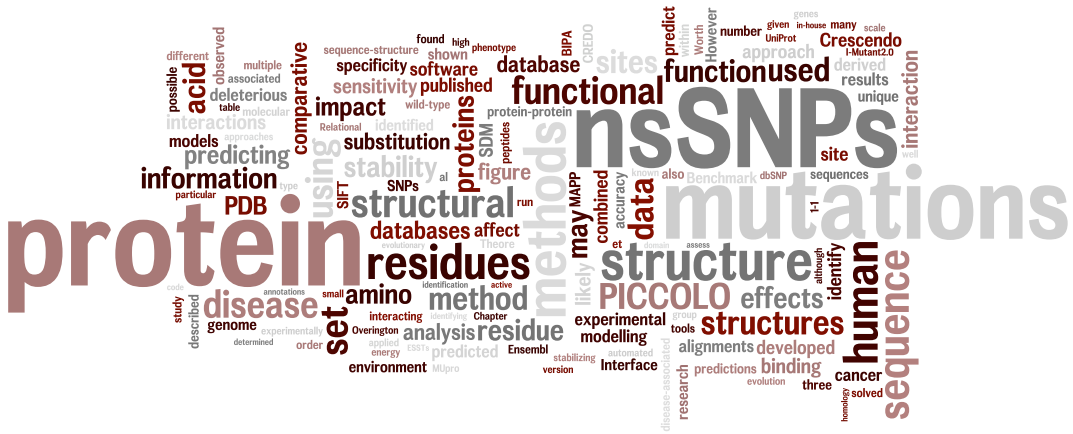
solution to this problem would be to generate *Fugue* profiles for each of the existent SCOP families. These could then be used as a search database to query against for all PDB polypeptide chains that are *not* classified by SCOP. *Fugue* could be used to delineate domains based on the best non-overlapping matches and classify the putative domains into the SCOP hierarchy as appropriate.

Use of TOCCATA for comparative modelling results in models of each structural domain being generated independently. This is the case for multidomain proteins even where suitable multidomain template structures exist, resulting in a loss of useful information with respect to the relative orientation of the domains. This is particularly pertinent to the genome-scale comparative modelling described in Chapter 5. To ameliorate this situation a second parallel set of BATON multiple structural alignments ought to be derived corresponding to clustered sets multidomain polypeptide chains. Preliminary analysis reveals that the number of unique, order-dependent combinations of domain families, of size greater than one, observed in the PDB is 1,106.

Aside from their utility in characterizing existing interactions the substitution tables could also have potential in the area of predicting novel protein-protein interactions through homology. Interologs are defined as homologous *pairs* of interacting proteins. That is, if proteins A and B bind to one another and a homologue of A (A') binds to a homologue B (B') then A' and B' are interologs of A and B. However not all homologous pairs of proteins are genuine interologs. Trivially, homologous pairs whose spatio-temporal expression patterns are distinct have no opportunity to interact. However, more interestingly, the effect of evolutionary divergence can be assessed using appropriate substitution tables. Naively, it might be expected that the likelihood of a homologous pair of proteins being genuine interologs is a simple function of the combined sequence similiarity of the two protein pairs. However, a more likely discriminator would be the level of similarity of the interface region, which may or may not correspond to the overall similarity. As such, examples can be found (e.g. Figure 1.2) of instances where relatively close homologues have interface regions that have diverged considerably, and conversely, of distant homologues with conserved interface regions. Interface-specific substitution tables, such as those described here, would be ideally suited to assess such interface similarity by specifically identifying those homologous pairs whose interface regions retain sufficient similarity to enable the mode of interaction to be maintained. This procedure would assess whether each constituent partner has retained some interface-like capacity through evolution, but says nothing regarding the possible compatibility of the putative interolog surfaces. One avenue to address this issue would be to extend the methodologies used to derive 20×20 amino acid substitution tables to investigate the patterns of substitutions of *pairs* of residues across homologous interfaces. The resultant 400×400 pairwise substitution matrix could be used in combination with PICCOLO and the two pairwise alignments to establish whether the putative interolog interaction residue pairs are likely to be compatible. However data sparsity is likely to be an even more significant issue than it is with the standard 20×20 substitution tables. A further significant obstacle to any such interaction-prediction methods is the lack of high quality true-negative data in order to accurately benchmark the method.

Chapter 5

Mutations



5.1 Introduction

Prediction of the effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on molecular function depends critically on exploiting all information available on the three-dimensional structures of proteins. A series of software and database methods have been developed for the analysis of nsSNPs that allow navigation from SNP to sequence to structure to function.

In both structure prediction and in the analysis of the effects of nsSNPs, information about protein evolution is exploited, in particular that derived from investigation of the relation of sequence to structure gained from the study of amino acid substitutions in divergent evolution. The techniques developed in the group allow fast and automated sequence-structure homology recognition to aid identification of templates in order to perform comparative modelling, as well as simple, robust and generally applicable algorithms to assess the likely impact of amino acid substitutions on structure and interactions. A strategy for approaching the relationship between SNPs and disease is described along with the results of benchmarking the approach on a set of human proteins of known structure and recognized mutation. The work described in this Chapter was originally published in 2007 (Worth *et al.* (2007a)) and is presented here with some of the results updated where indicated.

A major goal of current research on the human genome is to associate genetic variation with disease. In order to address this issue, many genome-wide association studies are being carried out to identify those genetic variants associated with disease phenotypes. Customarily, single nucleotide polymorphisms (SNPs) are the subset of single-base variants with a Minor Allele Frequency (MAF) greater than 1%, in a given population. A proportion of SNPs identified in these studies will alter protein sequences, termed non-synonymous SNPs. A nsSNP may affect the structure and/or the function of the encoded protein and where protein function is modified this may lead to disease. Disease can also be caused by gain in function which may result from either irregular/tighter binding with ligands or binding a wider range of ligands i.e. loss of specificity. However, modulation or disruption of protein structure or function may be necessary but not sufficient conditions for disease given the multiple redundancies of cellular pathways. Further, it should be remembered that non-synonymous mutations may also have any of the pre-translational effects normally associated with a synonymous SNP, for example by having an impact on transcriptional regulation, mRNA stability, splicing or translation rates (Kimchi-Sarfaty *et al.* (2007)).

nsSNPs can affect the function of a protein in many ways, four of which are of interest here. Firstly, nsSNPs may affect the functional residues of a protein i.e. the active site or a protein-protein interaction site, resulting in either loss or gain

of protein function and hence affecting the molecular pathway within which the protein operates (Dai *et al.* (2001)). Secondly, nsSNPs may affect the stability of a protein by either destabilizing it (increasing the ratio of unfolded protein to folded protein) or stabilizing it (decreasing the ratio of unfolded protein to folded protein) (Pakula & Sauer (1989)), this is also likely to affect function. A third effect of nsSNPs, related to protein stability, is that of causing protein aggregation (Palsdottir *et al.* (1988)). Lastly, nsSNPs may alter post-translational modifications, for example by inserting/deleting protease cleavage sites, glycosylation sites etc. Knowledge of the three-dimensional structure of a protein is useful for analyzing sequence variations by helping to identify the role that an amino acid may have in each of these four aspects.

Given the large and accelerating volume of human mutation data (Abecasis *et al.* (2007); Rocha *et al.* (2006)) being generated by high-throughput array-based genotyping methods (Gunderson *et al.* (2006)) and imminently by ultra-high-throughput sequencing techniques (Bennett *et al.* (2005); Sundquist *et al.* (2007)), robust, high-performance *in silico* tools are required that can enable prioritization of mutations; genome-wide association studies commonly identify large sets of candidate SNPs (Wang *et al.* (2005)) and often it is unclear which are causative. Automated methods of identifying those mutations most likely to confer susceptibility to or protection from complex diseases will enable a more rational approach to experimental verification of disease associations. Relational database tools are a pre-requisite to the storage, integration and analysis of such large and dynamic data sets.

5.1.1 Public domain methods

A variety of different approaches have been established for predicting the severity of non-synonymous mutation on proteins. Methods utilizing sequence information have the advantage of greater coverage than those that rely on protein structure data. The SIFT (Sorting Intolerant From Tolerant) program predicts whether a nsSNP will affect protein function by calculating a scaled probability for the substitution (Ng & Henikoff (2001)). The probability score is derived from the

observed frequencies of amino acids at the nsSNP position in a homologous sequence alignment of the protein of interest. The method was applied to 3,084 nsSNPs from the National Center for Biotechnology Information's (NCBI) dbSNP database (Sherry *et al.* (2001)) and predicted that 25% would affect protein function (Ng & Henikoff (2002)). Clifford *et al.* (Clifford *et al.* (2004)) also utilized position-specific scoring matrices to predict deleterious nsSNPs. However, their method used the HMMER2 (Eddy (1998)) software suite to predict whether a substitution will affect the fit of a protein sequence to its relevant PFAM (Finn *et al.* (2008)) motif model. They observed that the magnitude of the change in HMMER E-value caused by amino acid substitutions in HIV-1 protease and HIV-1 reverse transcriptase is a good predictor of whether it is deleterious. A different approach to predicting deleterious nsSNPs using sequence alone is used by the program MAPP (Multivariate Analysis of Protein Polymorphism) (Stone & Sidow (2005)). MAPP uses alignments of orthologous sequences to quantify the physiochemical characteristics of each position of the protein of interest and provides a continuous classification of nsSNPs. The method was shown to make slight improvements on the predictions made by SIFT.

A disadvantage of sequence-based methods is that they are unable to distinguish the evolutionary restraints that contribute to the conservation of a residue i.e. the functional and structural restraints. Sunyaev *et al.* (Sunyaev *et al.* (2001)) developed a set of rules to predict deleterious nsSNPs based on physical features (e.g. properties derived from crystal structures such as active sites, disulphide bonds etc) and comparative considerations (multiple sequence alignment profile scores). The method has been implemented as a web server, PolyPhen, for automated functional annotation of nsSNPs and has been used to annotate all SNPs deposited in the HGVbase database (Fredman *et al.* (2002)). Analysis of the structural characteristics of disease mutations indicated that various effects on protein stability are responsible for accumulation of deleterious nsSNPs in human genes (Ramensky *et al.* (2002)).

It has been estimated that up to 80% of disease-associated nsSNPs are caused by protein stabilization effects (Wang & Moult (2001)). Therefore, methods that predict the effect that nsSNPs will have on protein stability are useful for identifying possible disease-associations. This has previously been approached

by predicting the structural effects of mutations using: (1) molecular mechanics approaches (Bash *et al.* (1987); Funahashi *et al.* (2003); Kollman *et al.* (2000); Park & Lee (2005)); (2) empirical energy functions which are fitted to experimental data using weighted terms incorporating physical and statistical factors with structural knowledge (Bordner & Abagyan (2004); Guerois *et al.* (2002)); (3) machine learning methods, such as Support Vector Machines (SVMs) and neural networks (Capriotti *et al.* (2004, 2005a); Cheng *et al.* (2006); Frenz (2005)) and statistical potential energy functions which are derived using statistical analysis of information from protein databases (Gilis & Rooman (1997); Saraboji *et al.* (2006); Topham *et al.* (1997)). The SVM method I-Mutant2.0 published by Capriotti *et al.* (Capriotti *et al.* (2005b)), which incorporates information about thermodynamic experimental conditions, the wild-type and mutant amino acid types and the spatial environment of the residue, gave a high accuracy of predicting the sign of stability change and a high correlation between experimental and calculated thermodynamic data. A similar method was employed in developing MUpro, which implements three SVMs to predict stability changes for SNPs using just sequence, just structure and sequence with structure, combined with information of the wild-type and mutant residues (Cheng *et al.* (2006)). All three SVMs performed similarly in the task of predicting the sign of stability change and showed the highest correlation coefficients between predictions and experimental data. However, all three performed badly in the task of predicting stabilizing mutations, predicting more than 70% of stabilizing mutations as being destabilizing.

5.2 Methods

The approach developed involves applying in-house software to predict the effects that mutations have on protein structure and function and in-house relational databases to store (1) the results of running the in-house software (2) a comprehensive inventory of functional sites observed in solved structures and (3) accurate mapping of mutations to protein sequences and structures. The problem is explicitly broken into two different parts. First the effects of nsSNPs on known three-dimensional structures of individual proteins or complexes must be

predicted. The second challenge is to identify parts of the defined protein that might be involved in functional interactions. The software tools used to achieve this have been previously published (Burke *et al.* (2007); Chelliah *et al.* (2004); Worth *et al.* (2007b)); here their systematic application at the required genome scale is described. Together, the software and databases enable simple navigation from SNP to sequence to structure to function.

The approach taken to predicting the impact of nsSNPs depends critically on exploiting all information available on the 3D structure of proteins and combining it with information derived from observations of protein evolution. These observations are themselves derived from the careful study of amino acid substitutions in divergent evolution. As described in Chapter 4, these observations were originally compiled as a database of high-quality structural alignments of protein families in the form of the HOMSTRAD resource (Mizuguchi *et al.* (1998b)).

A key observation is that the likelihood that an amino acid substitution will be accepted through evolution depends strongly on the local environment of the amino acid sidechain (Overington *et al.* (1990, 1992)).This is illustrated by the fact that residues buried within the core of a protein tend to be more conserved than those on the surface on account of buried residues having a role in maintaining the structure of a protein (Overington *et al.* (1992); Shakhnovich *et al.* (1996); Zhou & Zhou (2004)). Functional residues, such as those in protein interaction and active sites, are also highly conserved, as they are required to make specific interactions on formation of a functional complex and although they are typically found on the surface of proteins, they are also likely to be buried upon complex formation. The idea that patterns of amino acid substitution are highly dependent on local structural environment has been used to derive a library of environment specific substitution tables (ESSTs) (Overington *et al.* (1990, 1992)), with local structural environments being defined on the basis of secondary structure, hydrogen bonding and solvent accessibility. These tables provide quantitative information about the existence of an amino acid in a particular structural environment and the probability of it being substituted by any other amino acid in that environment (Overington *et al.* (1992)). These principles have been applied to the areas of sequence-structure homology recognition (Shi *et al.* (2001)), structure prediction (Wako & Blundell (1994a,b)) as well as in the analysis of the

effects of nsSNPs. Our overall approach is comparable to previously published approaches at genome scale modelling and SNP analysis (Karchin *et al.* (2005)), although our approach is distinguished by our explicit use of evolutionary information throughout the modelling and nsSNP analysis as well as the robust and rapid performance of our nsSNP impact algorithms.

In order to maximize structural representation, experimentally determined protein structures are used where they are available and comparative models are built where they are not. Thus the first stage is an inventory of experimental structures available for proteins encoded in the human genome, followed by genome-wide automated sequence-structure homology recognition and comparative modelling of each of those proteins whose structure has not been experimentally determined.

5.2.1 Protein modelling pipeline

Experimentally determined structural information is restricted to 2,422 human genes in Ensembl release 45 (Hubbard *et al.* (2007)), or approximately 10% of the human genome (figure derived from PDB SIFTS (Velankar *et al.* (2005)) as described in Chapter 2). In order to extend our predictions to a wider set of genes, we employ a range of in-house programs to build comparative models of proteins (Figure 5.1).

The sequence structure-homology recognition program *Fugue* (Shi *et al.* (2001)), exploits ESSTs to identify distant homologues of a query sequence. A range of programs have been developed to build protein structural models using experimental and knowledge-based approaches. *Composer* (Sutcliffe *et al.* (1987)), *Modeller* (Sali & Blundell (1993)) and *Choral* (Montalvao *et al.* (2005)) build comparative models using restraint-based or fragment-based approaches. *Harmony* (Shi (2001)) validates models by comparing observed sequence amino acid substitution patterns with those predicted from ESSTs. More recent developments focus on *RAPPER* (de Bakker *et al.* (2006)) and *Rapper-tk* (Gore *et al.* (2007)), discrete conformational sampling tools that build ensembles of conformers under experimental and knowledge based restraints. Models of all mutant sequences are constructed using *Andante* (Smith *et al.* (2007)) which predicts

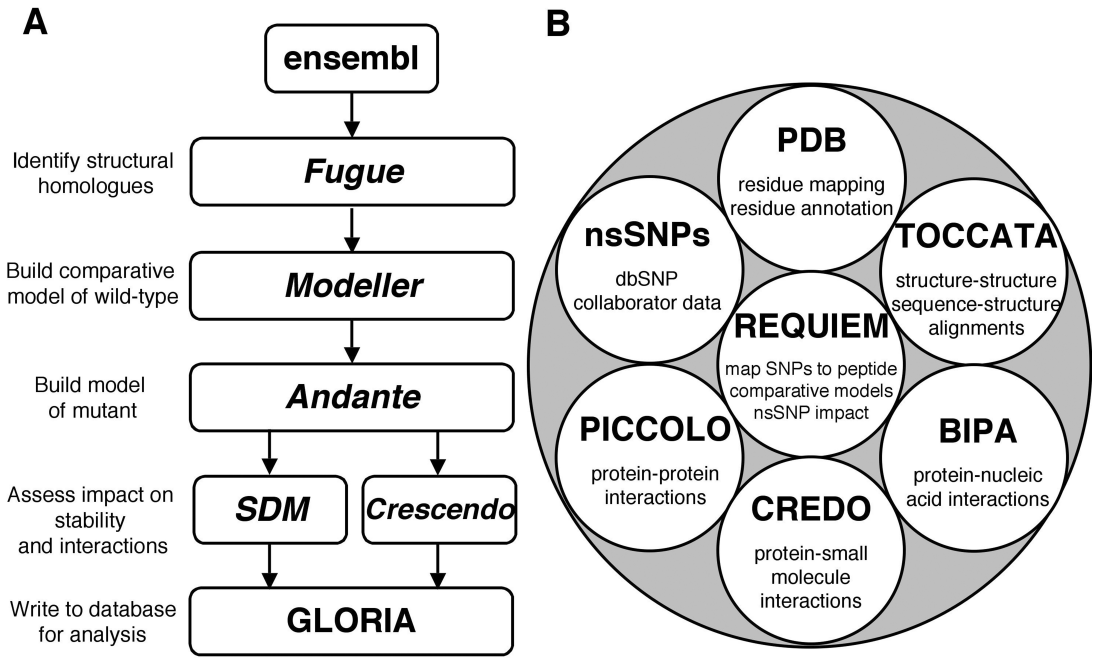


Figure 5.1: Modelling pipeline software and databases. (A) Automated tools for genome scale comparative modelling and analysis of impact of nsSNPs. (B) The platform comprises a federation of inter-connected databases integrating comprehensive structural annotations with the results of the automated modelling and nsSNP analysis.

side-chain conformations by use of environment-specific substitution probabilities and a high-quality rotamer library.

5.2.2 Software for predicting the effects of nsSNPs on protein structure

An example of a mutation that impacts protein stability is the L106R mutation in the enzyme aldehyde dehydrogenase 10 (FALDH10) that destabilizes the fold by introducing a positive charge into the hydrophobic core of the protein and is implicated in Sjogren-Larsson syndrome (Wang & Moult (2001)). Structural effects of nsSNPs can be estimated using Site Directed Mutator (*SDM*) (Topham *et al.* (1997); Worth *et al.* (2007b)), which incorporates a statistical potential energy function that predicts the effect that nsSNPs may have on the stability of proteins. *SDM*, originally developed by Chris Topham and extended by Catherine Worth, uses environment-specific amino acid substitution frequencies within homologous protein families to calculate a stability score, which is analogous to the free energy difference between a wild-type and mutant protein (Figure 5.2). The method performs comparably or better than other published methods in the task of classifying mutations as stabilizing or destabilizing (Worth *et al.* (2007b)). Additionally, *SDM* has much improved sensitivity in predicting stabilizing mutations compared to other published methods (five of the seven methods incorrectly classify >68% of the stabilizing mutations).

5.2.3 Software for predicting the effects of nsSNPs on protein interactions

Although many nsSNPs that are associated with disease are predicted to affect protein stability, there may also be a significant number that cause disease through affecting molecular function (Wang & Moult (2001)). One example of this is that of the G75D mutation that introduces a negative charge into a hydrophobic cavity in the centre of the beta barrel of retinol binding protein (RBP), thereby interfering with retinol binding both electrostatically and sterically, resulting in a night blindness phenotype. Methods that identify functional sites

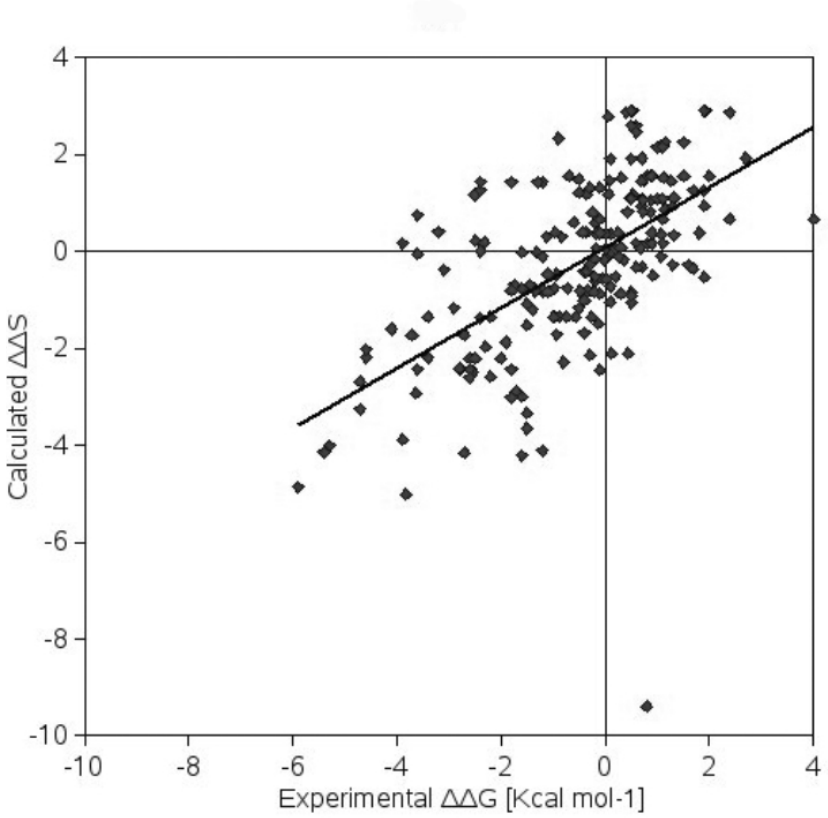


Figure 5.2: Experimentally measured energy changes versus predicted energy changes using our method, *SDM*, on a set of monomeric proteins with resolution <2Å. The correlation is 0.60 and the standard error is 1.36kcal/mol. Removal of the outlying data point increases the correlation to 0.66.

of proteins, such as enzyme active sites, receptor binding sites, specificity determining regions, co-factor binding residues and so forth, are therefore useful for estimating the effects of nsSNPs on protein function. Traditionally, sequence motif databases, such as PROSITE (Rocha *et al.* (2006)), have been used to identify specific residues likely to be involved in function. However, many functional regions are discontinuous in protein sequence. Attempts to predict functional interaction sites computationally have included identification of steric strain or other types of high-energy conformations that often occur at active sites (Heringa & Argos (1999); Rocha *et al.* (2006)) and identification of clefts that can accom-

modate ligands (Laskowski *et al.* (1996)). In *Crescendo* (Chelliah *et al.* (2004)), a technique has been developed for predicting interaction sites in proteins, which exploits an understanding of divergent evolution. A similar philosophy is exploited in evolutionary trace (Lichtarge & Sowa (2002)), in which conserved residues are phylogenetically partitioned and highlighted on the structure. In contrast, *Crescendo* exploits the fact that the conservation of amino acid residues is strongly dependent on the local structural environment in which they occur in the folded protein. By comparing expected amino acid substitutions from a library of ESSTs to those observed within alignments of homologous protein sequences, it is possible to identify those residues that are conserved more than would be expected for a given structural environment. These residues are therefore likely to be under some evolutionary restraint and have a functional role in the protein. Mapping the scores from this method onto the three-dimensional structure of the protein identifies clusters of residues potentially forming interaction sites. Where the nsSNP is found to fall within any such identified sites the impact of the mutation can be scored using an appropriate substitution table.

5.2.4 Protein databases

The results of applying all of these tools at genome scale are integrated into a suite of inter-connected relational databases (Figure 5.1), the components of which will be described here. Structural data from the Protein Data Bank (PDB) (Berman *et al.* (2000)) and sequence data from UniProt (Uniprot-Consortium (2009)) and Ensembl are integrated using the accurate residue-level mappings provided by MSD SIFTS (Velankar *et al.* (2005)) as described in Chapter 2. These sequence-structure mappings form the backbone of the inter-related databases and are supplemented by structural annotations (including secondary structure, solvent accessibility and hydrogen-bonding) generated by JoY (Mizuguchi *et al.* (1998a)), all of which are also stored in relational form.

TOCCATA was introduced in Chapter 4 as a relational database of structure-based alignments of homologous protein families, incorporating the principles of HOMSTRAD but extending the ideas further. An important feature is that the details of residue equivalences of the structural alignments are stored in relational

form. In this context the TOCCATA structural alignments are used to derive a library of sequence profiles. This library provides a query database against which the sequence-structure homology recognition program *Fugue* can be run for each Ensembl transcript, in order to identify putative templates for comparative modelling. Three further databases (PICCOLO, BIPA and CREDO) comprise an inventory of observed functional residues found in all experimentally determined protein structures in the PDB. They complement the generally-applicable predictive method *Crescendo* with a high confidence, low coverage set of characterized functional residue annotations. However, the annotations they provide can, in some circumstances, be extended through homology. Identifying the precise residues involved is straightforward given the sequence-structure alignments pre-calculated and stored in TOCCATA. Methods are being developed to identify cases where such homology-based annotation transfer is valid by applying appropriately derived substitution tables.

PICCOLO makes a useful contribution in this context as it comprises a comprehensive set of structurally characterized protein-protein interactions. The version of PICCOLO used in this work had 65,959 pairs of interacting chains, 6,371,711 pairs of interacting residues and 80,519,209 pairs of interacting atoms. When a nsSNP is found to correspond to a residue within a protein-protein interaction site it can be scored using substitution tables in a similar manner to *Crescendo*. CREDO (Schreyer & Blundell (2009)) is the equivalent database for interactions between proteins and small ligands from all experimentally determined structures in the PDB. The small molecules include metabolites, hormones, co-factors, drugs and inhibitors, both covalently and non-covalently bound, but with certain low-interest ligands, including crystallization buffers and certain modified residues, deliberately excluded. The latest version comprises contact information from 6,689 unique small molecules from 27,754 protein structures. BIPA is the equivalent database for protein-nucleic acid interactions. Protein-nucleic acid interfaces are identified from the PDB and clustered by their structural similarity in a manner analagous to that in PICCOLO. Atomic interactions have been identified from 1,193 protein-nucleic acid complexes and classified by their bonding type.

REQUIEM houses the large scale analysis of nsSNPs, including comparative models and estimates of the effects of nsSNPs on binding interactions and stability, from both the predictive methods (*SDM*, *Crescendo*) and the residue observations (PICCOLO, BIPA, CREDO) using the sequence-structure alignments pre-calculated and stored in TOCCATA. It links human genes from Ensembl to nsSNPs from the dbSNP database and disease-association data from other research groups. The information in REQUIEM can be used to derive useful summary information: Ensembl release 45 contains 43,570 human peptides, totaling 22,086,536 residues (many peptides overlap due to there being multiple transcripts per gene); of which 2,969 peptides (643,730 unique residues) are represented at least once in the PDB; of which 1,710 peptides (68,010 unique residues) can be found in PICCOLO, 1,335 peptides (30,410 unique residues) in CREDO and 186 peptides (5,599 unique residues) in BIPA.

5.2.5 Design of benchmark study

In order to assess objectively the performance of the various approaches in a quantitative manner, five locally developed tools (*SDM*, *Crescendo*, PICCOLO, CREDO and BIPA) and four published methods (SIFT, MUpro, MAPP and I-Mutant2.0) were assessed. The complementary nature of the locally developed software and databases means that although their performance can be assessed individually they are intended to be used in combination and so should be assessed as such. The combined result will therefore be evaluated as the union of each of the individual methods i.e. a positive result is recorded if at least one method predicts that mutation to be deleterious.

The benchmark set was chosen as all residues in human proteins of known 3D structure that occur in at least one of three classes of mutation data - Disease, Polymorphism and dbSNP. The Disease and Polymorphism sets, containing 3,966 and 1,366 individual mutations respectively, are provided by UniProt (http://www.expasy.org/cgi-bin/lists?humsavar.txt) as mappings of mutations to UniProt residues. The Disease set consists predominantly of Mendelian-type mutations from OMIM (Hamosh *et al.* (2005)) and as such are likely to be highly disruptive to protein structure and function. This set forms a positive control.

The Polymorphism set comprises sequence variants with no known association with disease. The dbSNP set, containing 4,982 unique mutations, is derived from residue-level mappings, provided by Ensembl-Variation database, of dbSNP mutations to Human Ensembl sequences. Identifying an appropriate negative control is notoriously difficult (Care *et al.* (2007)) - definitively classifying any mutation as not being involved in any disease is not possible as each mutation's phenotype is strongly dependent on its genotypic and environmental context. The Polymorphism and dbSNP sets were combined to provide a negative control, although in reality they are likely to be a mixture of neutral and deleterious mutations, dominated by neutral mutations. The two data sets overlap to a small degree. Mutations in the overlap are partitioned into the Disease set ahead of the combined Polymorphism-dbSNP set. Overall the benchmark set consists of 9,143 mutations (3,966 mutations with disease-association and 5,177 without), corresponding to 8,139 unique sequence positions from 1,477 Ensembl peptides.

Mappings of mutations to protein sequences and protein sequences to all of their respective solved structures are stored in the database. The same protein structure has often been solved several times under different conditions and frequently by different methods and as such the same residue can be observed several times. To run *SDM*, *Crescendo* and the published methods, a representative structure was selected for each mutation, prioritizing structures with high sequence coverage, lower resolution and X-ray over NMR structures. Importantly for PICCOLO, CREDO and BIPA no such prioritization takes place - all solutions of the protein are included thereby maximizing the opportunity that a particular residue may be annotated as being functional. *SDM* results were considered for buried residues only (percentage solvent accessibility less than 7% as calculated by JoY). Mutations in *Crescendo* and PICCOLO predicted sites were considered deleterious when their BLOSUM62 (Eddy (2004)) score was less than -1.

To run MAPP, an alignment of the sequence of interest with its orthologues is required. Such alignments were obtained using the Ensembl-Compara API (Application Programming Interface). Semphy (Friedman *et al.* (2002)) was used to reconstruct a phylogenetic tree for each alignment. SIFT version 2.1.1 was run on multiple sequence alignments created using the in-built PSI-BLAST functionality querying against the UniProt sequence database. A median conservation score of

2.75 was used. MUpro version 1.1 was run with the optional tertiary structure information included. I-Mutant2.0 was run in PDB mode, requiring the output of the secondary structure program DSSP (Kabsch & Sander (1983)), with a pH of 7.4 and temperature of 35°C. It has been shown that mutations that decrease the stability of a single domain by >2 kcal/mol result in severe disease phenotypes (Lindberg *et al.* (2005); Randles *et al.* (2006)). Therefore for each of the stability prediction methods we have used a 2 kcal/mol cutoff for classifying mutations as disease-associated.

Each method was evaluated using the following measures:

$$Sensitivity = 100 \times \frac{TP}{(TP + FN)} \tag{5.1}$$

$$Specificity = 100 \times \frac{TN}{(FP + TN)} \tag{5.2}$$

$$Accuracy = 100 \times \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{5.3}$$

where TP=True Positives, FP=False Positives, TN=True Negatives. All numbers refer to the number of unique mutations (there may be multiple mutants per sequence position).

5.3 Results

5.3.1 Benchmark study

Each of our methods has relatively low sensitivity in isolation, however when run in combination, as intended, the sensitivity is significantly increased while still maintaining good specificity (Table 5.1). The complementary nature of the methods is reflected in the relatively small overlap of our predictions (Figure 5.3).

A benefit of our combined approach is that it enables us to differentiate between structural and functional effects of nsSNPs thereby potentially aiding the identification of the causative mechanism underlying the disease. Figure 5.4 (page 171) describes the following examples of typical true positive predictions. In panel (A) *SDM* predicts that the nsSNP, P69S, in phosphomannomutase 2 (PMM2,

	TP	FP	TN	FN	Sum	Sens	Spec	Acc
SDM	535	274	4594	3025	8428	15.03	94.37	60.86
Crescendo	186	126	4482	3695	8489	4.79	97.27	54.99
PICCOLO	220	257	4920	3746	9143	5.55	95.04	56.22
CREDO	380	413	4764	3586	9143	9.58	92.02	56.26
BIPA	80	20	5157	3886	9143	2.02	99.61	57.28
COMBINED	1252	984	4193	2714	9143	31.57	80.99	59.55
SIFT	2709	2071	3011	1092	8883	71.27	59.25	64.39
MAPP	2659	1642	2395	1065	7761	71.40	59.33	65.12
I-Mutant2.0	1485	1677	2189	1061	6412	58.33	56.62	57.30
MUpro	175	146	5031	3791	9143	4.41	97.18	56.94

Table 5.1: TP=True Positives, FP=False Positives, TN=True Negatives. TP/FP/TN/FN are numbers of unique mutations. The Sum column shows the number of times the method succeeded and an observation was possible and therefore reflects the robustness of the method. Sens = Sensitivity, Spec = Specificity and Acc = Accuracy are defined in text.

PDB entry 2amy) will be damaging to protein structure. This mutation is associated with congenital disorder of glycosylation type 1A (Le Bizec *et al.* (2005); Matthijs *et al.* (2000)). In panel (B) *Crescendo* and PICCOLO identify Asp84 in Cyclin-dependent kinase inhibitor 2A (p16) as a protein-protein interaction site. The aspartate residue forms a side-chain hydrogen bond with an arginine residue in CDK6 (PDB entry 1bi7). Mutating the aspartate to tyrosine has been shown to reduce binding to CDK4 by 87% (Kubo *et al.* (1999)) and has been observed in cutaneous malignant melanoma 2 (CMM2) (Ruiz *et al.* (1999)). In panel (C) the mutation Asp201Tyr in hypoxanthine-guanine phosphoribosyltransferase (HGPRTASE, PDB entry 1bzy) is associated with Lesch-Nyhan syndrome (LNS) (Sculley *et al.* (1992)). PICCOLO identifies that the wild-type aspartate residue is within a protein-protein interface. The aspartate forms a side-chain hydrogen bond to a main-chain atom of the interacting chain, loss of which may disrupt the protein interface. In panel (D) the wild-type residue of the mutation, Met749Ile, in androgen receptor (AR) is identified by CREDO as forming a contact with the ligand, dihydrotestosterone (PDB entry 1xj7). The mutation has been detected

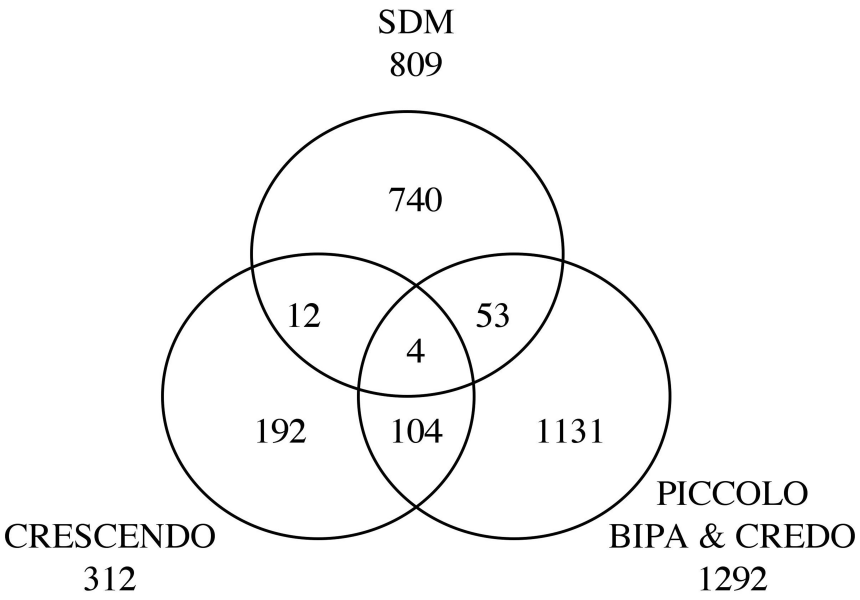


Figure 5.3: Venn diagram indicating the overlap of the results of three in-house methods for predicting the impact of nsSNPs.

in androgen-independent prostate cancer (Takahashi *et al.* (1995)). In panel (E) threonine 284 in cellular tumor antigen p53 is identified by BIPA as interacting with DNA (PDB entry 2ac0). UniProt annotates the nsSNP, Thr284Pro, as having been detected in lung tumor. This mutation may affect binding of DNA to p53.

Comparison of our combined method to each of the public domain methods indicates that our combined method has an overall accuracy superior to that of I-Mutant2.0 and MUpro and inferior to that of SIFT and MAPP (Table 5.1). The greater accuracy of SIFT and MAPP is largely due to their having a higher sensitivity score. However, comparison of the specificity scores reveals that our combined method performs significantly better than SIFT, MAPP and I-Mutant2.0 in this respect. Although SIFT and MAPP predict the majority of disease-associated nsSNPs correctly (as indicated by their high sensitivity scores), they predict 40% of the non-disease set as being disease-associated, limiting their utility in many real-world applications. The specificity of MUpro is higher than

our combined approach, however the sensitivity is so low as to limit its utility. Our combined approach has a comparable accuracy to the other methods tested but has the benefit of a much lower false-positive rate and therefore provides a high-quality set of predictions.

This benchmark was performed on a set of solved set of protein structures. When applying these methods at genome scale, solved structures will be a minority compared to the far larger number of comparative models. Given the unavoidable inaccuracy of even the best comparative models when compared to a correctly solved experimental structure, any method to assess the likely impact of nsSNPs that relies on structure will inevitably perform better on solved structures than on comparative models.

5.3.2 Benchmark study update

This study was initially published in summer 2007 and used an early version of PICCOLO. Since then a number of important improvements have been made to the database. These include the addition of interaction type definitions to give greater specificity than the original radial cutoff method and the use of PISA predicted quaternary structures as opposed to the original PDB ASU data. One further area for improvement would be in the area of predicting the impact of nsSNPs on functional sites. In the original methodology any nsSNP corresponding to these sites is assessed using a simple BLOSUM scoring matrix. This approach is somewhat naive as the actual impact is likely to depend on the likelihood of the substitution in that local structural environment, as well as its relative position in the functional site. A more accurate reflection of the likely severity of the substitution could be estimated using an appropriate set of ESSTs specifically derived for interface environments as described in Chapter 4. The benchmark study was repeated on the original data set with the addition of these new features. The performance of the updated PICCOLO is shown in Table 5.2.

Remarkably the accuracy of the updated PICCOLO results is precisely the same as that in the published benchmark. The sensitivity has increased by 2.01% (a proportional increase of 36%) whereas the specificity has dropped by 1.54% (a proportional decrease of 1.62%). Maintaining such high specificity is important as

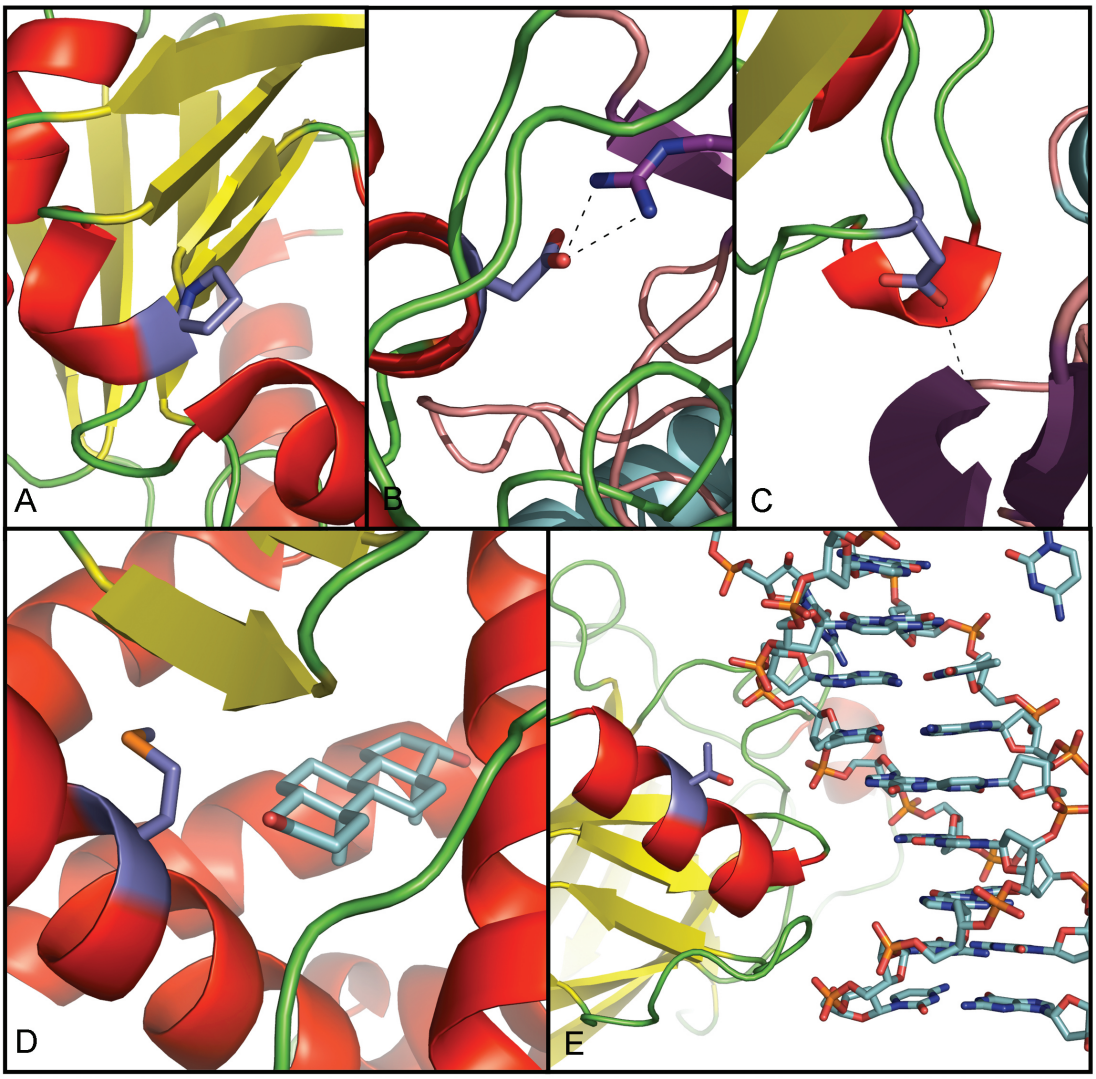


Figure 5.4: Examples of disease-associated mutations in protein structures that are predicted to be deleterious by our methods and not predicted to be deleterious by any of the public domain methods. For each case, wild-type side-chains are shown in mauve. Atoms are coloured by type. The secondary structure of the protein chain containing the nsSNP is shown in red (helix), yellow (strand) and green (coil). The secondary structure of interacting protein chains are shown in blue (helix), purple (strand) and pink (coil). Hydrogen bonds of wild-type residues are shown in black. See text for detailed description. Figure taken from Worth *et al.* (Worth *et al.* (2007a)) and produced using PyMOL (Delano (2002)).

	TP	FP	TN	FN	Sum	Sens	Spec	Acc
Published	220	257	4920	3746	9143	5.55	95.04	56.22
Latest	300	337	4840	3666	9143	7.56	93.49	56.22

Table 5.2: Results of benchmark update. Column headers are the same as in Table 5.1.

it can be argued that in most real-world applications a small set of high-confidence predictions is of more use than a larger set of low-confidence predictions. The low overall sensitivity of PICCOLO is largely due to the nature of the benchmark - only a subset of disease-associated nsSNPs will impact phenotype through modulation of protein-protein interactions. Due to the multiple redundancies in cellular signaling and metabolic networks many mutations that disrupt a particular protein-protein interaction may not have an observable phenotype, adding to the false positive rate. Furthermore, even if a disease-associated mutation corresponds to a protein interaction site it is possible that the mutation manifests its affect on phenotype through an independent mechanism (e.g. impacting the rate of transcription or translation).

5.3.3 nsSNP combinations

1,779 individual nsSNP positions from 617 different genes can be mapped to residues partaking in protein-protein interactions using the latest version of PICCOLO. However, of particular interest when assessing the likely impact of nsSNPs, is the possibility of identifying cases where pairs, or higher order combinations, of nsSNPs occur in the same interface, either within the same gene or in different genes. When such nsSNPs are structurally co-proximal it is possible they could have synergistic or compensatory effects, thereby providing a molecular basis for polygenetic disease. This principle can be used to prioritize particular pairs of associated nsSNPs for further analysis in association studies where genotyping of individuals could reveal whether any of these nSSNP combinations co-occur. Table 5.3 gives examples of homodimeric proteins where nsSNPs can be found on each side of the interface. A subset of the nsSNPs are self-interacting as they occur near an axis of cyclic symmetry.

Homodimer	Interacting nsSNPs	Self-interacting nsSNPs
Insulin (INS_HUMAN)	1-1	1
Prorelaxin H2 (REL2_HUMAN)	1-1	1
HLA class II histocompatibility antigen DR-1 β chain (HB2B_HUMAN)	5-5	1
Macrophage migration inhibitory factor (MIF_HUMAN)	2-2	0
Transthyretin (TTHY_HUMAN)	1-1	0
Ferritin light chain (FRIL_HUMAN)	2-2	1
Aquaporin-1 (AQP1_HUMAN)	1-1	1
Ubiquitin (UBIQ_HUMAN)	1-1	1
Tyrosine-protein phosphatase non-receptor type substrate 1 (SHPS1_HUMAN)	1-1	1
Epidermal growth factor receptor (EGFR_HUMAN)	1-1	1
Haemoglobin subunit β (HBB_HUMAN)	13-10	0
Haemoglobin subunit γ-1 (HBG1_HUMAN)	4-2	1
Haemoglobin subunit α (HBA_HUMAN)	14-9	1

Table 5.3: Homodimeric assemblies with the number residues interacting across the interface that correspond to nsSNPs.

Figures 5.5 and 5.6 give examples of heteromeric proteins exhibiting this phenomenon.

5.3.4 von Hippel-Lindau disease

In a separate study, Julia Forman (ex-TLB group) applied PICCOLO to help rationalize the molecular mechanism of how mutations in the von Hippel-Lindau disease gene (VHL) lead to a clinically heterogeneous, dominantly inherited familial cancer syndrome characterized by an increased risk of multifocal tumor development in multiple organs, particularly retinal angioma, central nervous system hemangioblastoma, renal cell carcinoma, and pheochromocytoma (Forman *et al.* (2009)). Missense mutations and phenotype data from Hes *et al.* (Hes *et al.* (2007)) and Ong *et al.* (Ong *et al.* (2007)), were analysed in detail to probe patterns of tumor development associated with individual missense mutations. PICCOLO was used to characterize known interaction sites, *Crescendo* to predict other functionally important residues and *SDM* to assess the effect on thermodynamic stability. Two crystal structures of pVHL in complex with elongin B, elongin C, and a HIF peptide were used (PDB entries 1lm8 and 1lqb). Known and predicted interaction sites and predictions of thermodynamic stability change upon mutation, were used to generate new hypotheses regarding the molecular aetiology of renal cell carcinoma (RCC) and pheochromocytoma. RCC was found to be caused by disruption of HIF binding or by mutations in the elongin B binding region, which act directly or through destabilizing the binding domain whereas Pheochromocytoma was shown to be triggered by mutations which disrupt interactions between pVHL and elongin C or a competing partner which binds at the same site.

5.4 Discussion and future directions

By harnessing structural and evolutionary information, a series of completely general methods for nsSNP impact assessment have been developed that are robust and of sufficiently rapid performance to be applied at genome-scale. Coupling the results of running these methods with the complete set of structurally-observed

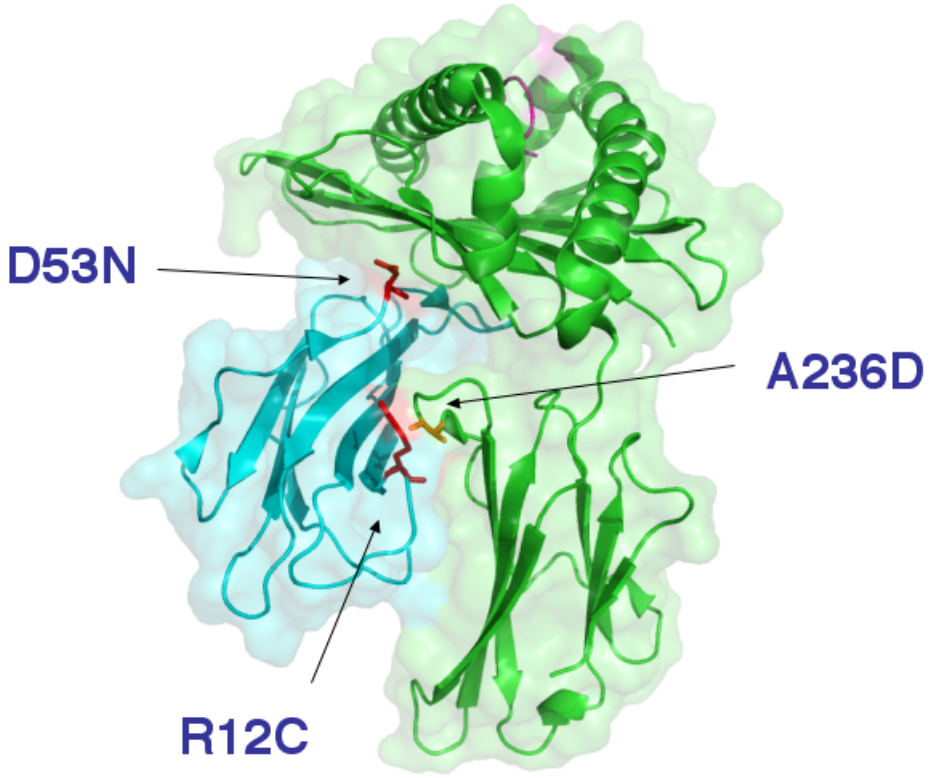


Figure 5.5: Three nsSNPs can be mapped to the interface between β2-microglobulin and the MHC CLass II molecule.

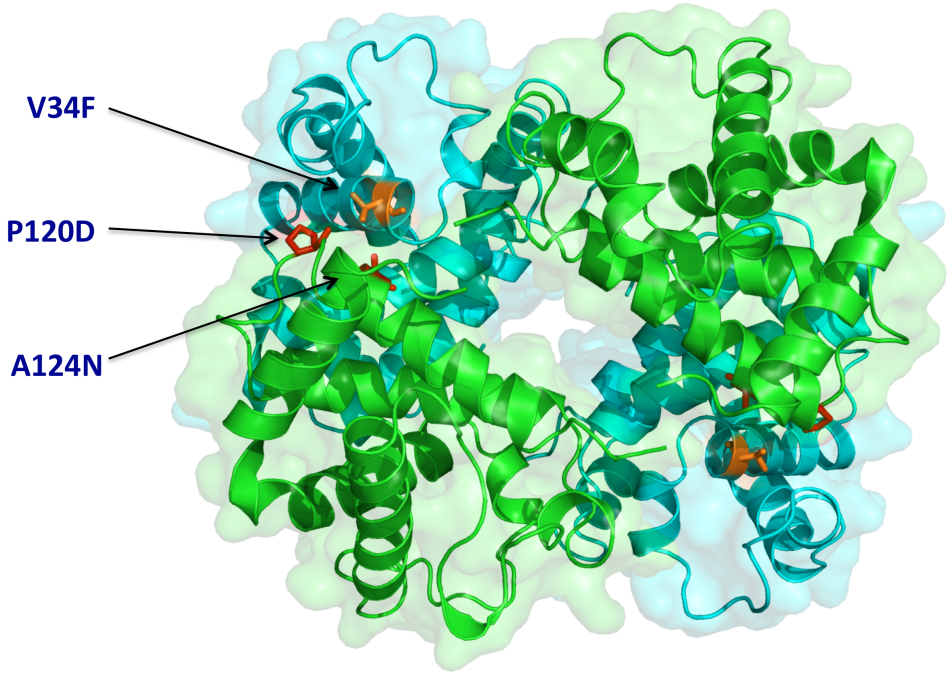


Figure 5.6: Three nsSNPs can be mapped to the interface between α and β haemoglobin molecule.

functional sites improves the likelihood of an accurate prediction being made. A key challenge for maintaining such a resource comes in ensuring it is responsive to the dynamic nature of the underlying data. As well as the novel SNPs being identified, novel transcripts continue to be added to the sequence databases and hundreds of new structures are deposited in the PDB each month, a number which will only increase as structural genomics projects continue, each of which carry valuable annotations.

One possible improvement to methods for predicting the effects of nsSNPs on mutations would be to take into account the relative position of the residue in the functional site by assessing the relative solvent accessibility of the residue in the apo- versus holo-complex, or alternatively by deriving connectivity graphs of all residues comprising the functional site and using graph properties to classify the residue as central or peripheral to the functional site.

Linking of these databases to those recording functional genomics information, such as the pathway databases REACTOME (Vastrik *et al.* (2007)) and KEGG (Ogata *et al.* (1999)), as well as experimental gene expression data, would enable components of the protein interaction network to be reconstructed, thereby aiding the identification of potentially linked nsSNPs that may confer susceptibility or resistance to polygenetic disease.

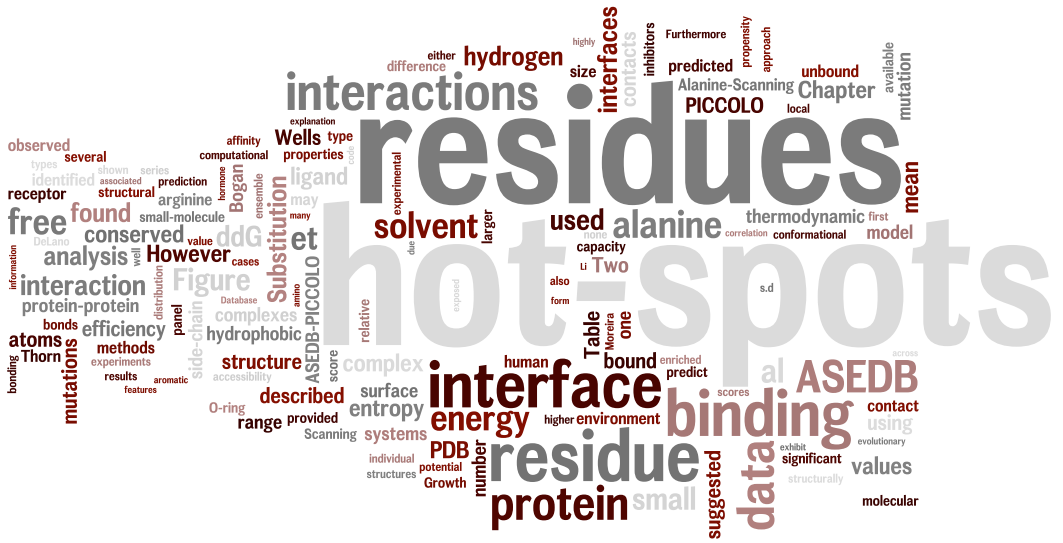
To date, the impact of previously observed nsSNPs has been examined. As all of the methodologies discussed are rapid and robust, in principle we could model and assess the impact of each of the 19 possible mutations of each residue of every human protein amenable to comparative modelling, resulting in pre-calculated data for every residue for every mutation, forgoing the requirement for an interactive web-server.

The group is collaborating with several researchers involved in human genetics that have identified nsSNPs implicated in disease. The software and databases are currently being applied as part of these collaborations to aid prioritization of their data. These collaborations include research teams studying type 1 diabetes (John Todd's group at the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory), breast cancer (Carlos Caldas' group at Cancer Research UK Cambridge Research Institute), lung cancer (Richard Houlston's group at the Institute of Cancer Research) and kidney cancer (Tim Eisen at

the Cancer Research UK Cambridge Research Institute). A further collaboration is ongoing with Mike Stratton and Andy Futreal on the Human Cancer Genome Project where our methods are being applied to somatic mutations identified in cancer cell lines.

Chapter 6

Hot-spots



6.1 Introduction

Historically protein-protein interactions have attracted little interest as therapeutic targets for small-molecule drugs. Indeed of the 248 distinct molecular targets for the 1,204 unique, marketed small-molecule drugs, none are *bona fide* protein-protein interactions (Overington *et al.* (2006)). This is despite a potentially huge

opportunity space - a recent estimate of the size of the human interactome suggested a value of ∼650,000 interactions (Stumpf *et al.* (2008)). Instead much of the focus of the pharmaceutical industry's efforts has been on the precedented target families - G-protein coupled receptors, ion channels, nuclear receptors, kinases and metabolic enzymes.

Unlike each of these target classes protein-protein interfaces lack endogenous small-molecule partner ligands that often provide a natural starting point for research and as such also typically lack the grooves and pockets that make the precedented target families so tractable. Whereas the contact surfaces of protein-small molecule interfaces tend to be in the range of ∼300-1,000Å^2, protein-protein interfaces tend to be much larger, in the range of ∼1500-3000Å^2 and more hydrophobic, properties that do not lend themselves to efficacious binding by small, drug-like chemical matter. However, mutational studies, described below, have revealed that only a subset of interface residues contribute most of the free energy of binding. This finding has led to a renewed interest in protein-protein interactions as therapeutic targets and a series of drug-discovery projects have been launched. Notable successes include the targeting of interleukin-2 (IL-2) and the α-chain of its receptor (IL-2Rα) (Thanos *et al.* (2003)); B7 and CD28 (Green *et al.* (2003)); B-cell lymphoma-2 (BCL-2) and BAK (Bcl-2-antagonist/killer) (Bruncko *et al.* (2007)); Lymphocyte function-associated antigen-1 (LFA-1) and intercellular adhesion molecule-1 (ICAM-1) (Sanfilippo *et al.* (1995)); inducible Nitric Oxide Synthase (iNOS) dimerization (McMillan *et al.* (2000)); Nerve Growth Factor (NGF) and its receptor (Niederhauser *et al.* (2000)); cytokine tumour-necrosis factor (TNF) and its receptors, TNFR1 and TNFR2 (He *et al.* (2005)); FtsZ and ZipA (Tsao *et al.* (2006)); human papilloma virus (HPV) E1 and E2 (Yoakim *et al.* (2003)); and human protein double minute 2 (HDM2) and p53 (Koblish *et al.* (2006))(Arkin & Wells (2004); Wells & Mcclendon (2007)). Furthermore, TIMBAL, a recently compiled hand-curated series of small molecule inhibitors of protein-protein interactions found in the literature, identified 105 small molecules from 40 publications targeting 21 multi-protein complexes (Higueruelo & Blundell (2009) in preparation).

The excellent recent review of Wells and McClendon (Wells & Mcclendon (2007)) suggested grounds for cautious optimism in prosecuting such targets.

Firstly, in many cases the interface exhibits sufficient adaptability for small-molecule inhibitors to bind in flexible grooves and pockets not observed in the static structure of either the free or the bound form. Screening procedures can be improved by enriching diverse chemotypes tailored for binding protein interactions or by screening with fragments of 150-250Da that may achieve higher ligand efficiency (the free energy of binding per non-hydrogen atom) (Hopkins (2004)) and more productive sampling of chemical space. Contrary to common perception, chemical matter can usually be found with binding affinities in the mid- to low-nanomolar range - comparable to that of the native protein binding partner with which the small molecule can compete. Protein interaction inhibitors tend to be in the range of 500-800Da - somewhat larger than the upper size threshold of 500Da suggested for good oral absorption and bioavailability by the widely-used "Rule-of-5" (Lipinski *et al.* (2001)) - but sufficiently close that properties may be optimized through medicinal chemistry. Furthermore the ligand efficiency of these inhibitors is comparable to that of kinase and protease inhibitors.

6.1.1 Alanine-Scanning mutagenesis

Alanine-scanning mutagenesis is the most commonly used experimental method for mapping functional epitopes on protein surfaces. Substitution of an amino acid residue with alanine removes the side-chain atoms beyond the β-carbon without adding further conformational flexibility. Substitution to glycine, which lacks a side-chain, would risk adding unwanted conformational variability. The procedure enables assessment of the energetic contribution of the side-chain of the substituted residue to protein binding through biophysical examination. Such experiments have revealed that individual residues exhibit a highly uneven distribution of energetic contributions across each interface.

The pioneering alanine-scanning experiments of Clackson and Wells on the interaction between the human growth hormone and its receptor (Clackson & Wells (1995); Wells (1996)) indicated that only a small subset of cooperatively-acting contact residues exhibit a significant drop in the binding free energy upon mutation to alanine. These residues have been termed "hot-spots". Bogan and Thorn compiled the results of several alanine-scanning experiments on protein

interfaces from the literature in the form of the Alanine Scanning Energetics Database (ASEDB) (Thorn & Bogan (2001)). From analysis of ASEDB the authors defined hot-spots as those residues that yield a change in the binding free energy of at least 2.0 kcal/mol upon mutation to alanine (Bogan & Thorn (1998)) - a threshold chosen empirically to give enough data for statistical analysis. The same value will be used in this work although other studies have used different thresholds (Kortemme & Baker (2002); Li *et al.* (2006, 2004); Ofran & Rost (2007b)).

The observation that protection from bulk solvent is a necessary, but not sufficient, criterion for a residue to have a significant effect on binding affinity, led Bogan and Thorn to the insightful postulation of the "O-ring" water-exclusion model, whereby hydrophobic residues surround the hot-spot. The hydrophobic residues themselves only provide weak contributions to the free energy of binding, but their main role is in occluding solvent, thereby increasing the strength of polar interactions between complementary hot-spot residues across the interface (Bogan & Thorn (1998)). The tight packing of these interaction hot-spots facilitates the exclusion of water molecules upon binding (Keskin *et al.* (2005a)). Keskin *et al.* (Keskin *et al.* (2005a,b)) analyzed the organization of 568 computationally predicted hot-spot residues from 44 interface clusters and found that 79% of the hot-spot residues were found to cluster into densely-packed "hot regions". The "coupling" hypothesis (Halperin *et al.* (2004)) suggests that experimentally predicted hot-spot residues on either side of the interface preferentially interact with one another. These observations have been used to refine the O-ring model with a "double water exclusion" model whereby the coupled hot-spots closely interact to give solvent-free hot-spots (Li & Liu (2009)). Analysis of ASEDB has revealed that hot-spots exhibit a non-random residue composition. Moreira *et al.* have suggested residue composition values of 21% for tryptophan, 13.3% for arginine, and 12.3% for tyrosine (Moreira *et al.* (2007c)). ASEDB has enabled several systematic studies into the nature and organization of hot-spots as well as their computational prediction.

6.1.2 Computational prediction of hot-spots

Experimental hot-spot identification requires a significant experimental effort making robust *in silico* methods of hot-spot prediction highly desirable. Kortemme *et al.* (Kortemme (2004)) describe an approach called computational alanine scanning which involves a simple free energy potential that includes a Lennard-Jones term, an implicit solvation model, an orientation-dependent hydrogen-bonding potential, probabilities of the backbone-dependent amino acid-type and rotamer, as well as an estimate of unfolded reference state energies. In their benchmark of 233 mutations from 19 protein complexes, 79% of hot-spots and 68% of neutral residues were correctly predicted. The approach of Moreira *et al.* (Moreira *et al.* (2007a)) involves a molecular dynamics simulation protocol performed in a continuum medium using the generalized Born solvent model with three different internal dielectric constants. Darnell *et al.* (Darnell *et al.* (2007, 2008)) used a machine learning approach called KFC which takes shape specificity features and biochemical contact features. Tong *et al.* (Tong *et al.* (2004)) predicted hot-spots by using side-chain modelling, energy minimization and binding free energy calculation. Landon *et al.* (Landon *et al.* (2007)) applied a computational solvent mapping algorithm (CS-Map) that involves moving small organic functional groups around the protein surface and determining the most energetically favorable binding positions. Grosdidier and Recio (Grosdidier & Fernández-Recio (2008)) predicted hot-spots by applying docking-derived normalized interface propensity (NIP) scores along with electrostatics and desolvation terms with which they obtain a positive predictive value of up to 80%. Li *et al.* (Li *et al.* (2006)) used solvent accessibility and residue contacts to identify hot-spot residues. ISIS, the method of Ofran and Rost (Ofran & Rost (2007a,b)) predicts hot-spots in protein sequences using a neural network trained on all interface residues found in structurally-characterized complexes using features including a conservation score, sequence environment and predicted solvent accessibility and secondary structure of each residue and its immediate neighbours. By representing proteins as small-world networks Del Sol and O'meara (del Sol & O'Meara (2005)) predict that residues that are highly central, conserved and buried in the protein complex, correspond to hot-spots or are in direct contact

with them. The HotSprint database (Guney *et al.* (2007)) systematically predicts hot-spots in protein interfaces from the PDB using an evolutionary conservation score and solvent accessibility terms.

6.1.3 Issues with alanine scanning

Site-directed mutagenesis is one of the most widely used approaches for probing the molecular determinants of macromolecular binding. However, it is not without its issues. In particular, the thermodynamic data are often interpreted with the inherent assumption that the only perturbations are with respect to specific interactions across the binding interface. By considering both the bound and unbound forms as conformational ensembles, DeLano (DeLano (2002)) enumerated a variety of molecular scenarios whereby the observed difference in free energy of binding could occur. Aside from those mechanisms that are dependent on the details of the intermolecular interaction in the bound ensemble, a mutation may perturb the unbound ensemble by inducing: local conformational rearrangements; local unfolding at the interface; global unfolding of the structural domain; increased entropy of unbound ensemble or aberrant aggregation of multiple protein molecules. Each of these effects could generate the observed thermodynamic phenomena, risking the possibility of false-positive assessment of the contribution of individual residues to the free energy of binding. In reality, mutations are unlikely to impact either the bound or unbound ensemble discretely and are more likely to have experimentally-indistinguishable effects on both ensembles.

As well as the risk of false positive predictions of a particular residue's contribution to binding, alanine scanning can give false negative results. Replacement of a residue's side-chain with alanine's methyl group may be compensated by local re-arrangements of neighbouring side-chains or solvent (Janin (1999)) giving misleadingly small values of $\Delta\Delta G$ for the wild type residue. Indeed, it has been suggested that the influential O-ring model may be a trivial result of the fact that side-chain atoms on the periphery of the interface surface have a greater capacity for non-disruptive replacement by solvent than atoms found towards the centre (DeLano (2002)). However, molecular mechanics simulation of the complex of lyzozyme (HEL) and antibody (FVD1.3) suggested that the hot-spot residues are

indeed kept sheltered from bulk solvent, supporting the O-ring model (Moreira *et al.* (2007b)). Multiple simultaneous alanine mutations can be used to detect co-operativity between interacting residues (Horovitz (1996)). Non-additivity between the free energy change from simultaneous mutations to that of the individual mutations is indicative of energetic coupling between the residues and in this way the energetics of co-operativity can be quantified. Such methods are known as Double Mutant Cycles (DMC) or alanine shaving and were discussed briefly in Chapter 2.

The example in Figure 6.1 indicates the difference between the PICCOLO definitions of interface core and periphery, and the phenomenon of a hot region with an O-ring. Solvent inaccessibility is taken as being a necessary but insufficient criterion for a hot-spot and although this is true of many hot-spot residues, examples of partially accessible residues can also be found.

The structural basis of the hot-spot phenomenon can be explored by combining the thermodynamic data from ASEDB with the detailed residue and atomic level information stored in PICCOLO. Dissection of the molecular properties of hot-spots, by harnessing some of the same analytical methods presented in Chapter 4 to describe the anatomy of interfaces, enables critical features that distinguish hot-spots from other residues to be identified, which, in the future, could provide the parametric basis for machine-learning methods for *in silico* hot-spot identification.

6.2 Methods

6.2.1 ASEDB cleanup

The ASEDB data is made available in the form of a MySQL dump comprising a simple schema of three tables (*reference*, *mutation* and *system*). The database comprises 3,010 mutations from 101 systems extracted from 74 references, although it has not been updated for some time (the most recent entry is from 2001). A system is defined here as an interface that has at least one residue that has been mutagenized to alanine (if both sides of an interface are subjected to alanine scanning this would be considered as two systems).

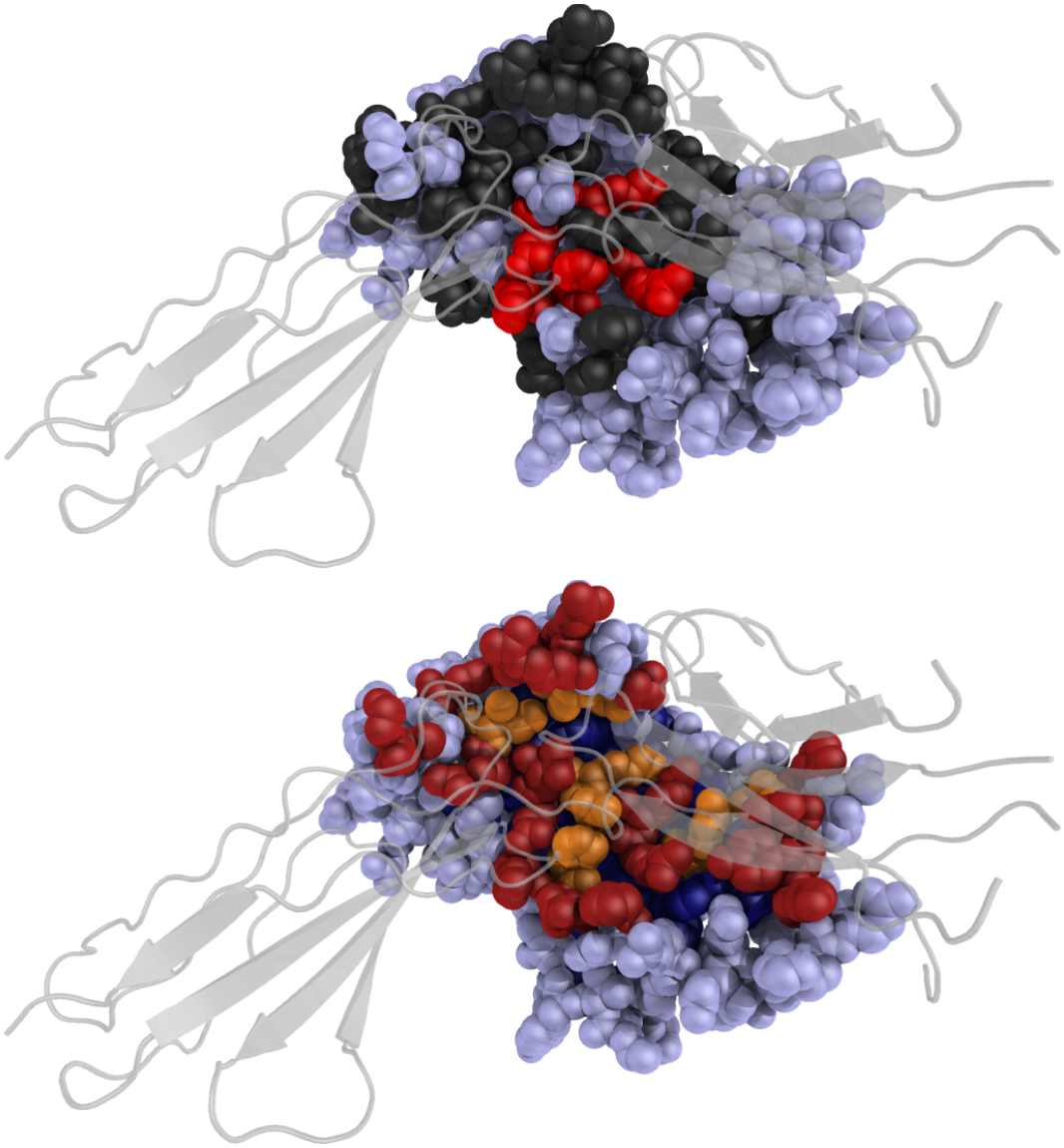


Figure 6.1: Two views of the complex of human growth hormone (spacefill) and one half of its dimeric receptor (transparent grey cartoons) (PDB entry 1bp3). In the first panel the residues of the ligand are coloured by their PICCOLO interfacial environment (Interface core in orange, interface periphery in dark red, exposed surface in light blue, buried in dark blue (see Figure 2.9 on page 63 for definitions)). In the second panel, taken from an identical viewpoint, residues are coloured by their ASEDB status (hot-spot residue with $\Delta\Delta G \geq 2$ kcal/mol are shown in red, other ASEDB residues with $\Delta\Delta G < 2$kcal/mol in black). Light blue residues are not considered by ASEDB. These figures were generated automatically by writing Python functions from Pymol to extract residue annotations from the MySQL database.

The data had several inherent issues and several steps of manual cleanup were required before they could be further analysed. Of the 101 available mutated interfaces, 91 concerned protein-protein interactions, and of these only 26 had been associated with a published structure of the co-complex from the PDB. However, since the last release of ASEDB, structures have been solved of some of the systems that had not been associated with PDB structures. Identifying whether a complex of two proteins has been solved experimentally is a non-trivial task as keyword searches are inherently unreliable. As an aside, PICCOLO proved particularly beneficial in this exercise - simply by performing a BLAST search with each of the two components of the complex against PICCOLO protein chains, and identifying the overlap of the two sets it was possible to identify all possible complexes of close homologues of the two proteins. This approach identified a further 11 systems that had both structural and alanine-scanning data. Of those that had been associated with a PDB structure, in several cases the chain identifiers were either incorrect or missing and had to be manually corrected. Furthermore the residue numbers and amino acid types provided often did not correspond to those observed in the associated structure. However, through careful curation, it was possible to use the relative spacing between residue types from incorrectly numbered residues from the same interface to match to the correct numberings observed in the PDB structure. A subset of the thermodynamic data was found to be duplicated and one instance was removed.

A further issue was that there was some redundancy inherent in the raw data. In particular 32 of the systems, corresponding to 1,799 mutations (or 59.7% of the data), correspond to 224 residues from human growth hormone bound to a series of monoclonal antibodies. This introduced considerable bias to ASEDB as individual residues were represented up to 20 times, a feature which may have skewed the published analysis (Thorn & Bogan (2001)). Most of the redundancy was confined to the human growth hormone system but in all cases, where multiple values of $\Delta\Delta G$ were provided, the value closest to zero was selected. After this cleanup process the data found in ASEDB were augmented with three further systems identified from the literature from published alanine-scanning mutation data series (the thrombin-thrombomodulin complex (PDB entry 1dx5)) (Pineda *et al.* (2002)) and both interfaces surfaces from the complex

between voltage-gated calcium channel $\beta_2\alpha$ subunit and the $\alpha_1 c$ subunit (PDB entry 1t0j)(Van Petegem *et al.* (2008)).

Hot-spot residue propensity and sequence entropy were calculated using the same methods as described previously in Chapter 4.

6.3 Results

The results of mapping the ASEDB data to the structurally characterized complexes in PICCOLO will be referred to as ASEDB-PICCOLO and in total comprises 764 mutated residues from 41 systems in 26 crystal structures. Of the 764 residues, 90 (11.8%) had a $\Delta\Delta$G $\geq$2.0Kcal/mol and were considered as hot-spots. Figure 6.2 shows the distribution of $\Delta\Delta$G values for the mutations in the ASEDB-PICCOLO set. Note that 280 of the residues (36.6%) map to regions of the structure external to a PICCOLO identified interface - largely due to systematic alanine scanning experiments, which left 484 residues (63.4%) that were in PICCOLO interfaces. Importantly, however, none of the residues outside a PICCOLO-identified interface had a $\Delta\Delta$G $\geq$2.0Kcal/mol. Note also that not all of these complexes were predicted by PISA as being stable in solution. Nonetheless, given biophysical binding data it is reasonable to assume that in these cases the complex observed in the ASU is physiological and the ASU complex was used in these cases.

6.3.1 Hot-spots are densely connected

Table 6.1 describes the counts of the number of interactions of various major interaction types for both hot-spot and non hot-spot residues from ASEDB-PICCOLO. For each interaction type hot-spots consistently have a higher mean number of interactions. On average a hot-spot residue mediates twice as any atomic contacts as a non-hot-spot residue.

6.3.2 Hot-spots are conserved

Two measures of sequence entropy were described in Chapter 4 (entropy and relative entropy). By either measure hot-spots are more conserved than none

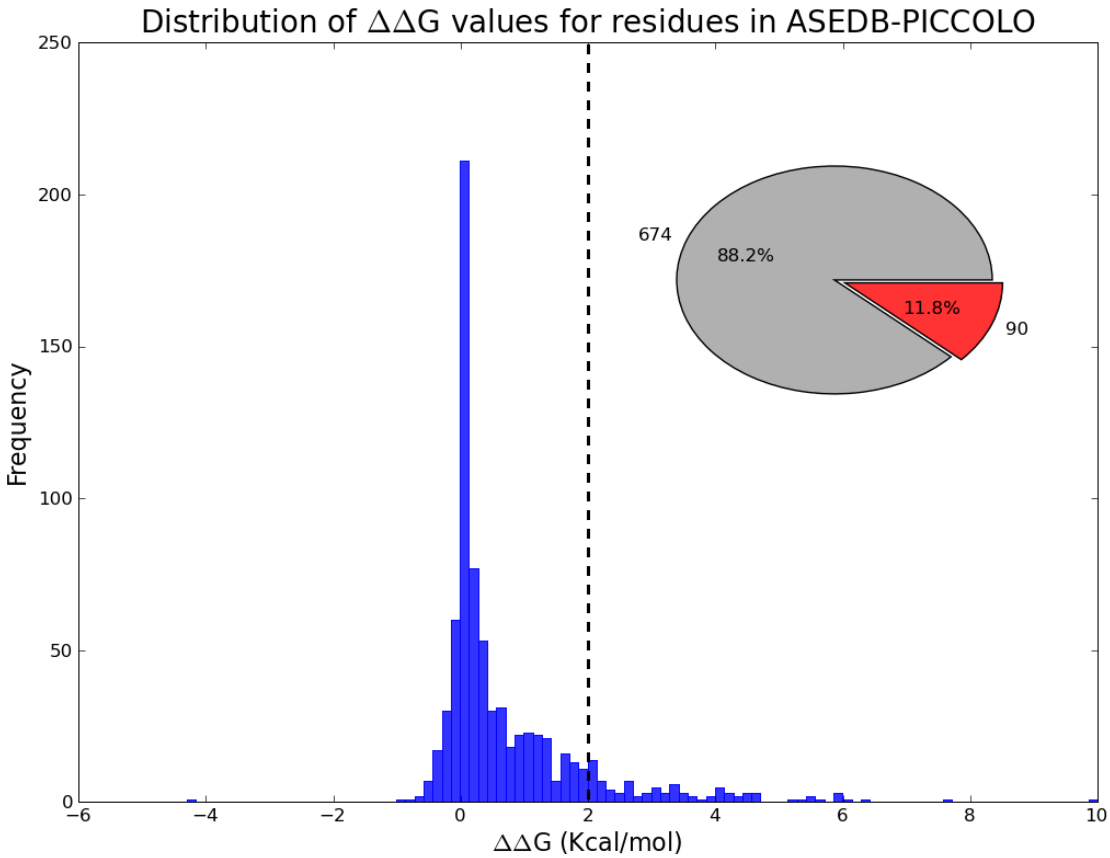


Figure 6.2: Distribution of ΔΔG values for the 764 mutations in the ASEDB-PICCOLO set.

	contacts	**van der Waals contacts**	**hydrogen bonds**	**water-mediated hydrogen** bonds	**ionic contacts**	**hydrophobic contacts**	**aromatic contacts**	**proximal contacts**
Hot-spots	19.64	7.63	0.75	0.52	2.34	6.23	2.82	82.10
	(17.11)	(7.55)	(1.04)	(1.23)	(5.87)	(7.76)	(7.97)	(68.23)
Non hot-spots	9.26	3.91	0.34	0.21	1.34	2.70	1.38	45.26
	(12.36)	(4.57)	(0.66)	(0.61)	(3.84)	(5.48)	(5.9)	(42.41)

Table 6.1: Mean number of interactions per residue of the major interaction types described in Chapter 2. Standard deviations are provided in brackets.

	Number of residues	Average entropy (s.d.)	Average relative entropy (s.d.)
Hot-spot residues	90	1.866 (1.187)	1.783 (1.442)
Non hot-spot residues	670	2.231 (0.974)	1.262 (1.049)

Table 6.2: Mean entropy and relative entropy for hot-spot and non hot-spot residues in ASEDB. The two measures have opposite directionality. Standard deviations are provided in brackets.

hot-spot residues, as shown in Table 6.2.

6.3.3 Hot spots show distinct propensities

Table 6.3 and Figure 6.3 show the relative enrichment of each amino acid type in hot-spots and ASEDB-PICCOLO. The residue propensity data for each structural environment from Figure 3.14 (page 104) in Chapter 3 is replicated in Figure 6.3 for context. This analysis was hampered by the limited availability of suitable data - only 90 data points were available where the mutation was characterized both thermodynamically and structurally. Such small samples are unlikely to be statistically significant (particularly when further partitioned by residue type), making assessment of the generality of pursuant observations difficult.

For the majority of residues, the overall distribution of ASEDB-PICCOLO residues corresponds to that of exposed surface residues. Glycine and proline are under-represented (glycine is seldom mutated in alanine-scanning experiments) whereas tryptophan, tyrosine and arginine are over-represented. With respect to the composition of hot-spots, rather than comparing hot-spot residues to ASEDB-PICCOLO residues a more meaningful analysis can be achieved through comparison of the distribution of residues in the interface core with residues in hot-spots. This analysis reveals that hydrophobic and small residues are significantly depleted in hot-spots, whereas the larger polar and charged residues are over-represented. Tyrosine, tryptophan, histdine, asparagine, glutamate, lysine and arginine are all enriched. Tryptophan's large size, aromatic nature and extensive hydrophobic surface mean that it can partake in aromatic π-interactions,

Residue	Entries in ASEDB	% of entries ASEDB	Hot-spots in ASEDB	% of hot-spots in ASEDB	Hot-spot enrich-ment
I	30	3.93	7	7.78	1.98
V	32	4.19	2	2.22	0.53
L	38	4.97	4	4.44	0.89
F	28	3.66	1	1.11	0.30
C	4	0.52	0	0	-
M	7	0.92	1	1.11	1.21
G	9	1.18	2	2.22	1.88
T	54	7.07	2	2.22	0.31
S	58	7.59	1	1.11	0.15
W	23	3.01	5	5.56	1.85
Y	50	6.54	18	20	3.06
P	14	1.83	0	0	-
H	24	3.14	4	4.44	1.41
Q	51	6.68	2	2.22	0.33
N	47	6.15	3	3.33	0.54
D	55	7.2	10	11.11	1.54
E	1	10.6	7	7.78	0.73
K	74	9.69	9	10.00	1.03
R	85	11.13	12	13.33	1.20

Table 6.3: Enrichment of each residue type in hot-spots.

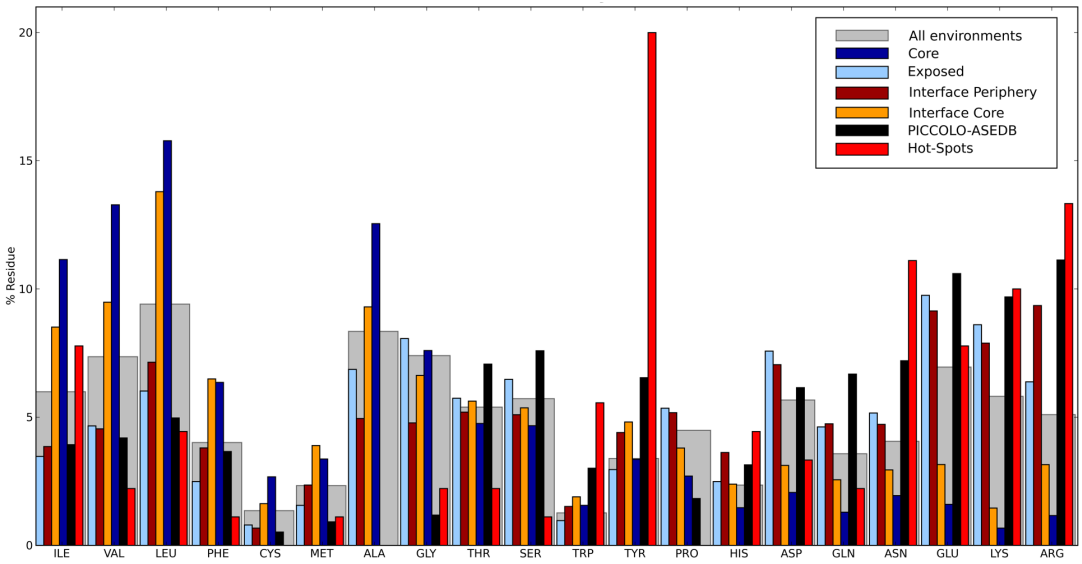


Figure 6.3: Enrichment of each residue type in hot-spots. Hot-spot data is shown in Table 6.3.

donate hydrogen bonds and protect its hydrogen bonds from water. Similarly, tyrosine's hydrophobic surface, capacity for both aromatic π-interactions and hydrogen bonding ability through its 4-hydroxyl group means it too can simultaneously mediate a diverse range of interactions. Arginine is also capable of engaging in multiple interactions with capacity for up to five hydrogen bonds as well as a salt bridge via its guanidinium motif. Comparison with the earlier analysis of Bogan and Thorn (Bogan & Thorn (1998)) suggests that arginine is not as significantly enriched in hot-spots as previously suggested (arginine was found to be two-fold enriched when compared to ASEDB). One explanation for this might be that the earlier study failed to adequately deal with the redundancies in the available experimental data. Overall, these propensities appear to support the O-ring model inasmuch as, in a solvent occluded environment, residues that are both hydrophobic and able to engage in hydrogen bonding ought to be favoured in hot-spots.

6.3.4 Hot-spots explored through substitution scores

The propensity analysis suggests that large residues are favoured at hot-spots. However, size alone cannot be the sole, phenylalanine is highly disfavoured, possibly due to its inability to mediate hydrogen bonds through its side-chain. Despite this, an appealing explanation for the hot-spot phenomenon resulting from the observed enrichment of the larger residues, is that upon mutation of these larger residues to alanine a large cavity is generated due to the significant difference in size. This would significantly destabilize the unbound conformational ensemble of the mutated protein leading to local, or global, structural re-arrangement reducing the capacity for binding in a manner analogous to that suggested by DeLano (DeLano (2002)). One way to explore such a possibility in a quantitative manner would be to harness the information provided in interface-specific substitution tables, whose generation was described in Chapter 5. Substitution scores can be obtained from the 48 ESSTs from the interface-dependent series by first identifying the appropriate substitution table, by matching terms for the secondary structure, interface solvent accessibility, intra-molecular hydrogen bonding and inter-molecular hydrogen bonding, and then looking up the log-odds score of substituting the interacting hot-spot residue with alanine. Table 6.4 and Figure 6.4 show the results of comparing the mean $\Delta\Delta$G of all hot-spots in ASEDB-PICCOLO with the mean substitution score.

The thermodynamic and evolutionary descriptors show a reasonably strong inverse correlation (correlation coefficient = -0.67, $R^2 = 0.45$). For comparison, the substitution scores from the equivalent *non-interacting* environments were examined. Here the appropriate environment is identified by matching the appropriate secondary structure and intra-molecular terms and fixing the solvent accessibility environment to be exposed (i.e. non-interface), and implicitly no inter-molecular hydrogen bonding (e.g. exchange HiWB with HAWb). With these non-interface environment definitions no significant correlation is found (correlation coefficient = -0.098, $R^2 = 0.0095$, data not shown). This result suggests that the impact of substitution of a hot-spot residue with alanine destabilizes the interface to a degree above and beyond that of the effect of any local structural rearrangement on the exposed surface.

Residue	Number of residues	Mean hot-spot ΔΔG (kcal/mol) (s.d.)	Mean substitution score (s.d.)
ARG	12	3.48 (1.16)	-2.33 (0.47)
LYS	9	4.34 (2.57)	-2.11 (0.31)
GLU	7	3.34 (0.72)	-1.57 (0.50)
GLN	2	2.70 (0.20)	-1.00 (0.00)
ASP	10	4.13 (1.54)	-3.10 (0.30)
ASN	3	2.57 (0.40)	-2.00 (0.00)
HIS	4	3.68 (1.35)	-1.75 (0.43)
TYR	18	3.72 (1.01)	-1.00 (0.00)
TRP	5	4.26 (1.19)	-2.00 (0.00)
SER	1	2.19 (0.00)	2.00 (0.00)
THR	2	2.15 (0.15)	0.00 (0.00)
GLY	2	2.12 (0.00)	2.00 (0.00)
MET	1	3.16 (0.00)	0.00 (0.00)
PHE	1	2.60 (0.00)	-1.00 (0.00)
LEU	4	2.80 (0.68)	-1.00 (0.00)
VAL	2	2.34 (0.24)	-1.00 (0.00)
ILE	7	2.55 (0.69)	-1.86 (0.35)

Table 6.4: Mean ΔΔG of all hot-spots in ASEDB-PICCOLO and mean substitution score taken from interface-specific substitution tables for each residue type.

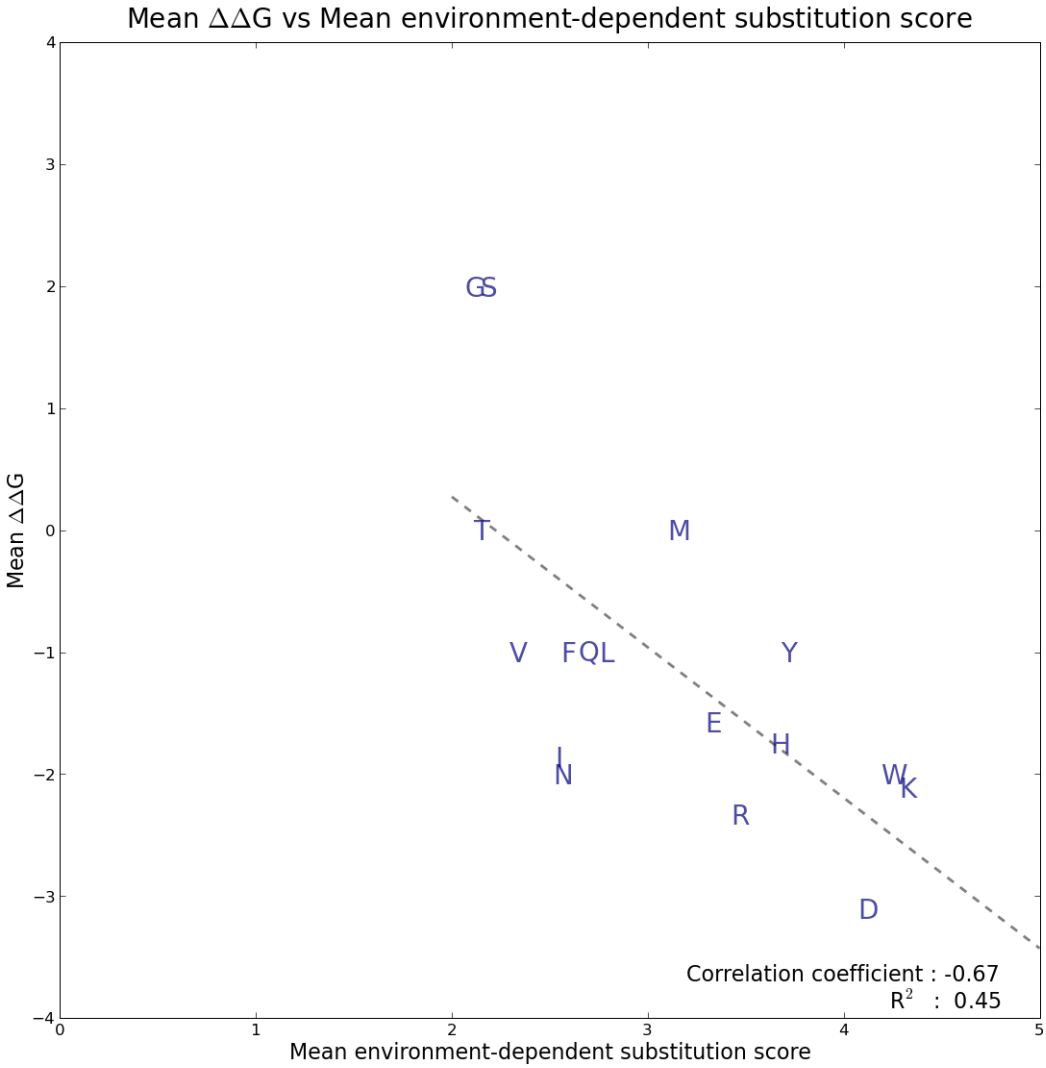


Figure 6.4: Scatter plot of mean $\Delta\Delta$G of all hot-spots in ASEDB-PICCOLO against mean substitution score taken from interface-specific substitution tables for each residue type.

6.3.5 Hot-spot Ligand Efficiency

The concept of ligand efficiency (Hopkins (2004)) has gained considerable traction in the arena of drug discovery in recent years. In principle it is the ratio of potency to size and is defined as

$$LE = \Delta G / N_{non-hydrogenatoms} \tag{6.1}$$

where ΔG is the free energy of binding and $N_{non-hydrogenatoms}$ is the number of "heavy" (i.e. non-hydrogen) atoms. It provides a valuable indicator in medicinal chemistry decision making as smaller ligands are believed to have preferable bioavailability properties to their larger counterparts with similar binding affinity. Wells and Mcclendon (Wells & Mcclendon (2007)) investigated the ligand efficiency of whole protein complexes by using published values for the binding affinity for several systems with known small-molecule inhibitors and counting the number of interface contact atoms as the denominator. Here a comparable analysis was performed at the higher resolution of individual residues using the thermodynamic and residue contact data from ASEDB-PICCOLO. The mean ligand-efficiency for hot-spots is shown for each residue type in Table 6.5.

Small molecule ligands naturally exhibit a wide range of ligand efficiencies, but values in the range -0.2 to -0.5 kcal/mol per non-hydrogen atom would be typical. The first observation from Table 6.5 would be that the ligand efficiency values for hot-spots are somewhat higher than this range. Factors contributing to this include the small sample size and the use of contact atoms only in the denominator. However, many of these residues are deeply buried in the interface where most of the side-chain atoms engage in some form of interaction. Furthermore, as hot-spots by definition contribute the greater part of the free energy of binding it would be expected that these values would be somewhat higher than those reported by Wells *et al.* for the whole interface. Interestingly, the maximal values reported here approach the figure of -1.5kcal/mol, a figure suggested by Kuntz *et al.* (Kuntz *et al.* (1999)) as the maximal possible affinity per non-hydrogen atom, suggesting that in some instances, residues are exquisitely evolved to bind their partners to their maximal potential. What remains most intriguing however, is that those residues that are most enriched in hot-spots (tyrosine, tryptophan,

Residue	Mean Ligand Efficiency (s.d.)
ARG	0.82 (0.58)
LYS	0.79 (0.52)
ASP	1.09 (0.48)
GLU	1.44 (1.09)
ASN	1.03 (0.37)
GLN	0.60 (0.24)
HIS	0.59 (0.17)
TYR	0.53 (0.43)
TRP	0.45 (0.20)
SER	1.10 (0.00)
THR	1.48 (0.82)
GLY	0.62 (0.09)
MET	0.45 (0.00)
PHE	0.87 (0.00)
LEU	0.74 (0.29)
VAL	0.67 (0.03)
ILE	1.09 (0.66)

Table 6.5: Mean ligand efficiency for hot-spots of each residue type.

histidine, asparagine, glutamate, lysine and arginine) are almost all (with the exception of glutamate) amongst the least ligand-efficient residues. One explanation might be that efficiency itself is not a key determinant for hot-spot residues, and that absolute affinity is more crucial, along with physico-chemical properties and the capacity engage in polar interactions in a solvent-excluded surface patch. Alternatively, it could be argued that the relative inefficiency of enriched residues is due to the ligand efficiency metric over-simplifying the complex surface-area to volume effects that dictate many residue properties. Whatever the explanation a clear theme is that the quantity of accessible experimental data is too small to generalize confidently.

6.4 Future directions

The capacity to identify hot-spot residues routinely and robustly *in silico* would be of tremendous value to burgeoning efforts at targeting protein-protein interfaces with small-molecule drugs. The analyses presented here provide the foundation for further development of such a method. Machine learning methods (including Support Vector Machines, Neural Networks and Random Forest Classifiers) have become standard tools in bioinformatics for classification problems involving supervised learning, where typically a small "true-positive" sample is available (which in this case would be provided by ASEDB-PICCOLO). The descriptors discussed here (sequence entropy, solvent exposure, residue propensity, substitution score as well as the number and type of interactions) would appear to have potential in discriminating between hot-spot and non hot-spot residues.

A second avenue for *in silico* hot-spot identification may stem from direct integration of PICCOLO interaction data with alignment information from TOCCATA. It has been proposed that interface modularity is an evolutionary conserved property (Rahat *et al.* (2008)). However, preservation of pairwise residue contacts across interfaces through evolutionarily conserved protein interactions has not yet been used to predict hot-spots. An example of the phenomenon of structurally conserved interactions is shown in Figure 6.5 where Fibroblast Growth Factors (FGFs) are shown bound to FGF Receptor 2 (FGFR2).

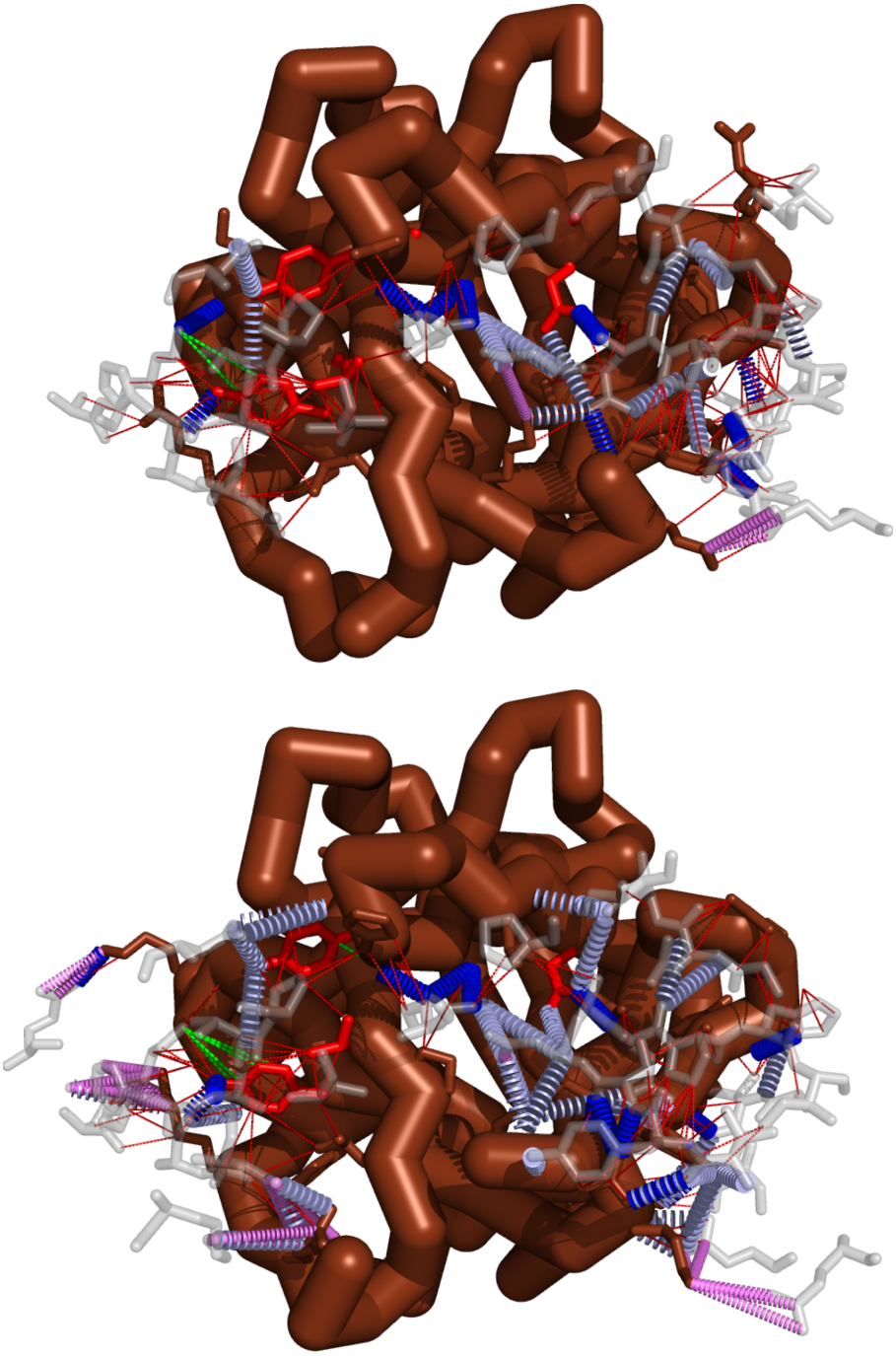


Figure 6.5: Structurally conserved interactions. The first panel depicts the interaction of Fibroblast Growth Factor (FGF) 2 bound to FGF Receptor 2 (FGFR2) (PDB entry 1ev2). The second panel depicts FGF1 bound to the same receptor FGFR2 (PDB entry 1djs). The receptor residues are shown as transparent sticks, the interaction types in the same format as described in Chapter 2. Experimentally identified hot-spot residues from FGF2 are shown in the first panel in red, as are their structurally equivalent conserved partners in FGF1 in the second panel.

Careful inspection reveals that some pairwise interactions are preserved between the two complexes. Furthermore, the experimentally-identified hot-spots of the residues mediating those structurally conserved pairwise interactions are a subset of the residues where the interactions are conserved. A broader study would be required to assess the sensitivity and specificity of using conserved interactions in hot-spot prediction. Such observations would be relatively straightforward to encode computationally given the available structured interaction and evolutionary information.

Another potential avenue for computational prediction of hot-spots would be to extend the analysis described above to use interface-specific ESSTs to predict computationally the difference in the free energy of binding upon mutation of an interface residue to alanine, in a manner directly analogous to that used by SDM to predict protein stability changes (Topham *et al.* (1997); Worth *et al.* (2007b)). By exploiting the thermodynamic cycle we can predict the difference in free energy of binding ($\Delta\Delta$G) as:

$$\Delta\Delta G = \Delta G_j^{U-B} - \Delta G_k^{U-B} = \Delta G_{jk}^{U} - \Delta G_{jk}^{B} \tag{6.2}$$

Figure 6.6 depicts the thermodynamic cycle that could by used to predict the difference in the stability scores for the bound and unbound state for the wild-type and mutant protein structures (i.e. alanine-substituted).

Note that this would be an entirely general method that could also be applied to the problem of predicting the effect of nsSNPs in protein-interfaces described in Chapter 5.

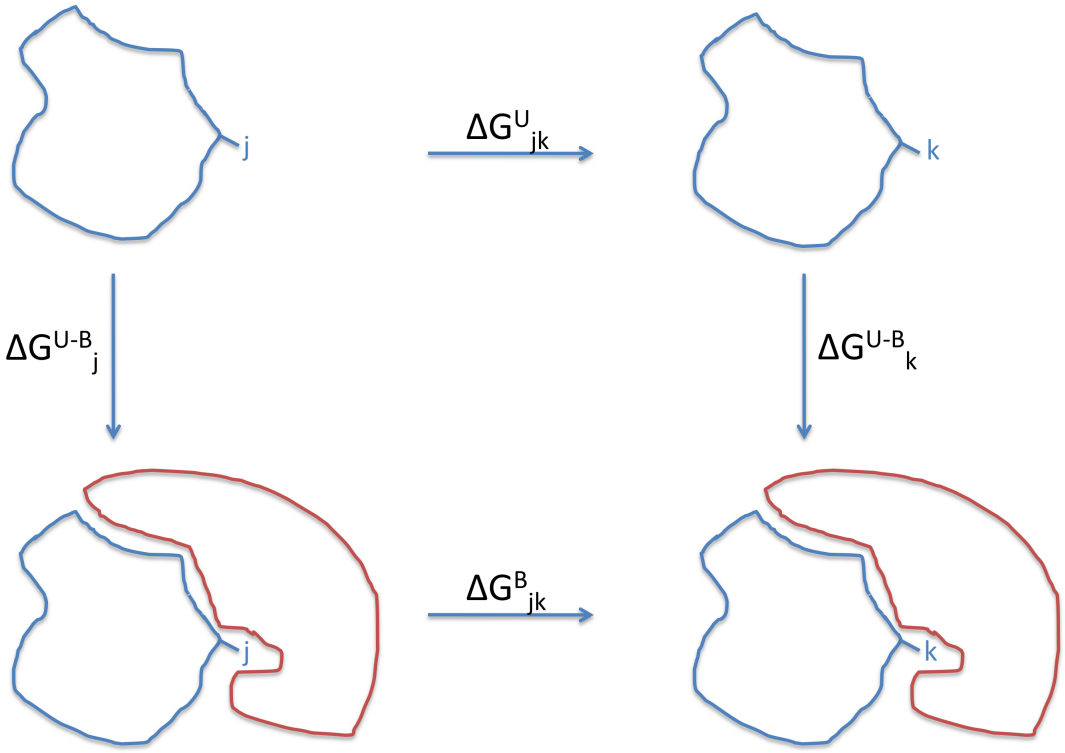


Figure 6.6: The thermodynamic cycle for for the bound and unbound state for the wild-type and mutant protein structures.

Chapter 7

Conclusion

7.1 Overview

The salient outcome of this work is the establishment of PICCOLO as a comprehensive database of structurally characterized protein interactions. To achieve this, issues of interface definition, interaction specificity, quaternary structure,

interface redundancy, interaction type and structural environment have been addressed. Development of the platform has enabled exploration of various physicochemical and evolutionary properties of protein interface to be explored, with respect to different classes of interface as well as different anatomical regions of interface surfaces. These properties include residue propensity, hydropathy, polarity, interface size, sequence entropy, residue contact preference and substitution profiles.

An emerging theme has been that while clear differences exist between different interface classes, the differences between anatomical regions of the interface, more specifically the core and periphery of the interface surface, are more striking. A further unfolding motif is that with respect to physico-chemical characterization, sequence-entropy and observed evolutionary patterns of substitutions, the interacting residues of the interface can be seen to be intermediate between the buried core of the protein domain and solvent exposed surface. Moreover, the interface periphery most resembles the exposed surface in most aspects, whilst the solvent inaccessible interface core most resembles the buried protein interior. This outcome reflects the dual lifestyle of (non-obligate) protein complexes; they must exist stably in solution without engaging in aberrant aggregations but also mediate short-lived, specific molecular recognition events.

TOCCATA, a relational database of almost 4,000 family-based structural alignments, was established with the initial aim of aiding the exploration of evolutionary aspects of protein-protein interaction surfaces, most notably sequence entropy and the identification of distinguishing patterns of substitutions accepted through evolution in the form of interface-specific substitution tables. The group is developing a platform of structure-based software and databases tools to facilitate the high-throughput analysis of nsSNPs to aid prioritization of those that are most likely to be deleterious to protein structure, function and interactions. A pre-requisite for this work is a procedure to maximize structural coverage of the genome. Fortuitously, the development of TOCCATA was also able to aid in this process, as it provides the necessary systematic collection and alignment of suitable template structures for comparative modelling procedures. PICCOLO was also able to make its own useful contribution to these efforts by identifying

mutations that could impair protein function by disrupting protein interaction sites.

7.2 Interaction druggability

Historically there has been little focus on protein-protein interactions as targets for small-molecule therapeutics. However, alanine-scanning mutagenesis studies have revealed that only a subset of residues contribute the greater part of free energy to binding - so-called "hot-spots". Molecular characterization of hot-spots, performed using PICCOLO and TOCCATA, probed the molecular basis underlying this important phenomenon with respect to their residue propensity, sequence entropy, number and type of interactions, evolutionary conserved interactions, ligand efficiency and their relationship to residue substitution scores. Such characterization provides the basis for the next phase of this work, which will be to apply machine-learning methods to the problem of *in silico* identification of hot-spot residues.

Should such work prove successful, this would raise the intriguing longer-term prospect of pathway-centric structure-based druggability assessment. Druggability assessment is increasingly being applied to prioritize putative targets and focus valuable resources on those targets most likely to be chemically tractable to drug-like small molecules (Agüero *et al.* (2008); Hopkins & Groom (2002, 2003)). To date such approaches have been applied to individual proteins - not their complexes. Given a cellular pathway whose activity we wish to modulate (through some *a priori* knowledge of disease association), druggability assessment could be achieved through three stages. First structural coverage of the individual cellular components would have to be maximized. Efforts in genome-scale modelling could be applied to those components whose structure has not been solved experimentally. Docking methods, combined with advances in the comparative modelling of complexes, could then be applied to gain some structural representation of the details of the interfaces between each of the components. Integration of any available experimental data would be invaluable at this stage. Finally, the hot-spot prediction approaches, in combination with other established druggability assessment methods (precedence-based, site-tractability assessment,

chemogenomic analysis (Agüero *et al.* (2008)) would identify those points in the pathway most likely to be a chemically tractable point of intervention. There are prominent issues with the current reliability of some of the individual procedures in this scheme, however as these computational methods mature, complemented by further experimental data, such schemes become increasingly realistic, and indeed necessary to focus valuable resources.

7.3 Interaction dynamics

Specific and sensitive signal transduction cannot be sustained solely through weak, transient binary interactions. Tightly bound, enduring pairwise complexes would forego the opportunity for sensitive regulation. Moreover, it appears fidelity is ensured through co-operative assembly of multi-protein complexes; each binary interaction is weak but collectively they are strong. Nonetheless, in order to gain a thorough understanding of the structural determinants of such higher-order complexes, the nature of the constituent pairwise interfaces must be examined. There is growing evidence that specific pairwise association of interaction partners is preceded by non-specific association of in the so-called "encounter-complex", providing an opportunity for them to rapidly reorient to a discrete complementary arrangement due to the reduced degrees of conformational freedom (Blundell & Fernandez-Recio (2006)). Once assembled, further regulation can be achieved through adjustment of subcellular localization through post-translational modification and cytoskeletal transport.

7.4 Protein Flexibility

One fundamental aspect of protein-protein interactions that has conspicuously not been addressed here is that of flexibility. Proteins are not rocks. A variety of rearrangements are seen to occur upon association, ranging from side-chain conformation alterations, local backbone movements or large conformational changes involving entire secondary structure units or even complete domains (Goh *et al.* (2004)). An important distinction can be made between mobile but ordered regions versus intrinsically unstructured regions (Radivojac *et al.* (2004)). X-ray

crystallographic structures capture a consensus snapshot of a frozen ensemble, gathering little information on mobility, whereas NMR structures reflect protein flexibility to some degree. Aside from simple comparisons of the bound and unbound forms of members of a complex, flexibility of interactions can be assessed through a variety of methods (whose accuracy broadly corresponds to computational cost) including B-factors (Radivojac *et al.* (2004)), conformational variability across evolutionary families (Velazquez-Muriel & Carazo (2009)), normal modes (Demirel & Keskin (2005)), resolving structures to generate conformational ensembles (Furnham *et al.* (2006)) and molecular dynamics (Smith *et al.* (2005)). PICCOLO is well placed to undertake a systematic review of the types of alterations that occur upon binding. The interaction data are organized and clustered by equivalence and secondary structure, local environment and interaction annotations are pre-calculated.

7.5 PICCOLO Availability

A simple web interface to display the contents of PICCOLO has been made available through the following URL:

http://www-cryst.bioc.cam.ac.uk/piccolo/piccolo.php

It provides a simple query form where a PDB entry can be entered and interface summaries are provided at the level of chain pairs, residue pairs and atom pairs. At the time of writing work is ongoing to add functionality to the web interface, in particular to address the issue of visualization, possibly with images such as those in Figure 2.15. Future developments being considered include making the data available in other forms, possibly as a webservice (Papazoglou (2007)) or in the form of a PyMOL plug-in (Delano (2002)). Alicia Higueruelo has kindly agreed to take a role in helping to maintain and update the PICCOLO database. However, at the current time the long term future of PICCOLO remains uncertain. Efforts are being made to further automate data generation.

7.6 Outlook

The dawn of the genomics era has yielded rapid progress in experimental determination of protein structures. Concomitantly, new experimental and computational techniques have begun to generate comprehensive protein-protein interaction maps. The combination of these factors creates the opportunity for new efforts in genome-wide structural modeling of protein-protein interactions. For such endeavors to become practical, significant efforts at organizing underlying data applied are required, along with new advances in high-throughput approaches to docking and modelling of interactions. Neither the recent efforts at achieving a systems level understanding of cellular processes, nor the component-by-component reductionist approach, can offer complete insight into the phenomena of cellular processes in isolation. Rather, the mutually-informing synthesis of the complementary "top-down" and "bottom-up" approaches offers the best hope of providing true insight.

Appendix A

Amino acid atom properties

Table A.1: Atomic properties for each residue used in generation of PICCOLO interaction fingerprints.

Residue	Atom	Hydrophobic	Aromatic	Cationic	Anionic	H-bond donor	H-bond acceptor	Amino-aromatic h-bond acceptor	vdW radius (Å)	Covalent radius (Å)
ALA	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
ARG	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+							1.88	0.77
	CD								1.88	0.77
	NE			+		+			1.64	0.70
	CZ			+					1.61	0.77
	NH1			+		+			1.64	0.70
	NH2			+		+			1.64	0.70
ASN	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG								1.61	0.77
	OD1						+		1.42	0.66
	ND2					+			1.64	0.70
ASP	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG				+				1.61	0.77
	OD1				+		+		1.42	0.66
	OD2				+		+		1.42	0.66
CYS	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	SG					+	+		1.77	1.04
GLN	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+							1.88	0.77
	CD								1.61	0.77
	OE1						+		1.42	0.66
	NE2					+			1.64	0.70
GLU	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+							1.88	0.77
	CD				+				1.61	0.77
	OE1				+		+		1.42	0.66
	OE2				+		+		1.42	0.66
GLY	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
HIS	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG		+	+					1.61	0.77

Table A.1: Atomic properties for each residue used in generation of PICCOLO interaction fingerprints.

Residue	Atom	Hydrophobic	Aromatic	Cationic	Anionic	H-bond donor	H-bond acceptor	Amino-aromatic h-bond acceptor	vdW radius (Å)	Covalent radius (Å)
	ND1		+	+		+	+		1.64	0.70
	CD2		+	+					1.76	0.77
	CE1		+	+					1.76	0.77
	NE2		+	+		+			1.64	0.70
ILE	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG1	+							1.88	0.77
	CG2	+							1.88	0.77
	CD1	+							1.88	0.77
LEU	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+							1.88	0.77
	CD1	+							1.88	0.70
	CD2	+							1.88	0.77
LYS	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+							1.88	0.77
	CD	+							1.88	0.77
	CE								1.88	0.77
	NZ			+		+			1.64	0.70
MET	N					+			1.64	0.7
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+							1.88	0.77
	SD						+		1.77	1.04
	CE	+							1.88	0.77
PHE	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+	+					+	1.61	0.77
	CD1	+	+					+	1.76	0.77
	CD2	+	+					+	1.76	0.77
	CE1	+	+					+	1.76	0.77
	CE2	+	+					+	1.76	0.77
	CZ	+	+					+	1.76	0.77
PRO	N								1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB	+							1.88	0.77
	CG	+							1.88	0.77
	CD								1.88	0.77
SER	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB								1.88	0.77
	OG					+	+		1.46	0.66
THR	N					+			1.64	0.70
	CA								1.88	0.77
	C								1.61	0.77
	O						+		1.42	0.66
	CB								1.88	0.77

Table A.1: Atomic properties for each residue used in generation of PICCOLO interaction fingerprints.

Residue	Atom	Hydrophobic	Aromatic	Cationic	Anionic	H-bond donor	H-bond acceptor	Amino-aromatic	h-bond acceptor	vdW radius (Å)	Covalent radius (Å)
	OG1					+	+			1.46	0.66
	CG2	+								1.88	0.77
TRP	N					+				1.64	0.70
	CA									1.88	0.77
	C									1.61	0.77
	O						+			1.42	0.66
	CB	+								1.88	0.77
	CD1		+						+	1.76	0.77
	CD2	+	+						+	1.61	0.77
	NE1		+			+			+	1.64	0.70
	CE2	+	+						+	1.61	0.77
	CE3	+	+						+	1.76	0.77
	CG	+	+						+	1.61	0.77
	CZ2	+	+						+	1.76	0.77
	CZ3	+	+						+	1.76	0.77
	CH2	+	+						+	1.76	0.77
TYR	N					+				1.64	0.70
	CA									1.88	0.77
	C									1.61	0.77
	O						+			1.42	0.66
	CB	+								1.88	0.77
	CG	+	+						+	1.61	0.77
	CD1	+	+						+	1.76	0.77
	CD2	+	+						+	1.76	0.77
	CE1	+	+						+	1.76	0.77
	CE2	+	+						+	1.76	0.77
	CZ	+	+						+	1.61	0.77
	OH					+	+			1.46	0.66
VAL	N					+				1.64	0.70
	CA									1.88	0.77
	C									1.61	0.77
	O						+			1.42	0.66
	CB	+								1.88	0.77
	CG1	+								1.88	0.77
	CG2	+								1.88	0.77

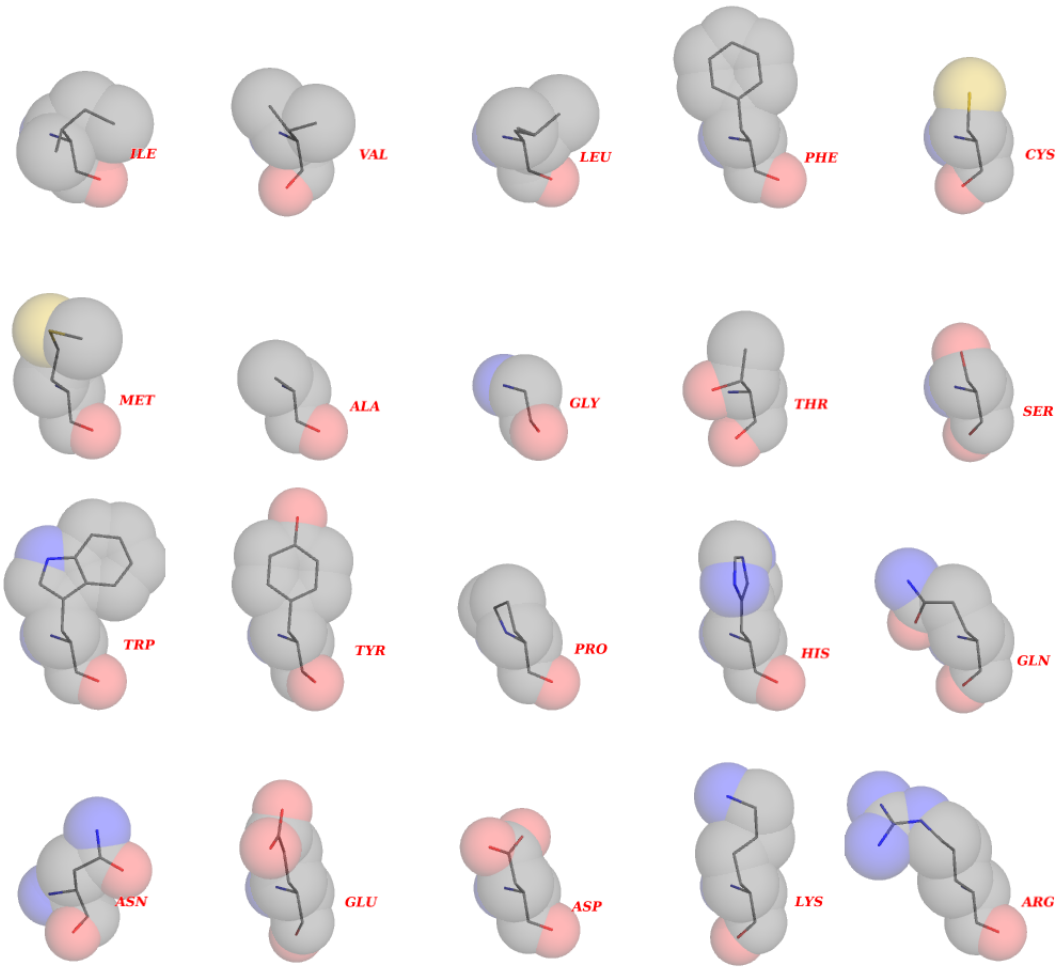


Figure A.1: van der Waals radius for atoms from the 20 canonical residues.

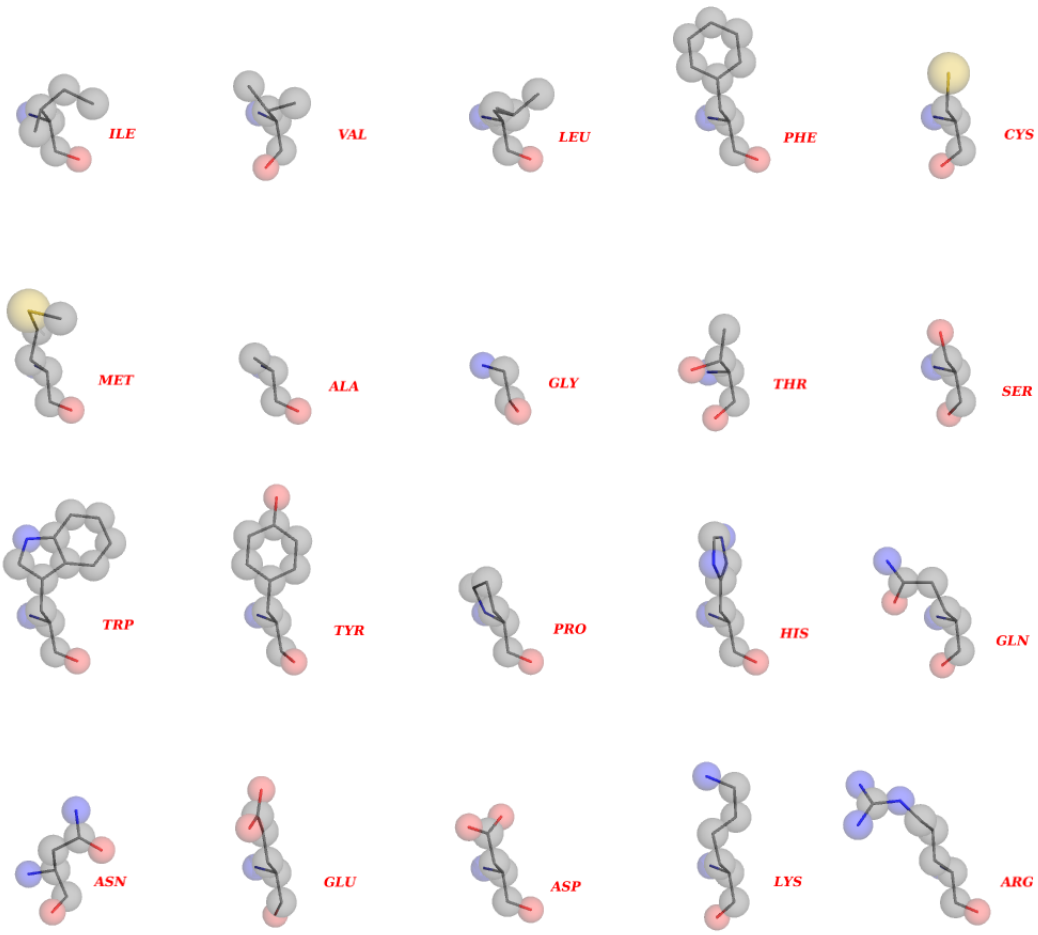


Figure A.2: Covalent radius for atoms from the 20 canonical residues.

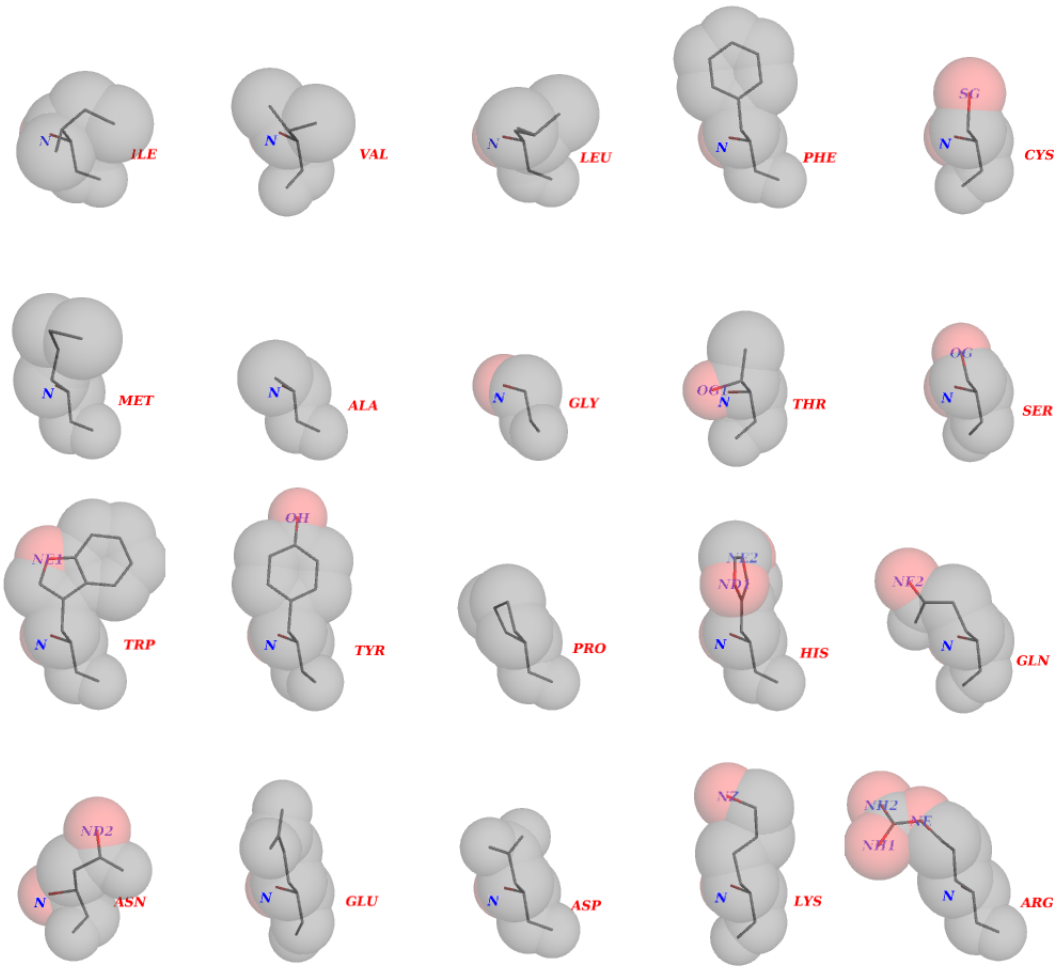


Figure A.3: Hydrogen bond donors from the 20 canonical residues.

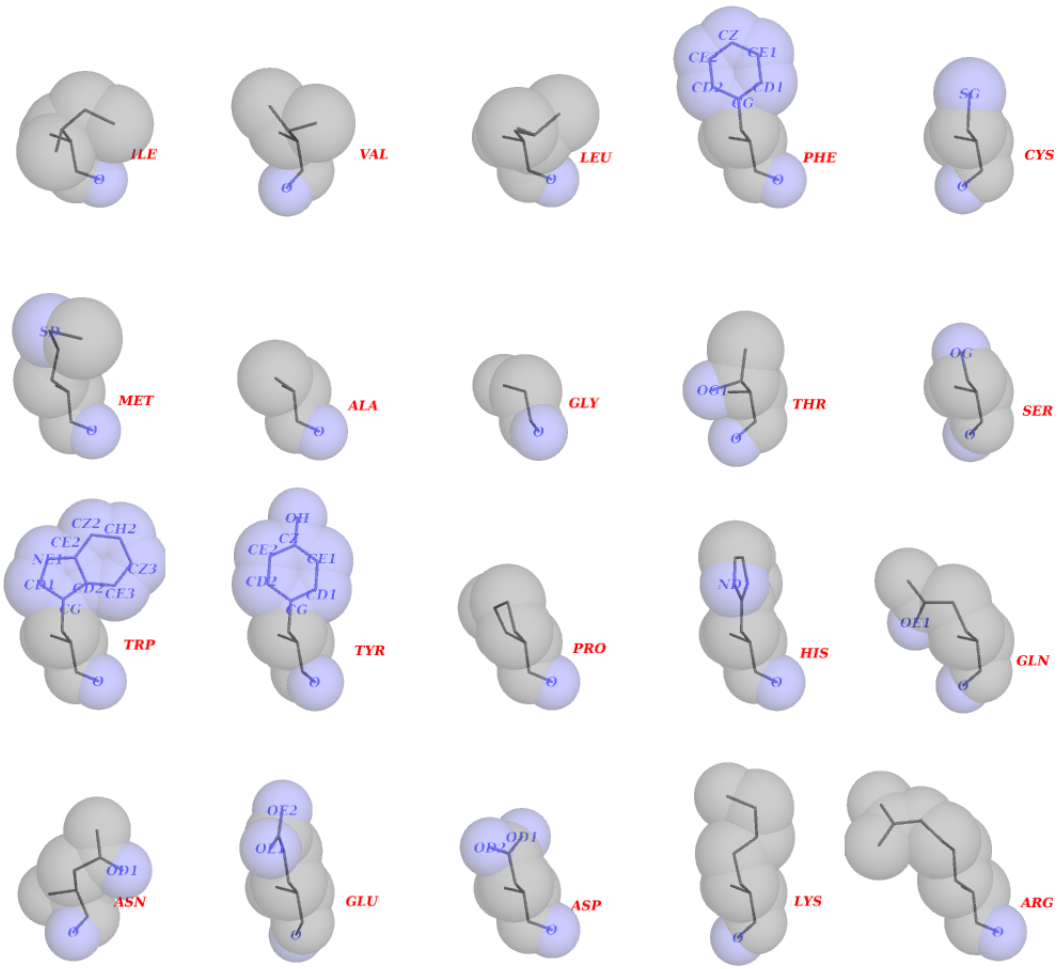


Figure A.4: Hydrogen bond acceptors from the 20 canonical residues.

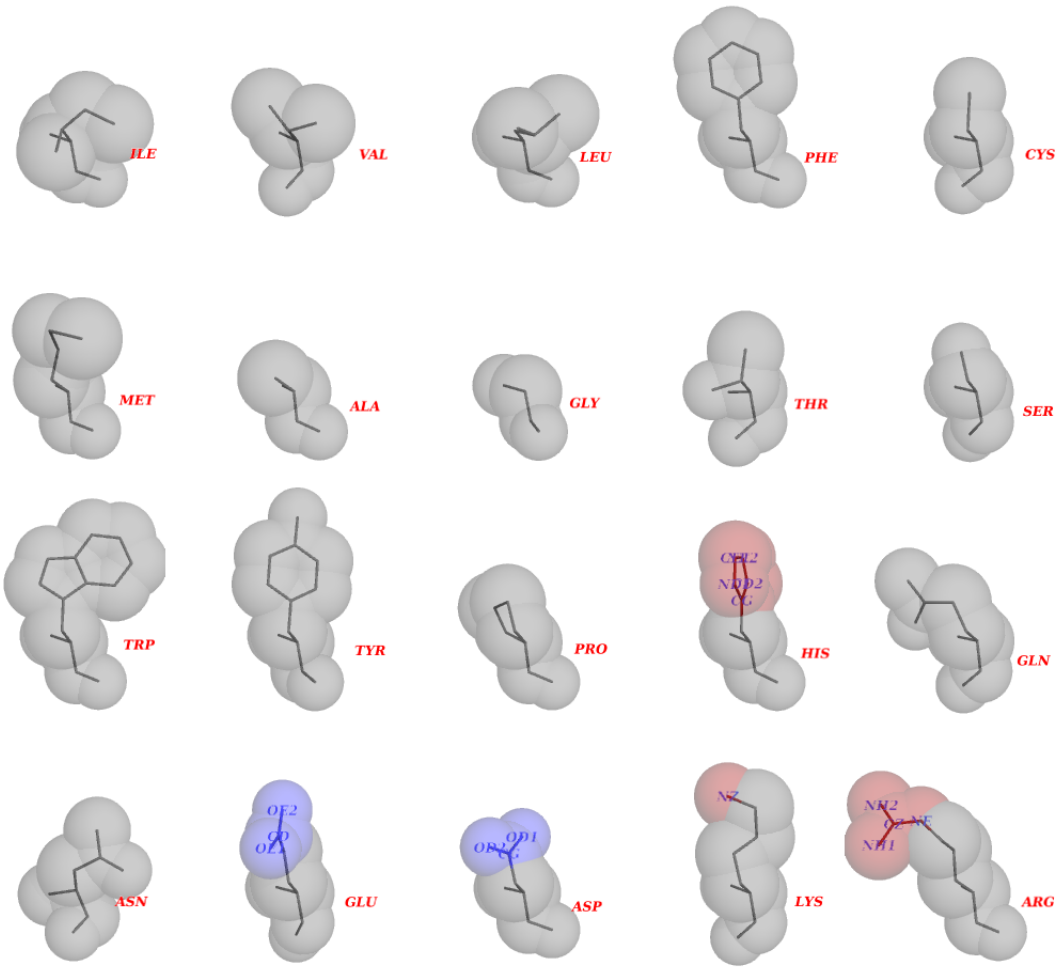


Figure A.5: Ionizable atoms from the 20 canonical residues.

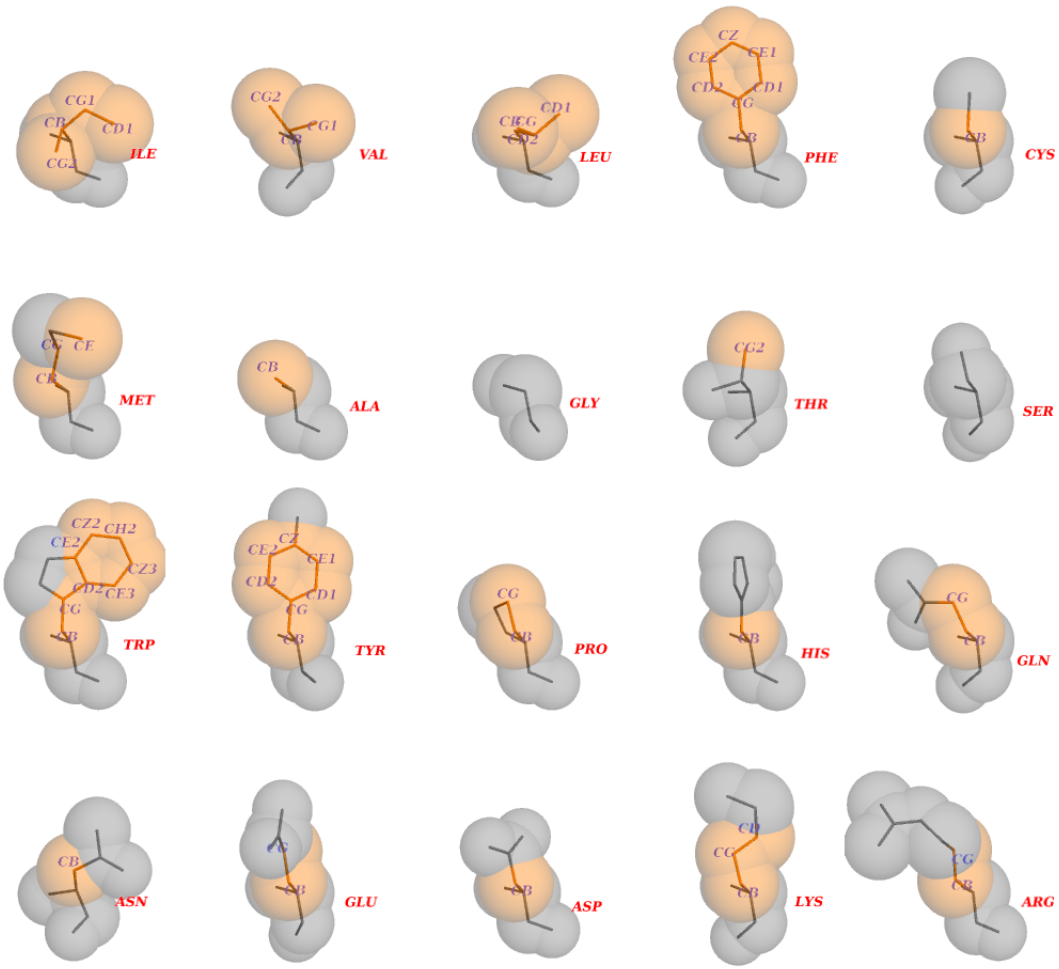


Figure A.6: Hydrophobic atoms from the 20 canonical residues.

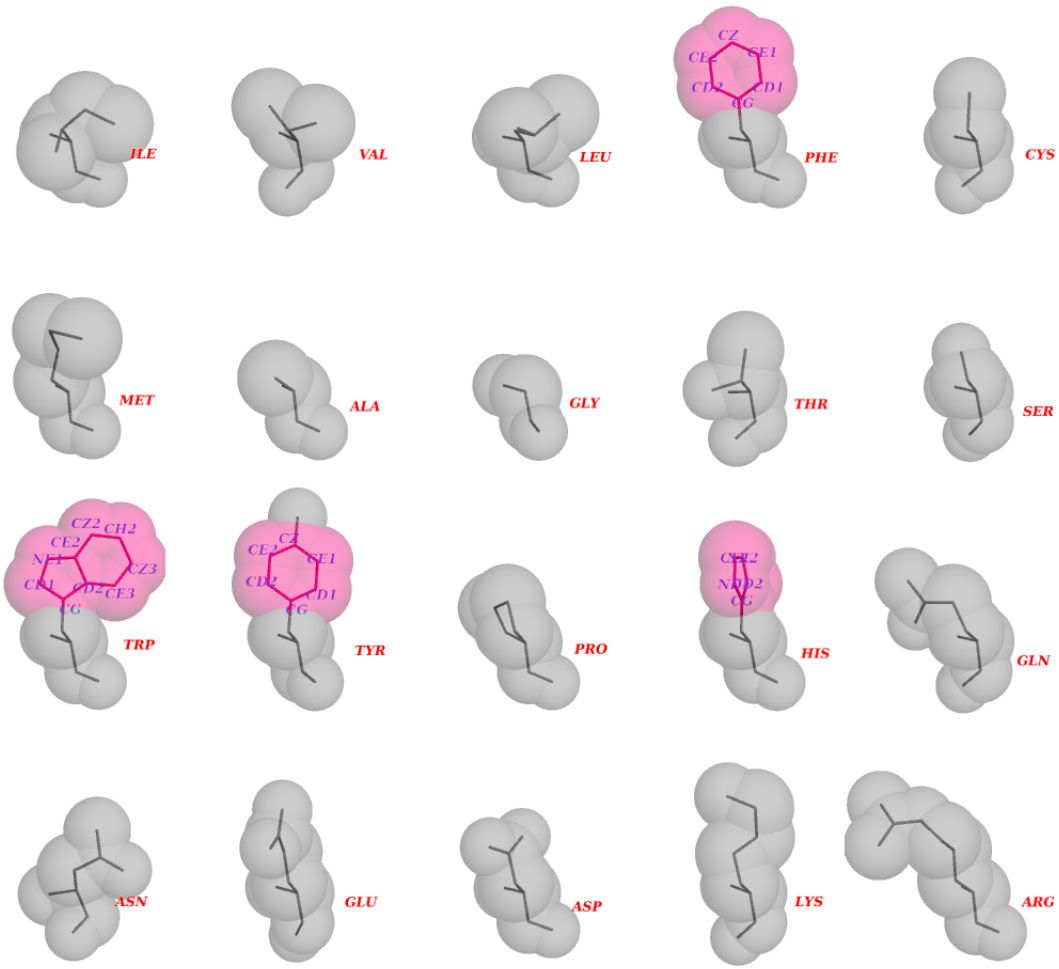


Figure A.7: Aromatic atoms from the 20 canonical residues.

Appendix B

Double Mutant Cycle.

Table B.1: Double mutant cycle data used in interaction identification benchmark.

PDB entry	P1 chain identifier	P1 residue number	P1 residue type	P2 chain identifier	P2 residue number	P2 residue type	$\Delta\Delta G$ (kcal/mol)
1dqj	A	32	N	C	96	K	4.4
1dqj	A	91	S	C	21	R	-0.5
1dqj	A	91	S	C	20	Y	1.1
1dqj	A	96	Y	C	21	R	-1.1
1dqj	A	96	Y	C	100	S	1
1dqj	B	32	D	C	97	K	3
1dqj	B	53	Y	C	62	W	0.7
1dqj	B	53	Y	C	63	W	0.3
1dqj	B	53	Y	C	75	L	1.5
1dqj	B	53	Y	C	101	D	-0.2
1dqj	B	98	W	C	100	S	0.4
1dqj	B	98	W	C	97	K	1.8
1dqj	B	98	W	C	20	Y	3.1
1brs	C	27	K	F	42	T	1.5
1brs	C	27	K	F	39	D	4.8
1brs	C	27	K	F	35	D	0.4
1brs	C	27	K	F	76	E	0.1
1brs	C	27	K	F	29	Y	0.2
1brs	C	27	K	F	80	E	0.4
1brs	C	59	R	F	42	T	0.2
1brs	C	59	R	F	35	D	3.4
1brs	C	59	R	F	76	E	1.7
1brs	C	59	R	F	29	Y	0.6
1brs	C	59	R	F	80	E	0.6
1brs	C	83	R	F	76	E	0.1
1brs	C	87	R	F	38	W	0.2
1brs	C	87	R	F	39	D	6.1
1brs	C	87	R	F	29	Y	1
1brs	C	87	R	F	42	T	0.4
1brs	C	87	R	F	76	E	0.1
1brs	C	87	R	F	80	E	0
1brs	C	102	H	F	39	D	4.9
1brs	C	102	H	F	42	T	-0.1
1brs	C	102	H	F	76	E	0
1brs	C	102	H	F	80	E	0.1
1brs	C	102	H	F	29	Y	3.3
1brs	C	73	E	F	39	D	2.9
1a4y	A	434	Y	B	5	R	1.4
1a4y	A	435	D	B	5	R	0.9
1vfb	A	32	Y	C	121	Q	2
1vfb	A	50	Y	C	18	D	-0.4
1vfb	A	50	Y	C	119	D	0.3
1vfb	A	92	W	C	121	Q	2.7
1vfb	A	32	Y	C	124	I	0
1vfb	A	92	W	C	124	I	0.7

Table B.1: Double mutant cycle data used in interaction identification benchmark.

PDB entry	P1 chain identifier	P1 residue number	P1 residue type	P2 chain identifier	P2 residue number	P2 residue type	$\Delta\Delta G$ (kcal/mol)
1vfb	A	92	W	C	125	R	1.7
1vfb	A	92	W	C	129	L	0.2
1vfb	B	32	Y	C	116	K	0.2
1vfb	B	52	W	C	119	D	-0.3
1vfb	B	54	D	C	118	T	0.6
1vfb	B	100	D	C	24	S	0.3
3hfm	L	31	N	Y	96	K	4.7
3hfm	L	50	Y	Y	96	K	3.8
3hfm	L	96	Y	Y	21	R	-1.9
3hfm	L	50	Y	Y	21	R	-0.7
3hfm	L	50	Y	Y	97	K	3.5
3hfm	H	50	Y	Y	21	R	0.5
3hfm	H	98	W	Y	96	K	4.8
3hfm	H	32	D	Y	97	K	3.5
3hfm	H	33	Y	Y	97	K	5
1lfd	A	29	N	B	225	Q	0.9
1lfd	A	32	K	B	225	Q	0.2
1lfd	A	53	H	B	229	V	-0.8
1lfd	A	51	D	B	231	E	-1.3
1lfd	A	53	H	B	231	E	-0.4
1lfd	A	51	D	B	233	D	0.6
1lfd	A	53	H	B	233	D	-0.1
1lfd	A	20	R	B	237	E	0.6
1lfd	A	27	N	B	237	E	0.1
1lfd	A	31	Y	B	237	E	1.2
1lfd	A	33	S	B	237	E	0.8
1lfd	A	31	Y	B	238	D	2.7
1lfd	A	33	S	B	238	D	1.7
1lfd	A	52	K	B	238	D	1.4
1lfd	A	29	N	B	239	S	-0.3
1lfd	A	31	Y	B	239	S	-1.3
1lfd	A	48	K	B	239	S	-0.2
1lfd	A	53	H	B	239	S	-0.1
1lfd	A	27	N	B	241	R	0.3
1lfd	A	29	N	B	241	R	-0.2
1lfd	A	48	K	B	263	E	0.4
1lfd	A	32	K	B	238	D	0.7
1lfd	A	33	S	B	238	D	0.9
1lfd	A	51	D	B	238	D	0
1lfd	A	52	K	B	238	D	3.1
1gua	A	21	V	B	68	T	0.1
1gua	A	21	V	B	69	V	0.3
1gua	A	21	V	B	88	V	0.7
1gua	A	27	I	B	88	V	-0.4
1gua	A	31	E	B	59	R	0

Table B.1: Double mutant cycle data used in interaction identification benchmark.

PDB entry	P1 chain identifier	P1 residue number	P1 residue type	P2 chain identifier	P2 residue number	P2 residue type	$\Delta\Delta G$ (kcal/mol)
1gua	A	31	E	B	68	T	0.4
1gua	A	31	E	B	84	K	0.7
1gua	A	33	D	B	59	R	0.3
1gua	A	33	D	B	84	K	0.6
1gua	A	36	I	B	59	R	-0.9
1gua	A	36	I	B	68	T	0.2
1gua	A	36	I	B	69	V	0.1
1gua	A	37	E	B	59	R	1.2
1gua	A	37	E	B	68	T	0.4
1gua	A	37	E	B	69	V	0.1
1gua	A	37	E	B	84	K	0.5
1gua	A	38	D	B	59	R	0.5
1gua	A	38	D	B	68	T	1.6
1gua	A	39	S	B	66	Q	-0.1
1gua	A	39	S	B	67	R	0.1
1gua	A	39	S	B	68	T	-0.2
1gua	A	41	R	B	64	N	0.1
1gua	A	41	R	B	66	Q	-0.2
1gua	A	46	V	B	64	N	-0.5
1gua	A	46	V	B	66	Q	-0.1
1gua	A	46	V	B	67	R	-0.1
1gua	A	46	V	B	68	T	0.1
1gua	A	37	E	B	59	R	1
1gua	A	37	E	B	69	V	-0.3

References

Abecasis, G., Tam, P.K., Bustamante, C.D., Ostrander, E.A., Scherer, S.W., Chanock, S.J., Kwok, P.Y. & Brookes, A.J. (2007). Human genome variation 2006: emerging views on structural variation and large-scale snp analysis. *Nat Genet*, **39**, 153–5+. 155

Agüero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F.S., Campbell, R.K., Carmona, S., Carruthers, I.M., Chan, E.A.W., Chen, F., Crowther, G.J., Doyle, M.A., Hertz-Fowler, C., Hopkins, A.L., Mcallister, G., Nwaka, S., Overington, J.P., Pain, A., Paolini, G.V., Pieper, U., Ralph, S.A., Riechers, A., Roos, D.S., Sali, A., Shanmugam, D., Suzuki, T., Van Voorhis, W.C. & Verlinde, C.L.M.J. (2008). Genomic-scale prioritization of drug targets: the tdr targets database. *Nature Reviews Drug Discovery*, **7**, 900–907. 204, 205

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., Rout, M.P. & Sali, A. (2007). Determining the architectures of macromolecular assemblies. *Nature*, **450**, 683–694. 12

Allen, F.H. (2002). The cambridge structural database: a quarter of a million crystal structures and rising. *Acta crystallographica. Section B, Structural science*, **58**, 380–388. 54

Aloy, P. & Russell, R.B. (2003). Interprets: protein interaction prediction through tertiary structure. *Bioinformatics*, **19**, 161–162. 12, 25

Aloy, P., Ceulemans, H., Stark, A. & Russell, R.B. (2003). The relationship between sequence and interaction divergence in proteins. *Journal of molecular biology*, **332**, 989–998. 11

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. 121, 122, 130

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. 127

Andrusier, N., Mashiach, E., Nussinov, R. & Wolfson, H.J. (2008). Principles of flexible protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, **73**, 271–289. 14

Ansari, S. & Helms, V. (2005). Statistical analysis of predominantly transient protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, **61**, 344–355. 17, 18, 81, 111

Arkin, M.R. & Wells, J.A. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat Rev Drug Discov*, **3**, 301–317. 21, 179

Bader, G.D., Betel, D. & Hogue, C.W. (2003). Bind: the biomolecular interaction network database. *Nucleic acids research*, **31**, 248–250. 19

Bahadur, R.P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, **336**, 943–955. 40

Bairoch, A. (1991). Prosite: a dictionary of sites and patterns in proteins. *Nucleic acids research*, **19 Suppl**, 2241–2245. 122

Barlow, D.J. & Thornton, J.M. (1983). Ion-pairs in proteins. *J Mol Biol*, **168**, 867–885. 57

Bash, P.A., Singh, U.C., Langridge, R. & Kollman, P.A. (1987). Free energy calculations by computer simulation. *Science*, **236**, 564–568+. 157

Bennett, S.T., Barnes, C., Cox, A., Davies, L. & Brown, C. (2005). Toward the 1,000 dollars human genome. *Pharmacogenomics*, **6**, 373–82+. 155

Berman, H., Henrick, K., Nakamura, H. & Markley, J.L. (2007). The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic acids research*, **35**, D301–D303. 2

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. *Nucl. Acids Res.*, **28**, 235–242. 163

Bernauer, J., Bahadur, R.P.P., Rodier, F., Janin, J. & Poupon, A. (2008). Dimovo: a voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics (Oxford, England)*, **24**, 652–658. 35, 40

Birney, E. (2006). Ensembl 2006. *Nucleic acids research*, **34**, D556–D561. 46

Blundell, T.L. & Fernandez-Recio, J. (2006). Cell biology: brief encounters bolster contacts. *Nature*, **444**, 279–280. 205

Bogan, A.A. & Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, **280**, 1–9. 21, 181, 192

Bordner, A.J. & Abagyan, R.A. (2004). Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins*, **57**, 400–13+. 157

Bordner, A.J. & Gorin, A.A. (2008). Comprehensive inventory of protein complexes in the protein data bank from consistent classification of interfaces. *BMC bioinformatics*, **9**, 234+. 40

Bradford, J.R. & Westhead, D.R. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494. 8, 80

Brewer, J.W., Randall, T.D., Parkhouse, R.M. & Corley, R.B. (1994). Mechanism and subcellular localization of secretory igm polymer assembly. *The Journal of biological chemistry*, **269**, 17338–17348. 42

Bruncko, M., Oost, T.K., Belli, B.A., Ding, H., Joseph, M.K., Kunzer, A., Martineau, D., Mcclellan, W.J., Mitten, M., Ng, S.C., Nimmer, P.M., Oltersdorf, T., Park, C.M., Petros, A.M., Shoemaker, A.R., Song, X., Wang, X., Wendt, M.D., Zhang, H., Fesik, S.W., Rosenberg, S.H. & Elmore, S.W. (2007). Studies leading to potent, dual inhibitors of bcl-2 and bcl-xl. *J. Med. Chem.*, **50**, 641–662. 179

Burke, D.F., Worth, C.L., Priego, E.M., Cheng, T., Smink, L.J., Todd, J.A. & Blundell, T.L. (2007). Genome bioinformatic analysis of nonsynonymous snps. *BMC Bioinformatics*, **8**, 301+. 121, 158

Burley, S.K. & Petsko, G.A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, **229**, 23–28. 33

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J. & Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein science : a publication of the Protein Society*, **13**, 190–202. 17, 35, 109

Calloni, G., Zoffoli, S., Stefani, M., Dobson, C.M. & Chiti, F. (2005). Investigating the effects of mutations on protein aggregation in the cell. *The Journal of biological chemistry*, **280**, 10607–10613. 98

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. 14

Capriotti, E., Fariselli, P. & Casadio, R. (2004). A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **20 Suppl 1**, I63–I68+. 157

Capriotti, E., Fariselli, P., Calabrese, R. & Casadio, R. (2005a). Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **21 Suppl 2**, ii54–ii58+. 157

Capriotti, E., Fariselli, P. & Casadio, R. (2005b). I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–10+. 157

Care, M.A., Needham, C.J., Bulpitt, A.J. & Westhead, D.R. (2007). Deleterious snp prediction: be mindful of your training data! *Bioinformatics (Oxford, England)*, **23**, 664–672. 166

Carugo, O. & Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Sci*, **6**, 2261–2263. 40

Cazals, F., Proust, F., Bahadur, R.P. & Janin, J. (2006). Revisiting the voronoi description of protein-protein interfaces. *Protein Sci*, **15**, 2082–2092. 35

Chandonia, J.M. & Brenner, S.E. (2006). The impact of structural genomics: Expectations and outcomes. *Science*, **311**, 347–351. 123

Chandonia, J.M.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S.E. (2004). The astral compendium in 2004. *Nucleic acids research*, **32**, D189–D192. 48

Chatr-Aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. & Cesareni, G. (2007). Mint: the molecular interaction database. *Nucleic Acids Res*, **35**, D572–D574. 19

Chelliah, V., Chen, L., Blundell, T.L. & Lovell, S.C. (2004). Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J Mol Biol*, **342**, 1487–1504. 10, 21, 123, 158, 163

Chelliah, V., Blundell, T.L. & Fernandez-Recio, J. (2006). Efficient restraints for protein-protein docking by comparison of observed amino acid substitution patterns with those predicted from local environment. *Journal of Molecular Biology*, **357**, 1669–1682. 15

Chen, C.Z. & Shapiro, R. (1999). Superadditive and subadditive effects of "hot spot" mutations within the interfaces of placental ribonuclease inhibitor with angiogenin and ribonuclease a. *Biochemistry*, **38**, 9273–9285. 38

Chen, R., Mintseris, J., Janin, J. & Weng, Z. (2003). A protein-protein docking benchmark. *Proteins*, **52**, 88–91. 28

Chen, Y.C., Chen, H.C. & Yang, J.M. (2006). Dapid: a 3d-domain annotated protein-protein interaction database. *Genome informatics. International Conference on Genome Informatics*, **17**, 206–215. 25

Cheng, J., Randall, A. & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–32+. 157

Clackson, T. & Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, **267**, 383–386. 180

Clifford, R.J., Edmonson, M.N., Nguyen, C. & Buetow, K.H. (2004). Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics*, **20**, 1006–14+. 156

Codd, E.F. (1983). A relational model of data for large shared data banks. *Commun. ACM*, **26**, 64–69. 43

Collins, S.R., Kemmeren, P., Zhao, X.C.C., Greenblatt, J.F., Spencer, F., Holstege, F.C., Weissman, J.S. & Krogan, N.J. (2007). Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP*, **6**, 439–450. 4

Crowley, P.B. & Golovin, A. (2005). Cation-pi interactions in protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, **59**, 231–239. 34

Dai, D., Tang, J., Rose, R., Hodgson, E., Bienstock, R.J., Mohrenweiser, H.W. & Goldstein, J.A. (2001). Identification of variants of cyp3a4 and characterization of their abilities to metabolize testosterone and chlorpyrifos. *J. Pharmacol. Exp. Ther.*, **299**, 825–31+. 155

Dall'Acqua, W., Goldman, E.R., Lin, W., Teng, C., Tsuchiya, D., Li, H., Ysern, X., Braden, B.C., Li, Y., Smith-Gill, S.J. & Mariuzza, R.A. (1998). A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry*, **37**, 7981–7991. 38

Darnell, S.J., Page, D. & Mitchell, J.C. (2007). An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, **68**, 813–823. 182

Darnell, S.J.J., Legault, L. & Mitchell, J.C.C. (2008). Kfc server: interactive forecasting of protein interaction hot spots. *Nucleic acids research*, W265–W269. 182

Dasgupta, S., Iyer, G.H., Bryant, S.H., Lawrence, C.E. & Bell, J.A. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, **28**, 494–514. 40

Davies, M.N., Toseland, C.P., Moss, D.S. & Flower, D.R. (2006). Benchmarking pka prediction. *BMC Biochemistry*, **7**, 18+. 57

Davis, F.P. & Sali, A. (2005). Pibase: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907. 26

De, S., Krishnadev, O., Srinivasan, N. & Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Structural Biology*, **5**, 15+. 18

de Bakker, P.I., Furnham, N., Blundell, T.L. & Depristo, M.A. (2006). Conformer generation under restraints. *Curr Opin Struct Biol*, **16**, 160–5+. 159

de Berg, M., van Kreveld, M., Overmars, M. & Schwarzkopf, O. (1997). *Computational Geometry: Algorithms and Applications*. Springer, 1st edn. 52

de Vries, S.J. (2006). Intramolecular surface contacts contain information about protein-protein interface regions. *Bioinformatics*, **22**, 2094–2098. 15

de Vries, S.J., van Dijk, A.D. & Bonvin, A.M. (2006). Whiscy: What information does surface conservation yield? application to data-driven docking. *Proteins: Structure, Function, and Bioinformatics*, **63**, 479–489. 10

del Sol, A. & O'Meara, P. (2005). Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, **58**, 672–682. 182

Delano, W.L. (2002). The pymol molecular graphics system. xv, xxi, 62, 78, 171, 206

DeLano, W.L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Current opinion in structural biology*, **12**, 14–20. 21, 183, 193

Demirel, M.C. & Keskin, O. (2005). Protein interactions and fluctuations in a proteomic network using an elastic network model. *J Biomol Struct Dyn*, **22**, 381–386. 206

Dolinsky, T.J., Czodrowski, P., Li, H., Nielsen, J.E., Jensen, J.H., Klebe, G. & Baker, N.A. (2007). Pdb2pqr: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, **35**, W522–W525. 57

Dou, Y., Baisnée, P.F., Pollastri, G., Pécout, Y., Nowick, J. & Baldi, P. (2004). Icbs: a database of interactions between protein chains mediated by beta-sheet formation. *Bioinformatics*, **20**, 2767–2777. 25

Douguet, Dominique, Chen, Huei-Chi, Tovchigrechko, Andrey, Vakser & Ilya, A. (2006). Dockground resource for studying proteinprotein interfaces. *Bioinformatics*, **22**, 2612–2618. 25

Drummond, D.A. (2005). A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution*, **23**, 327–337. 7

Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. 121

Eddy, S.R. (1998). Profile hidden markov models. *Bioinformatics*, **14**, 755–63+. 156

Eddy, S.R. (2004). Where did the blosum62 alignment score matrix come from? *Nature biotechnology*, **22**, 1035–1036. 166

Eisenstein, M. & Katchalski-Katzir, E. (2004). On proteins, grids, correlations, and docking. *C R Biol*, **327**, 409–420. 14

Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2002). Soft protein-protein docking in internal coordinates. *Protein Sci*, **11**, 280–291. 15

Fernandez-Recio, J., Totrov, M. & Abagyan, R. (2003). Icm-disco docking by global energy optimization with fully flexible side-chains. *Proteins*, **52**, 113–117. 15

Fernandez-Recio, J., Totrov, M., Skorodumov, C. & Abagyan, R. (2005). Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins*, **58**, 134–143. 8

Finn, R.D., Marshall, M. & Bateman, A. (2005). ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412. 28

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L.L. & Bateman, A. (2008). The pfam protein families database. *Nucl. Acids Res.*, **36**, D281–288. 122, 156

Fischer, T.B., Arunachalam, K.V., Bailey, D., Mangual, V., Bakhru, S., Russo, R., Huang, D., Paczkowski, M., Lalchandani, V., Ramachandra, C., Ellison, B., Galer, S., Shapley, J., Fuentes, E. & Tsai, J. (2003). The binding interface database (bid): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, **19**, 1453–1454. 28

Fischer, T.B., Holmes, J.B., Miller, I.R., Parsons, J.R., Tung, L., Hu, J.C. & Tsai, J. (2006). Assessing methods for identifying pair-wise atomic contacts across binding interfaces. *J Struct Biol*, **153**, 103–112. 37

Forman, J.R., Worth, C.L., Bickerton, G.R., Eisen, T.G. & Blundell, T.L. (2009). Structural bioinformatics mutation analysis reveals genotype-phenotype correlations in von hippel-lindau disease and suggests molecular mechanisms of tumorigenesis. *Proteins: Structure, Function, and Bioinformatics*, **77**, 84–96. 174

Fredman, D., Siegfried, M., Yuan, Y.P., Bork, P., Lehvaslaiho, H. & Brookes, A.J. (2002). Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.*, **30**, 387–91+. 156

Frenz, C.M. (2005). Neural network-based prediction of mutation-induced protein stability changes in staphylococcal nuclease at 20 residue positions. *Proteins*, **59**, 147–51+. 157

Friedman, N., Ninio, M., Pe'er, I. & Pupko, T. (2002). A structural em algorithm for phylogenetic inference. *J Comput Biol*, **9**, 331–53+. 166

Fukuhara, N. & Kawabata, T. (2008). Homcos: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Research*, **36**, W185–W189. 12

Funahashi, J., Sugita, Y., Kitao, A. & Yutani, K. (2003). How can free energy component analysis explain the difference in protein stability caused by amino acid substitutions? effect of three hydrophobic mutations at the 56th residue on the stability of human lysozyme. *Protein Eng.*, **16**, 665–71+. 157

Furnham, N., Blundell, T.L., Depristo, M.A. & Terwilliger, T.C. (2006). Is one solution good enough? *Nature Structural & Molecular Biology*, **13**, 184–185. 206

Gabb, H.A., Jackson, R.M. & Sternberg, M.J. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology*, **272**, 106–120. 14

Gallivan, J.P. & Dougherty, D.A. (1999). Cation-pi interactions in structural biology. *PNAS*, **96**, 9459–9464. 33, 58

Galperin, M.Y. & Koonin, E.V. (2000). Who's your neighbor? new computational approaches for functional genomics. *Nat Biotechnol*, **18**, 609–613. 6

Gao, Y., Douguet, D., Tovchigrechko, A. & Vakser, I.A. (2007). Dockground system of databases for protein recognition studies: Unbound structures for docking. *Proteins: Structure, Function, and Bioinformatics*, **69**, 845–851. 25

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. & Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636. 5

Gilis, D. & Rooman, M. (1997). Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol.*, **272**, 276–90+. 157

Glaser, F., Steinberg, D.M., Vakser, I.A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, **43**, 89–102. 35, 84, 117, 119

Goh, C.S. & Cohen, F.E. (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. *Journal of molecular biology*, **324**, 177–192. 7

Goh, C.S., Milburn, D. & Gerstein, M. (2004). Conformational changes associated with protein-protein interactions. *Current Opinion in Structural Biology*, **14**, 104–109. 205

Gong, S. & Blundell, T.L. (2008). Discarding functional residues from the substitution table improves predictions of active sites within three-dimensional structures. *PLoS Comput Biol*, **4**, e1000179+. 124

Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J., Bolser, D.M., Oh, D., Kim, D.S. & Bhak, J. (2005a). A protein domain interaction interface database: Interpare. *BMC Bioinformatics*, **6**, 207+. 26

Gong, S., Yoon, G., Jang, I., Bolser, D., Dafas, P., Schroeder, M., Choi, H., Cho, Y., Han, K., Lee, S., Choi, H., Lappe, M., Holm, L., Kim, S., Oh, D. & Bhak, J. (2005b). Psibase: a database of protein structural interactome map (psimap). *Bioinformatics*, **21**, 2541–2543. 26

Gore, S.P., Burke, D.F. & Blundell, T.L. (2005). Provat: a tool for voronoi tessellation analysis of protein structures and complexes. *Bioinformatics*, **21**, 3316–3317. 35

Gore, S.P., Karmali, A.M. & Blundell, T.L. (2007). Rappertk: a versatile engine for discrete restraint-based conformational sampling of macromolecules. *BMC Struct Biol*, **7**, 13+. 159

Gottschalk, K.E., Neuvirth, H. & Schreiber, G. (2004). A novel method for scoring of docked protein complexes using predicted protein-protein binding sites. *Protein Eng Des Sel*, **17**, 183–189. 14

Green, N.J., Xiang, J., Chen, J., Chen, L., Davies, A.M., Erbe, D., Tam, S. & Tobin, J.F. (2003). Structure-activity studies of a series of dipyrazolo[3,4-b:3',4'-d]pyridin-3-ones binding to the immune regulatory protein b7.1. *Bioorganic & medicinal chemistry*, **11**, 2991–3013. 179

Greene, L.H., Lewis, T.E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J.M. & Orengo, C.A. (2007). The cath domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic acids research*, **35**, D291–D297. 127

Grosdidier, S. & Fernández-Recio, J. (2008). Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC bioinformatics*, **9**, 447+. 182

Gruber, J., Zawaira, A., Saunders, R., Barrett, C.P. & Noble, M.E.M. (2006). Computational analyses of the surface properties of protein-protein interfaces. *Acta Crystallographica Section D Biological Crystallography*, **63**, 50–57. 17

Guerois, R., Nielsen, J.E. & Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–87+. 157

Guharoy, M. & Chakrabarti, P. (2005). Conservation and relative importance of residues across protein-protein interfaces. *Proc Natl Acad Sci U S A*, **102**, 15447–15452. 62, 109

Gunderson, K.L., Steemers, F.J., Ren, H., Ng, P., Zhou, L., Tsan, C., Chang, W., Bullis, D., Musmacker, J., King, C., Lebruska, L.L., Barker, D., Oliphant, A., Kuhn, K.M. & Shen, R. (2006). Whole-genome genotyping. *Methods Enzymol*, **410**, 359–76+. 155

Guney, E., Tuncbag, N., Keskin, O. & Gursoy, A. (2007). Hotsprint: database of computational hot spots in protein interfaces. *Nucleic Acids Research*, **36**, D662–D666. 183

Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, **47**, 409–443. 14

Halperin, I., Wolfson, H. & Nussinov, R. (2004). Protein-protein interactions: Coupling of structurally conserved residues and of hot spots across interfaces. implications for docking. *Structure*, **12**, 1027–1038. 181

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. 7

Hamelryck, T. & Manderick, B. (2003). Pdb file parser and structure class implemented in python. *Bioinformatics*, **19**, 2308–2310. 50

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & Mckusick, V.A. (2005). Online mendelian inheritance in man (omim), a knowledge-base of human genes and genetic disorders. *Nucleic Acids Res*, **33**, D514–7+. 165

He, M.M., Smith, A.S., Oslob, J.D., Flanagan, W.M., Braisted, A.C., Whitty, A., Cancilla, M.T., Wang, J., Lugovskoy, A.A., Yoburn, J.C., Fung, A.D., Farrington, G., Eldredge, J.K., Day, E.S., Cruz, L.A., Cachero, T.G., Miller, S.K., Friedman, J.E., Choong, I.C. & Cunningham, B.C. (2005). Small-molecule inhibition of tnf-alpha. *Science*, **310**, 1022–1025. 179

Henrick, K. & Thornton, J.M. (1998). Pqs: a protein quaternary structure file server. *Trends Biochem Sci*, **23**, 358–361. 40, 41

Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C.L., Markley, J.L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E.L., Vranken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M. & Berman, H.M. (2007). Remediation of the protein data bank archive. *Nucleic Acids Research*, **36**, D426–D433. 50

Heringa, J. & Argos, P. (1999). Strain in protein structures as viewed through nonrotameric side chains: Ii. effects upon ligand binding. *Proteins*, **37**, 44–55+. 162

Hes, , van der Luijt, , Janssen, , Zewald, , Jong, D., , Lenders, , Links, , Luyten, , Sijmons, , Eussen, , Halley, , Lips, , Pearson, , den Ouweland, V., , Majoor-Krakauer & (2007). Frequency of von hippel-lindau germline mutations in classic and non-classic von hippel-lindau disease identified by dna sequencing, southern blot analysis and multiplex ligation-dependent probe amplification. *Clinical Genetics*, **72**, 122–129. 174

Higueruelo, A.P. & Blundell, T.L. (2009). Timbal: small molecules disrupting protein-protein interactions. 179

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. & Tyers, M. (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183. 5

Hopkins, A. (2004). Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today*, **9**, 430–431. 180, 196

Hopkins, A.L. & Groom, C.R. (2002). The druggable genome. *Nat Rev Drug Discov*, **1**, 727–730. 204

Hopkins, A.L. & Groom, C.R. (2003). Target analysis: a priori assessment of druggability. *Ernst Schering Res Found Workshop*, 11–17. 204

Horovitz, A. (1996). Double-mutant cycles: a powerful tool for analyzing protein structure and function. *Folding and Design*, **1**, R121–R126. 33, 184

Hoskins, J., Lovell, S. & Blundell, T.L. (2006). An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Protein Sci*, **15**, 1017–1029. 8

Huang, B. & Schroeder, M. (2008). Using protein binding site prediction to improve protein docking. *Gene*, **422**, 14–21. 9

Hubbard (1993). 'naccess', computer program. xii, 34, 36, 51, 79, 84

Hubbard, T.J., Ailey, B., Brenner, S.E., Murzin, A.G. & Chothia, C. (1999). Scop: a structural classification of proteins database. *Nucleic acids research*, **27**, 254–256. 42, 122

Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A. & Birney, E. (2007). Ensembl 2007. *Nucleic Acids Res*, **35**, D610–7+. 159

Hunter, C.A. & Sanders, J.K.M. (1990). The nature of π-π interactions. *Journal of the American Chemical Society*, **112**, 5525–5534. 33

Hunter, C.A., Singh, J. & Thornton, J.M. (1991). π-π interactions: the geometry and energetics of phenylalanine-phenylalanine interactions in proteins. *J Mol Biol*, **218**, 837–846. 33

Hwang, H., Pierce, B., Mintseris, J., Janin, J. & Weng, Z. (2008). Protein-protein docking benchmark version 3.0. *Proteins: Structure, Function, and Bioinformatics*, **73**, 705–709. 28

Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein-protein and protein-dna recognition. *Structure (London, England : 1993)*, **7**, R277–R279. 183

Janin, J. (2005). Assessing predictions of protein-protein interaction: the capri experiment. *Protein Sci*, **14**, 278–283. 16

Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *The Journal of biological chemistry*, **265**, 16027–16030. 17, 87

Janin, J. & Rodier, F. (1995). Protein-protein interaction at crystal contacts. *Proteins*, **23**, 580–587. 40

Jefferson, E.R., Walsh, T.P., Roberts, T.J. & Barton, G.J. (2007). Snappi-db: a database and api of structures, interfaces and alignments for protein-protein interactions. *Nucleic Acids Research*, **35**, D580–D589. 27, 42, 54

Johnson, M.S., Overington, J.P. & Blundell, T.L. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *J Mol Biol*, **231**, 735–752. 122, 138

Jones, S. & Thornton, J.M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13–20. 17, 18, 28, 41, 81, 87, 106

Jones, S. & Thornton, J.M. (1997). Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*, **272**, 133–143. 8

Jünger, M. (2003). *Graph Drawing Software (Mathematics and Visualization)*. Springer, 1st edn. xiv, 73

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–637+. 167

Karchin, R., Diekhans, M., Kelly, L., Thomas, D.J., Pieper, U., Eswar, N., Haussler, D. & Sali, A. (2005). Ls-snp: large-scale annotation of coding non-synonymous snps based on multiple information sources. *Bioinformatics*, **21**, 2814–20+. 159

Katchalski-Katzir, E., Shariv, Eisenstein, M., Friesem, A.A., Aflalo, C. & Vakser, I.A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, **89**, 2195–2199. 14

Keskin, O., Tsai, C.J., Wolfson, H. & Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, **13**, 1043–1055. 26, 54

Keskin, O., Ma, B. & Nussinov, R. (2005a). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, **345**, 1281–1294. 21, 181

Keskin, O., Ma, B., Rogale, K., Gunasekaran, K. & Nussinov, R. (2005b). Proteinprotein interactions: organization, cooperativity and mapping in a bottom-up systems biology approach. *Physical Biology*, **2**, S24–S35. 181

Kiel, C., Serrano, L. & Herrmann, C. (2004). A detailed thermodynamic analysis of ras/effector complex interfaces. *Journal of Molecular Biology*, **340**, 1039–1058. 38

Kimchi-Sarfaty, C., Oh, J.M., Kim, I., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V. & Gottesman, M.M. (2007). A "silent" polymorphism in the mdr1 gene changes substrate specificity. *Science*, **315**, 525–528. 154

Kirk, D. (1994). *Graphics Gems III (IBM Version) (Graphics Gems - IBM) (No. 3)*. Morgan Kaufmann, har/dis edn. 58

Koblish, H.K., Zhao, S., Franks, C.F., Donatelli, R.R., Tominovich, R.M., LaFrance, L.V., Leonard, K.A., Gushue, J.M., Parks, D.J.,

Calvo, R.R., Milkiewicz, K.L., Marugán, J.J., Raboisson, P., Cummings, M.D., Grasberger, B.L., Johnson, D.L., Lu, T., Molloy, C.J. & Maroney, A.C. (2006). Benzodiazepinedione inhibitors of the hdm2:p53 complex suppress human tumor cell proliferation in vitro and sensitize tumors to doxorubicin in vivo. *Molecular cancer therapeutics*, **5**, 160–169. 179

Kollman, P., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. & Cheatham (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, **33**, 889–97+. 157

Korbel, J.O., Jensen, L.J., von Mering, C. & Bork, P. (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotech*, **22**, 911–917. 7

Korkin, D., Davis, F.P., Alber, F., Luong, T., Shen, M.Y., Lucic, V., Kennedy, M.B. & Sali, A. (2006). Structural modeling of protein interactions by analogy: Application to psd-95. *PLoS Computational Biology*, **2**, e153+. 11

Kortemme, T. (2004). Computational alanine scanning of protein-protein interfaces. *Science's STKE*, **2004**, 2pl–2. 182

Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 14116–14121. 181

Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, **372**, 774–797. 41

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine,

A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A. & Greenblatt, J.F. (2006). Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, **440**, 637–643. 5

Kryshtafovych, A., Venclovas, v., Fidelis, K. & Moult, J. (2005). Progress over the first decade of casp experiments. *Proteins: Structure, Function, and Bioinformatics*, **61**, 225–236. 9, 16

Kubo, A., Nakagawa, K., Varma, R.K., Conrad, N.K., Cheng, J.Q., Lee, W.C., Testa, J.R., Johnson, B.E., Kaye, F.J. & Kelley, M.J. (1999). The p16 status of tumor cell lines identifies small molecule inhibitors specific for cyclin-dependent kinase 4. *Clin Cancer Res*, **5**, 4279–86+. 168

Kufareva, I., Budagyan, L., Raush, E., Totrov, M. & Abagyan, R. (2007). Pier: Protein interface recognition for structural proteomics. *Proteins: Structure, Function, and Bioinformatics*, **67**, 400–417. 8

Kumar, M.D.S. (2006). Pint: Protein-protein interactions thermodynamic database. *Nucleic Acids Research*, **34**, D195–D198. 28

Kundrotas, P.J. & Alexov, E. (2007). Protcom: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Research*, **35**, D575–D579. 26

Kuntz, I.D., Chen, K., Sharp, K.A. & Kollman, P.A. (1999). The maximal affinity of ligands. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 9997–10002. 196

Kyte, J. & Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, **157**, 105–132. 32, 80

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. & Ben-Tal, N. (2005). Consurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Research*, **33**, W299–W302. 9

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K. & International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921. 2

Landon, M.R., Lancia, D.R., Yu, J., Thiel, S.C. & Vajda, S. (2007). Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. *Journal of medicinal chemistry*, **50**, 1231–1240. 182

Laskowski, R.A. (2009). Pdbsum new things. *Nucleic acids research*, **37**, D355–359. 28

Laskowski, R.A., Luscombe, N.M., Swindells, M.B. & Thornton, J.M. (1996). Protein clefts in molecular recognition and function. *Protein Sci*, **5**, 2438–52+. 163

Latif, R., Graves, P. & Davies, T.F. (2002). Ligand-dependent inhibition of oligomerization at the human thyrotropin receptor. *The Journal of biological chemistry*, **277**, 45059–45067. 42

Le Bizec, C., Vuillaumier-Barrot, S., Barnier, A., Dupre, T., Durand, G. & Seta, N. (2005). A new insight into pmm2 mutations in the french population. *Hum Mutat*, **25**, 504–5+. 168

Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, **55**, 379–400. 34

Lesk, A.M., Brändén, C.I. & Chothia, C. (1989). Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet. *Proteins*, **5**, 139–148. 121

Levy, E.D. (2007). Piqsi: protein quaternary structure investigation. *Structure (London, England : 1993)*, **15**, 1364–1367. 43

Levy, E.D., Pereira-Leal, J.B., Chothia, C. & Teichmann, S.A. (2006). 3d complex: A structural classification of protein complexes. *PLoS Computational Biology*, **2**, e155+. 42

Li, H. & Li, J. (2005). Discovery of stable and significant binding motif pairs from pdb complexes and protein interaction datasets. *Bioinformatics*, **21**, 314–324. 28

Li, H., Robertson, A.D. & Jensen, J.H. (2005). Very fast empirical prediction and rationalization of protein pk a values. *Proteins: Structure, Function, and Bioinformatics*, **61**, 704–721. 57

Li, J. & Liu, Q. (2009). 'double water exclusion': a hypothesis refining the o-ring theory for the hot spots at protein interfaces. *Bioinformatics*, **25**, 743–750. 181

Li, L., Zhao, B., Cui, Z., Gan, J., Sakharkar, M.K. & Kangueane, P. (2006). Identification of hot spot residues at protein-protein interface. *Bioinformation*, **1**, 121–126. 181, 182

Li, X., Keskin, O., Ma, B., Nussinov, R. & Liang, J. (2004). Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking. *J Mol Biol*, **344**, 781–795. 181

Li, Y., Urrutia, M., Smith-Gill, S.J. & Mariuzza, R.A. (2003). Dissection of binding interactions in the complex between the anti-lysozyme antibody hyhel-63 and its antigen. *Biochemistry*, **42**, 11–22. 38

Liang, S., Zhang, C., Liu, S. & Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. *Nucl. Acids Res.*, **34**, 3698–3707. 9

Lichtarge, O. & Sowa, M.E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol*, **12**, 21–7+. 163

Lindberg, M.J., Bystrom, R., Boknas, N., Andersen, P.M. & Oliveberg, M. (2005). Systematically perturbed folding patterns of amyotrophic lateral sclerosis (als)-associated sod1 mutants. *Proc Natl Acad Sci U S A*, **102**, 9754–9+. 167

Lipinski, C.A., Lombardo, F., Dominy, B.W. & Feeney, P.J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*, **46**, 3–26. 180

Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol*, **285**, 2177–2198. 17, 35

Lu, H., Lu, L. & Skolnick, J. (2003a). Development of unified statistical potentials describing protein-protein interactions. *Biophys J*, **84**, 1895–1901. 13

Lu, L., Lu, H. & Skolnick, J. (2002). Multiprospector: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, **49**, 350–364. 12

Lu, L., Arakaki, A.K., Lu, H. & Skolnick, J. (2003b). Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the saccharomyces cerevisiae proteome. *Genome Res*, **13**, 1146–1154. 13

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753. 6

Matthijs, G., Schollen, E., Bjursell, C., Erlandson, A., Freeze, H., Imtiaz, F., Kjaergaard, S., Martinsson, T., Schwartz, M., Seta, N., Vuillaumier-Barrot, S., Westphal, V. & Winchester, B. (2000). Mutations in pmm2 that cause congenital disorders of glycosylation, type ia (cdg-ia). *Hum Mutat*, **16**, 386–394+. 168

McDonald, I.K. & Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, **238**, 777–793. 51, 56

McMillan, K., Adler, M., Auld, D.S., Baldwin, J.J., Blasko, E., Browne, L.J., Chelsky, D., Davey, D., Dolle, R.E., Eagen, K.A., Erickson, S., Feldman, R.I., Glaser, C.B., Mallari, C., Morrissey, M.M., Ohlmeyer, M.H., Pan, G., Parkinson, J.F., Phillips, G.B., Polokoff, M.A., Sigal, N.H., Vergona, R., Whitlow, M., Young, T.A. & Devlin, J.J. (2000). Allosteric inhibitors of inducible nitric oxide synthase dimerization discovered via combinatorial chemistry. *Proc Natl Acad Sci U S A*, **97**, 1506–1511. 179

Mintseris, J. & Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins*, **53**, 629–639. 40

Mintseris, J. & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 10930–10935. 18, 80, 109

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J. & Weng, Z. (2005). Protein-protein docking benchmark 2.0: An update. *Proteins: Structure, Function, and Bioinformatics*, **60**, 214–216. 28

Mitchell, J. (1994). Amino/aromatic interactions in proteins: Is the evidence stacked against hydrogen bonding? *Journal of Molecular Biology*, **239**, 315–331. 56

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. & Overington, J.P. (1998a). Joy: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623. 51, 62, 122, 133, 163

Mizuguchi, K., Deane, C.M., Blundell, T.L. & Overington, J.P. (1998b). Homstrad: a database of protein structure alignments for homologous families. *Protein Sci*, **7**, 2469–2471. 122, 158

Montalvao, R.W., Smith, R.E., Lovell, S.C. & Blundell, T.L. (2005). Choral: a differential geometry approach to the prediction of the cores of protein structures. *Bioinformatics*, **21**, 3719–25+. 159

Moont, G., Gabb, H.A. & Sternberg, M.J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**, 364–373. 84, 111

Moreira, I.S., Fernandes, P.A. & Ramos, M.J. (2007a). Computational alanine scanning mutagenesis–an improved methodological approach. *J Comput Chem*, **28**, 644–654. 182

Moreira, I.S., Fernandes, P.A. & Ramos, M.J. (2007b). Hot spot occlusion from bulk water: a comprehensive study of the complex between the lysozyme hel and the antibody fvd1.3. *J Phys Chem B*, **111**, 2697–2706. 184

Moreira, I.S.S., Fernandes, P.A.A. & Ramos, M.J.J. (2007c). Hot spots-a review of the protein-protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812. 181

Murakami, Y. & Jones, S. (2006). Sharp2: protein-protein interaction predictions using patch analysis. *Bioinformatics*, **22**, 1794–1795. 8

Murphy, J., Gatchell, D.W., Prasad, J.C. & Vajda, S. (2003). Combination of scoring functions improves discrimination in protein-protein docking. *Proteins*, **53**, 840–854. 14

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., Masi, F.d., Gibson, T.J., Lewis, J., Serrano, L. & Russell, R.B. (2005). Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biology*, **3**, e405+. 7

Negi, S.S., Schein, C.H., Oezguen, N., Power, T.D. & Braun, W. (2007). Interprosurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics*, **23**, 3397–3399. 8

Neuvirth, H., Raz, R. & Schreiber, G. (2004). Promate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, **338**, 181–199. 9, 80

Ng, P.C. & Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–74+. 155

Ng, P.C. & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Research*, **12**, 436–446. 156

Niederhauser, O., Mangold, M., Schubenel, R., Kusznir, E.A., Schmidt, D. & Hertel, C. (2000). Ngf ligand alters ngf signaling via p75(ntr) and trka. *Journal of neuroscience research*, **61**, 263–272. 179

Nimrod, G., Glaser, F., Steinberg, D., Ben-Tal, N. & Pupko, T. (2005). In silico identification of functional regions in proteins. *Bioinformatics*, **21 Suppl 1**. 10

Nooren, I.M. & Thornton, J.M. (2003). Diversity of protein-protein interactions. *EMBO J*, **22**, 3486–3492. 18, 109

Ofran, Y. & Rost, B. (2003). Analysing six types of protein-protein interfaces. *J Mol Biol*, **325**, 377–387. 17, 35, 81, 103, 111

Ofran, Y. & Rost, B. (2007a). Isis: interaction sites identified from sequence. *Bioinformatics (Oxford, England)*, **23**, e13–16. 182

Ofran, Y. & Rost, B. (2007b). Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol*, **3**. 181, 182

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **27**, 29–34+. 176

Ogmen, U., Keskin, O., Aytuna, S.S., Nussinov, R. & Gursoy, A. (2005). Prism: protein interactions by structural matching. *Nucleic acids research*, **33**, W331–W336. 26

Ong, K.R., Woodward, E.R., Killick, P., Lim, C., Macdonald, F. & Maher, E.R. (2007). Genotype-phenotype correlations in von hippel-lindau disease. *Human mutation*, **28**, 143–149. 174

Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-Aryamontri, A., Salwinski, L., Nerothin, J. & Hermjakob, H. (2007). Submit your interaction data the imex way: a step by step guide to trouble-free deposition. *Proteomics*, **7 Suppl 1**, 28–34. 20

Overington, J., Johnson, M.S., Sali, A. & Blundell, T.L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.*, **241**, 132–45+. 158

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 Sci*, **1**, 216–226. 10, 29, 123, 158

Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. (2006). How many drug targets are there? *Nature Reviews Drug Discovery*, **5**, 993–996. 178

Pakula, A.A. & Sauer, R.T. (1989). Genetic analysis of protein stability and function. *Annu. Rev. Genet.*, **23**, 289–310+. 155

Pal, D. & Chakrabarti, P. (2001). Non-hydrogen bond interactions involving the methionine sulfur atom. *Journal of biomolecular structure & dynamics*, **19**, 115–128. 34

Palma, P.N., Krippahl, L., Wampler, J.E. & Moura, J.J. (2000). Bigger: a new (soft) docking algorithm for predicting protein interactions. *Proteins*, **39**, 372–384. 15

Palsdottir, A., Abrahamson, M., Thorsteinsson, L., Arnason, A., Olafsson, I., Grubb, A. & Jensson, O. (1988). Mutation in cystatin c gene causes hereditary brain haemorrhage. *Lancet*, **2**, 603–4+. 155

Papazoglou, M. (2007). *Web Services: Principles and Technology*. Prentice Hall, 1st edn. 206

Park, H. & Lee, S. (2005). Prediction of the mutation-induced change in thermodynamic stabilities of membrane proteins from free energy simulations. *Biophys. Chem.*, **114**, 191–7+. 157

Park, S.Y., Beel, B.D., Simon, M.I., Bilwes, A.M. & Crane, B.R. (2004). In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11646–11651. 11

Pazos, F. & Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, **14**, 609–614. 7

Pazos, F. & Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**, 219–227. 7

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 4285–4288. 6

Pineda, A.O., Cantwell, A.M., Bush, L.A., Rose, T. & Di Cera, E. (2002). The thrombin epitope recognizing thrombomodulin is a highly cooperative hot spot in exosite i. *The Journal of biological chemistry*, **277**, 32015–32019. 186

Pons, J., Rajpal, A. & Kirsch, J.F. (1999). Energetic analysis of an antigen/antibody interface: alanine scanning mutagenesis and double mutant cycles on the hyhel-10/lysozyme interaction. *Protein science : a publication of the Protein Society*, **8**, 958–968. 38

Ponstingl, H., Kabir, T., Thornton & J.M. (2003). Automatic inference of protein quaternary structure from crystals. *Journal of Applied Crystallography*, **36**, 1116–1122. 41

Ponstingl, H., Henrick, K. & Thornton, J.M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57. 40

Ponstingl, H., Kabir, T., Gorse, D. & Thornton, J.M. (2005). Morphological aspects of oligomeric protein structures. *Progress in biophysics and molecular biology*, **89**, 9–35. 81

Porollo, A. & Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. *Proteins: Structure, Function, and Bioinformatics*, **66**, 630–645. 8

Pugalenthi, G., Shameer, K., Srinivasan, N. & Sowdhamini, R. (2006). Harmony: a server for the assessment of protein structures. *Nucleic Acids Research*, **34**, W231–W234. 123

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002). Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18 Suppl 1**. 9

Qin, S. & Zhou, H.X. (2007). meta-ppisp: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, **23**, 3386–3387. 9

Radivojac, P., Obradovic, Z., Smith, D.K., Zhu, G., Vucetic, S., Brown, C.J., Lawson, J.D. & Dunker, A.K. (2004). Protein flexibility and intrinsic disorder. *Protein Sci*, **13**, 71–80. 205, 206

Rahat, O., Yitzhaky, A. & Schreiber, G. (2008). Cluster conservation as a novel tool for studying protein-protein interactions evolution. *Proteins: Structure, Function, and Bioinformatics*, **71**, 621–630. 198

Rajagopala, S.V., Goll, J., Gowda, N.D., Sunil, K.C., Titz, B., Mukherjee, A., Mary, S.S., Raviswaran, N., Poojari, C.S., Ramachandra, S., Shtivelband, S., Blazie, S.M., Hofmann, J. & Uetz, P. (2008). Mpi-lit: a literature-curated dataset of microbial binary protein–protein interactions. *Bioinformatics*, **24**, 2622–2627. 20

Ramensky, V., Bork, P. & Sunyaev, S. (2002). Human non-synonymous snps: server and survey. *Nucleic. Acids. Res.*, **30**, 3894–900+. 156

Randles, L.G., Lappalainen, I., Fowler, S.B., Moore, B., Hamill, S.J. & Clarke, J. (2006). Using model proteins to quantify the effects of pathogenic mutations in ig-like proteins. *J. Biol. Chem.*, **281**, 24216–26+. 167

Reid, K.S.C., Lindley, P.F. & Thornton, J.M. (1985). Sulphur-aromatic interactions in proteins. *FEBS Letters*, **190**, 209–213. 34

Reš, I. & Lichtarge, O. (2005). Character and evolution of proteinprotein interfaces. *Physical Biology*, **2**, S36–S43. 17, 18

Reynolds, C., Damerell, D. & Jones, S. (2008). Protorp: a protein-protein interaction analysis server. *Bioinformatics*, **25**, 413–414. 28

Richards, F.M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol*, **82**, 1–14. 35

Ringer, A.L., Senenko, A. & Sherrill, C.D. (2007). Models of s/pi interactions in protein structures: comparison of the h2s benzene complex with pdb data. *Protein science : a publication of the Protein Society*, **16**, 2216–2223. 34

Robert, C.H. & Janin, J. (1998). A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J Mol Biol*, **283**, 1037–1047. 52

Rocha, D., Gut, I., Jeffreys, A.J., Kwok, P.Y., Brookes, A.J. & Chanock, S.J. (2006). Seventh international meeting on single nucleotide polymorphism and complex genome analysis: 'ever bigger scans and an increasingly variable genome'. *Hum Genet*, **119**, 451–6+. 155, 162

Rodier, F., Bahadur, R.P., Chakrabarti, P. & Janin, J. (2005). Hydration of protein-protein interfaces. *Proteins*, **60**, 36–45. 40

Rohl, C., Price, Y., Fischer, T.B., Paczkowski, M., Zettel, M.F. & Tsai, J. (2006). Cataloging the relationships between proteins: a review of interaction databases. *Mol Biotechnol*, **34**, 69–93. 19

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, **12**, 85–94. 121

Ruiz, A., Puig, S., Malvehy, J., Lazaro, C., Lynch, M., Gimenez-Arnau, A.M., Puig, L., Sanchez-Conejo, J., Estivill, X. & Castel, T. (1999). Cdkn2a mutations in spanish cutaneous malignant melanoma families and patients with multiple melanomas and other neoplasia. *J Med Genet*, **36**, 490–3+. 168

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M. & Sali, A. (2004). A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, **14**, 313–324. 29

Saha, R.P., Bahadur, R.P., Pal, A., Mandal, S. & Chakrabarti, P. (2006). Proface: a server for the analysis of the physicochemical features of protein-protein interfaces. *BMC Structural Biology*, **6**, 11+. 28

Sali, A. & Blundell, T.L. (1990). Definition of general topological equivalence in protein structures. a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, **212**, 403–428. 122, 132

Sali, A. & Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, **234**, 779–815. 11, 12, 126, 127, 159

Salwinski, L. (2004). The database of interacting proteins: 2004 update. *Nucleic Acids Research*, **32**, 449D–451. 19

Sanfilippo, P.J., Jetter, M.C., Cordova, R., Noe, R.A., Chourmouzis, E., Lau, C.Y. & Wang, E. (1995). Novel thiazole based heterocycles as inhibitors of lfa-1/icam-1 mediated cell adhesion. *Journal of medicinal chemistry*, **38**, 1057–1059. 179

Saraboji, K., Gromiha, M.M. & Ponnuswamy, M.N. (2006). Average assignment method for predicting the stability of protein mutants. *Biopolymers*, **82**, 80–92+. 157

Schierz, A.C., Soldatova, L.N. & King, R.D. (2007). Overhauling the pdb. *Nature biotechnology*, **25**, 437–442. 50

Schreiber, G. & Fersht, A.R. (1995). Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *Journal of molecular biology*, **248**, 478–486. 37, 38

Schreyer, A. & Blundell, T. (2009). Credo: A protein–ligand interaction database for drug discovery. *Chemical Biology & Drug Design*, **73**, 157–167. 164

Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. & Baker, D. (2005). Progress in modeling of protein structures and interactions. *Science*, **310**, 638–642. 15

Schultz, J., Milpetz, F., Bork, P. & Ponting, C.P. (1998). Smart, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, **95**, 5857–5864. 122

Schulz, G.E. & Schirmer, R.H. (1996). *Principles of Protein Structure (Springer Advanced Texts in Chemistry)*. Springer. 33

Sculley, D.G., Dawson, P.A., Emmerson, B.T. & Gordon, R.B. (1992). A review of the molecular basis of hypoxanthine-guanine phosphoribosyltransferase (hprt) deficiency. *Hum Genet*, **90**, 195–207+. 168

Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature*, **379**, 96–8+. 158

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. & Sirotkin, K. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, **29**, 308–11+. 156

Shi, J. (2001). *Towards fully automated structure prediction: homology recognition and alignment validation.*. Ph.D. thesis, University of Cambridge, Department of Biochemistry. 159

Shi, J., Blundell, T.L. & Mizuguchi, K. (2001). Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**, 243–257. 122, 123, 125, 126, 158, 159

Shoemaker, B.A. & Panchenko, A.R. (2007). Deciphering proteinprotein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS computational biology*, **3**, e43+. 6

Sippl, M.J. (1990). Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, **213**, 859–883. 126

Smith, G.R., Sternberg, M. & Bates, P.A. (2005). The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *Journal of Molecular Biology*, **347**, 1077–1101. 15, 206

Smith, R.E., Lovell, S.C., Burke, D.F., Montalvao, R.W. & Blundell, T.L. (2007). Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities. *Bioinformatics (Oxford, England)*, **23**, 1099–1105. 159

Smith, T.F. & Waterman, M.S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147**, 195–197. 132

Sprinzak, E. & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, **311**, 681–692. 7

Stebbings, L.A. (2004). Homstrad: recent developments of the homologous protein structure alignment database. *Nucleic Acids Research*, **32**, 203D–207. 10

Stein, A. (2004). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Research*, **33**, D413–D417. 25

Stone, E.A. & Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res*, **15**, 978–86+. 156

Stumpf, M.P., Thorne, T., de Silva, E., Stewart, R., An, H.J.J., Lappe, M. & Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6959–6964. 179

Sundquist, A., Ronaghi, M., Tang, H., Pevzner, P. & Batzoglou, S. (2007). Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE*, **2**, e484+. 155

Sunyaev, S., Ramensky, V., Koch, I., Lathe, Kondrashov, A.S. & Bork, P. (2001). Prediction of deleterious human alleles. *Hum. Mol. Genet.*, **10**, 591–7+. 156

Sutcliffe, M.J., Hayes, F.R. & Blundell, T.L. (1987). Knowledge based modelling of homologous proteins, part ii: Rules for the conformations of substituted sidechains. *Protein Eng*, **1**, 385–92+. 159

Takahashi, H., Furusato, M., Allsbrook, W.C., Nishii, H., Wakui, S., Barrett, J.C. & Boyd, J. (1995). Prevalence of androgen receptor gene mutations in latent prostatic carcinomas from japanese men. *Cancer Res*, **55**, 1621–4+. 169

Teyra, J., Doms, A., Schroeder, M. & Pisabarro, M.T. (2006). Scowlp: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, **7**, 104+. 27

Teyra, J., Paszkowski-Rogacz, M., Anders, G. & Pisabarro, M.T. (2008). Scowlp classification: Structural comparison and analysis of protein binding regions. *BMC Bioinformatics*, **9**, 9+. 27

Thanos, C.D., Randal, M. & Wells, J.A. (2003). Potent small-molecule binding to a dynamic hot spot on il-2. *Journal of the American Chemical Society*, **125**, 15280–15281. 179

Thorn, K.S. & Bogan, A.A. (2001). Asedb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285. 181, 186

Thornton, J.M. (1981). Disulphide bridges in globular proteins. *Journal of molecular biology*, **151**, 261–287. 34, 58

Tina, K.G., Bhadra, R. & Srinivasan, N. (2007). Pic: Protein interactions calculator. *Nucleic Acids Research*, **35**, W473–W476. 28, 57

Tjong, H., Qin, S. & Zhou, H. (2007). Pi2pe: protein interface/interior prediction engine. *Nucleic Acids Research*, **35**, W357–W362. 8

Todd, A.E., Marsden, R.L., Thornton, J.M. & Orengo, C.A. (2005). Progress of structural genomics initiatives: an analysis of solved target structures. *J Mol Biol*, **348**, 1235–1260. 3

Tong, W., Li, L. & Weng, Z. (2004). Computational prediction of binding hotspots. *Conf Proc IEEE Eng Med Biol Soc*, **4**, 2980–2983. 182

Topham, C.M., Mcleod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S. & Blundell, T.L. (1993). Fragment ranking in modelling of protein structure : Conformationally constrained environmental amino acid substitution tables. *Journal of Molecular Biology*, **229**, 194–220. 126

Topham, C.M., Srinivasan, N. & Blundell, T.L. (1997). Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng*, **10**, 7–21. 123, 125, 157, 161, 200

Tress, M., de Juan, D., Graña, O., Gómez, M.J., Gómez-Puertas, P., González, J.M., López, G. & Valencia, A. (2005). Scoring docking models with evolutionary information. *Proteins*, **60**, 275–280. 15

Tsai, C.J., Lin, S.L., Wolfson, H.J. & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*, **6**, 53–64. 32

Tsai, J. & Gerstein, M. (2002). Calculations of protein volumes: sensitivity analysis and parameter database. *Bioinformatics*, **18**, 985–995. 35, 54

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes1. *Journal of Molecular Biology*, **290**, 253–266. 35, 54

Tsao, D.H., Sutherland, A.G., Jennings, L., Li, Y., Rush, Alvarez, J.C., Ding, W., Dushin, E.G., Dushin, R.G., Haney, S.A., Kenny, C.H., Karl, Nilakantan, R. & Mosyak, L. (2006). Discovery of novel inhibitors of the zipa/ftsz complex by nmr fragment screening coupled with structure-based design. *Bioorganic & Medicinal Chemistry*, **14**, 7953–7961. 179

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. 5

Uniprot-Consortium (2009). The universal protein resource (uniprot) 2009. *Nucl. Acids Res.*, **37**, D169–174. 40, 46, 163

Vajda, S. & Camacho, C.J. (2004). Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol*, **22**, 110–116. 14, 17

Vakser, I.A. & Kundrotas, P. (2008). Predicting 3d structures of protein-protein complexes. *Current pharmaceutical biotechnology*, **9**, 57–66. 14

Valdar, W.S. (2002). Scoring residue conservation. *Proteins*, **48**, 227–241. 83

Valdar, W.S. & Thornton, J.M. (2001). Conservation helps to identify biologically relevant crystal contacts. *Journal of molecular biology*, **313**, 399–416. 40, 109

Valencia, A. & Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, **12**, 368–373. 6

van Dijk, A.D., Boelens, R. & Bonvin, A.M. (2005a). Data-driven docking for the study of biomolecular complexes. *FEBS J*, **272**, 293–312. 15

van Dijk, A.D., de Vries, S.J., Dominguez, C., Chen, H., Zhou, H.X. & Bonvin, A.M. (2005b). Data-driven docking: Haddock's adventures in capri. *Proteins*, **60**, 232–238. 15

Van Petegem, F., Duderstadt, K.E., Clark, K.A., Wang, M. & Minor, D.L. (2008). Alanine-scanning mutagenesis defines a conserved energetic hotspot in the cavalpha1 aid-cavbeta interaction site that is critical for channel modulation. *Structure (London, England : 1993)*, **16**, 280–294. 187

Van Rossum, G. (2003). *The Python Language Reference Manual*. Network Theory Ltd. 62

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E. & Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biol*, **8**, R39+. 176

Velankar, S., McNeil, P., Mittard-Runte, V., Suarez, A., Barrell, D., Apweiler, R. & Henrick, K. (2005). E-msd: an integrated data resource for bioinformatics. *Nucleic acids research*, **33**, D262+. 47, 159, 163

Velazquez-Muriel & Carazo, J.M. (2009). Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Structural Biology*, **9**, 6+. 206

Venter, J.C. (2001). The sequence of the human genome. *Science*, **291**, 1304–1351. 2

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403. 5

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B. & Bork, P. (2007). String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, **35**. 19

Wako, H. & Blundell, T.L. (1994a). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. ii. secondary structures. *J Mol Biol*, **238**, 693–708. 123, 158

Wako, H. & Blundell, T.L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. i. solvent accessibility classes. *J Mol Biol*, **238**, 682–92+. 158

Wang, K. & Samudrala, R. (2006). Incorporating background frequency improves entropy-based residue conservation measures. *BMC Bioinformatics*, **7**, 385+. 83

Wang, W.Y., Barratt, B.J., Clayton, D.G. & Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–18+. 155

Wang, Z. & Moult, J. (2001). Snps, protein structure, and disease. *Hum Mutat*, **17**, 263–270. 156, 161

Welch, B.L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, **34**, 28–35. 107

Wells, J.A. (1996). Binding in the growth hormone receptor complex. *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 1–6. 180

Wells, J.A. & Mcclendon, C.L. (2007). Reaching for high-hanging fruit in drug discovery at proteinprotein interfaces. *Nature*, **450**, 1001–1009. 21, 179, 196

Widenius, M., Axmark, D. & Mysql, A.B. (2002). *MySQL Reference Manual*. O'Reilly Media, Inc., 1st edn. 43

Winter, C. (2006). Scoppi: a structural classification of protein-protein interfaces. *Nucleic Acids Research*, **34**, D310–D314. 27

Worth, C.L. & Blundell, T.L. (2009). Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. *Proteins: Structure, Function, and Bioinformatics*, **75**, 413–429. 111, 123, 147

Worth, C.L., Bickerton, G.R., Schreyer, A., Forman, J.R., Cheng, T.M., Lee, S., Gong, S., Burke, D.F. & Blundell, T.L. (2007a). A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nssnps) and their relation to disease. *Journal of bioinformatics and computational biology*, **5**, 1297–1318. xxi, 121, 154, 171

Worth, C.L., Burke, D.F. & Blundell, T.L. (2007b). Estimating the effects of single nucleotide polymorphisms on protein structure: how good are we at identifying likely disease associated mutations? *Proceedings of Molecular Interactions*, 11–26+. 158, 161, 200

Xu, Q., Canutescu, A., Obradovic, Z. & Dunbrack, R.L. (2006). Protbud: a database of biological unit structures of protein families and superfamilies. *Bioinformatics*, **22**, 2876–2882. 26, 42

Yan, C., Wu, F., Jernigan, R.L., Dobbs, D. & Honavar, V. (2008). Characterization of protein-protein interfaces. *The protein journal*, **27**, 59–70. 17, 81, 111, 117

Yao, H., Kristensen, D.M., Mihalek, I., Sowa, M.E., Shaw, C., Kimmel, M., Kavraki, L. & Lichtarge, O. (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of molecular biology*, **326**, 255–261. 9

Yoakim, C., Ogilvie, W.W., Goudreau, N., Naud, J., Haché, B., O'Meara, J.A., Cordingley, M.G., Archambault, J. & White, P.W. (2003). Discovery of the first series of inhibitors of human papillomavirus type 11: inhibition of the assembly of the e1-e2-origin dna complex. *Bioorganic & medicinal chemistry letters*, **13**, 2539–2541. 179

Yu, H., Luscombe, N.M., Lu, H.X.X., Zhu, X., Xia, Y., Han, J.D.D., Bertin, N., Chung, S., Vidal, M. & Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. *Genome research*, **14**, 1107–1118. 11

Zauhar, R.J., Colbert, C.L., Morgan, R.S. & Welsh, W.J. (2000). Evidence for a strong sulfur-aromatic interaction derived from crystallographic data. *Biopolymers*, **53**, 233–248. 34, 58

Zhou, H. & Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **23**, 2203–2209. 9

Zhou, H. & Zhou, Y. (2004). Quantifying the effect of burial of amino acid residues on protein stability. *Proteins*, **54**, 315–22+. 158

Zhu, H., Domingues, F., Sommer, I. & Lengauer, T. (2006). Noxclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, **7**, 27+. 79, 80

Zhu, Z.Y., Sali, A. & Blundell, T.L. (1992). A variable gap penalty function and feature weights for protein 3-d structure comparisons. *Protein Eng*, **5**, 43–51. 132

Zollner, F., Neumann, S., Kummert, F. & Sagerer, G. (2005). Database driven test case generation for protein-protein docking. *Bioinformatics*, **21**, 683–684. 28