Characterization, Classification and Alignment of Protein-Protein Interfaces

Zhu, Hongbo

A Dissertation
Submitted to the Faculty of

Natural Sciences and Technology I
Mathematics and Computer Science
Saarland University

In Partial Fulfillment of the
Requirements for the Degree of

Doktor der Naturwissenschaften (Dr. rer. nat.)

Saarbrücken
2010

Date of Colloquium:	24-06-2010
Dean of the Faculty:	Prof. Dr. Holger Hermanns
Chairman:	Prof. Dr. Hans-Peter Lenhof
Doctorate Committee Members:	Prof. Dr. Dr. Thomas Lengauer
	Prof. Dr. Peter Lackner

To my parents.

Abstract

Protein structural models provide essential information for the research on protein-protein interactions. In this dissertation, we describe two projects on the analysis of protein interactions using structural information. The focus of the first is to characterize and classify different types of interactions. We discriminate between biological obligate and biological non-obligate interactions, and crystal packing contacts. To this end, we defined six interface properties and used them to compare the three types of interactions in a hand-curated dataset. Based on the analysis, a classifier, named *NOXclass*, was constructed using a support vector machine algorithm in order to generate predictions of interaction types. NOXclass was tested on a non-redundant dataset of 243 protein-protein interactions and reaches an accuracy of 91.8%. The program is beneficial for structural biologists for the interpretation of protein quaternary structures and to form hypotheses about the nature of protein-protein interactions when experimental data are yet unavailable.

In the second part of the dissertation, we present *Galinter*, a novel program for the geometrical comparison of protein-protein interfaces. The Galinter program aims at identifying similar patterns of different non-covalent interactions at interfaces. It is a graph-based approach optimized for aligning non-covalent interactions. A scoring scheme was developed for estimating the statistical significance of the alignments. We tested the Galinter method on a published dataset of interfaces. Galinter alignments agree with those delivered by methods based on interface residue comparison and backbone structure comparison. In addition, we applied Galinter on four medically relevant examples of protein mimicry. Our results are consistent with previous human-curated analysis. The Galinter program provides an intuitive method of comparative analysis and visualization of binding modes and may assist in the prediction of interaction partners, and the design and engineering of protein interactions and interaction inhibitors.

Acknowledgements

The work presented in the dissertation had been carried out in the group of Computational Biology and Applied Algorithmics at the Max-Planck-Institute for Informatics in Saarbrücken. First, I would like to express my gratitude to the group leader and my supervisor Professor Dr. Dr. Thomas Lengauer for his guidance during my research studies. He had provided me with generous support, advice and encouragement, which are indispensable for me to finish my work.

The research has been performed under the guidance of Francisco Domingues and Ingolf Sommer. I would like to thank them for their invaluable advice to my work. The discussions with them were always inspiring and fruitful, the time spent with them was indeed great and unforgettable.

The work had benefited from helpful discussions with many colleagues. Thanks to Oliver Sander, Gabriele Mayr, Jörg Rahnenführer, Andreas Steffen, and Tobias Sing for their expertise and the discussions with them on work.

I want to thank all group members for the unique stimulating and friendly working atmosphere. The discussions and conversations with them make my stay in the group a colorful memory.

I want to thank Dr. Joachim Büch and Ruth Schneppen-Christmann for their support and help with many technical, administrative and other problems.

Finally, I would like to thank my family and friends for their constant support, especially my wife Dongmei and my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Our knowledge about macromolecules in the cell is growing extremely fast, as genome sequencing has provided nearly complete lists of the macromolecules that are present in many organisms (Lander *et al.*, 2001; Venter *et al.*, 2001). However, little information about the function of the biological systems is revealed by these lists, because the functional units of the cell are usually complex assemblies of macromolecules. Proteins are the most versatile macromolecules found in all living systems. Most of the functionalities of proteins are realized via interactions with other proteins, DNA, RNA, or ligands, which play crucial roles in almost all biological processes in the cell. On average, a protein is estimated to have five different interaction partners (Piehler, 2005).

Protein complexes are often described as "molecular machines" because of their resemblance to real-world machines in terms of their characteristic features, such as modularity, complexity, cyclic function and energy consumption (Alberts, 1998; Nogales and Grigorieff, 2001). These macromolecule assemblies are of widely variable sizes and activities. For example, the enzyme *helicase*, which initiates the DNA replication process by unwinding the DNA double helix (Fass *et al.*, 1999), is a ring-shaped protein complex composed of six identical subunits. In eukaryotes, the transcription process is activated by a protein assembly *preinitiation complex*, formed from several heterogeneous proteins including TFIID, a RNA polymerase II, and several associated transcription factors (Lee and Young, 2000; Dvir *et al.*, 2001; Reese, 2003). Proteins can also participate in the formation of larger assemblies like ribosomes. *Ribosomes*, the "workhorses" for protein biosynthesis, are about 200 Å in diameter and are composed of a few different ribosomal RNAs (rRNAs) and more than 50 proteins (Lodish *et al.*, 1999). The *nuclear pore complexes* (NPCs) are 50–100 MDa protein assemblies responsible for the transport and regulation of the bidirectional trafficking of macromolecules through the nuclear envelope (Davis, 1995; Rout *et al.*, 2000). Macromolecular complexes participate in many immune response. For example, the *T-cell receptors* recognize and bind to short fragments of antigens that are complexed with MHC molecules on the surface of other cells (Garcia and Teyton, 1998). Macromolecular complexes are also involved in cell signaling, as illustrated by

the G proteins. *G proteins* (guanine nucleotide-binding proteins), which are made up of α, β, and γ subunits, are important signal transduction molecules involved in second messenger cascades (Gilman, 1987; Neves *et al.*, 2002). A structural description of protein interactions within such macromolecule assemblies is a fundamental step toward the elucidation of the interaction mechanism and the understanding of biochemical, cellular and higher order biological processes (Edwards *et al.*, 2002; Sali *et al.*, 2003; Russell *et al.*, 2004).

In this chapter, we introduce relevant background knowledge about protein-protein interactions and protein complexes, as well as methods for the detection and characterization of protein interactions. In Section 1.1, we first introduce proteins and protein structures. Then we review important experimental and computational methods for the detection of protein-protein interactions. After that, we address the classification of protein-protein interactions based on different criteria. In Section 1.2, we focus on the stucture models of protein complexes. We first summarize methods for the determination of protein structure models. We also address the interpretation of structure models. Following this, we discuss the identification of interfaces in protein complex structures. In the end, we describe the outline of the dissertation in Section 1.3.

1.1 Protein-Protein Interactions

A large portion of the research about protein complexes is devoted to the detection and characterization of protein-protein interactions, which play central roles in the elucidation of protein functions. In this section, we first introduce the nature of protein structures. We also describe experimental methods for the detection and characterization of protein-protein interactions. Finally, we discuss the classification of protein-protein interactions.

1.1.1 Introduction to Proteins

A protein is an organic macromolecule made of amino acids, which are linked together as a polypeptide chain by peptide bonds between the carboxyl group and the amino group of adjacent amino acid residues. The protein sequence is defined by the DNA sequence of a gene. The triplets of nucleotides in the gene code for the 20 essential amino acids, which exhibit diverse physicochemical properties. The physicochemical characters of the amino acids are determined by their side chains. Depending on the polarity of the side chain, the 20 amino acids may be classified as hydrophobic or hydrophilic. The side chains of a few amino acids are positively or negatively charged. The side chains of amino acids exhibit different tendencies to interact with each other and with water.

The amino acid sequences of proteins are termed the *primary structures* of proteins. At physiological temperature in aqueous solution and neutral pH, the polypeptide chains of proteins usually fold into special and characteristic three-dimensional

(3D) conformations. Such 3D structures of protein folds are referred to as the *tertiary structures* of proteins. Very frequently, different regions of a protein sequence form local *secondary structures* through regular hydrogen-bonding interactions, such as α-helices, β-strands, or coils. Many proteins contain more than one polypeptide chain. In such proteins, the tertiary structures of their polypeptide chains associate by non-covalent interactions into *quaternary structures* (IUPAC, 2005).

According to the number of polypeptide chains contained in a protein molecule, proteins can be classified as *monomeric* if they contain only one chain, and *multimeric* or *oligomeric* if they consist of multiple chains. The individual chains are also termed the *subunits* of the assembly. Multimeric protein complexes may contain two, three, four, five, six or more subunits, known as *dimers*, *trimers*, *tetramers*, *pentamers*, *hexamers*, and so on. For example, a viral capsid, the protein shell of a virus, may be composed of 60, 120, or even 240 subunits (Branden and Tooze, 1999).

In this context, we use the term "protein structure" to refer to both the tertiary structures of monomeric proteins and the quaternary structures of multimeric proteins, unless otherwise specified. It is widely accepted that it is the protein structure that dictates the protein biological function (Petsko and Ringe, 2003).

1.1.2 Identification of Protein-Protein Interactions

The identification of protein-protein interactions is a key topic in the research of life science due to the vital importance of protein interactions in life. Elucidating both associations of individual proteins and networks of protein interactions is a fundamental goal of functional genomics. It is important for understanding the molecular basis of diseases, for the discovery of new therapies, for molecular engineering and biotechnology. For example, identifying interacting partners for a certain protein of unknown function can provide valuable insights into the actual function of the protein and provide a platform for further research.

A variety of experimental methodologies have been established in recent years for identifying protein-protein interactions. The IntAct database[1], an open source database of molecular interactions (Kerrien *et al.*, 2007), lists more than 170 experimental methods for the detection of molecular interactions (Panchenko and Przytycka, 2008). These methods are divided into four groups. See Table 1.1 for the classification with some important example methods in each group.

Alternatively, these detection approaches may be divided into high-throughput methods and low-throughput methods. Some of these experimental technologies focused on the high-throughput study of protein-protein interactions and generated a vast amount of interaction data, including yeast two-hybrid (Y2H) system, tandem affinity purification (TAP), and protein complex purification combined with mass spectrometry. The contributions of Y2H and TAP methods to the total number of interactions in the IntAct database are 45.9% (89575/195058) and

[1] `http://www.ebi.ac.uk/intact/`

15.2% (29663/195058), respectively[2]. Such methods are often exploited for screening protein-protein interactions at a large scale. Most other methods are mainly utilized at a small scale for detecting and characterizing specific interactions between proteins.

These techniques may be distinguished according to whether they are established for monitoring protein-protein interactions *in vivo* or *in vitro*. Representative methods for detecting interactions *in vivo* include two hybrid techniques, fluorescent resonance energy transfer, and protein complementation techniques (Morell *et al.*, 2009). Experimental techniques that are suitable for screening interactions *in vitro* include tandem affinity purification, co-immunoprecipitation, and protein array (Piehler, 2005).

In the following paragraphs, we review several important experimental methods for identifying interactions between proteins. These methods are categorized into four groups: protein complementation assays, biophysical approaches, biochemical approaches, and imaging techniques (see Table 1.1).

Protein Complementation Assays

In protein complementation assay (PCA), a bait protein and a prey protein are fused with two complementary fragments of a reporter protein or a fluorescent protein. Upon the association between the bait and prey proteins, the function of the reporter protein or the fluorescent protein is restored by the complementation between the two fragments.

Ubiquitin was first implemented as the proximity reporter of interactions (Johnsson and Varshavsky, 1994). It is split into two inactive fragments N_{ub} and C_{ub}. The C-terminal fragment of ubiquitin C_{ub} is expressed as a fusion to a reporter protein. The association of N_{ub} and C_{ub} leads to the cleavage of the reporter protein by ubiquitin-dependent proteases. Alternatively, a PCA system based on green fluorescent proteins (GFPs) uses two fragments of a GFP (Remy and Michnick, 2004). Fluorescence is detected when the two GFP fragments assemble with each other upon the interaction between the bait and prey proteins. The latter system is a highly sensitive method as the background of fluorescence is nearly zero in the case of non-binding. In comparison to fluorescence resonance energy transfer (see the following section on biophysical approaches), this approach has a higher dynamic range of the assay. Furthermore, PCA may be employed to identify protein-protein interactions in different compartments of the cell (Remy *et al.*, 1999), or to monitor interactions between membrane proteins (Blakely *et al.*, 2000).

- **Yeast two-hybrid system.** The Y2H system was developed by Fields and Song (1989). The system is based on the modular structures of transcription factors containing a *DNA-binding domain* (DBD) and a *transcription activation domain* (AD). The transcription factor is inactivated if the two domains

[2]Data collected from http://www.ebi.ac.uk/intact/ as of 18 Nov, 2009

Table 1.1: Approaches for identifying protein-protein interactions[a]. Methods in *italic* are addressed in the text. The number of interactions detected by each class/method is listed in parentheses.

- **Protein complementation assay** (51.5%, 100360/195058)
  - transcriptional complementation assay (89627)
    - *two hybrid* (89575)
  - cytoplasmic complementation assay (10506)
    - dihydrofolate reductase reconstruction (10197)
    - ubiquitin reconstruction (229)
    - green fluorescence protein complementation assay (10)
  - bimolecular fluorescence complementation (157)
- **Biochemical** (46.4%, 90478/195058)
  - affinity technology (85343)
    - affinity chromatography technology (80763)
      - *tandem affinity purification* (29663)
      - *co-immunoprecipitation* (26375)
      - pull down (22740)
    - array technology (2956)
      - *protein array* (1493)
      - peptide array (1374)
- **Biophysical** (1.5%, 2899/195058)
  - x-ray crystallography (1160)
  - fluorescence technology (550)
    - bimolecular fluorescence complementation (157)
    - *fluorescent resonance energy transfer (FRET)* (134)
  - nuclear magnetic resonance (65)
  - detection by mass spectrometry (6)
- **Imaging technique** (0.6%, 1258/195058)
  - fluorescence microscopy (600)
  - confocal microscopy (446)
  - electron microscopy (121)

[a]The classification and examples of methods are taken from the IntAct database at `http://www.ebi.ac.uk/intact/`. The number of interactions detected by each class/method is collected from `http://www.ebi.ac.uk/intact/main.xhtml` as of 18 Nov, 2009

are separated. But when the DBD and AD domains are fused with two interacting proteins, which are named bait and prey respectively, the function of the transcription factor can be restored though the two domains are indirectly linked. The binding of the prey to the bait protein subsequently activate the transcription of a reporter gene and the interaction between the proteins is inferred from the gene product.

Many improvements and variations of the method have been reported as reviewed in Toby and Golemis (2001). The most powerful application of the Y2H system is the screening of protein-protein interaction on a genome-wide scale, e.g., for *S. cerevisiae* (Schwikowski *et al.*, 2000) and *C. elegans* (Li *et al.*, 2004).

Biochemical Approaches

While protein complementation techniques are mostly used for detecting interactions *in vivo*, biochemical approaches are often applied for *in vitro* analysis of protein-protein interactions. Widely employed biochemical approaches include tandem affinity purification, co-immunoprecipitation, and protein array (Miernyk and Thelen, 2008).

- **Tandem affinity purification.** The TAP method was introduced by Rigaut *et al.* (1999). The protein of interest (the *target protein*, or the *bait*) is fused with a *TAP tag*. The TAP tag consists of a *calmodulin binding peptide* (CBP) and a IgG binding unit *protein A*, linked together by a specific protease cleavage sequence (*TEV protease cleavage site*). The complex of the target protein and its interacting partners are recovered by two consecutive steps of affinity purifications. First the complexes are captured by affinity selection on an IgG matrix. After washing, the TEV protease is added to release the complexes. Then, a second affinity purification step is performed to remove the TEV protease and other remaining contaminants.

  An advantage of the TAP method is that prior knowledge of protein complex composition or function is not required. In addition, the TAP method has the feature of automation (Puig *et al.*, 2001). Thus it may be used for the large-scale exploration of proteome. However, using the TAP method transient interactions usually cannot be captured (Piehler, 2005).

- **Co-immunoprecipitation.** Co-immunoprecipitation is a technique to precipitate a target protein (antigen) out of a solution of protein mixture by using an antibody that binds to the antigen specifically. The target protein might interact with one or more other proteins in a large protein complex. After binding to the antibody, the entire protein complex may be isolated or pulled down from the solution using immobilized antibody binding proteins such as protein A or protein G. Co-immunoprecipitation experiments are often repeated by targeting different members of the protein complex to verify the results.

The co-immunoprecipitation is usually considered the gold standard approach for the detection of protein-protein interactions and provides the most convincing evidence that two or more proteins physically interact (Monti *et al.*, 2005; Miernyk and Thelen, 2008). It is often used to confirm Y2H interactions. A major limitation of the co-immunoprecipitation approach is the requirement of specific antibodies. Furthermore, co-immunoprecipitation works only when subunit proteins in the complex bind to each other tightly.

- **Protein array.** The protein array is conceptually comparable to the DNA array. In protein arrays, different proteins are immobilized onto solid supports and the interaction partners to each of these proteins are detected in parallel. Usually the read-out of protein arrays is performed indirectly using fluorescent or chemiluminescent probes. Recently, direct read-out methods based upon mass spectrometry or solid phase detection have been developed (Espina *et al.*, 2004). By combining with direct read-out strategies like surface plasmon resonance (SPR), protein array has the potential to integrate both the identification of interaction partners, and the characterization of interactions (e.g., interaction equilibrium and kinetics) (Karlsson, 2004).

Biophysical Approaches

A multitude of physical/biophysical techniques have been used for identifying protein-protein interactions, including nuclear magnetic resonance, mass spectrometry, fluorescence technology. Fluorescence resonance energy transfer (FRET) is one of common methods for detecting binary interaction between proteins. As we introduced in the previous sections, Y2H and TAP methods provide information of protein-protein interactions that might occur. However, it is definitely much more useful to precisely detect the time and the location of a protein-protein interaction in the cell. This is of particular importance for interactions involved in the highly dynamic and dramatically variable signaling processes. Using approaches based on FRET, protein-protein interactions can be visualized directly in living cells.

- **Fluorescence resonance energy transfer.** The principle of FRET is similar to that of the protein complementation assay. FRET is an energy transition between a donor and an acceptor fluorophore that locates within 10 nm distance. Hence, FRET can be used as a probe to discover a proximity between proteins upon interaction. The detection of FRET is achieved by monitoring the change in emission intensity of the donor and the acceptor, or a change in the fluorescence life-time (Stryer, 1978; Yan and Marriott, 2003). The identification of protein-protein interactions using FRET can be done in a high-throughput manner in living cells conveniently. The application of FRET has been greatly boosted with the progress in fluorescence microscopy techniques (Kenworthy, 2001).

Imaging Techniques

Advances in imaging techniques such as fluorescence microscopy and electron microscopy have provided tools for analyzing dynamics of molecular interactions in real time. Fluorescence microscopy is based on the phenomena of fluorescence or phosphorescence. In fluorescence microscopy, the protein under study is labeled with a fluorescent molecule (*fluorophore*) such as GFP. The protein is then illuminated with light of a specific wavelength, which is absorbed by the fluorophores and causes them to emit light of a different wavelength. The emitted fluorescence is detected by a fluorescence microscope. The introduction of GFP has allowed systematic imaging studies of the structures and functions of proteins in living cells using fluorescence microscopy (Yuste, 2005). There are various applications of the fluorescence microscopy technique. The intensity and fluorescence lifetime (FLIM) variants of FRET have been used in the study of protein-protein interactions, and the scanning cysteine accessibility method (SCAM) have been employed for the determination of protein conformation (Jensen-Smith *et al.*, 2009). An introduction to electron microscopy can be found in Section 1.2.1.

1.1.3 Inference of Protein-Protein Interactions

In parallel to the rapid development and massive application of experimental techniques for the determination of protein-protein interaction, computational approaches have also been proposed for the prediction of interactions between proteins. At present, experimental techniques are still expensive, time-consuming and labor-intensive. Therefore, computational methods are attractive complements to experimental techniques for finding interacting proteins. Computational methods utilize the structural, genomic, and biological context of genes and proteins to predict functional linkages or physical interactions between proteins (Shoemaker and Panchenko, 2007; Skrabanek *et al.*, 2008).

It is worthwhile to mention that the term "protein-protein interactions" has been used in a very broad sense in the context of interaction prediction. It refers to both "functional associations" and "physical interactions". Many computational approaches have been proposed for inferring "functional interactions" between proteins (Huynen *et al.*, 2000). Such proteins might be involved in the same biological processes and have related functions, but do not necessarily interact physically. Meanwhile, there are also computational approaches that have been developed to predict direct interactions between proteins, or to predict interaction partners. In addition, multiple methods have been invented for the prediction of binding sites.

Prediction of Functional Associations

Essentially, most of the methods in this group exploit evolutionary information derived from genomic context of genes for the prediction of related functions between proteins (Valencia and Pazos, 2002). There are mainly three types of genomic con-

texts used in these methods: the co-occurrence of genes across genomes, the conservation of gene order, and the fusion of genes.

- **Phylogenetic profile.** A phylogenetic profile is the pattern of the presence or absence of a particular gene in a set of genomes. The similarity of phylogenetic profiles of genes may indicate that the corresponding proteins are desired to be present simultaneously in order to carry out their functions. The application of the phylogenetic profile method is mainly restricted by the availability of genomic data, because complete genomes are required to derive a phylogenetic profile. This method is not applicable to the essential proteins common to most organisms either. In addition, similar phylogenetic profiles do not necessarily imply direct physical interaction between the corresponding proteins (Pellegrini *et al.*, 1999).

- **Conservation of gene neighborhood.** It has been observed that the coding genes for functionally related proteins are often organized into neighboring regions in bacterial genomes (Tamames *et al.*, 1998). The conservation of adjacency for genes in multiple bacterial genomes is a strong signal for the related functions between corresponding proteins. Thus, the conservation of gene neighborhood can be used for inferring physical interactions or similar functions between proteins in bacteria (Dandekar *et al.*, 1998; Overbeek *et al.*, 1999a,b).

- **Gene fusion.** Gene fusion refers to the event that two separate genes in a genome are joined to form a single hybrid gene in another genome. The corresponding protein domains coded by the separate genes in one organism are merged into a multi-domain protein in another organism. The gene fusion events have been exploited to predict interactions between proteins that are found to form parts of a single multi-domain protein in other organism (Marcotte *et al.*, 1999; Enright *et al.*, 1999; Sprinzak and Margalit, 2001), especially for metabolic proteins (Tsoka and Ouzounis, 2000).

The three genomic context-based approaches have been applied to the genome of *Mycoplasma genitalium*, and their performances have been evaluated systematically (Huynen *et al.*, 2000). Conservation of gene neighborhood has been shown to be the method with the highest coverage, applying to 37% of the genes. Phylogenetic profile and gene fusion are found to cover 11% and 6% of the genes, respectively. It has also been observed that spatial proximity of genes on the genome correlates with the directness of interaction between the proteins encoded by the genes.

Prediction of Physical Interactions

Many approaches have taken advantage of information derived from both protein sequences and protein structures in order to predict physical interactions between proteins. In addition to the physical interaction state, some of these approaches are

also capable of determining the residues participating in interactions, and characterizing protein-protein interfaces. One of the disadvantages of structural context approaches is that they are particularly restricted in terms of scale, because the amount of proteins with accurate known 3D structures is still limited.

- **Using coevolution data.** Coevolution of interacting proteins like insulin and its receptors have been previously reported based on the observation that their corresponding phylogenetic trees exhibit higher degree of resemblance than those of non-interacting proteins (Fryxell, 1996). The similarity between phylogenetic trees have been quantified via the linear correlation between the matrices used to construct the phylogenetic trees, and used as an approximation for the probability of interactions between proteins (Goh *et al.*, 2000; Pazos and Valencia, 2001). This method is named *mirrortree*. The mirrortree method is similar to the phylogenetic profile method introduced before, since both methods are based on the analysis of phylogenetic information of proteins. The mirrortree approach requires also complete multiple alignments of sequences from the same species for the two proteins.

  Coevolution between interacting proteins can also be analyzed by considering the correlated mutation between pairs of residues of the proteins. Correlated mutations refer to mutations in one protein being compensated by mutations in its partner protein such that the interaction between the proteins is still stable. These positions involved in correlated mutations are mainly residues participating in interactions. Therefore, the methods using information about correlated mutations can be used to predict residues in proximity to interaction partner (Olmea and Valencia, 1997). Pazos and Valencia (2002) developed *in silico two-hybrid* system for inferring physical interactions between proteins from the predicted contacts between residues that mutate in a correlated manner. Again, to apply the in silico two-hybrid method, multiple sequence alignments with a good coverage of the species for both the proteins under investigation are required.

  The methods based on the coevolution of proteins require only protein sequences. Therefore, these methods are particularly useful for predicting interactions between proteins without known structures. However, the methods using coevolution data have a high requirement on the availability of homologous sequences to the query proteins. The performance of the methods relies heavily on the coverage of diverse species of the homologous sequences. Obviously, these methods cannot be applied to proteins without homologous sequences.

- **Integration of multiple high-throughput experimental data.** Many of the high-throughput experimental methods for determining gene or protein interactions are inherently noisy (see Section 1.1.4). Several computational approaches combine multiple datasets related to the biological context of genes or proteins obtained from high-throughput experiments to predict protein-protein interactions (Jansen *et al.*, 2003; Ben-Hur and Noble, 2005). In these methods,

each individual source of evidence for interactions was first validated against known positive (proteins known to form complexes) and negative (proteins with different cellular localizations) interactions to derive the statistical reliability of the source. Predictions of protein-protein interactions were then made by combining the different sources of evidence according to the calculated reliability. Unlike the approaches based on protein coevolution, these methods do not require a good coverage of homologous sequences of different species. They may also be applied to proteins without known structures. By combining evidence collected from multiple sources of experimental data, the performance of the methods is better than that of using individual data source.

- **Interface motif methods.** Pitre et al. have developed an interaction prediction engine based on re-occurring short polypeptide sequences between known interacting protein pairs in yeast (Pitre *et al.*, 2006, 2008). The approach predicts the interaction between proteins by searching short polypeptide sequences that occur repeatedly in pairs of proteins known to interact. These sequences are used in different proteins and in different contexts within the cell to mediate protein interactions. The method reaches an overall accuracy of 75% for the detection of protein interactions in yeast by using only protein primary structures.

  Betel *et al.* (2007) developed a method called *domain-motif interactions from structural topology* (D-MIST) based on conserved binding profiles derived from protein structures and protein interaction data. First, domain-binding motifs were extracted from structure templates of protein domains. These domain-binding motifs were then converted to sequence profiles as position-specific scoring matrices (PSSMs). Such domain-binding profiles were then employed to predict novel protein-protein interactions in yeast using only protein sequence information. Part of the predicted interactions were experimentally confirmed by the authors.

  Interface motif methods are suitable for proteins without known structures or homologous sequences. However, the identification of interface motifs depends heavily on known protein interaction data. Furthermore, two proteins may still interact without the presence of interface motifs.

- **Homology modeling methods.** Homology modeling is often exploited to extend the interaction data in the known 3D structures of protein complexes. In 2002, Aloy and Russell (2002) have suggested an empirical potential for identifying the most probable interactions generated by homology modeling on known 3D structures. The empirical potential is based on the observed propensity of residues to be on protein surfaces. Using known 3D structures as templates, Davis *et al.* (2006) modeled 3387 binary and 1234 higher order interactions in yeast. The predicted protein complexes were also filtered using functional annotation and sub-cellular localization information and were deposited in MODBASE (Sánchez *et al.*, 2000). Web tools like 3D-partner (Chen

et al., 2007) and HOMCOS (Fukuhara and Kawabata, 2008) predict potential interacting partners and generate models of complex structures for a query protein based on known interactions in which the homologous proteins to the query are involved.

Homology modeling methods not only predict whether two proteins interact, but also yield potential 3D model for the predicted interaction, which maybe used in further analysis about the function of the proteins. Nevertheless, methods in this category are restricted by the availability of known structures for protein complexes. In addition, it has been reported that homologous proteins may interact with different orientations (Aloy *et al.*, 2003).

- **Classification methods.** Aiming at distinguishing between truly interacting protein pairs and non-interacting protein pairs, classification methods have been applied in the prediction of interactions (Shoemaker and Panchenko, 2007). Various sources of data are often considered in the training of classifiers for separating positive and negative interactions. Machine learning algorithms like kernel methods are particularly helpful in this regard because they provide a vector representation encoding information from heterogeneous data sources in the feature space through a set of pairwise comparisons (Schölkopf *et al.*, 2004). Qi *et al.* (2005) and Chen and Liu (2005) employed random forests algorithms (Breiman, 2001) to predict whether two proteins interact. Qi *et al.* (2005) tested their algorithm on 4,000 interaction data in yeast collected from DIP (Xenarios *et al.*, 2000) and demonstrated that the method reaches a coverage of 20% of interacting pairs with a false positive rate of 50%. Chen and Liu (2005) showed that their approach predicts protein-protein interactions with a sensitivity of 79.78% and a specificity of 64.38% on a yeast protein interaction dataset containing 5,000 interactions derived from DIP and two other sources. Ben-Hur and Noble (2005) presented a kernel method for predicting interactions using a combination of multiple data sources. Their classifier retrieves 80% of a set of known interactions at a false positive rate of 1% on a refined dataset of 750 interactions that is expected to have low false positives derived from BIND (Bader *et al.*, 2001). The considerable difference in the performance of the methods resulted from not only the different underlying algorithms, but also the different benchmark datasets used for assessing the methods.

  These methods take advantage of data from multiple sources in order to improve prediction accuracy. Missing data are well handled by machine learning methods (Chen and Liu, 2005). One disadvantage of these methods is that the interpretability of the prediction results is not straightforward.

- **Prediction of binding sites.** A wide range of sequence, structural and physical attributes have been investigated to gain insights into the characteristics of protein binding regions and interactions. Based on these attributes, most prediction methods distinguish interface residues from non-interface surface residues in proteins, and infer potential binding sites based on the classification.

A number of interface properties are widely exploited in these prediction methods, including both sequence-based and structure-based properties. Sequence-based properties include amino acid composition, hydrophobicity and residue conservation. Examples of structure-based properties are solvent accessible surface area, geometric shape and electrostatic potential. Different classification procedures have been used mainly for the binary classification of binding site and non-binding site residues, including support vector machines (SVMs), neural networks and Bayesian networks. For example, Bradford and Westhead (2005) have trained a SVM classifier by using the aforementioned surface properties to predict protein binding sites with an accuracy of 76%. Sander *et al.* (2008) proposed a structural descriptor encoding the spatial arrangement of physicochemical properties of binding sites. The descriptor is based on the distance distribution between five types of functional atoms (Shulman-Peleg *et al.*, 2004). Combined with a SVM algorithm, this descriptor was successfully applied for the prediction of HIV-1 coreceptor usage (Sander *et al.*, 2007). Zhou and Qin (2007) and Ezkurdia *et al.* (2009) have provided comprehensive reviews on the prediction of binding sites. The inference of binding sites on protein surface provides valuable information on potential interaction partners for the protein and complexes the protein may form.

1.1.4 Management and Analysis of Interaction Data

The advance of new experimental techniques contributes enormously to the generation of the ever-increasing volume of high-throughput protein-protein interaction data. Consequently, the study on protein-protein interactions has been broadened from focusing on only binary interactions to a system-wide level. Entire networks of protein-protein interactions (*interactomes*) start to enter the scope of many researchers' view. On the one hand, the large volume interaction data have become very important foundation for new biological discoveries. On the other hand, a great challenge is posed to bioinformatics researchers for the management and analysis of the interaction data. This holds particularly true for the data obtained from high-throughput techniques. In this section, we discuss the management and analysis of protein-protein interaction data.

Interaction Databases

The vast amount of interaction data available today has been collected and organized in a number of publicly available databases. We list some of them as follows:

- 3did[3] (3D interacting domains (Stein *et al.*, 2005)) is a collection of physical domain-domain interactions that involve direct atomic contacts between domains in proteins with known structures from the Protein Data Bank (PDB) (Berman *et al.*, 2000). Pfam (Bateman *et al.*, 2000) domain definitions are used

[3] `http://gatealoy.pcb.ub.es/3did/`

in 3did. The 3did database provides detailed information about atomic contacts between domains derived from protein 3D structures and Gene Ontology-based (Ashburner *et al.*, 2000) functional annotations for the interactions. Both intra-chain and inter-chain domain-domain interactions are included. In addition, 3did also contains a hand-curated set of transient peptide-mediated interactions (Stein *et al.*, 2009). There are 133,071 domain-domain interactions of known 3D structures in 3did as of March 2010.

- BioGRID[4] (general repository for interaction datasets (Stark *et al.*, 2006)) is a database of both physical and genetic interactions for 22 organisms. BioGRID provides a flexible visualization tool named Osprey for producing graphical representations of interaction networks. It contains 114,506 non-redundant physical and 122,176 non-redundant genetic interactions as of March 2010.

- HPRD[5] (Human Protein Reference Database (Prasad *et al.*, 2009)) integrates curated comprehensive annotation for proteins in human proteome. Not only protein-protein interactions, but also interactions between proteins and nucleic acids or small molecules are reported. In addition to interactions, information such as post-translational modification, protein domain architecture, and association with human diseases is provided for proteins as well. As of March 2010, there are 27,081 proteins and 38,806 interactions in the database.

- IntAct[6] (an open source molecular InterAction database (Hermjakob *et al.*, 2004)) provides interaction data derived from the literature curation or user submission. For each interaction, information including a brief description of the interaction, the related experimental method and the literature citation for interacting proteins is presented. In total, there are 208,593 protein interactions stored in IntAct as of March 2010.

- iPfam[7] (Protein Domain Interactions Database (Finn *et al.*, 2005)) contains physical interactions between protein domains as defined in the Pfam database (Bateman *et al.*, 2000). The domain-domain interactions in iPfam are derived from structure models deposited in the PDB. The current iPfam database corresponding to Pfam release 21.0 contains 6,081 domain-domain interactions.

- MINT[8] (Molecular INTeractions Database (Chatr-aryamontri *et al.*, 2007)) consists of experimentally verified interaction data curated from published work with special emphasis on proteomes of mammalian organisms. Both low- and high-throughput detection methods have been considered in the collection of molecular interactions. Besides proteins, some non-protein entities like promoter regions are also featured in MINT. In addition, MINT provides a

[4] http://www.thebiogrid.org/
[5] http://www.hprd.org/
[6] http://www.ebi.ac.uk/intact/
[7] http://ipfam.sanger.ac.uk/
[8] http://mint.bio.uniroma2.it/

separate annotation of human proteome data named HomoMINT. Users may visualize and manipulate interaction networks using a built-in viewer. As of March 2010, there are 82,816 interactions in MINT.

- MPPI[9] (MIPS Mammalian Protein-Protein Interaction Database (Pagel *et al.*, 2005)) focuses on mammalian interaction data curated from scientific literature. In MPPI, only interactions collected from individual experiments are included. Data resulting from large-scale surveys like Y2H experiments are not considered. For each interaction, relevant experiment type and literature evidence, along with a description of binding regions of interaction partners are provided. The MPPI database currently contains 1,814 entries for protein-protein interactions involving >900 proteins from 10 mammalian species.

- STRING[10] (Search Tool for the Retrieval of INteracting Genes/Proteins (Snel *et al.*, 2000)) includes not only direct (physical) but also indirect (functional) interactions. STRING maps interaction evidence, which is collected from a variety of sources, including high-throughput experiments, computational prediction methods and public literature collections, onto a common set of genomes and proteins. Currently (March 2010), as many as 2,590,259 proteins from 630 organisms are covered in the database.

Most of the databases contain experimentally detected interactions, including both high-throughput methods (e.g., Y2H) and low-throughput method (e.g., methods based on PDB structures). Some databases provide both experimentally and computationally derived interaction data, e.g., STRING. Literature curation is employed by several databases to provide interaction evidence (MIPS).

Computational Analysis of High-throughput Interaction Data

There exist intrinsic limitations to the experimental technologies for detecting interactions, especially high-throughput methods. Such limitations leads to the low agreement between the interaction data obtained by different methods. Ito *et al.* (2001) reported that their interaction data obtained from a Y2H screening have a very small overlap (less than 20%) with the data derived using the same method by Uetz *et al.* (2000). Possible explanations for the disagreement are mainly the various differences in experimental conditions, which in turn lead to distinguishing interaction properties of proteins.

The evaluation of the accuracy of high-throughput interaction data is difficult mainly owing to the incompleteness of the data sets. Nevertheless, various approaches have been proposed for the assessment of the interaction data (Deane *et al.*, 2002; Bader *et al.*, 2004; Ramírez *et al.*, 2007; Lin *et al.*, 2009). Analysis of high-throughput interaction data suggests that they have a low coverage of complete interactome maps (Bader and Hogue, 2002; Bader *et al.*, 2004). It is estimated

[9] http://mips.helmholtz-muenchen.de/proj/ppi/
[10] http://string-db.org/

that there are in total 20,000–30,000 specific protein interactions in yeast, with the majority remaining uncovered (Bader and Hogue, 2002). It has also been suggested that protein-protein interaction data inferred from high-throughput methods contain not only false negatives but also false positives. Analysis of high-throughput interaction data with a focus on yeast suggests that more than 50% of the interactions are biologically irrelevant (Deane *et al.*, 2002; Bader *et al.*, 2004). Only 21% of proteins linked by high-throughput interactions are found to belong to the same functional category (von Mering *et al.*, 2002). Interestingly, the predicted protein-protein interaction data are found to provide as reliable information as high-throughput datasets (Ramírez *et al.*, 2007). Therefore, predicted interaction data may be combined with experimental data to increase the coverage of interactome.

1.1.5 Classification of Protein-Protein Interactions

Protein-protein interactions can be classified into different categories based upon various criteria as detailed below (Nooren and Thornton, 2003a).

- **Homo- vs. Hetero-.** Protein-protein interactions may be classified according to the composition of the complexes. If a protein complex is built from identical subunits, it is called *homo*-multimeric, otherwise *hetero*-multimeric. A survey of *E. coli.* proteins showed that homo-multimeric proteins are predominant with 79% of multimers composed of from 2 to 12 subunits being homo-multimeric (Goodsell and Olson, 2000). The identical subunits in homo-multimer may be associated in either *isologous* or *heterologous* manner (Monod *et al.*, 1965). With respect to homodimers, an isologous arrangement of subunits implies that the subunits interact with each other using the same binding region on their surfaces and are related by two-fold symmetry (Figure 1.1a and c). In a heterologous association, subunits use different binding regions.

- **Obligate vs. Non-obligate.** Based on whether the subunits in the complex also exist as stable structures *in vivo*, protein-protein interactions can be classified as *obligate* and *non-obligate*. For an obligate complex, the subunits only exist in the complex and cannot be found as separate stable structures *in vivo*. While for a non-obligate complexes, the subunits also exist independently as stable structures.

- **Permanent vs. Transient.** From the lifetime perspective of protein complexes, they can be classified as *transient* or *permanent*. Subunits in transient complexes may dissociate from each other. Permanent complexes are very stable and the subunits do not separate.

In addition to the terms introduced here based on Nooren and Thornton (2003a), there are other nomenclatures used for describing these protein-protein interaction types from different perspectives. Tsai *et al.* (1997a) divide protein-protein interactions into two classes: *two-state* and *three-state*. The polypeptide chains of a

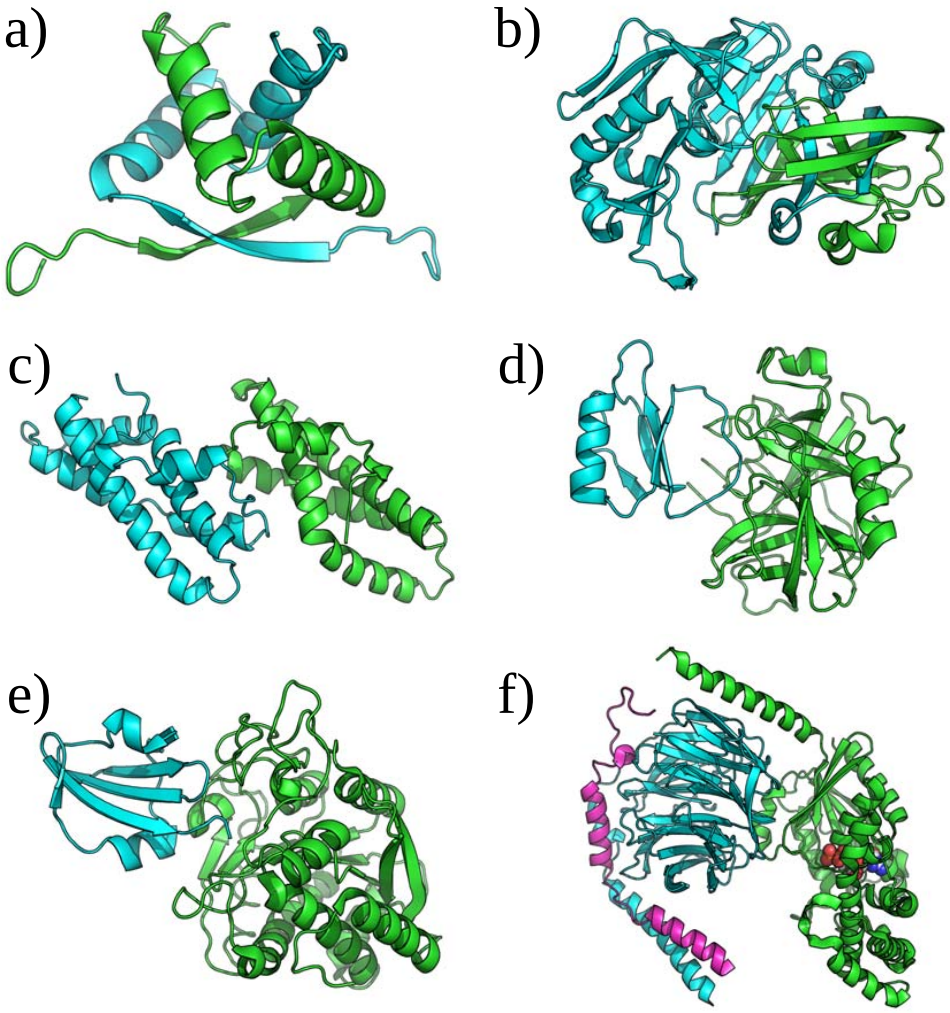


Figure 1.1: Examples of different types of protein-protein interactions. **a)** Obligate homodimer: Arc repressor (PDB ID: 1arr); **b)** Obligate heterodimer: Human cathepsin D (PDB ID: 1lya); **c)** Non-obligate homodimer: Sperm lysin (PDB ID: 3lyn); **d)** Non-obligate heterodimer: Chymotrypsin and inhibitor (PDB ID: 1acb); **e)** Non-obligate permanent heterodimer: Thrombin and Rodniin complex (PDB ID: 1tbr); **f)** Non-obligate transient heterotrimer: The complex formed between the α subunit (green) and the $\beta\gamma$ subunits (cyan and magenta) of bovine G protein (PDB ID: 1gg2). All figures of protein structure models in this dissertation are generated using PyMOL (DeLano, 2002).

two-state protein complex exist in only two states: they either exist unfolded, or folded together in a complex. Namely, the chains of a two-state protein complex do not exsit as folded monomers. In contrast, the chains of a three-state protein may fold independently as monomers. Thus, there are three states for these chains: unfolded, folded as monomers, or folded together in the complex. Obviously, two-state and three-state protein complexes are equivalent to obligate and non-obligate complexes, respectively. Goodsell and Olson (2000) classify interactions into two classes: interactions between globular subunits and interlocked interactions. According to their definition, interactions between globular subunits form after the folding of the globular subunits. But in interlocked interactions, the subunits adopt their folded structure only after the complexation. Therefore, the interactions between globular subunits are non-obligate interactions, and interlocked interactions are equivalent to obligate interactions. Gunasekaran *et al.* (2004) define a protein complex to be *ordered* if its component subunits remain stable when separated from their partners. Otherwise a complex is considered to be *disordered*. The concept of ordered and disordered protein complexes are in fact the extension of three-state and two-state proteins. Ordered protein complexes include three-state complexes plus crystal packing dimers, which are not biologically relevant. Disorder complexes contain not only two-state complexes, but also ribosomal proteins that help stabilize specific RNA structures but become disordered when separated from the ribosome (Moore, 1998; Draper and Reynaldo, 1999). In addition, natively unfolded proteins are also included in the disordered complexes. These proteins lack ordered well-defined structures and undergo the transition of disorder–order states during or prior to their biological functions (Uversky, 2002). Mintseris and Weng (2003) put forward the terms of *folding complexes* and *recognition complexes*. For a folding complex, its component chains fold and form the complex simultaneously. But for a recognition complex, the two processes of folding and binding may happen independently. Although the authors use permanent for folding complex interchangeably, as well as transient for recognition complexes, the folding and recognition complexes actually correspond to the obligate and non-obligate types, respectively.

Different protein-protein interactions fall into different types of the three classifications. Here we present a few exemplary protein complexes. Some of them have been discussed in Nooren and Thornton (2003a). Both obligate and non-obligate protein-protein interactions can be involved in either homo- or hetero-multimeric protein complexes. For example, Arc repressor is an obligate homodimeric protein molecule (Figure 1.1a) important for DNA binding (Breg *et al.*, 1990; Smith and Sauer, 1995). Human cathepsin D is an obligate heterodimer (Figure 1.1b), which works as a protease to break down proteins into peptide fragments (Minarowska *et al.*, 2008). Non-obligate protein-protein interactions are often involved in hetero-multimeric complexes, such as enzyme-inhibitor, receptor-ligand, antibody-antigen. Nevertheless, non-obligate interactions may also be found in homodimers. For instance, sperm lysin is a non-obligate homodimer (Figure 1.1c), which is formed via the interaction between the hydrophobic patch of two lysin monomers (Shaw

et al., 1995). The protease chymotrypsin and its inhibitor form a non-obligate heterodimer (Frigerio *et al.*, 1992) (Figure 1.1d). Permanent complexes involve not only obligate but also non-obligate interactions. Non-obligate interactions may be either permanent or transient. Thrombin is a protease that initiates the deactivation of the clotting cascade. Rodniin is a very specific inhibitor to thrombin and this non-obligate enzyme-inhibitor complex is permanent (van de Locht *et al.*, 1995) (Figure 1.1e). Non-obligate transient interactions are very common. The interaction between the α subunit and the $\beta\gamma$ subunits of bovine G protein is one of such examples (Wall *et al.*, 1995).

Many protein-protein interactions cannot be classified easily into distinct types. Moreover, the stability of protein complexes depend heavily on the physiological conditions and environments such as the local physiochemical conditions or the concentration of protein components.

1.2 Structure Models of Protein Complexes

Valuable information about the biochemical and biological role of proteins may be extracted from structural data. The structure reveals the overall organization of the residues in proteins in 3D space. Based on this, residues exposed to the solvent on the surface of the structure are distinguished from those buried in the core, and the shape and molecular composition of the surface can be investigated. The contacts between residues are disclosed, including both covalent and non-covalent contacts. The interaction between proteins and other molecules including ligands are illustrated in the structure. The protein function may be postulated from the nature of the binding site. Combining structral data with information such as experimental data, the molecular basis of protein function, catalytic mechanisms, conformational changes, or functional impact of mutations may be explored (Thornton *et al.*, 2000).

In this part, we present important methods for determining protein structure models. We focus on X-ray crystallography, which has been most widely employed and thus generated the vast majority of protein structure models. We also discuss the biologically irrelevant contacts between proteins present in such models determined by using X-ray crystallography. Furthermore, we review the approaches used for identifying binding regions on the protein surface using structure models.

1.2.1 Determination of Protein Structures

Many experimental techniques for the determination of protein structure models have been established (Branden and Tooze, 1999; Russell *et al.*, 2004). Among them *X-ray crystallography*, *nuclear magnetic resonance spectroscopy* (NMR spectroscopy), and *electron microscopy* (EM) are most frequently used. When using the term protein structure models, we usually refer to the atomic coordinates available from the models of protein tertiary or quaternary structures. To date, more than 86% (54,869 out of 63,559) of all the deposited structure models in the PDB are determined by

Table 1.2: Biological macromolecular structures in the PDB[a]

Exp. Method	Molecule Type				
	Proteins	Nucleic Acids	Protein/NA[b] Complexes	Other	**Total**
X-ray	51,291	1,193	2,368	17	54,869
NMR	7,206	891	152	7	8,256
EM	184	17	71	0	272
Hybrid	18	1	1	1	21
Other	120	4	4	13	141
Total	58,819	2,106	2,596	38	63,559

[a]Data collected on 23 February 2010, according to a statistics by the PDB at `http://www.rcsb.org/pdb/statistics/holdings.do`.
[b]NA: Nucleic Acid.

X-ray crystallography. In contrast, only around 13% and less than 1% of them are determined by using NMR spectroscopy and EM, respectively. See Table 1.2 for a statistic of biological macromolecules deposited in the PDB.

Protein Crystallography in a Nutshell

Crystallography is the most frequently used experimental technique for determining the structure models of biological macromolecules (Russell *et al.*, 2004) (also see Table 1.2). The X-ray crystallography technique for determining protein structure models can be split into five steps as depicted in Figure 1.2. The following introduction is based on Branden and Tooze (1999) and Rhodes (2000).

- **Protein crystallization.** Well-ordered protein crystals that are large enough to strongly diffract X-ray beams is the first prerequisite for solving protein 3D structures by X-ray crystallography. Given a target protein, the normal procedure of protein crystallography starts with the purification of the protein sample, because a pure and homogeneous protein sample is crucial for successful crystallization. The purified protein sample is then crystallized, most frequently by using a *hanging-drop* method. In this method, a drop of protein solution is hung on a glass plate and sealed on the top of a container, in which precipitant like salt solution is added. The drop of protein solution becomes supersaturated gradually by loss of water from the droplet to the precipitant. Crystals form when protein molecules are precipitated very slowly from the supersaturated solutions.

- **Collecting diffraction data.** Appropriate protein crystals are then mounted between an X-ray source and an X-ray detector. A narrow and parallel X-ray

beam from the source is directed toward the crystal, and is then diffracted into many discrete beams upon striking on the crystal. Diffracted X-ray beams strike the detector and leave a pattern of spots, named the *reflections* of the X-ray. Only one image of the reflections is insufficient for the reconstruction of the whole crystal. The X-ray beam must hit the crystal from various directions during the experiment, such that all possible diffraction patterns are produced. Usually this is achieved by rotating the crystal in the X-ray beam. The positions and intensities of these reflections produced on the detector are collected as the X-ray diffraction pattern for the protein crystal.

- **Processing diffraction data.** When the X-ray beam strikes a protein crystal, it is the clouds of electrons of the protein molecules that diffract the beam. Therefore, the diffraction pattern reveals the *electron density*. The electron density of the protein molecules is computed from the collected X-ray diffraction data by applying a Fourier transformation. The positions of the "spots" of the diffraction pattern are dependent on the shape and the size of the unit cells, the smallest representative unit in a protein crystal (see Section 1.2.2), and the inherent symmetry of the crystal. The intensities of the "spots" are proportional to the square of the wave amplitude. To compute the electron density from the diffraction pattern, both the amplitude and the phase of the diffracted X-ray waves must be obtained. However, from the diffraction pattern only the amplitude can be derived from the intensity of the spots. Several methods have been proposed to measure the phase of the diffracted spots, and solve the "phase problem" (Rhodes, 2000). Some approaches take advantage of determined structures of homologous proteins, other approaches rely on introducing heavy atoms (atoms that have a large atomic number) to the protein crystal. After the initial phase is estimated, the initial electron density map can be built.

- **Determining structure model.** In this step, the electron density map is interpreted as a polypeptide chain of amino acids weaving its way through the map. First, the trace of the protein backbone is inferred from the electron density map (*chain tracing*). Because an electron density map rarely shows a clear continuous trace of electron density from the N-terminus to the C-terminus, usually many possible traces are brought up. Then the positions of the atoms in the amino acids are determined in an initial model such that the atoms fit best the electron density (*model building*). All possible knowledge about the protein under investigation is utilized for the interpretation of the map. In most cases, the primary structure of the protein molecule is the most important information, given it is available independent of the X-ray crystallography experiment. In addition, the general physicochemical principles of molecular structure and stereochemistry should not be violated by the model.

- **Refining the structure model.** The building of a protein structure model is carried out in an iterative manner. The primary hypothetical model is im-

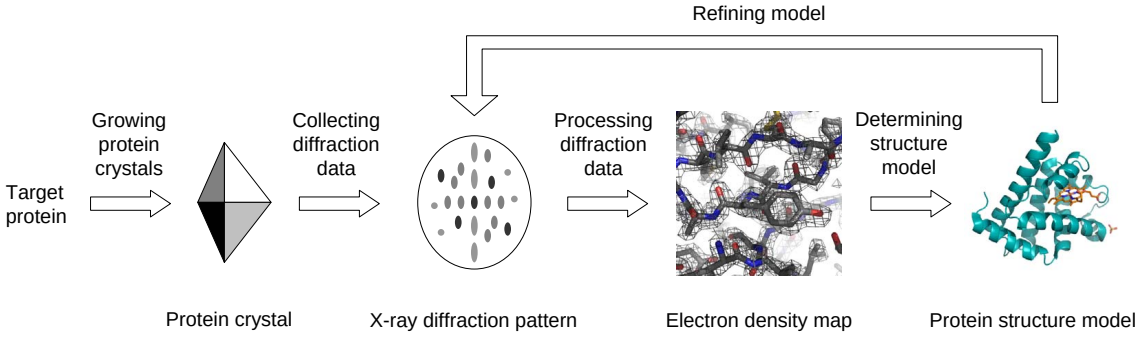


Figure 1.2: X-ray crystallography in a nutshell.

proved repeatedly, by minimizing the difference between the diffraction pattern observed in the experiment and the one calculated from the model. The agreement between the overall diffraction patterns is measured by the *residual index*, or the *R-factor*. In theory, the R-factor is in the range of 0.0 for perfect agreement, and 1.0 for total disagreement. In practice, an R-factor of 0.5 or greater indicates very poor agreement between the diffraction patterns, and the models are often discarded unless other supporting data are available. Another commonly used criterion of model quality is the *free R-factor*, or R_{free}. To compute R_{free}, a small set of randomly chosen diffraction data is set aside in the beginning and then used as the "test set" in a *cross-validation* process for assessing the agreement between observed and calculated diffraction data. Usually, both factors are reported together with the structure model deposited in the PDB. They are important guides for monitoring the convergence of the refinement process. Furthermore, like in the step for determining the initial structure model, a set of chemical, stereochemical, and conformational properties of the model are examined for controlling the model's quality.

Other Techniques for Determining Protein Structures

In addition to protein crystallography, two further techniques for determining protein structures are NMR spectroscopy and electron microscopy.

- **NMR Spectroscopy.** Nuclear magnetic resonance is a physical phenomenon based on a magnetic property (magnetic moment, or *spin*) of the nuclei in certain atoms, such as ^{1}H, ^{13}C, ^{15}N, and ^{31}P. When protein molecules are placed in a strong magnetic field, the spin of the nuclei in these atoms aligns with the magnetic field and reaches an equilibrium. If a radio frequency pulse is then applied to the protein molecules, the nuclei are perturbed and change from the equilibrium state to an excited state. When they are relaxed back to the equilibrium state, the nuclei emit radio frequency radiation. The properties of the radiation such as its frequency and magnitude are determined by the

molecular environment of the nuclei. The technique NMR spectroscopy[11] is employed to record the emitted radiation. The relative positions of the atoms in the proteins are derived from the recorded radiation data and are used as constraints to derive the structure models of the proteins.

NMR spectroscopy is considered a complementary technique to X-ray crystallography. As we have pointed out in Section 1.2.1, the crystallization of protein molecules is often the most difficult step in the whole process to determine protein structures by X-ray crystallography. Using NMR, structure models can be built for protein molecules in solution. This has several benefits for the study of protein molecules. For instance, for small proteins that are difficult to crystallize, NMR can be employed alternatively. Furthermore, NMR spectroscopy is also suitable for the investigation of dynamic processes like protein folding via the characterization of the structures of unfolded and partly folded proteins (Dyson and Wright, 2005).

The NMR spectroscopy technique has a few limitations in practice. First of all, usually a set of possible structure models instead of a single model is built based on the distance constraints. These models should not be considered as different conformations of protein molecules in solution. Rather, they are possible models that satisfy the distance constraints equally well (Branden and Tooze, 1999). The primary reason for the ambiguity is that the number of distance constraints is not sufficient for concluding a unique model. In addition, the application of NMR spectroscopy is limited to small protein molecules, traditionally with a molecular weight of up to 25 kDa. With the progress of related techniques, the upper limit of the molecular weight has been increased substantially (Fiaux *et al.*, 2002).

- **Electron Microscopy.** Electron microscopy allows the visualization of the structure and dynamics of biological macromolecule assemblies at relatively low resolutions (3–30 Å) (Volkmann and Hanein, 2003). In the imaging process of EM, an electron beam passes the specimen of the target molecule prepared in a very thin film. The emerging electrons, either scattered or unscattered by the specimen, are collected and focused by the imaging optics of the microscope. The focused diffraction pattern of the electrons is recorded. The electron micrographs are the two-dimensional (2D) projections of the target macromolecule. To reconstruct the 3D structure of the molecule, electron micrographs of the molecule are taken from different views and of various conformations. The 3D structure model of the molecule is then constructed by aligning and combining thousands of such 2D projection images (Volkmann and Hanein, 2003). Cryogenic methods are often employed for the fixation of the target molecules and the corresponding field is called cryo-EM. The application of cryogenic methods enables the observation of the protein molecules in their native aqueous environment.

[11] NMR spectroscopy is often simply mentioned as NMR.

The resolution of the electron density determined by EM is relatively low (about 5 Å using Cryo-EM) (Chiu *et al.*, 2005), and it needs to be combined with other methods in order to build models of atomic resolutions. The main advantage of EM is that large molecular assemblies (molecular weight greater than 200–500 kDa) can be visualized and their different conformational states can be analyzed. Similar to NMR spectroscopy, the samples under study by EM do not need to be presented in crystalline forms. Furthermore, there is generally no upper limit for the size of the proteins to be studied by EM.

1.2.2 Crystal Packing in Protein Structure Models

Most structure models of biological macromolecules deposited in PDB are determined by X-ray crystallography (see Table 1.2). The term *X-ray crystallography* implies that the protein samples under examination are in their crystalline states. A protein crystal consists of many identical protein molecules arranged in a regular and repeating array. These protein molecules in the crystal are highly packed. Generally, the more closely the protein molecules pack in the crystal, the better the diffraction pattern is. Because of the tight arrangement of protein molecules in the crystal, some interactions present in crystals are not biologically relevant. In other words, certain interactions between protein molecules observed in the structure determined by X-ray crystallography do not happen in physiological environment and are the result of *crystal packing*.

In this section, we describe the crystalline lattice of proteins in detail. We also discuss several key concepts that are relevant to non-biological interactions. These concepts are important for the understanding and utilization of structure models generated by using X-ray crystallography.

Asymmetric Unit, Unit Cell, and Crystal

The whole crystal structure can be imagined as many identical units stacked orderly beside or on top of each other resulting in a 3D array. Each unit is called a *unit cell*. The unit cell is the smallest representative element in protein crystal, because the entire crystal structure can be generated by translating unit cells in 3D space.

The smallest volume element of a crystal structure to which crystallographic symmetry can be applied to generate the whole crystal is called an *asymmetric unit* (ASU). The content of a unit cell can be built from the ASU by applying symmetry operations (rotations and translations). The relationship between the ASU, unit cell, and crystal is illustrated in Figure 1.3. By convention, most of the deposited structure models in the PDB are the ASUs of protein crystals.

Biological Unit, and its Relationship with Asymmetric Unit

Although most of structure models deposited in the PDB derived from X-ray crystallography are the ASUs of the protein crystals, most users of protein structure models

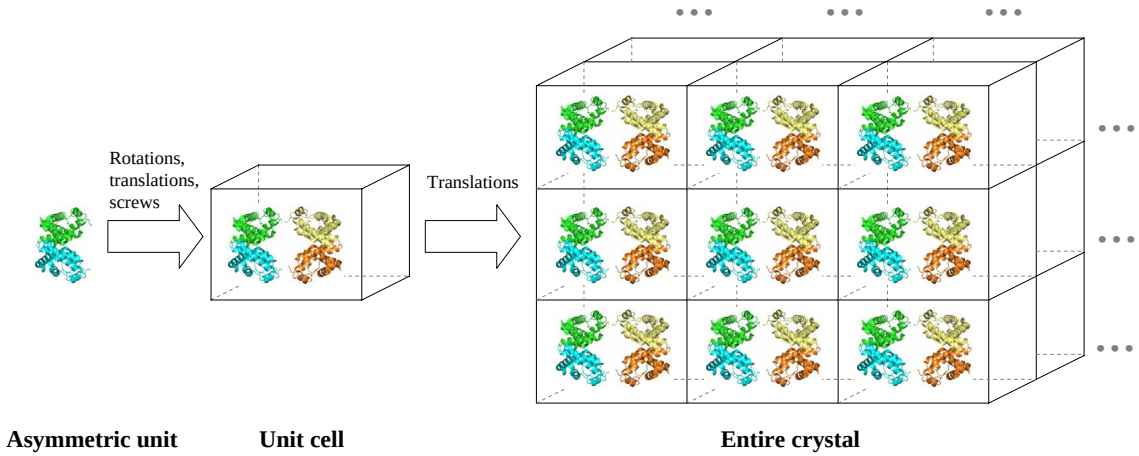


Figure 1.3: Relationship between asymmetric unit, unit cell, and crystal. Here an ASU is rotated 180° around an axis vertical to the paper plane to produce a second copy of the ASU, and the two ASUs together comprise a unit cell. The unit cell is then translated in the 3D space to build the whole crystal.

are interested in the so-called *biological units* (BU) of proteins. Protein BUs represent the quaternary structures of protein functional forms. Frequently, the ASUs stored in the PDB are the same as the BUs of the proteins. However, in a considerable number of PDB entries, the protein structure models are not equivalent to the protein quaternary structures of their functional states. In fact, an asymmetric unit might contain

a) part of a BU, or
b) exactly one BU, or
c) multiple BUs.

We take a tetrameric protein example *hemoglobin* to illustrate this relationship (see Figure 1.4):

a) In PDB entry `1hho`, only two chains are reported in the deposited ASU, corresponding to one half of the biological quaternary structure of the hemoglobin (ASU = $\frac{1}{2}$ BU) (Figure 1.4 a).
b) In PDB entry `2hhb`, the ASU is equivalent to the biological unit of the hemoglobin molecule (ASU = BU) (Figure 1.4 b).
c) In PDB entry `1hv4`, two hemoglobin molecules are deposited in one ASU (ASU = 2 BU) (Figure 1.4 c).

For many cases, the relationship between the BU and the ASU is more complicated than for the hemoglobin instances. For example, one ASU might comprise a part of one BU and some other part of a neighboring BU in the crystal. The contacts between proteins observed in the ASUs that are not present in the protein BUs are the crystal packing contacts and need to be distinguished from the biological interactions that are present within the BUs. PDB users who are interested in the functional forms

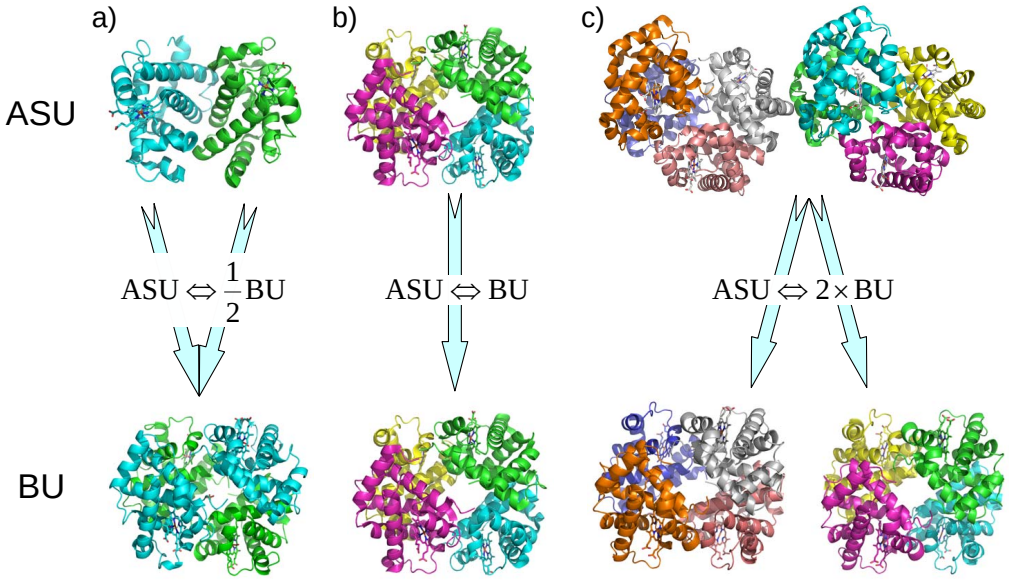


Figure 1.4: Illustration of the relationship between the asymmetric unit and the biological unit. a) PDB entry `1hho` contains only half of the biological unit; b) PDB entry `2hhb` corresponds exactly to one biological unit; c) PDB entry `1hv4` comprises two biological units.

of proteins should pay particular attention to the relationship between the ASU and the BU, and avoid taking for granted that each PDB file contains the model for the biological unit of certain protein.

Detection of Biological Units

Since 1999, the PDB has introduced several new records in the PDB file format to deal with the intricate relationship between the ASU and the BU. Specifically, `REMARK 300` in PDB files gives the description of the corresponding biologically functional molecule in free text, `REMARK 350` cites the transformation matrices needed to build the BU from the ASU deposited in PDB files.

Nowadays, the PDB provides also the structure model of the BU for each PDB entry. These models are constructed based on the details provided by depositors in `REMARK 300` and `REMARK 350` for entries deposited since 1999. For the PDB entries deposited prior to 1999, the models are built based on either details provided by the depositors or on supporting information obtained from the Swiss-Prot (Boeckmann *et al.*, 2003) or the PQS databases (Henrick and Thornton, 1998).

In addition to the PDB, several other databases also offer access to the atomic coordinate data of the biological molecules corresponding to the PDB entries using information other than that provided in `REMARK 300` and `REMARK 350`. These

include PQS, PITA database[12], PISA database[13], ProtBuD and PiQSi. PQS (Protein Quaternary Structure) applies crystal symmetry to the ASU deposited in each PDB file to build up the potential quaternary assembly of the protein. Monomeric chains are progressively added to the potential assembly and only suitable ones are retained based mainly on the number of inter-chain atomic contacts. All the pairwise interfaces between protein chains in each PQS entry are examined and scored based on mainly the interface area and a solvation energy in order to distinguish specific interactions from non-specific ones (crystal packing contacts) (Henrick and Thornton, 1998). The accuracy of the PQS annotations reaches 81% in the identification of homodimers on a dataset of 7,001 proteins. The PITA (Protein InTerfaces and Assemblies) approach starts with the largest protein complex constructed based on the crystal symmetry and bi-partitions it iteratively until a chosen threshold for the stability score of the assembly is achieved (Ponstingl *et al.*, 2003). Each bi-partition is performed such that the sum of scores between chains across the two parts is minimal. This minimum-cut score is taken as the stability score of the assembly. In PITA, inter-chain interfaces are scored based on buried surface area and a statistical pair potential of chemical complementarity, which is derived from atom-pair frequency across interfaces (Moont *et al.*, 1999). The PITA program reached an accuracy ranging from 74% to 84% depending on the oligomeric state of the proteins on a dataset of 218 protein structures using a threshold of 67.3 for bi-partitions. PISA (Protein Interfaces, Surfaces and Assemblies) differs from both PQS and PITA as it does not use iterative procedures but a graph-exploring approach to detect all possible protein assemblies that may be formed in the crystal (Krissinel and Henrick, 2007). A thermodynamic energy function integrating protein binding energy and complex dissociation entropy was introduced to evaluate the stability of interfaces observed in the crystal. The PISA approach may identify protein biological assemblies missed by PQS or PITA. In a benchmark using the same dataset as used by PITA, the PISA program reaches an accuracy between 84% and 93% depending on the oligomeric state of proteins, an improvement of 5%–8% compared with PITA. ProtBuD (Protein Biological unit Database) (Xu *et al.*, 2006) reports BU information of proteins based on both PDB and PQS. Interestingly, the survey of ProtBuD shows that the BUs provided in the PDB and the PQS differ on 18% of the PDB entries. PiQSi (Protein Quaternary Structure Investigation) provides protein quaternary structure data by large-scale mining of biochemical literature (Levy, 2007). Benchmark results of the underlying method for PiQSi demonstrate that PiQSi annotations agree with a curated dataset very well. Of the over 10,000 BU structures that have been annotated by using this approach, around 15% disagree with the PDB BU annotations.

[12]the underlying approach is also named PITA.
[13]the underlying approach is also named PISA.

1.2.3 Binding Sites and Interfaces

In 3D structures of protein complexes, *binding sites* are the regions on protein surfaces that are invovled in the interactions. Protein-protein *interfaces* are normally defined to be the group of residues or atoms at the interacting binding sites.

Identification of Interfaces in Protein Complex Structures

Protein-protein interfaces may be defined using certain criteria on the known 3D structures of the corresponding protein complexes. The most popular criteria include distance and solvent accessible surface area. These criteria are usually utilized both for detecting protein-protein interactions in macromolecular assemblies, and for identifying protein-protein interfaces.

- **Distance Criteria.** All non-covalent interactions stabilizing protein-protein interactions take place between atoms at a close distance (see Section 1.2.3). It is thus straightforward to define interface atoms or residues as those that are close in space to their partner subunit(s).

  Using this type of criteria, the protein interface is defined based upon the proximity of certain entities of the two interacting subunits. The proximate entities under consideration can be C_β atoms (C_α for Gly) (Glaser *et al.*, 2001), any two non-hydrogen heavy atoms (Tsai *et al.*, 1996; Ofran and Rost, 2003; Dafas *et al.*, 2004; Davis and Sali, 2005), or some pseudopoints defined according to the physicochemical properties of amino acid side chains (Shulman-Peleg *et al.*, 2004). The distance cutoff value can be either uniform (Glaser *et al.*, 2001; Dafas *et al.*, 2004; Davis and Sali, 2005), or dependent on the van der Waals radii of the proximate entities (Tsai *et al.*, 1996). Based on a set of 621 interfaces from 440 PDB entries, Glaser *et al.* (2001) derived the relationship between the distance cutoff and the frequency densities of the 20 standard amino acids. The frequency density was defined to be the frequency of a certain residue type at the interface normalized by the square of the distance cutoff. They concluded that when the distance cutoff between C_β atoms (C_α for glycine) is around 6 Å, the frequency density is maximal for all residue types except Trp and Phe. Namely, the average number of contacts across interfaces for most residue types is maximized using this cutoff. Based on the observation, a distance cutoff of 6 Å was chosen to define pairs of interacting residues. Tsai *et al.* (1996) defined two residues to be interacting if any pair of their non-hydrogen atoms are closer than the sum of the corresponding van der Waals radii (Chothia, 1975) plus 0.5 Å tolerance. Jones and Thornton (1995) used the same definition but enlarged the tolerance value to 1.0 Å for recognizing interacting atoms. The main reason for including a tolerance in the distance threshold between heavy atoms is to compensate for missing hydrogen atoms in protein structure models[14]. The distance criteria are very simple,

[14]The position of hydrogen atoms cannot be resolved in most protein crystals by using X-ray

since they require only the coordinates of the entities under consideration as input. Data structures such as a *kd-tree* can be used for accelerating the search for neighboring entities (de Berg *et al.*, 2000).

- **Solvent Accessible Surface Area Criteria.** The second group of definitions focuses on the reduction of the *solvent accessible surface area* (SASA) of residues or atoms upon the complexation of protein subunits. Consequently, an interface is defined as the buried surface region between interacting subunits. The definition is sometimes formulated as the increase in SASA after the separation of the subunits in the protein complexes. The two expressions are equivalent and the change in SASA is denoted as ΔSASA.

  Either interface atom or interface residue can be identified using this criterion. For instance, Jones and Thornton (1996) defined interface residues and atoms as those having absolute reductions in their SASAs of more than $1\,\text{Å}^2$ (interface residue) and $0.01\,\text{Å}^2$ (interface atom) upon the formation of a complex. Chakrabarti and Janin (2002) defined interface residues and atoms using a cutoff of $0.1\,\text{Å}^2$.

  SASA criteria may be constructed based on the relative change in SASAs, because the SASAs of the 20 amino acids differ greatly. For instance, the size in terms of SASA for glycine (Gly) is $85\,\text{Å}^2$ but $259\,\text{Å}^2$ for tryptophan (Trp) (Miller *et al.*, 1987). The total and side-chain SASAs for the 20 amino acids are compared in Figure 1.5. Miller *et al.* (1987) defined exposed residues in monomeric proteins as those residues with SASA $> 5\%$ of their total sizes. Jones and Thornton (1997a) followed the same cutoff to identify surface residues of a protein. Similarly, interface residues can also be defined based upon the relative reduction of SASA on complexation. De *et al.* (2005) defined a residue to be at the core of an interface if its SASA shows large variation between the exposed ($> 10\%$) and buried ($< 7\%$) state.

  The contributions of the component subunits to the total ΔSASA are not necessarily equal, especially in heterodimeric complexes. Although the discrepancy is mostly very small ($\sim$3%), the distribution of ΔSASA over the components may be as variant as 46:54 in protease-inhibitor complexes (Lo Conte *et al.*, 1999).

  The SASA criteria identify atoms/residues that are partly or fully buried between the component subunits in protein complexes from the solvent. Therefore, the interface region recognized by SASA criteria is more continuous on the protein surface than by distance criteria. However, using SASA criteria only the binding sites or interfaces are identified, but not the pairs of interacting atoms/residues. To study interface features relevant to inter-atom or

crystallography. Thus in most PDB files, hydrogen atoms are absent. Only in a few structure models of a high resolution (1.2 Å or less), hydrogen positions are partially assigned (Rhodes, 2000).

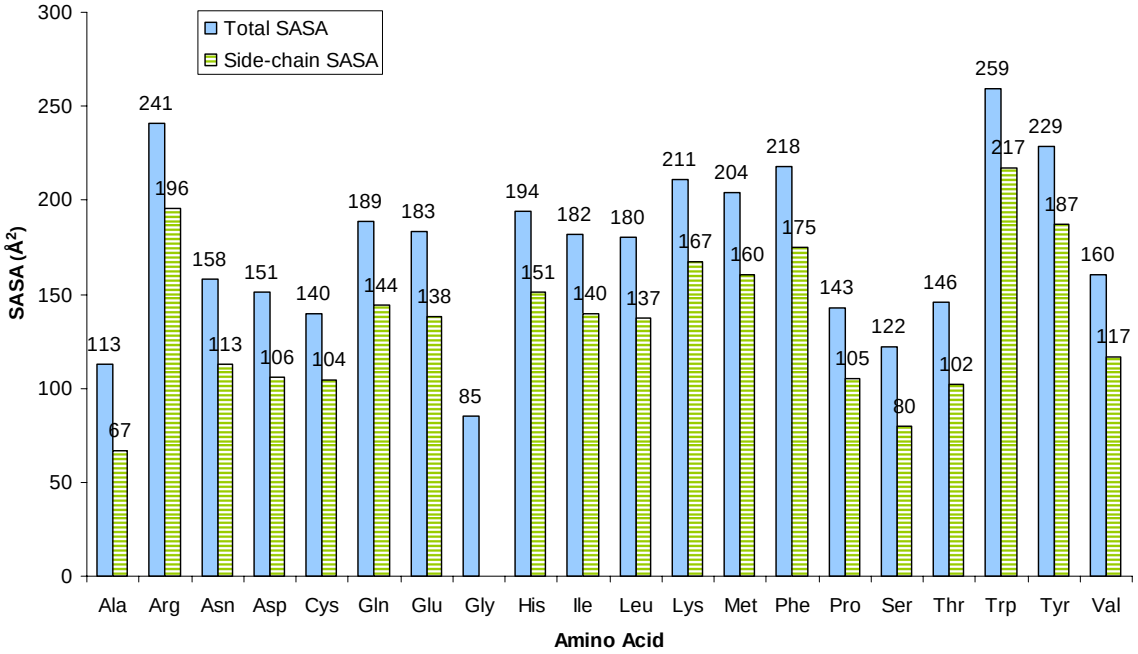


Figure 1.5: Solvent accessible surface area for 20 amino acids. The plot is based on the data taken from Miller *et al.* (1987).

inter-residue contacts (e.g., residue-residue contact preference), extra steps are required to infer interactions between atoms/residues.

In general, the interface defined by the two types of definitions overlap to a large extent (Headd *et al.*, 2007). Nevertheless, the two criteria may lead to different definitions of interfaces. For example, when there are interior cavities between interacting subunits, the SASA criteria may not be able to detect such cavities (depending on the algorithm and the probe radius) and report larger interface area than distance criteria. In addition, the cutoff values used in the criteria, both distance and SASA criteria, may have great impact on the size and the composition of interfaces.

The application of these criteria is limited by the number of structure models available at present. As of February of 2010, there are 63,559 biological macromolecular structures stored in the PDB in total (see Table 1.2), including monomeric structures. Aloy and Russell estimated that there are approximately 10,000 types of protein-protein interactions in Nature and only about 2,000 of them were known to us by then (Aloy and Russell, 2004). All pairs of proteins that interact similarly in the 3D structures of their complexes were grouped into the same interaction type. Proteins sharing more than 25% sequence identity on both sides of the interfaces were assumed to interact in a similar way in the 3D structures. Given the rate of structure determination for macromolecules at the time, it would take more than 20 years for us to apprehend the full set of protein-protein interaction types. Nevertheless, in contrast to the data derived from high-throughput experimental methods, the interaction data derived from protein complexes with known 3D structures have a very

low error rate and provide also detailed structural information of the interactions.

Non-covalent Interactions at Protein-Protein Interfaces

Non-covalent interactions, or non-covalent bonds, are weak chemical bonds that do not involve the sharing of electrons between atoms (Alberts *et al.*, 2002). In general, a non-covalent interaction has a lower energy than a covalent bond (see table 1.3) and is reversible under standard conditions. However, due to the enormous amount of occurrences, non-covalent interactions are the basis for the stability of folded protein structures, protein-protein interactions, protein-DNA or protein-RNA interactions (Lodish *et al.*, 1999; Petsko and Ringe, 2003).

There are various types of non-covalent interactions. We discuss in the following section electrostatic interactions, including hydrogen bonds and van der Waals interactions.

- **Electrostatic Interactions.** *Electrostatic interactions* are formed between oppositely charged atoms due to the electrostatic attraction between the atoms. They are also referred to as *ionic interactions* or *ion-ion interactions*. In electrostatic interactions, the electrons are not shared between the interacting atoms as in covalent bonds. The strength of an electrostatic interaction is described by *Coulomb's law*,

$$F = k\frac{Q_1 Q_2}{r^2}$$

  where k is *Coulomb's constant*, Q_1 and Q_2 are the charges of the two atoms, and r is the distance between the two atoms. The term "electrostatic interaction" is often used for interactions due to full charges. However, in principle, all polar interactions are electrostatic attractions between oppositely charged atoms.

  - **Hydrogen Bonds.** *Hydrogen bonds* are the non-covalent interactions formed between an electronegative atom and a hydrogen atom, which is covalently bonded to another electronegative atom. The electronegative atoms to which the hydrogen atom is bonded is called the hydrogen bond *donor*, and the other electronegative atom is called the hydrogen bond *acceptor* (commonly oxygen or nitrogen in proteins). The distance between the hydrogen bond donor and acceptor atoms varies largely, depending on the surrounding environment in which the hydrogen bonded is formed. Jeffery categorized hydrogen bonds as strong, moderate, and weak interactions (Jeffrey, 1997). For strong hydrogen bonds, e.g., $\mathrm{F{-}H\cdots F^-}$, they are "mostly covalent" and the donor-acceptor distance can be as short as 2.2–2.5 Å. Most hydrogen bonds in proteins fall into the moderate category, in which the donor-acceptor distance is in the range of 2.5 and 3.2 Å, on average 3.0 Å. In weak hydrogen bonds, the donor-acceptor distance is usually from 3.2 to 4.0 Å. Their binding energies also differ greatly.

They can be as high as 60–170 kJ/mol for strong hydrogen bonds as in HF_2^- (Emsley, 1980), 15–60 kJ/mol for moderate hydrogen bonds, and less than 15 kJ/mol for weak hydrogen bonds. In proteins and nucleic acids, the energy of hydrogen bonds is only 4 to 8 kJ/mol on average (Lodish *et al.*, 1999). If the donor or the acceptor is fully charged, the hydrogen bond is stronger than when both are uncharged. When both the donor and the acceptor are fully charged, the hydrogen-bonding ion pairs is called a *salt bridge* and the binding energy is significantly higher (Petsko and Ringe, 2003). Hydrogen bonds play important roles in the stabilization of protein 3D structures and the formation of base pairs in nucleic acids.

– **Van der Waals Interactions.** The electron distribution around an atom undergoes constant random fluctuation. This fluctuation causes asymmetric distribution of the electrons, resulting in a transient electric dipole. The dipole will induce a momentary opposite dipole in a nearby non-covalently bonded atom. The attraction between the two transient dipoles of the two atoms is called the *van der Waals interaction*. The strength of van der Waals interactions decreases rapidly with the increase of the interatomic distance, but two atoms will be repelled by their negative electron clouds if they get too close to each other. The attractive and repulsive forces are balanced when the distance between the two atoms equals the sum of their van der Waals radii. Van der Waals interactions are very weak and the energy is less than 4 kJ/mol (Lodish *et al.*, 1999). However, the van der Waals interactions between macromolecules can be appreciable because usually a large quantity of van der Waals interactions are formed between complementary surfaces.

Physicochemical Properties of Protein-Protein Interfaces

It has been shown that protein-protein interfaces exhibit many common physicochemical properties. We discuss several of such properties in this section. The data are taken from Petsko and Ringe (2003), Ponstingl *et al.* (2005), and Janin *et al.* (2008).

- **Binding regions are complementary.** Macromolecular complexes are stabilized by non-covalent interactions. These non-covalent interactions are usually weak bonds that may break easily at room temperature (see Section 1.2.3). The strength of binding between subunits in protein complexes needs to exceed a certain threshold (15–20 kJ/mol) for the protein complex to stay stable. The tight binding is commonly achieved by the occurrence of a large number of non-covalent bonds between complementary binding regions. At the same time, protein surfaces are normally observed to be irregular. This is a very important property for the specificity of protein-protein interaction. The

Table 1.3: Chemical bonds stabilizing proteins[a].

Interaction	Example	Typical distance (Å)	Free energy (kJ/mol)
Covalent bond	$-C_\alpha-C-$	1.5	356
Disulfide bond	$-Cys-S-S-Cys-$	2.2	167
Salt bridge	$-COO^-\cdots{}^+H_3N-$	2.8	12–17
Hydrogen bond	$>N-H\cdots O=C<$	2.4 – 3.5	2–6[b] or 12.5–21[c]
Long range electrostatic interaction	$-COO^-\ \cdot\ \cdot\ {}^+H_3N-$	Variable	Depends on distance and environment
Van der Waals interaction	$-H_3C\cdots CH_3-$	3.5	2–4

[a]Adapted from Figure 1.10 in Petsko and Ringe (2003).
[b]in water
[c]if either donor or acceptor is charged

interaction between proteins does not only depend on the complementarity between the shapes of the subunits. At binding regions, hydrogen bond donors and acceptors occupy opposite positions, positively charged residues are opposite to negatively charged residues, and nonpolar groups face other nonpolar groups. Such complementarity is not only observed between the binding regions of protein-protein interactions, but also protein-ligand interactions.

- **Interface size is related to protein size.** The interface size depends highly on the size of protein. The fraction of the protein surface involved in the interaction varies. On average, about 18% of a protein's surface is involved in binding. Usually this fraction is correlated with the number of subunits in the protein complex. In general, the fraction is larger for proteins composed of more subunits.

- **Atomic packing density is similar to that of the protein core.** Atoms buried at protein-protein interfaces are commonly closely packed. The atomic packing density at interfaces is similar to that at the interior of the subunits (Lo Conte *et al.*, 1999). This also reflects that the binding regions of protein surfaces have complementary shapes.

- **Secondary structure content shows no preference.** At protein-protein interfaces, all the secondary structure motifs occur frequently. The content of the secondary structure types are greatly variable. None of the secondary

structure motifs is preferred at interfaces compared to the rest of the protein surfaces.

- **Binding region is more hydrophobic than the rest of protein surface.** The hydrophobic portion of protein molecules tend to aggregate and hide in the core of the protein fold, leaving polar and charged groups being enriched on the exterior of the protein fold. This hydrophobic effect is a main driving force for protein folding (Dill, 1990) and protein-protein interaction (Chothia and Janin, 1975; Tsai *et al.*, 1997b). Binding regions on protein surfaces are more hydrophobic than the rest of protein surfaces, although the distribution of individual hydrophobic groups may vary. On average, the hydrophobicity of the interface as a whole lies between the core and the surface.

- **Amino acid composition differs from the non-interacting protein surface.** The amino acid composition of interfaces is different from that of the rest of protein surface. This discrepancy is also the consequence of the hydrophobic effect. Residues with polar and charged groups generally tend to occur on the surface exposed to the solvent. Residues with hydrophobic groups like aromatic residues and aliphatic residues are observed to be abundant at interfaces. Residues totally buried at interfaces are found to be more hydrophobic than those that are partially accessible to the solvent.

1.3 Outline of the Dissertation

There has been constant progress in the analysis and understanding of the mechanism underlying protein-protein interactions, especially by exploiting structural information of known 3D structural models of proteins. The main goal of this dissertation was to contribute to the understanding of the underlying physicochemical rules governing interactions between proteins through the comparison of protein-protein interactions. The comparison of interactions may be carried out in different ways. In our work, we have focused on the comparison of interactions on two different levels. In the first part of the dissertation, we aimed at identifying distinct interface properties for various types of interactions and applying them in the classification of the interactions. In the second part of our work, we compared the binding modes of non-covalent interactions at protein-protein interfaces.

In Chapter 2, we describe the characterization and classification of three types of protein-protein interactions . We defined six interface properties for the analysis of biological obligate and biological non-obligate interactions, and crystal packing contacts using a set of hand-curated protein complex data. Based on this analysis, we developed a prediction approach NOXclass using statistical learning methods for inferring protein interaction types. This approach was tested for the classification of the three types of interactions and the results were discussed and compared to those produced by similar methods.

In Chapter 3, we introduce an approach Galinter for aligning non-covalent interactions between different protein-protein interfaces. The method aligns the vector representations of van der Waals interactions and hydrogen bonds based on their geometry. We applied the method to a dataset that comprises a variety of diverse protein-protein interfaces and compared our results to those obtained using two other complementary approaches. In addition, we applied the method to several examples of protein mimicry. Furthermore, a scoring function for measuring statistical significance of the alignments was developed and tested.

In Chapter 4, we present conclusions and give an outlook.

Chapter 2

Characterization and Prediction of Protein-Protein Interaction Types

In this chapter, we describe the characterization and classification of three different types of protein-protein interactions: biological obligate, biological non-obligate, and crystal packing contacts (Zhu *et al.*, 2006). A set of six protein interface properties was compared and analyzed for the three types of interactions. We constructed a classifier based on the analysis of the interface properties using support vector machine (SVM) algorithms. The classifier was trained on a non-redundant dataset and is able to distinguish the three types of protein-protein interactions automatically.

In Section 2.2, we present the characterization of the three types of interactions by using six interface properties. We then describe the classification of the three interaction types in Section 2.3 using SVM algorithms.

2.1 Introduction

2.1.1 Background

Structure models of the protein complexes are necessary for the understanding of biological processes. However, not all interactions observed in protein complex structures are biologically relevant. Many of them are formed during the crystallization process and would not appear *in vivo*. Such crystal packing contacts are non-specific and are not associated with any biological functions (Janin and Rodier, 1995). The annotation for the biological units of proteins are missing in many protein structure models. Therefore, the determination and analysis of the quaternary structure of protein complexes remains a field of active research (Janin *et al.*, 2008).

Meanwhile, there are diverse types of biological interactions. For instance, subunits from obligate complexes do not exist as stable structures *in vivo*, whereas subunits of non-obligate complexes may dissociate from each other and stay as stable and functional units. Similarly, protein complexes have been classified as permanent or transient according to their lifetime. Section 1.1.5 provides more details about the

classification of protein-protein interactions. The interaction type of protein complexes contains important information about protein functions. For example, most transient interactions perform a regulatory role, and non-obligate permanent interactions are often formed in receptor-ligand, enzyme-inhibitor and antibody-antigen complexes (Nooren and Thornton, 2003a). Hence, the determination of protein interaction type is also in the focus of intensive research.

2.1.2 Related Work

Distinguishing Biological and Non-Biological Interactions

A number of studies examined properties of protein-protein interfaces in order to discriminate biologically relevant interactions and non-biological interactions resulting from crystal packing contacts. It has been shown that biological interactions tend to have larger interface sizes than non-biological interactions (Janin and Rodier, 1995; Janin, 1997; Carugo and Argos, 1997; Dasgupta *et al.*, 1997; Henrick and Thornton, 1998; Ponstingl *et al.*, 2000). PQS (Henrick and Thornton, 1998), which uses interface size as one of its main discriminants, identifies true homodimers with an accuracy of 78% on a non-redundant dataset (Valdar and Thornton, 2001). In PQS, a 400 Å^2 cutoff for the interface size between biological interactions and non-biological interactions is used. Ponstingl and coworkers reported an optimal cutoff of 856 Å^2 for differentiating homodimers and monomers (Ponstingl *et al.*, 2000). However, counterexamples were also observed for which this criterion failed (Janin, 1997; Ponstingl *et al.*, 2000). Amino acid composition of the interface is another well-analyzed property for identifying biological interactions (Carugo and Argos, 1997; Jones and Thornton, 1997a; Lo Conte *et al.*, 1999; Bahadur *et al.*, 2004). It has been reported that the amino acid composition of biological interfaces is different from that of the rest of protein surface (Jones and Thornton, 1997a; Lo Conte *et al.*, 1999; Bahadur *et al.*, 2004). On the other hand, Carugo and collaborators showed that the chemical composition of crystal packing contacts is very similar to that of the rest of the surface as a whole (Carugo and Argos, 1997). The importance of residue conservation in the identification of the oligomeric state of protein complexes has been investigated. Using a neural network algorithm for combining the size and conservation measures of the interface, biological homodimeric interactions and crystal packing contacts were successfully classified with an accuracy of 98.3% (Valdar and Thornton, 2001). Zhang et al. introduced machine learning methods to predict protein quaternary structures based on protein sequence information (Zhang *et al.*, 2003). SVMs were trained to separate the primary structures of homodimeric complexes from those of the monomeric proteins. The highest classification accuracy of their SVMs reaches 87.5%.

A related topic of research is the identification of protein-protein interaction sites. The aim of the research is to predict potential biologically relevant interaction sites on the surface of proteins by distinguishing them from the rest part of the surface. Similar interface properties were employed for identifying protein-protein interaction

sites. Jones and Thornton analyzed six physicochemical interface properties (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area) and used them for predicting interaction sites (Jones and Thornton, 1997a,b). In their work, each protein surface was first split into patches. Then, for each surface patch a simple score averaging the six properties was calculated to give the probability of it to form interactions with other proteins. The prediction accuracy of the method reaches 66% on a test set of 59 protein complexes. On the one hand, the analysis results demonstrate that the physicochemical properties are clearly dissimilar between interaction and non-interaction sites on protein surface. On the other hand, it seems that the prediction accuracy might be further improved by combining the interface properties using a more sophisticated algorithm. Gallet et al. identified residues involved in protein interaction sites by analyzing the distribution of hydrophobicity in protein sequences (Gallet *et al.*, 2000). Zhou and Shan (2001) used a neural network algorithm to combine the information about sequence profiles of neighboring residues and solvent accessibility of a target residue to predict protein interaction sites. The prediction accuracy of Zhou's work is 70%. This work again illustrates that to a large degree it is possible to detect the residues involved in protein-protein interactions based on their physicochemical properties. Moreover, the analysis of residue conservation was employed to infer functional hot spots on the protein surface (Lichtarge *et al.*, 1996; Lockless and Ranganathan, 1999; Armon *et al.*, 2001; Ma *et al.*, 2003). The approaches are based on the assumption that key residues involved in biologically relevant interactions are more strongly conserved in evolution than the rest of the protein surfaces. Therefore, the methods distinguish protein interaction sites from the remainder of the protein surface by identifying structurally conserved residues, or functional hot spots on the protein surface. Though several conservation scores have proven useful, there is still room for improvement (Valdar, 2002). Furthermore, it has been shown that machine learning algorithms are effective for improving prediction accuracy of protein interaction sites. For instance, different properties were combined using a SVM implementation in order to predict protein-protein binding sites (Bordner and Abagyan, 2005; Bradford and Westhead, 2005). In the work of Bradford and Westhead (2005), the authors succeeded in predicting protein interaction sites with an accuracy of 76% by combining seven interface properties using a SVM algorithm. Although the interface properties are very similar to those used in Jones and Thornton (1997b), the prediction accuracy has been greatly improved.

Distinguishing Different Types of Biological Interactions

Many efforts have been made to discriminate different types of biological interactions. Transient protein-protein interactions, including both homodimers and heterodimers, were characterized on the structural level (Nooren and Thornton, 2003b). This work revealed that interface properties of transient complexes correlate with their binding affinity. In comparison to "stable" transient dimers (dissociation constant in the nanomolar range), the interfaces of "weak" transient dimers (dissociation constant

in the micromolar range) have a smaller area, and are more planar and polar on average. In addition, interface residues of transient homodimers were found to be more conserved than the other surface residues. Gunasekaran and coworkers divided protein-protein interactions into "ordered" and "disordered" groups. Ordered interactions correspond to non-obligate interactions and crystal packing contacts, whereas disordered interactions include mostly obligate interactions (see Section 1.1.5). They reported that both the per-residue surface area and the interface area of ordered proteins are much smaller than those of disordered proteins (Gunasekaran *et al.*, 2004). De et al. performed a statistical analysis of the interface properties for obligate and non-obligate interactions (De *et al.*, 2005). They reported that obligate interfaces have more contacts than non-obligate interfaces, and these contacts are mainly nonpolar. In addition, involvement of secondary structure elements at interfaces were reported to be significantly different. Mintseris and Weng investigated the difference between obligate and transient complexes from an evolutionary point of view (Mintseris and Weng, 2005). In obligate interactions, interface residues were reported to be significantly more conserved than those in transient interactions. It has also been shown that for obligate complexes, the mutation of interface residues are much more correlated than for transient complexes. In general, obligate and non-obligate proteins have been shown to have distinct interaction preferences. Nevertheless, there is no single interface property with a clear cutoff on whose basis one can discriminate between the different protein interaction types. This is not surprising given the complexity and diversity of protein interactions. Mintseris and Weng used atomic contact vectors (ACVs) to discriminate obligate from non-obligate interactions (Mintseris and Weng, 2003). They achieved respectable accuracy (91%) in such a classification problem.

Related Work Published after NOXclass

There had been considerable progress in the analysis and classification of the different types of interactions when our work was carried out. However, no automatic method had been made available by then for the prediction of protein-protein interaction types, especially for distinguishing obligate, non-obligate and crystal packing interactions. We published our work for the characterization and classification of protein-protein interactions in 2006 (Zhu *et al.*, 2006). Since then, more work has been published on this topic, which we will discuss in Section 2.4.2.

2.2 Characterization of Protein-Protein Interactions

A protein complex may consist of several subunits, resulting in several interactions. For our work, when investigating a protein-protein interaction, we consider only the two subunits involved in the interaction, which normally refer to two polypeptide chains in the protein complex..

As we reviewed in Section 1.1.5, there are different classifications of protein-

protein interactions based on various criteria. In this study, we focused on the effect of the complex formation on the stability of the protein subunits. If the subunits only exist in the protein complex and are not found as stable structures separately, the interaction is obligate, otherwise non-obligate. Together with crystal packing contacts, we considered three types of protein-protein interactions (Figure 2.1). Two types of them are biological (obligate and non-obligate interactions), while the third type is non-biological (crystal packing).

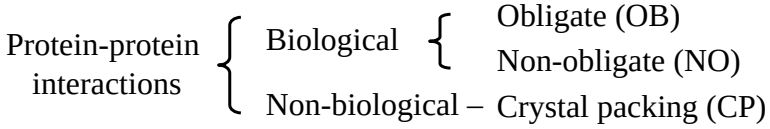


Figure 2.1: Three types of protein-protein interactions considered in the characterization and classification.

2.2.1 Dataset

We compiled a non-redundant data set with three types of protein-protein interactions from several sources. Obligate interactions were taken from a previously compiled set (Bradford and Westhead, 2005). Non-obligate interactions were obtained from both a set of non-obligate interactions (Bradford and Westhead, 2005) and a set of transient interactions (Neuvirth *et al.*, 2004), which are non-obligate by definition. To remove redundancies (Aloy *et al.*, 2003), these interactions were first divided into groups. Each group is defined by the two SCOP families to which the two interaction subunits belong. Then, we selected within each group the interaction whose complex had the highest AEROSPACI score (Chandonia *et al.*, 2004). The AEROSPACI score is a measure of the quality of the structure models available in the PDB (Berman *et al.*, 2000). After removing redundancy, we have 94 obligate interactions and 88 non-obligate interactions. Through visual inspection, we identified problematic cases and removed them from the set. For example, small ligands were found in some interfaces, or there was an interaction between two different parts of the same protein that was cleaved into two chains as a result of proteolysis. In total we removed eight cases from the obligate set (1bbh, 1bft, 1g4y, 1mka, 1nsy, 1scf, 1vfr and 5hvp) and six entries from the non-obligate set (1bpl, 1noc, 1fap, 1bmq 1ef1 and 2kau). Conservation scores of protein sequences were used as one of the interface features. Only for a subset of these interactions we could obtain conservation scores for the subunits involved. In this subset, there are 75 obligate interactions and 62 non-obligate interactions.

Enzyme homodimers predominate in the obligate set, but the set also includes other types of proteins, such as transcription regulators and membrane receptors. The non-obligate set includes many interactions between enzymes and inhibitors, but it also includes other types of interactions, such as different examples of receptor-ligand interactions and transient signaling complexes.

Crystal packing contacts are biologically irrelevant interactions in protein structure models. As described in Section 1.2.2, all contacts present in protein crystalline lattice that are not identified as biological interactions are crystal packing contacts. Using this criterion, we compiled a set of crystal packing contacts from the PDB in two steps. First, we collected a non-redundant set of biological dimers from the PDB, including both homo- and heterodimers. We selected all dimeric complexes as defined in the PDB file sections REMARK 300 and REMARK 350. A similar procedure as described above for reducing redundancy in the obligate and non-obligate datasets was also used to eliminate the redundancy in the crystal packing contact set. Specifically, the dimers were grouped according to the pair of SCOP families to which they belong. For each group, the complexes with AEROSPACI scores below 0.5 were removed. The biological units for the remaining dimers were confirmed by inspecting the relevant literature. Then, for each group the dimer with the highest AEROSPACI score was selected. In total we collected 120 dimers, for which we rebuilt unit cells and chose the largest non-biological interface in each unit cell for our final set of crystal packing contacts. In these 120 crystal packing contacts, only 106 of them, for which we could obtain conservation scores using ConSurf (Armon *et al.*, 2001), were retained.

In total, we gathered 243 protein-protein interactions of which 75 are obligate interactions, 62 are non-obligate interactions and 106 are crystal packing contacts. We will refer to this final dataset as BNCP-CS. The PDB ids and the interacting chain names are listed in Table 2.1.

2.2.2 Definition of Interfaces Properties

In this work, a residue is defined as being part of the interface if its solvent accessible surface area (SASA) decreases by more than 1 Å^2 upon the formation of the complex (Jones and Thornton, 1996). Solvent accessible surface areas for residues were calculated using NACCESS (Hubbard and Thornton, 1993), with a probe sphere of radius 1.4 Å. A protein-protein interface is defined to be the ensemble of all interface residues from both subunits.

Many physicochemical properties and their derivatives or combinations were used to study protein interfaces. We chose the set of interface properties based on two criteria: 1) previous relevant studies have shown that such properties are characteristic for different types of protein-protein interactions; 2) the properties have clear physicochemical meanings. We selected the following six interface properties and included them in the characterization of the interactions:

1. interface area (IA)
2. ratio of interface area to protein surface area (IAR)
3. amino acid composition of the interface (AAC)

Table 2.1: Dataset **BNCP-CS**[a]

Obligate Interactions (75)[b]

1ahjAB	1b34AB	1dceAB	1efvAB	1guxAB	1h2aLS	1lucAB
1pnkAB	1reqAB	1tcoAB	2aaiAB	1a0fAB	1a4iAB	1afwAB
1aj8AB	1ajsAB	1aomAB	1aq6AB	1at3AB	1b3aAB	1b5eAB
1b7bAC	1b8aAB	1b8jAB	1b9mAB	1bjnAB	1bo1AB	1brmAB
1byfAB	1bykAB	1c7nAB	1cliAB	1cmbAB	1cnzAB	1cozAB
1cp2AB	1dorAB	1f6yAB	1gpeAB	1hgxAB	1hjrAC	1hssAB
1isaAB	1jkmAB	1kpeAB	1mspAB	1nseAB	1oneAB	1pp2LR
1qaeAB	1qaxAB	1qbiAB	1qfeAB	1qfhAB	1qorAB	1qu7AB
1smtAB	1soxAB	1spuAB	1trkAB	1vltAB	1vokAB	1wgjAB
1xikAB	1xsoAB	1ypiAB	1yveIJ	2ae2AB	2hdhAB	2hhmAB
2nacAB	2pflAB	2utgAB	3tmkAB	4mdhAB		

Non-obligate Interactions (62)

1avaAC	1avwAB	1bvnTP	1cseIE	1eaiCA	1f34AB	1fssAB
1glaFG	1kxqHA	1smpIA	1tabIE	1tgsIZ	2ptcIE	2sicIE
4sgbIE	1agrEA	1atnAD	1b6cAB	1bkdRS	1buhAB	1dowAB
1euvAB	1i2mAB	1i8lAC	1kacAB	1pdkAB	1qavAB	1tx4AB
1c0fSA	1zbdAB	1ak4AD	1d09AB	1cqiAB	1finAB	1dhkAB
1bi7AB	1wq1RG	1rrpAB	1cc0AE	1eg9AB	1avzBC	1frvAB
3hhrAB	1ycsAB	1cvsAC	1aroLP	1cmxAB	1bmlAC	2pcbAB
1f60AB	1stfEI	1emvAB	1ueaAB	1qbkBC	1hluAP	1itbAB
1ethAB	1jtdAB	1lfdAB	1dn1AB	1tmqAB	1a4yAB	

Crystal Packing Contacts (106)[c]

1k55	1ual	1mxr	1j98	1e9g	1iup	1is3	1gy7	1jzl	1jke
1km1	1ihr	2btc	1eq9	1qf8	1k8u	1m7g	1p5z	1e19	1k75
1iat	1m9f	1ht9	1hqs	1b8z	1lc5	1gs5	1gve	1k20	1i4u
1k9u	1e58	1es9	1qkm	1j8b	1kli	1eyv	1j24	1h1y	1ijy
1exq	1lw6	1m7y	1n3l	1nms	1pe0	1f6b	1jp3	1kqp	1j79
1mxi	1my7	1k4i	1jat	1f1m	1jd0	1nrv	1mvo	1m2d	1f7z
1gyo	1fs8	1b67	1kzk	1nxm	1k94	1i0r	1euv	1ql0	1g2y
1mh9	1ed9	1dtd	1ld8	1jlt	1ct4	1nsz	1iq6	1i2m	1lqp
1lqv	1n2e	1i12	1ubk	1g8q	1e87	1jl0	1jr8	1qip	1nf9
1g60	1uaq	1ozu	1dmh	1eye	1i52	1fjj	1b16	1e4m	3lyn
1ock	1icr	1i0d	1jtg	1elu	1kic				

[a]One PDB entry can contain several interfaces of different types. Therefore, the same PDB entry can appear in different subsets. For example, 1i2m has a non-obligate interaction between chains A and B. At the same time, the contact between chains B and D is included under the crystal packing contact subset.

[b]For obligate and non-obligate interactions, the PDB codes and the names of the interacting chains are given.

[c]For crystal packing contacts, only the PDB codes are given. From each of these PDB structures, the largest interface that does not belong to any BU is chosen as the crystal packing contact in the PDB structure.

4. correlation between the amino acid compositions of the interface and the protein surface (COR)
5. gap volume index (GVI)
6. conservation score of the interface (CS)

In addition, we studied further interface features, including interface hydrophobicity, secondary stucture composition at the interface, the number of hydrogen bonds at the interface. However, these features were not found to be discriminating based on our dataset and were thus not included in the final analysis and classifier.

Interface area

The interface area is defined as one half of the total decrease of SASA (ΔSASA) of the two subunits upon the formation of the interaction:

$$\mathrm{IA} = \frac{1}{2}\left(\mathrm{SASA}_a + \mathrm{SASA}_b - \mathrm{SASA}_{ab}\right)$$

where a and b are two subunits in the complex ab; SASA_a, SASA_b and SASA_{ab} are the SASA values for a, b, and ab, respectively.

Interface area ratio

Protein interactions that involve a small subunit cannot have large interface areas. This applies to some enzyme-inhibitor complexes, for instance. Therefore, we defined interface area ratio, in which the interface area is normalized by the SASA of the smaller subunit in the complex:

$$\mathrm{IAR} = \frac{\mathrm{IA}}{\min(\mathrm{SASA}_a, \mathrm{SASA}_b)}$$

where SASA_a and SASA_b are the SASA values for subunits a and b, respectively.

Amino acid composition of the interfacey

We calculated both a number-based amino acid composition and an area-based amino acid composition of interfaces (Bahadur *et al.*, 2004). The number-based amino acid composition (AACn) is defined as the frequency of each type of the 20 standard amino acids in the protein-protein interface. By weighting each residue with its ΔSASA, the area-based amino acid composition (AACa) is computed:

$$\mathrm{AACa}[i] = \frac{1}{2\,\mathrm{IA}} \sum_{\mathrm{type}(r)=i} \Delta\mathrm{SASA}(r), \quad i = 1 \ldots 20$$

where $\mathrm{type}(r)$ returns the type of the amino acid of residue r, encoded as integers from 1 to 20; $\Delta\mathrm{SASA}(r)$ returns the ΔSASA value of residue r.

In order to compare the AACn and AACa of the three types of interactions, we used two different measures to assess the similarity. Given two AACs u and v (either AACn or AACa), the Δv distance[1] between them is defined as (Lo Conte *et al.*, 1999; Bahadur *et al.*, 2004):

$$\Delta v\_\text{dist}(u, v) = \sqrt{\frac{1}{19}\sum_{i=1}^{20}(u_i - v_i)^2}$$

The second measure is the Pearson's correlation coefficient between two AACs u and v. It is defined as

$$\text{Cor}(u, v) = \frac{\sum_{i=1}^{20}(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{20}(u_i - \overline{u})^2}\sqrt{\sum_{i=1}^{20}(v_i - \overline{v})^2}}$$

where $\overline{u}$ and $\overline{v}$ are the means of u and v, respectively.

Correlation between amino acid compositions of interface and protein surface (CORn and CORa)

The AAC of the biological interface was shown to be significantly different from that of the rest of the protein surface (Ofran and Rost, 2003). It is reasonable to expect the AAC of the crystal packing interface to be similar to that of the rest of the protein surface. To measure this effect, the Pearson's correlation coefficients between the amino acid compositions of the interface and the surface were calculated. These correlations were calculated for both AACn and AACa, resulting in a number-based correlation (CORn) and an area-based correlation (CORa).

Gap volume index

The gap region between two interacting proteins was detected using the SURFNET program (Laskowski, 1995). The volume of the gap region was calculated as the gap volume (GV) of the interaction using the same program. It has been shown that the protein-protein interfaces are more complementary in obligate complexes than those in non-obligate complexes (Jones and Thornton, 1996; Bahadur *et al.*, 2004). The GVI is one of the measurements for interface complementarity (Bahadur *et al.*, 2004). Since the gap volume is related to interface area, we normalized the GV between subunits by the corresponding IA and define the gap volume index as:

$$\text{GVI} = \frac{\text{GV}}{\text{IA}}$$

[1]This measure is named "a Euclidean distance" between AACs in Lo Conte *et al.* (1999). However, the formula given in Lo Conte *et al.* (1999) does not agree with the well accepted definition for Euclidean distance $d = \sqrt{\sum_i (u_i - v_i)^2}$. It is more like the definition for the unbiased estimator of the variance (DeGroot and Schervish, 2001). Therefore, we do not use the misleading term "Euclidean distance" for the measure in our work. At the same time, to follow the convention we still use the same formula and define it as a "distance".

The smaller the GVI, the more complementary the interface shapes are. The GV was computed using the SURFNET program. The minimum and maximum radius for gap spheres were set to 1.0 and 5.0 Å, respectively. The grid separation was set to 2.0 Å.

Conservation score of the interface

We calculated the conservation scores for residues in the interface as determined by the ConSurf method (Armon *et al.*, 2001). The conservation score calculated by ConSurf is a normalized measure. The average score of all residues in a target protein is zero, and the standard deviation is one. Negative conservation scores are indicative of slowly evolving, conserved sites, and positive conservation scores suggest rapidly evolving, variable sites. The ConSurf conservation score is a relative measure of evolutionary conservation for each position of the amino acid sequence of the target chain. The lowest score represents the most conserved position in the target protein.

The conservation score for an interface (CSn) was defined as the average value of the conservation scores for all the residues at the protein-protein interface. In a similar way to the AACa, we weighted the conservation score for each residue by its ΔSASA upon the formation of the interaction. The average of these weighted residue conservation scores was used as the area-based conservation score of the interface (CSa).

Table 2.2: List of interface properties

AACa	Amino Acid Composition of the interface, area-based
AACn	Amino Acid Composition of the interface, number-based
CORa	CORrelation between amino acid compositions of the interface and the surface, area-based
CORn	CORrelation between amino acid compositions of the interface and the surface, number-based
CSa	Conservation Score of the interface, area-based
CSn	Conservation Score of the interface, number-based
GV	Gap Volume
GVI	Gap Volume Index
IA	Interface Area
IAR	Interface Area Ratio
SASA	Solvent Accessible Surface Area

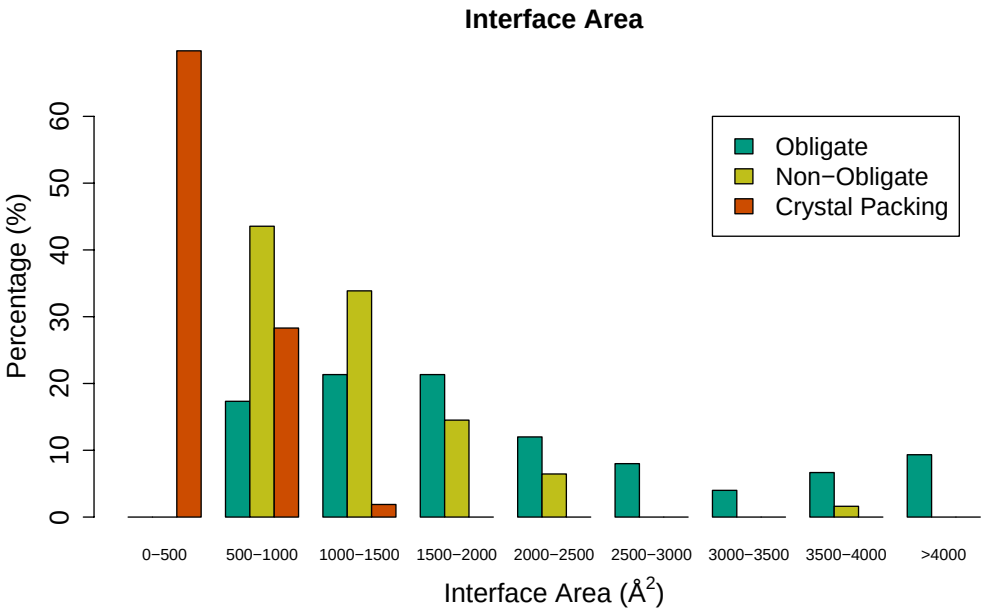


Figure 2.2: Distribution of interface area for the three types of interactions in the BNCP-CS dataset.

2.2.3 Analysis of Interface Properties

The six interface properties for the 243 interactions in the BNCP-CS dataset were calculated and presented in the following paragraphs.

Interface Area

The histogram of IAs for the three types of interactions in the BNCP-CS dataset is shown in Figure 2.2. The average values of IA for obligate, non-obligate and crystal packing interactions are 2156.5 Å^2, 1170.7 Å^2, and 435.9 Å^2, respectively. The three types of interactions exhibit considerable differences regarding this property. The distribution of obligate IAs has the largest variance among the three sets with a spread from 500 Å^2 to more than 4000 Å^2. Biological interactions, including both obligate and non-obligate interactions, exhibit clearly a larger IA than non-biological interactions. Biological interfaces can have an area as large as 4000 Å^2. Most non-obligate interactions have an IA between 500 and 2500 Å^2. In contrast, all crystal packing IAs are smaller than 1500 Å^2, with most of them being smaller than 500 Å^2. We observed that the majority of biological interactions have an IA of more than 650 Å^2, and most crystal packing contacts have less than 650 Å^2 interface area. In total, only 7% of the interactions in the BNCP-CS dataset do not follow the rule. Similar distributions of interface area values have been observed in previously published work (Ponstingl *et al.*, 2000; Bahadur *et al.*, 2004).

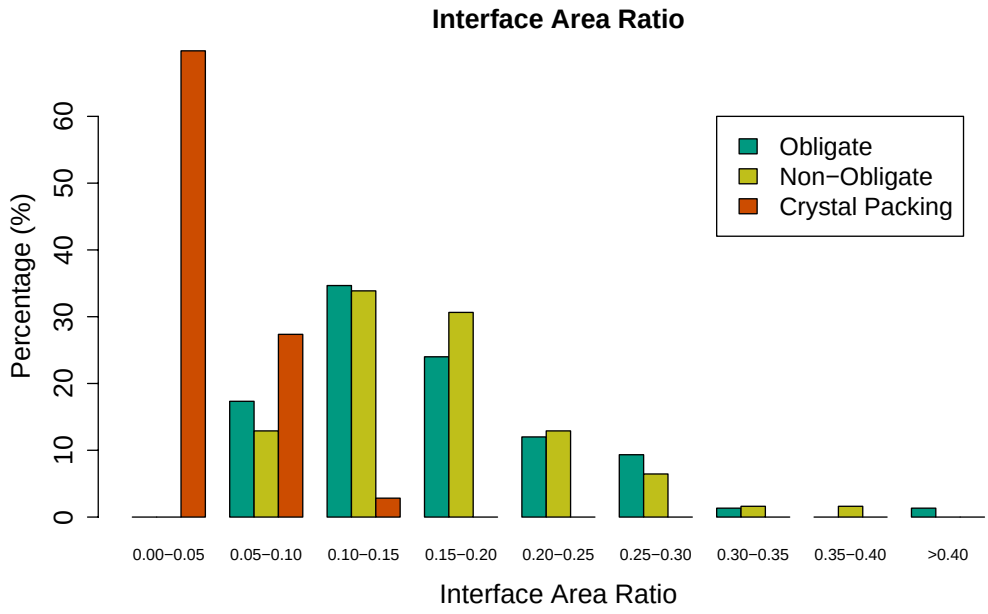


Figure 2.3: Distribution of interface area ratio for the three types of interactions in the BNCP-CS dataset.

Interface Area Ratio

The distribution of IARs for the BNCP-CS dataset is shown in Figure 2.3. The average values of IAR for obligate, non-obligate and crystal packing interactions are 0.16, 0.17, and 0.05, respectively. While the distributions of obligate and non-obligate interactions are similar, both are considerably different from the distribution of the crystal packing contacts. Compared to the distribution of IA values, the majority of IAR values of biological interactions have shifted away from the IAR values of crystal packing contacts. This phenomenon suggests that the large IAs observed in some of the crystal packing contacts result mostly from the large subunits in the protein complexes. Similarly, the reason for the reduced difference in the distributions of IAR values between obligate and non-obligate interactions in Figure 2.3 in contrast to Figure 2.2 is that the large IAs of obligate interactions are mainly caused by subunits with large surface areas.

Amino Acid Composition of the Interface

The overall AACa of the interfaces for the three types of complexes in the BNCP-CS dataset is reported in Figure 2.4. Hydrophobic residues (FILV) contribute twice as much area to obligate interfaces as to crystal packing contacts. For instance, on average each of the amino acid leucine contributes 46.1 Å^2 and 39.5 Å^2 to the interface area in obligate and non-obligate interactions, respectively. In contrast, in crystal packing interfaces leucine contributes only around 25.9 Å^2 to the interface area. Charged residues (EKR) also show different distributions in the obligate and

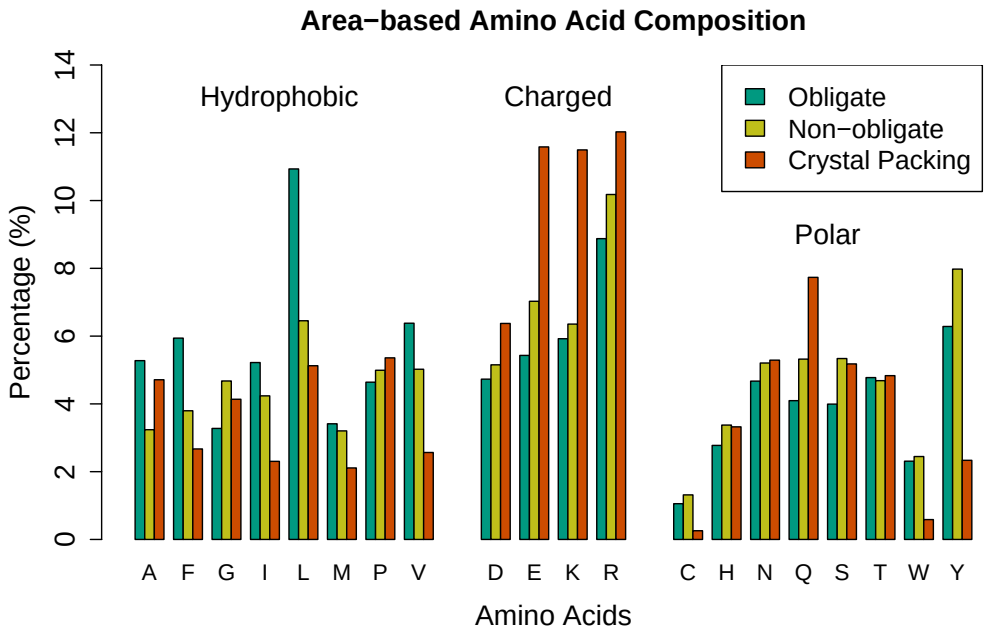


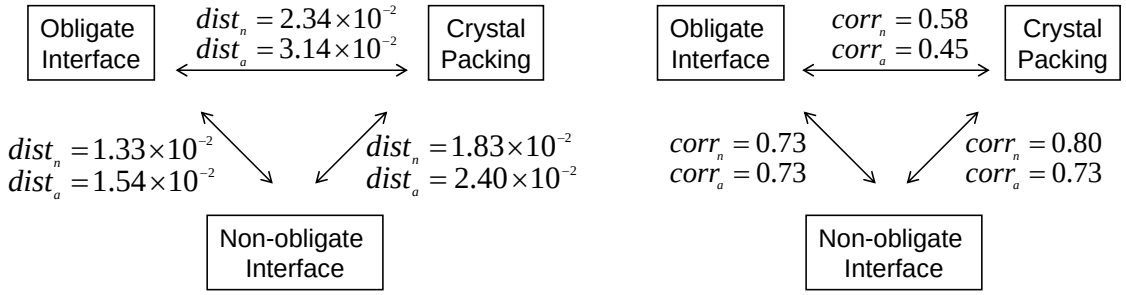
Figure 2.4: Area-based amino acid composition for the three types of interactions in the BNCP-CS dataset.

crystal packing interfaces. Aromatic residues (FWY) tend to be more abundant in biological interfaces. We observed that cysteine occurs more often in the biological interfaces than in crystal packing contacts. The difference between the AACs of the three types of interactions have been compared in terms of Δv distances and correlation coefficients (Figure 2.5). Both the AACn and AACa have been used. The lower correlation values and the larger Δv distance values of AACa indicate that it is a better discriminant than AACn for differentiating the three types of interactions in our study.

Similar conclusions have been reported previously (Jones and Thornton, 1997a; Bahadur *et al.*, 2004). These results also indicate that non-obligate interfaces exhibit intermediate values with respect to AACn and AACa between obligate interactions and crystal packing contacts. This is particularly true for the sets of hydrophobic and charged residues.

Correlation between Amino Acid Compositions of Interface and Protein Surface

Correlation coefficients calculated using both AACn and AACa are reported in Figure 2.6. The average correlation coefficients for obligate, non-obligate and crystal packing interactions from the BNCP-CS dataset are 0.35, 0.47, and 0.49, respectively, using number-based composition. These average values are 0.39, 0.48, and 0.59 when using area-based composition. Previous investigations have reported similar results (Lo Conte *et al.*, 1999; Bahadur *et al.*, 2004; De *et al.*, 2005). Again,



a) Δv Distance b) Correlation Coefficient

Figure 2.5: Measures of similarity between amino acid compositions. Both Δv distances (a) and Pearson's correlation coefficients (b) are calculated for every pair of the three interaction types. Measures $dist_n$ and $corr_n$ are Δv distance and correlation coefficient calculated based on AACn, while $dist_a$ and $corr_a$ are calculated based on AACa.

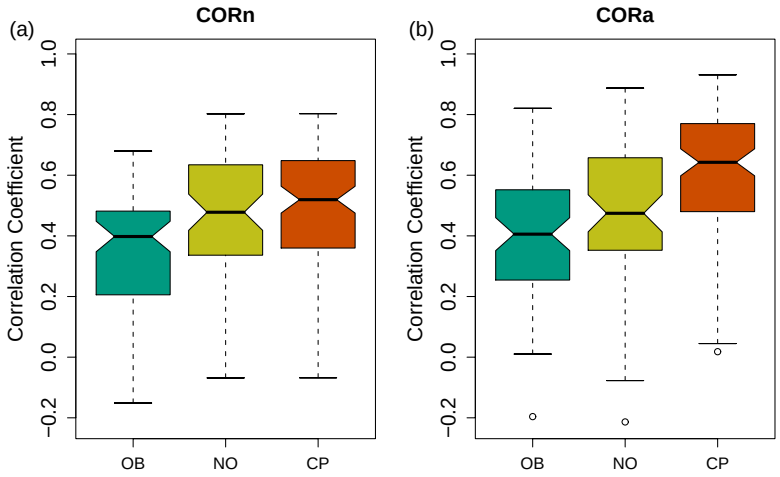


Figure 2.6: Correlation coefficients between amino acid compositions of interface and protein surface for the three types of interactions in the BNCP-CS dataset, calculated using number-based composition (a) and area-based composition (b).

non-obligate interactions exhibit intermediate characteristics. The discrimination is more pronounced for area-based correlation.

Gap Volume Index

GV and GVI values are presented in Figure 2.7. It is shown in Figure 2.7a that obligate and non-obligate interactions tend to have larger GV values than crystal packing contacts. Obligate and non-obligate interactions have much smaller GVI values than crystal packing contacts (Figure 2.7b). On average, the GVIs are 4.0, 5.3, and 13.8 for obligate, non-obligate interactions, and crystal packing contacts, respectively. The property GVI discriminates better the three kinds of interactions

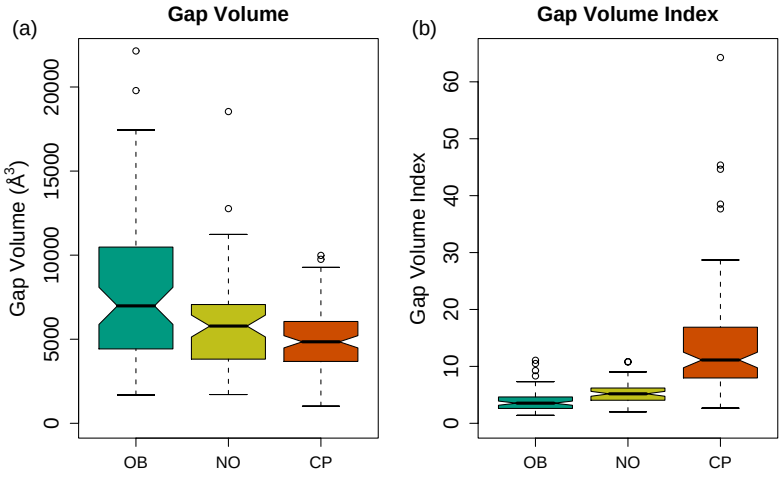


Figure 2.7: Gap volumes (a) and gap volume indices (b) for the three types of interactions in the BNCP-CS dataset

than the property GV.

Conservation Score of the Interface

Figure 2.8 illustrates that interface residues in obligate and non-obligate interactions show a higher degree of conservation than those in crystal packing contacts. Average CSa values for obligate and non-obligate interfaces are -0.07 and 0.02, respectively. In contrast, the average CSa for crystal packing interfaces is 0.44. These results agree with previous observations that interface residues in biological interactions are conserved more strongly (Lichtarge *et al.*, 1996; Lockless and Ranganathan, 1999; Armon *et al.*, 2001; Ma *et al.*, 2003).

As shown in Figure 2.9, conserved residues in biological interfaces are more involved in the formation of protein interfaces (high ΔSASA) than those in crystal packing contacts that exhibit the same degree of conservation. The effect is more pronounced with an increasing degree of conservation. On average, the ΔSASA for the most conserved residues (discretized conservation score equals 9) is 37.6 Å^2 and 32.6 Å^2 for obligate and non-obligate interactions, respectively, but for crystal packing contacts this value is only 18.6 Å^2.

Relationship between Interface Properties

Scatter plots comparing different interface properties are presented in Figure 2.10. In general, the values for the properties associated with crystal packing contacts tend to be distinct from the values of the two types of biological interactions, especially in the plots where IA and IAR are considered (see also Figure 2.2 and 2.3). These observations suggest that crystal packing contacts are often dissimilar from biological interactions in terms of IA and IAR features. For obligate and non-obligate inter-

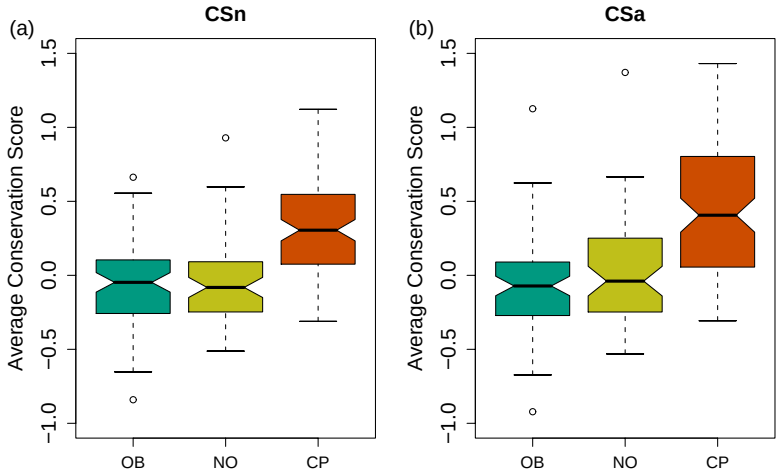


Figure 2.8: Conservation scores of the interfaces for the three types of interactions in the BNCP-CS dataset, calculated using number-based strategy (a), and area-based strategy (b). Lower conservation scores indicate higher degrees of conservation.

actions, the corresponding data points scatter in similar religions in most plots and are not so clearly separated. Data points representing obligate interactions occupy wider area than non-obligate interactions on the Y-axes in Figure 2.10 a, b, c and d.

We also noticed that the IA and IAR values are correlated in all types of interactions (Figure 2.10 a). Normally, interactions with larger interface area also exhibit larger IAR values, or also have a larger portion of protein surfaces involved in the interactions. However, when the IA values are similar, obligate interactions tend to have smaller IAR values than non-obligate interactions. For crystal packing contacts, the ratio between their IA and IAR values is similar to that of non-obligate interactions, though the absolute values of IA and IAR are both smaller. One reason is that the subunits in the obligate interactions have larger surface areas than those involved in the other two types of interactions. Another explanation is that the proteins involved in obligate interactions have a significantly larger per-residue interface area than the proteins involved in non-obligate interactions or crystal packing contacts (Gunasekaran *et al.*, 2004). Namely, the interface residues of obligate interactions are more buried than those of non-obligate interactions or crystal packing contacts.

The property GVI shows a distinct distribution for biological interactions and crystal packing contacts (see column 4 of Figure 2.10). But for obligate and non-obligate interactions, the difference between their GVI values are very small.

In the sub-figures from column 3 and 5 of Figure 2.10, points representing obligate and non-obligate interactions are hardly distinguishable with respect to their X coordinates. This suggests that CORa and CSa are poor features for discriminating the two types of biological interactions. In general, the data points corresponding to crystal packing contacts in these sub-figures have slightly larger CORa and CSa values.

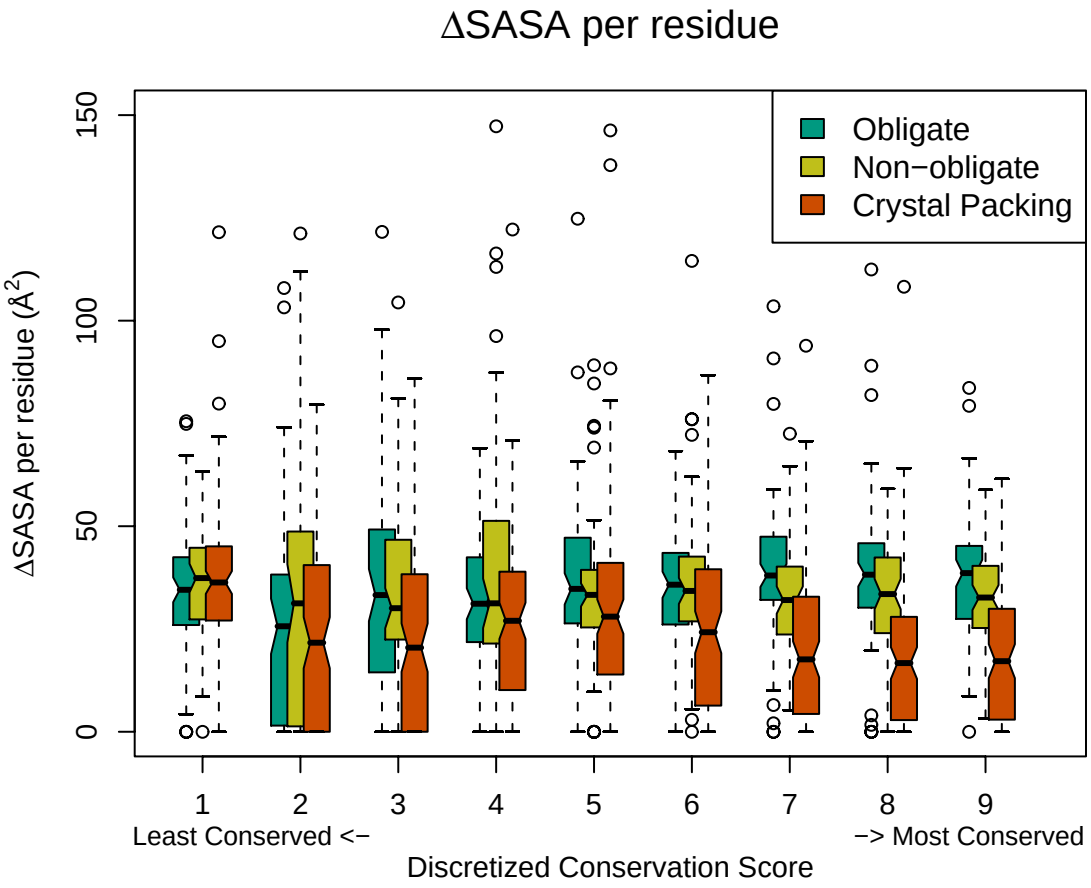


Figure 2.9: Average ΔSASA per residue for different degrees of conservation. Conservation scores from ConSurf are discretized using the same coloring scheme as that used in (Glaser *et al.*, 2003). The larger the discretized ConSurf scores, the more conserved the residues in evolution. The conserved residues tend to be more strongly involved in the biological interfaces.

2.3 Classification of Protein-Protein Interactions

In this section, we discuss the classification of the three types of interactions based on the analysis of their interface properties in the previous section. As shown in Figure 2.10, no single feature or simple combination of features is capable of separating the three types of interactions from the BNCP-CS dataset. We thus employed machine learning methods to combine the features for classifying these interactions.

2.3.1 Machine Learning Techniques related to Classification

With the increasing amount of data needed to be processed, machine learning techniques are more and more effective in many research fields. With the rapid progress in algorithmics and computational power, machine learning techniques have become

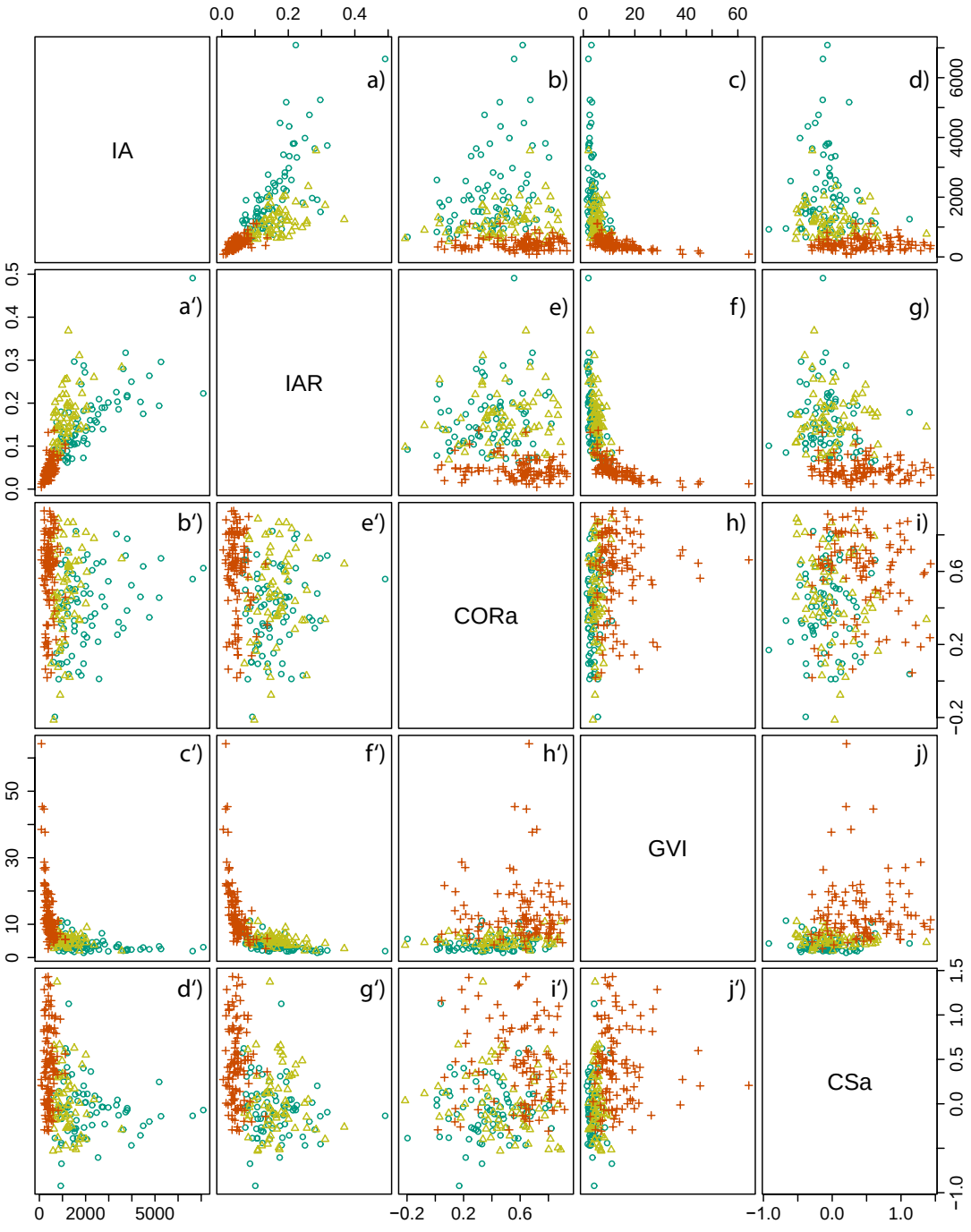


Figure 2.10: Scatter plots for the three types of interactions in the BNCP-CS dataset. All the 243 protein-protein interactions are displayed in each scatter plot. Each point stands for one interaction with respect to the two features considered in the plot. Blue circles stand for obligate interactions; yellow triangles stand for non-obligate interactions; and red crosses represent crystal packing contacts.

indispensable in many areas. Bioinformatics is one of these areas where machine learning techniques have been widely applied to various topics such as genomics, proteomics, microarray analysis, text mining, etc. (Larrañaga *et al.*, 2006).

One of the most important applications of machine learning techniques is *classification*. In classification, individual items are assigned to different categories based on the the characteristics inherent in the items (sometimes referred to as featuers, traits or characters) (Duda *et al.*, 2001). Classification is usually divided into two types, *unsupervised classification* and *supervised classification*. Normally, unsupervised classification methods like clustering are employed when *labels/classes* for individual items are unknown. Otherwise, supervised classification methods can be applied for inducing classification rules. In this study, the labels, namely, the types for individual protein-protein interactions are known. Thus, we employed only supervised classification methods. At the beginning of this section, we briefly introduce some relevant techniques for supervised classification. We have tested two machine learning algorithms, which were SVM and random forests, for the classification of interaction types.

Support Vector Machine

SVMs are a set of supervised learning methods for classification or regression (Vapnik, 1995, 1998). In classification problems, a linear classifier assigns an item into a class based on the value of a linear combination of the item's features. Binary class data are not always linearly separable in their input space. SVMs will project the data points (training data) into a higher dimensional space (feature space) by using a mapping function Φ. In the feature space, a hyperplane with the maximum *margin* is chosen to separate the two classes of data. The margin is defined as twice the distance from the separating hyperplane to the nearest data point (see Figure 2.11). The data points on the border of the margin are called *support vectors* because these points determine the orientation of the separating hyperplane.

In practice, a *kernel function* $K(u, v) = \phi(u)^T\phi(v)$, where u and v are two feature vectors, is commonly used so that the computation of the separating hyperplane is possible without explicitly mapping the training date into the feature space. Four commonly used basic kernel functions are:

- linear: $K(u, v) = u^T v$
- polynomial: $K(u, v) = (\gamma u^T v + r)^d,\ \gamma > 0$
- radial basis function (RBF): $K(u, v) = e^{-\gamma\|u-v\|^2},\ \gamma > 0$
- sigmoid: $K(u, v) = \tanh(\gamma u^T v + r)$

where γ, r and d are kernel parameters.

The final SVM classifier is dependent on the training data and is able to separate them without error. However, a perfect separating hyperplane might result in a classifier of a too high complexity, i.e., the hyperplane is only found in a very high dimensional feature space. In addition, the final classifier might have a high generalization error when applied to unseen data. This is called *overfitting* in machine

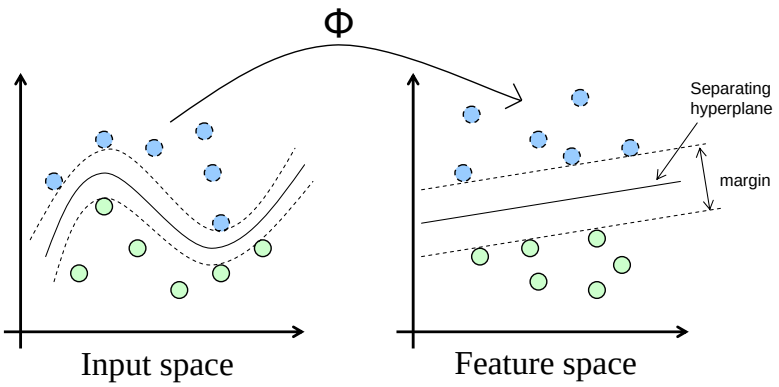


Figure 2.11: Principle of SVM. Data points are projected from input space into higher dimensional feature space by certain mapping function Φ. A hyperplane is chosen to separate the data points in the feature space.

learning, which results from fitting the statistical model with too many parameters. Such an overfitted model might perfectly explain the training data. But its ability to generalize beyond the training data is often reduced.

To address these issues, a *soft margin* is commonly used in the construction of SVMs. With soft margin, misclassifications are tolerated to some extent in the training of the SVMs. The final classifier is optimized for a tradeoff between maximizing the margin and minimizing the misclassification error.

For more than two classes of data, multi-class techniques are required. These techniques include "one-against-one" and "one-against-all" approaches (Hsu and Lin, 2002). In these approaches, several binary SVM classifiers are constructed and the appropriate class is determined using a majority voting scheme. An alternative approach is a multi-stage classifier that separates the data in a progressive manner. The classification is performed in several stages, and in each stage one class of data is separated.

Random Forests

The random forests method is another popular algorithm for classification and regression. The method constructs a number of decision trees (thus a *forest*), each of which is constructed based a random sample of input training data. The output is the aggregation of the predictions from all individual trees in the forest (majority vote for classification, or average for regression) (Ho, 1995; Breiman, 2001). The random forests algorithm has been shown to present good performance compared to other methods including SVM, and to be robust against overfitting (Breiman, 2001). Furthermore, it is usually possible to obtain measures of importance for predictor variables, which is very helpful for the interpretation of the whole model. In random forests, there are two ways to generate variable importance. To estimate the importance of a variable m, the first way is to permute the values of variable m. Then the decrease in prediction accuracy is collected from the decisions trees in the random

forest as the estimation for the importance of variable m (*permutation importance*). The other way is based on the Gini impurity index, which measures the impurity of a set (Duda *et al.*, 2001). In each split of a node in a decision tree in the random forests, the Gini impurity index decreases in the descendant nodes. By adding up for variable m the decrease of the Gini impurity index over all trees in the random forests, the importance of variable m is estimated (*Gini importance*) and is usually consistent with the permutation importance measure.

Cross-Validation

Cross-validation is a widely used method for estimating the prediction error. Ideally, when there are enough data, we can use part of them as the training set for fitting a classification model, and use the rest as the test data for assessing the performance of the model. In practice, the size of data is often too small. To handle this, K-fold cross-validation is usually used for estimating the prediction error of the model. This method operates by first splitting given data into K parts of equal sizes. The kth part is then set aside and the remaining $K-1$ parts are used to train the model. The prediction error of the trained model is assessed by applying the model to the kth part of the data. This process is repeated for $k = 1, 2, ..., K$ (Figure 2.12). The combination of all the K assessments is reported as the estimation of the prediction error. When K equals to the size of the data, the method becomes *leave-one-out cross-validation* (LOOCV).

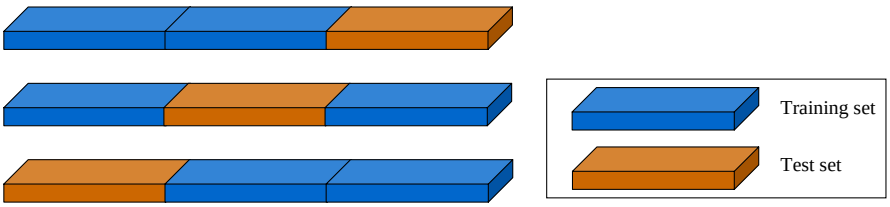


Figure 2.12: Illustration of a K-fold cross-validation ($K = 3$ here).

Ruschhaupt *et al.* (2004) described a more sophisticated approach for estimating misclassification rate (see Figure 2.13). An example protocol of the approach is summarized as follows:

1. Divide the whole data into three parts (say, part A, B, and C) using stratified sampling, so that each class of data is roughly evenly distributed in the three parts;
2. Take part A and B of the data, train the models and optimize parameters using 10-fold cross-validation;
3. Test the models on data part C;
4. Repeat step 2 and 3 twice by selecting different training and test parts;
5. Repeat step 1–4 five times.

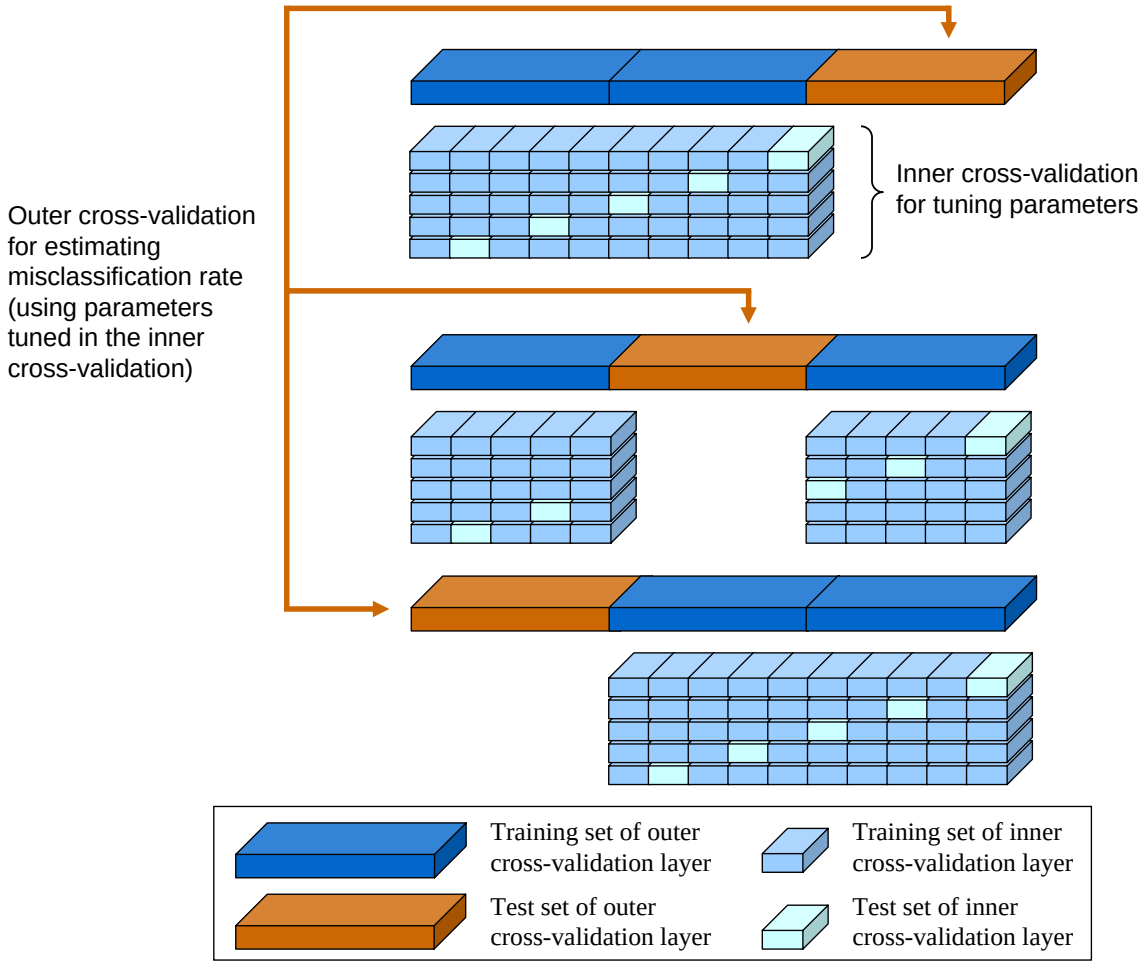


Figure 2.13: Nested cross-validation. Nested cross-validation is composed of two layers: inner cross-validation for tuning parameters, outer cross-validation for estimating misclassification rate. Figure adapted from Ruschhaupt *et al.* (2004)

Feature Selection

Feature selection is a technique commonly used in machine learning for selecting a subset from a set of candidate features such that the learning models based on the subset performs best under some classification system. Feature selection is an important step in machine learning because it can not only reduce the complexity of models, but also improve the generalization ability and interpretability of models. Ideally, the optimal subset of features can be selected via an exhaustive search of all possible combinations of candidate features. However, this is normally computationally intractable in practice since the number of feature combinations to be tested is too large. In such cases, alternative search strategies have to be employed for selecting a well-performing feature subset without carrying out an exhaustive search. A wide range of search strategies has been proposed, including *complete*, *heuristic* and

randomized approaches (Dash and Liu, 1997; Kohavi and John, 1997). Examples of complete search strategies are branch and bound and best first search. In the class of heuristic approaches, popular methods include sequential forward selection and sequential backward selection. Randomized approaches include genetic algorithm and simulated annealing.

Principal Component Analysis

Principal component analysis (PCA) is a technique for revealing the internal structure of given data. This method reveals the directions along which the variances of data are maximal in data space. The vectors representing these directions are called the principal components of the data. PCA is often used to reduce the dimensionality of multidimensional data. It operates by considering only the low-ordered (or the largest) principal components, which contribute most to the variance of the data.

2.3.2 Classification Methods

We mainly applied SVM algorithms on our multi-class data to classify the three types of interactions. In addition, we also tested a random forests algorithm for the same classification purpose.

Support Vector Machine Classifiers

In this work, we used both a multi-class SVM classifier based on "one-against-one" technique and a multi-stage SVM classifier.

Multi-Class SVM Classifier (MCC) The three types (OB, NO, and CP) of interactions are presented to SVM as labeled data with six interface features. The multi-class SVM classifier produces classification using several binary "one-against-one" classifiers and a majority voting scheme.

Multi-Stage SVM Classifier (MSC) The three types of interaction data can be organized into two categories: biological (including OB and NO) and non-biological (CP) as depicted in Figure 2.1. Based on this property of the three interaction types, we have designed a multi-stage SVM classifier, or a two-stage classifier in our work, which is composed of two binary SVMs, one at each stage. In the first stage, the first SVM (SVM1) separates non-biological contacts (CP) from biological interactions. Then, putative biological interactions were passed to the second stage (SVM2), where obligate and non-obligate complexes were distinguished (Figure 2.14).

Implementation of SVMs The R package `e1071` (R Development Core Team, 2005; Dimitriadou *et al.*, 2005) interfacing to *libsvm* (Chang and Lin, 2005) was

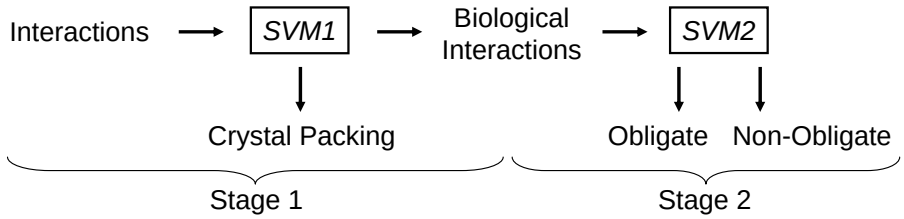


Figure 2.14: Schematic plot of the two-stage SVM. If an interaction is classified as crystal packing by SVM1, it will not be considered by SVM2; otherwise it is classified by SVM2 as either obligate or non-obligate interaction.

used to perform the SVM classification. RBF kernel was adopted in the SVM classifiers due to its superior performance compared to the other three kernels (linear, polynomial and sigmoid kernel).

Parameter Optimization To achieve best performance, we optimized parameters of both the multi-class and two-state SVN classifiers. The parameter *gamma* in the RBF and the cost parameter C were tuned using the build-in function `tune` in R package `e1071`. The kernel parameter *gamma* in RBF regulates the width of the Gaussian kernel. The cost parameter C controls the amount of penalty to misclassification errors. The larger the C value, the more penalized the misclassifications are. We performed a recursive grid search for the best parameters using a leave-one-out cross-validation procedure. The parameter search stops when the improvement of accuracy is less than 0.1%.

Classification Probability We obtained posterior probabilities for our classification with the same R package. It fits a logistic distribution to the pairwise classification decision values using a maximum likelihood algorithm (Chang and Lin, 2005). With this fitted distribution the posterior pairwise class probabilities are estimated for each prediction.

Random Forests Classifiers

We tested the random forests (RF) method in the classification of the three types of interactions. Different feature combinations have been used in the construction of random forests classifiers. We used the random forests program provided in the R package `randomForest` (Liaw and Wiener, 2002) to carry out the test. LOOCV was employed to assess the performance of the classifiers.

2.3.3 Performance Measures

For a classification problem of two classes of data, each instance in the data has a label that is one element of the set $\{p, n\}$ of *positive* and *negative* class labels. A classification model (or a classifier) is a mapping from input data to predicted classes.

We use the labels $\{Y, N\}$ for the predictions of *positive* and *negative* classes produced by the classifier. Given a classifier and an instance in the data, there are four possible outcomes. If the instance is positive (p) and it is classified as positive (Y) by the classifier, it is considered a *true positive* (TP); if it is classified as negative (N), it is considered a *false negative* (FN). Similarly, the outcome may be *true negative* (TN) or *false positive* (FP) (see Table 2.3).

Table 2.3: Definitions of notions TP, FN, FP, and TN

		Predicted Class	
		Y[a]	**N**
True Class	**p**	TP	FN
	n	FP	TN

[a]In this study, *positive* class can be any of the three types of interactions (OB, NO, CP).

For assessing the prediction performance of a classifier, four measures *precision*, *sensitivity*, *specificity*, and *accuracy* are commonly used (Fawcett, 2006). The definition of performance measures are:

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad Sensitivity = \frac{\text{TP}}{\text{TP+FN}} \qquad Specificity = \frac{\text{TN}}{\text{TN+FP}}$$

and

$$Accuracy = \frac{\text{Sum of correct predictions}}{\text{Sum of total predictions}}$$

These notions and performance measures were originally proposed for classification problems involving two classes of data (*positive* and *negative*). However, there are three classes of data (OB, NO, CP) in our classification problem. To calculate precision, sensitivity and specificity, we dynamically regrouped our data into two categories for each class. For instance, considering class OB, we separated all our data into class *OB* and class *none OB* (NO+CP). The overall accuracy for the three-way classification was calculated as the correct prediction rate for all three classes of data.

2.3.4 Classification Results

We first investigated the performances of the multi-class SVM and the two-stage SVM classifiers by considering all possible combinations of the six interface features. Then, detailed results of the classifiers are presented using the best performing feature combinations. We have tested four common kernel functions, namely, linear

kernel, polynomial kernel, radial basis function (RBF) kernel, and sigmoid kernel. Best results were obtained when the RBF kernel was chosen for both the multi-class SVM and the two-stage SVMs.

Feature Selection

We investigated the best performances of the multi-class SVM and the two-stage SVM in terms of cross-validation accuracy when using all possible combinations of the six individual features: IA, IAR, AACa, CORa, GVI, and CSa. Here an exhaustive feature selection was carried out and the performances of all possible combinations of interface features are reported in Table 2.4.

For the multi-class SVM, the most discriminative feature is IA with a LOOCV accuracy of 76.1%. The overall best performance of 90.9% is achieved when using four features IA, IAR, AACa, and GVI. When considering all the six features, the accuracy decreases to 88.5%. For the two-stage SVM, the best single feature is IA with an accuracy of 76.5%. The best combination of two features is IA and AACa, yielding 86.0%. Using the three features IA, IAR, and AACa, yields 91.8%. This is the highest overall classification accuracy we have reached. With the four features, IA, IAR, AACa, and GVI (or CSa), the classification accuracy is 91.4%. The best accuracy is 90.5% when using five features with IA, IAR, AACa, GVI, and CSa. When using all six features, the accuracy is 89.7%. In general, the two-stage SVM classifier performs better than the multi-class classifier, though the absolute difference in the LOOCV accuracies of the two SVM classification methods are marginal.

The SVM classifier did not benefit from including conservation scores. We investigated whether confidence measures for the conservation score improve performance. To this end, we tested the number of sequences used to calculate the ConSurf score. Improvement was only observed when the number of sequences was combined with the conservation score feature in comparison to only using the ConSurf score as a single feature (improvement in accuracy from 56.4% to 60.0% using multi-class SVM). No significant improvement was observed when using the number of sequences in addition to any of the five other features.

Multi-Class SVM Classifier

The best performing multi-class SVM uses four interface properties (IA, IAR, AACa, and GVI), with *gamma* set to 0.0008 and C set to 1278.3. With a leave-one-out cross-validation procedure we obtained a best accuracy of 90.9% when using four properties, IA, IAR, AACa, and GVI on the BNCP-CS dataset. Detailed results are shown in Table 2.5 and Table 2.6. The classifier performs best in the identification of crystal packing contacts as the precision, sensitivity, and specificity of the classifier is the highest for discriminating crystal packing contacts. The performance of the classifier for obligate interactions is similar to that for non-obligate interactions. Out of the 243 interactions, 221 are correctly classified by the multi-class classifier.

Table 2.4: Prediction results (LOOCV) using all feature combinations.

Interface Properties						Multi-class SVM	Two-stage SVM
IA	IAR	AACa	CORa	GVI	CSa	Acc.(%)	Acc.(%)
+						76.1	76.5
	+					67.9	67.9
		+				74.9	74.9
			+			53.1	51.4
				+		72.0	72.0
					+	56.4	32.5
+	+					83.5	84.8
+		+				82.7	86.0
+			+			78.2	78.6
+				+		79.4	79.8
+					+	77.8	77.4
	+	+				79.0	80.7
	+		+			72.8	72.4
	+			+		81.9	82.3
	+				+	71.2	72.4
		+	+			74.9	77.0
		+		+		76.1	77.4
		+			+	76.5	77.4
			+	+		68.7	70.0
			+		+	62.1	62.1
				+	+	74.9	76.1
+	+	+				90.1	**91.8**
+	+		+			84.8	87.7
+	+			+		87.7	89.3
+	+				+	83.5	84.0
+		+	+			84.4	85.2
+		+		+		82.3	84.8
+		+			+	84.0	87.7
+			+	+		81.5	83.5
+			+		+	79.8	81.1
+				+	+	79.4	83.5
	+	+	+			79.4	81.9
	+	+		+		78.6	81.1
	+	+			+	79.4	80.7
	+		+	+		79.4	81.1
	+		+		+	73.3	74.5
	+			+	+	83.5	81.9
		+	+	+		77.0	78.2
		+	+		+	77.4	79.0
		+		+	+	79.0	81.1
			+	+	+	74.1	74.0
+	+	+	+			89.3	90.9
+	+	+		+		**90.9**	91.4
+	+	+			+	90.5	91.4
+	+		+	+		85.2	89.7
+	+		+		+	86.8	87.7
+	+			+	+	86.8	88.9
+		+	+	+		81.1	84.8
+		+	+		+	86.8	87.7
+		+		+	+	85.2	87.2
+			+	+	+	81.9	85.2
	+	+	+	+		79.8	81.9
	+	+	+		+	80.7	81.5
	+	+		+	+	80.2	81.9
	+		+	+	+	80.7	82.7
		+	+	+	+	79.8	82.3
+	+	+	+	+		87.7	90.1
+	+	+	+		+	89.7	89.3
+	+	+		+	+	88.1	90.5
+	+		+	+	+	88.1	89.3
+		+	+	+	+	85.6	86.4
	+	+	+	+	+	80.2	83.1
+	+	+	+	+	+	88.5	89.7

Two-Stage SVM Classifier

In the best performing two-stage SVM using three interface properties (IA, IAR, and AACa), *gamma* and C were set to 0.004 and 128.0 for the SVM in the first stage, and 0.00085 and 512.0 for the SVM in the second stage. Table 2.7 and Table 2.8 list the leave-one-out cross-validation results and performances of the two-stage SVM classifiers for the BNCP-CS datasets using three feature combination with highest accuracy (IA, IAR, AACa). The classifier identifies crystal packing contacts more accurately than it did for the other two types of interactions. The performance for obligate and non-obligate interactions is similar. The two stages SVM1 and SVM2, as depicted in Figure 2.14, have leave-one-out cross-validation accuracies 97.9% and 86.4%, respectively for the BNCP-CS dataset. In total, the accuracy is 91.8% (=223/243) for the two-stage SVM classifiers.

Test for Overfitting with Nested Cross-Validation

By selecting parameters for the SVMs after cross-validation, we followed a standard procedure applied when limited data are available. Ideally, the data should be split into training, parameter optimization, and validation sets. Since our dataset is of limited size, we maximized the size of the training dataset to get the best-performing SVM classifiers. However, the drawback of this strategy is that the accuracy estimates are possibly too optimistic. In order to test for overfitting, we estimated the misclassification rate following a previously described nested cross-validation protocol (Ruschhaupt *et al.*, 2004) as depicted in Figure 2.13.

We have computed the average classification accuracies based on the protocol described in Section 2.3.1. The average accuracies and standard deviations are 81.4±1.46% (multi-class SVM using four features IA, IAR, AACa, and GVI), 83.1±1.16% (two-stage SVM using three features IA, IAR, and AACa). In addition, for the two-stage SVM, the accuracies for the first and second stage are 94.5±0.92% and 75.2±2.52%, respectively. These results are summarized and compared to the leave-one-out cross-validation results in Table 2.9. Apparently, there is no considerable difference between the two average accuracy values for the best performing multi-class and two-stage SVMs. The low standard deviations indicate that our classification methods are quite robust. Because of the small size of the training dataset, the accuracy estimates from the nested cross-validation might be overly pessimistic.

Testing on Bahadur's Dataset

We have applied our best performing SVM, which is the two-stage SVM trained using three features (IA, IAR, and AACa), to the dataset used by Bahadur et al. (Bahadur *et al.*, 2004). This dataset includes 188 crystal packing contacts, 122 homodimers, and 70 other protein-protein complexes. This dataset has some overlap with the BNCP-CS dataset. Between the two sets there are 36 homodimers and 19 other biological complexes with more than 40% sequence identity. In total, the accuracy of the first stage SVM is 80.0%, which is considerably less than the performance of the

Table 2.5: Prediction results (LOOCV) using the multi-class SVM[a]

		Predicted			
		OB	**NO**	**CP**	**Total**
Actual	**OB**	68	7	0	75
	NO	9	51	2	62
	CP	3	1	102	106
	Total	80	59	104	243

[a]Four out of the six interface properties (IA, IAR, AACa, GVI) are used in the SVM classification for the BNCP-CS dataset.

Table 2.6: Performance of the multi-class SVM[a]

	OB	**NO**	**CP**	**Combined**
Precision	85.0%	86.4%	98.1%	-
Sensitivity	90.7%	82.3%	96.2%	-
Specificity	92.9%	95.6%	98.5%	-
Accuracy	-	-	-	90.9%

[a]The same properties are used as in Table 2.5.

Table 2.7: Prediction results (LOOCV) using the two-stage SVM[a]

		Predicted			**Total**
		OB	**NO**	**CP**	
Actual	**OB**	69	6	0	75
	NO	9	52	1	62
	CP	3	1	102	106
Total		81	59	103	243

[a]Three out of the six properties (IA, IAR, and AACa) are used in the SVM classification for the BNCP-CS dataset.

Table 2.8: Performance of the two-stage SVM classifier[a]

	OB	**NO**	**CP**	**Stage 1**	**Stage 2**	**Combined**
Precision	85.2%	88.1%	99.0%	-	-	-
Sensitivity	92.0%	83.9%	96.2%	-	-	-
Specificity	92.9%	96.1%	99.3%	-	-	-
Accuracy	-	-	-	97.9%	86.4%	91.8%

[a]The same properties are used as in Table 2.7.

Table 2.9: Nested cross-validation results of SVM classifiers

	Multi-class SVM	Two-stage SVM		
		Stage 1	Stage 2	Combined
LOOCV	90.9%	97.9%	86.4%	91.8%
Nested CV	81.4±1.46%	94.5±0.92%	75.2±2.52%	83.1±1.16%

first stage SVM on the nested cross-validation (94.5±0.92%). This can be explained by the fact that the crystal packing dataset used by Bahadur et al. is heavily biased toward crystal packing contacts with large contacting area ($> 400\,\text{Å}^2$).

We can reasonably expect that in this dataset the subset of homodimers mostly includes obligate interactions. In addition, inspecting the descriptions of the 70 other protein-protein complexes in the PDB files, one can expect that this subset mostly contains non-obligate interactions. The second stage SVM predicts 84.4% of the homodimers to be obligate, and 78.6% of the remaining complexes to be non-obligate. Although these results do not represent an actual validation, they do agree with our expectations.

Application of Random Forests Method

For each random forests classifier, we have performed LOOCV 20 times and obtained average accuracies and standard deviations (Table 2.10). In addition, we have extracted the importance of the interface features in the classification estimated in the construction of the random forests (Figure 2.15).

Table 2.10: Performance (LOOCV) of random forests classifiers

Features	Accuracy (RF)	Accuracy [a] (multi-class SVM)	Accuracy [a] (two-stage SVM)
IA, IAR, AACa, CORa, GVI, CSa	84.6±0.77%	88.5%	89.7%
IA, IAR, AACa, GVI	83.9±0.57%	90.9%	91.4%
IA, IAR, AACa	83.3±0.79%	90.1%	91.8%
IA, IAR, GVI	81.4±0.67%	87.7%	89.3%

[a] LOOCV accuracies of NOXclass multi-class and two-stage SVMs are taken from Table 2.4.

When all six interface features are considered, the random forests classifier reaches the highest classification accuracy (84.6%). If the four features that are used in the best-performing MCC (IA, IAR, AACa, GVI) are employed, the performance of random forests classifier decreases slightly to 83.9%. The accuracy of the random forests classifier using only the three features that are exploited in the best-performing MSC (IA, IAR, AACa) is only 83.3%. However, the difference in the LOOCV accuracies

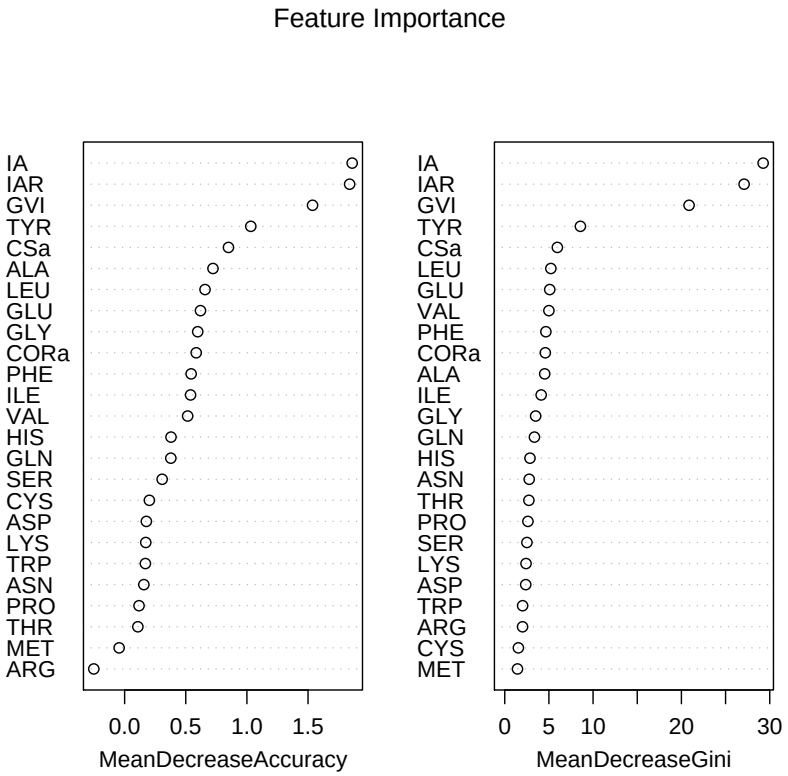


Figure 2.15: Importance of interface properties in the classification using the random forests method. Interface properties are listed in descending order according to their importance values (Left: permutation importance; Right: Gini importance). The names of interface properties are abbreviated as given in Table 2.2. The importance of the area-based composition for the 20 standard amino acids (AACa) is listed separately.

using different feature combinations is marginal. If we consider the three most important features IA, IAR and GVI (see Figure 2.15) , the LOOCV accuracy of random forests reaches 81.4%.

Figure 2.15 lists the interface features in descending order of feature importance. We observed that IA, IAR, and GVI are the most discriminative features in the classification of the three types of interactions. The results agree with what we concluded by using the SVM method. As reported in Table 2.4, when only one feature is used, the LOOCV accuracy of SVMs using IA, IAR, and GVI are among the highest. The area-based composition of tyrosine (Tyr) is the most discriminative in AACa. The distinction of tyrosine in the composition may be also noticed in Figure 2.4. On average, the compositions are 6.28%, 7.98%, and 2.33% for Tyr at OB, NO, and CP interfaces, respectively. Tyrosine has been discovered to be likely contained in hot spots (Bogan and Thorn, 1998). It had been suggested that because aromatic residues like tyrosine have few rotatable bonds, they contribute to the binding en-

ergy through the hydrophobic effect without a large entropic penalty. Furthermore, tyrosine is also capable of forming hydrogen bonds or π-stacking (Schalley, 2006) across interfaces.

According to our test, the performance of the random forests classifiers is always worse than that of the multi-class or two-stage classifiers constructed using SVM. Although random forests methods have appealing theoretical and practical characteristics such as being robust against overfitting and being capable of estimating the importance of variables, the random forests classifiers exhibit larger classification errors than the SVM classifiers for separating different protein interaction types.

2.3.5 Classification Using Atomic Contact Vectors

In 2003, Mintseris and Weng (2003) developed the method of atomic contact vectors (ACVs) for distinguishing transient complexes from permanent complexes, as well as to separate homodimers from crystal contacts. The authors reached favorable accuracies of 91% and 93% for the two classifications compared to other methods at the time. Due to its respectful performance, we also tested this method and compared the results to those of the NOXclass program. We computed the ACVs for the interactions in the BNCP-CS dataset and utilized them in the classification problem.

Atomic Contact Vectors

The ACV method works as follows: First, atomic contacts between non-hydrogen atoms from the two interacting subunits are identified using a distance criterion (distance cutoff = 6 Å). Then, each of these contacting atoms in the 20 standard amino acids is assigned one of 18 atoms types (see Table 2.11 for details), which have been designated according to their estimated contact energies (Zhang *et al.*, 1997). Consequently, each atomic interaction is assigned an atom type pair. There are in total $\binom{18}{2} + 18 = 171$ different types of atom pairs. In the final atomic contact vector for every interaction, each element denotes the number of contacts formed between the corresponding pair of atom types at the protein-protein interfaces.

Classification of Interaction Types using Atomic Contact Vectors

Atomic Contact Vectors for Interactions in the BNCP-CS Dataset We computed the atomic contact vectors for all the interactions in the BNCP-CS dataset. The average percentage of each atom type pair in all ACVs for the three types of interactions were calculated and organized as matrices in Figure 2.16. We can observe in the heat maps that atoms of type 6 and 16 have formed more interactions than the other atom types. This is not surprising since these two atom types contain more atoms than the rest (Table 2.11). We computed also the differences between the ACVs of each pair of interactions types (Figure 2.17). It seems the most significant

Table 2.11: The definition of the 18 atom types

Atom type index	Amino acid	Atom	Atom type index	Amino acid	Atom
1	Backbone	N	13	Ser	C^{β}
2	Backbone	C^{α}			O^{γ}
3	Backbone	C		Thr	$O^{\gamma 1}$
4	Backbone	O		Tyr	O^{η}
5	Gly	C^{α}	14	His	C^{γ}
6	Ala	C^{β}			$N^{\delta 1}$
	Arg	C^{β}			$C^{\delta 2}$
	Asn	C^{β}			$C^{\epsilon 1}$
	Asp	C^{β}			$N^{\epsilon 2}$
	Cys	C^{β}		Trp	$N^{\epsilon 1}$
	Gln	C^{β}	15	Tyr	$C^{\epsilon 1}$
	Glu	C^{β}			$C^{\epsilon 2}$
	His	C^{β}			C^{ζ}
	Ile	C^{β}	16	Arg	C^{γ}
	Leu	C^{β}		Gln	C^{γ}
	Lys	C^{β}		Glu	C^{γ}
	Met	C^{β}		Ile	$C^{\gamma 1}$
	Phe	C^{β}		Leu	C^{γ}
	Pro	C^{β}		Lys	C^{γ}
		C^{γ}		Met	C^{γ}
		C^{δ}			S^{δ}
	Thr	C^{β}		Phe	C^{γ}
	Trp	C^{β}			$C^{\delta 1}$
	Tyr	C^{β}			$C^{\delta 2}$
	Val	C^{β}			$C^{\epsilon 1}$
7	Lys	C^{ϵ}			$C^{\epsilon 2}$
		N^{ζ}			C^{ζ}
8	Lys	C^{δ}		Thr	$C^{\gamma 2}$
9	Asp	C^{γ}		Trp	C^{γ}
		$O^{\delta 1}$			$C^{\delta 1}$
		$O^{\delta 2}$			$C^{\delta 2}$
	Glu	C^{δ}			$C^{\epsilon 2}$
		$O^{\epsilon 1}$			$C^{\epsilon 3}$
		$O^{\epsilon 2}$			$C^{\zeta 2}$
10	Arg	C^{ζ}			$C^{\zeta 3}$
		$N^{\eta 1}$			$C^{\eta 2}$
		$N^{\eta 2}$		Tyr	C^{γ}
11	Asn	C^{γ}			$C^{\delta 1}$
		$O^{\delta 1}$			$C^{\delta 2}$
		$N^{\delta 2}$	17	Ile	$C^{\gamma 2}$
	Gln	C^{δ}			C^{δ}
		$O^{\epsilon 1}$		Leu	$C^{\delta 1}$
		$N^{\epsilon 2}$			$C^{\delta 2}$
12	Arg	C^{δ}		Met	C^{ϵ}
		N^{ϵ}		Val	$C^{\gamma 1}$
					$C^{\gamma 2}$
			18	Cys	S^{γ}

Table adapted from Table 1 in Zhang *et al.* (1997).

difference still involves atom type 16. Most atoms of type 16 originate from aromatic residues and C^{γ} atoms. By re-examining the area-based amino acid composition plot (Figure 2.4), we discovered consistent results. The three aromatics residues (Phe, Trp, Tyr) display large differences in their composition in the three types of interactions. For example, there are on average 6.28% and 7.98% of Tyr at OB and NO interfaces, respectively. But there are only 2.33% of Tyr at CP interfaces. In addition, we observed that the atomic contacts involving atoms of type 17 are more abundant in OB interactions than in NO interactions (see the heap map OB-NO in Figure 2.17). Atom type 17 consists of atoms at the end of side chains of hydrophobic residues. Therefore, the enrichment of atomic contacts involving atoms in type 17 reveals that there are more hydrophobic residues involved in OB interactions. Furthermore, we noticed that compared to CP, both OB and NO interfaces are depleted with atomic contacts involving atom type 9 and 11, that is, atoms in charged residues Asp, Glu, and polar residues Asn and Gln. At the same time, there are slightly less atomic contacts involving atom type 9 and 11 in OB interactions than in NO interactions.

Feature Selection The number of features used in the ACV method is 171. But the number of data in the BNCP-CS dataset is only 243. This may result in a model with too many parameters, or, too high complexity, thus lead to the overfitting of the classification model. We decided to select a subset of the ACVs for the classification purpose. To avoid an infeasible exhaustive search, we have employed the PCA technique for identifying the most important combinations of ACVs. In order to determine the number of principal components (PCs) to be used in the model, we have tested the MCC and MSC SVM classifiers by using all possible number of PCs. An overview of the performance of the MCC and MSC SVMs has been depicted in Figure 2.18 for $n = 1, 2, ..., 171$, where n is the number of principal components. The performances of the three classifiers (two from the two-stage SVM and the multi-class SVM) all reach their peaks around $n = 26$. The cumulative proportions of variance these 26 PCs account for is 97.7%. Therefore, we chose the first 26 PCs of the ACVs as the features to distinguish the three types of interactions using the SVM methods. In addition, the dimensionality of this feature vector (26) is also close to the dimensionality of the feature vectors used in the best performing NOXclass classifiers (23 for MCC, and 22 for MSC).

Classification Results The LOOCV accuracies by using the 26 PCs extracted from the ACVs are reported in Table 2.12. At the same time, we combined the ACVs with the features used in NOXclass and the performances are reported in the same table. The performance of the NOXclass SVM classifiers are also listed in the same table as comparison. We observed that the features used NOXclass are more discriminative than the 26 PCs of ACVs in the discrimination of biological interactions and crystal packing contacts, but less discriminative in the separation of obligate from non-obligate interactions in terms of the LOOCV accuracies. The

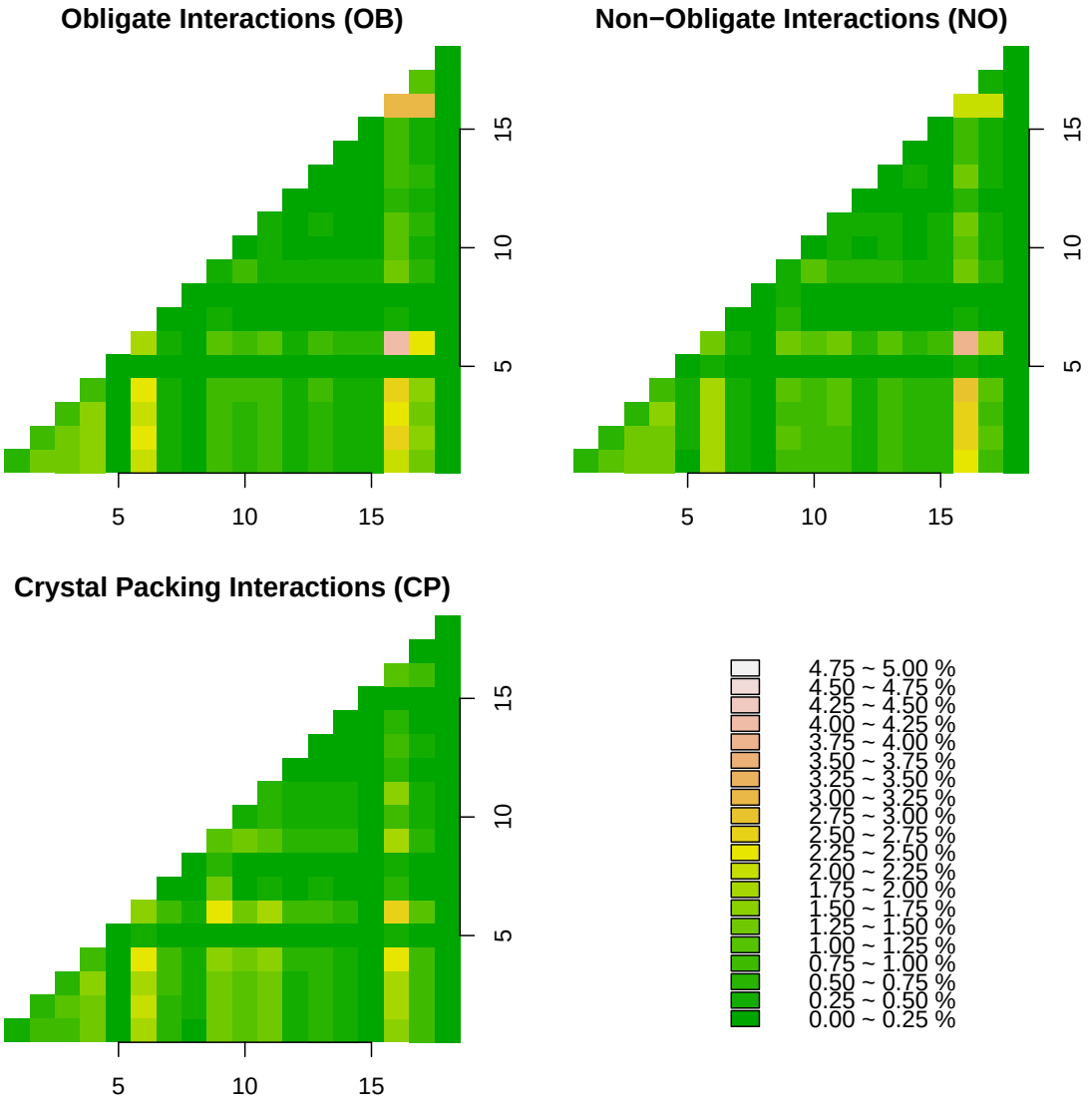


Figure 2.16: Overview of the atomic contact vectors for the interactions in the BNCP-CS dataset. On X and Y axes, the 18 atom type indices are listed as defined in Table 2.11. The three heat maps correspond to the average percentages for the 171 atom type pairs of the three types of interactions OB, NO and CP.

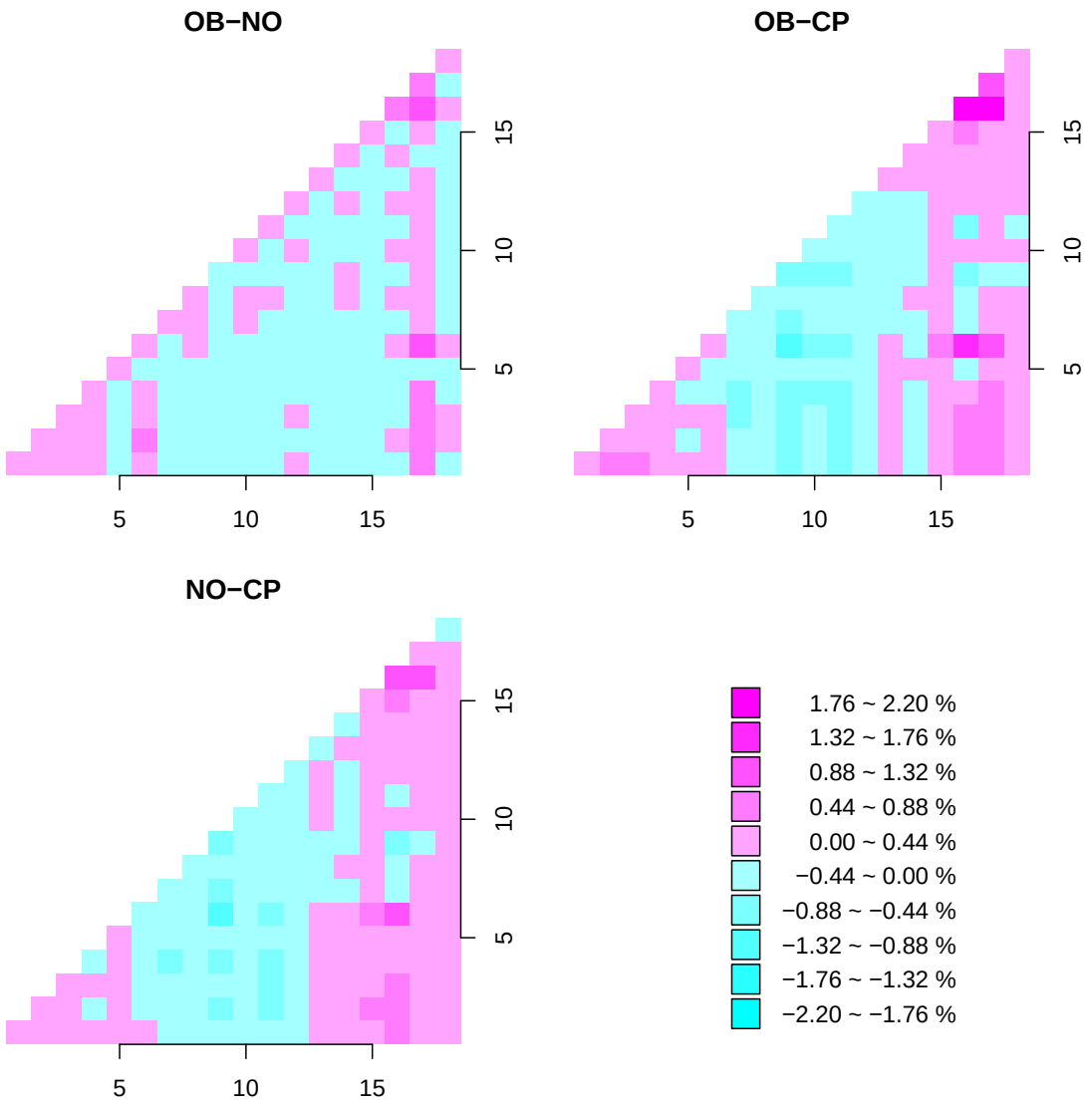


Figure 2.17: Overview of the differences between atomic contact vectors for the interactions in the BNCP-CS dataset. On X and Y axes, the 18 atom type indices are listed as defined in Table 2.11. The three heat maps correspond to the differences between each pair of heat maps (OB-NO, OB-CP, NO-CP) depicted in Figure 2.16.

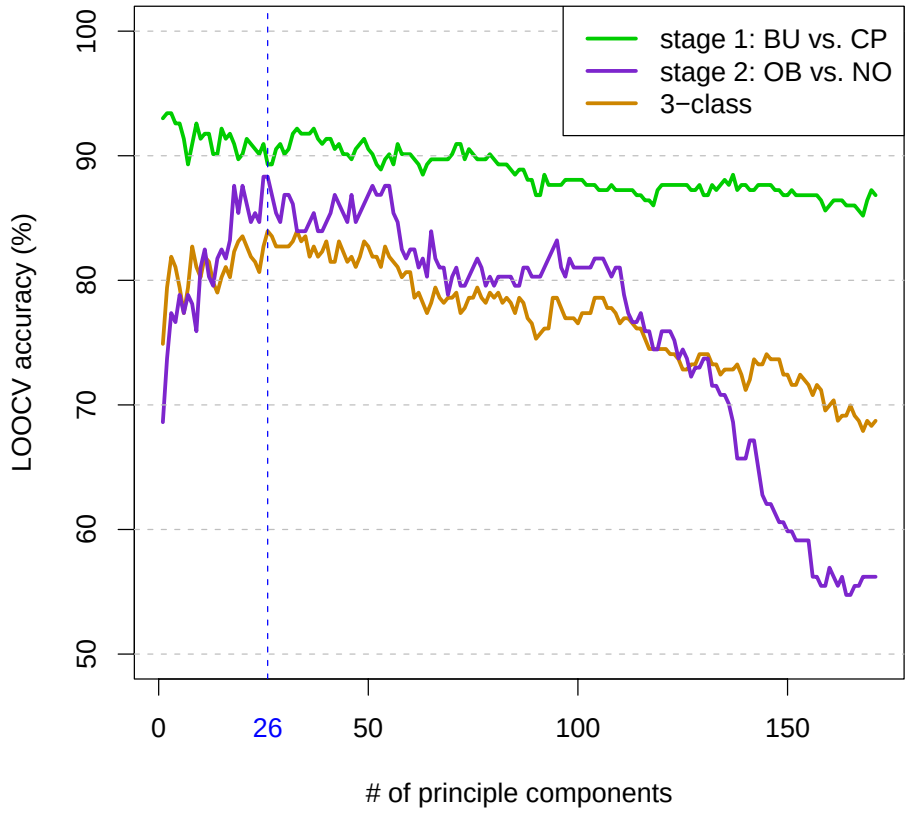


Figure 2.18: Performance of SVM classifiers using different numbers of principal components extracted from the 171-dimensional atomic contact vectors in the classification of protein-protein interaction types. Parameter optimization is not performed for the SVM classifiers.

NOXclass MSC has higher LOOCV accuracy (97.9%) than the MSC using the 26 PCs of ACVs (94.2%) at stage 1. But the MSC using the 26 PCs of ACVs performs better at stage 2 with an accuracy of 87.8%, which is slightly higher than the NOXclass MSC (86.4%). In the end, the NOXclass MSC for the classification of the three types of interactions outperformed the MSC using the 26 PCs of the ACVs (91.8% vs. 89.3%).

Based on the observation that the SVM classifiers using NOXclass features and ACVs perform differently at the two classification stages, we constructed a hybrid two-stage SVM classifier. The first stage of the hybrid MSC is the best performing NOXclass SVM classifier for separating biological interactions from crystal packing contacts. The second stage of the hybrid MSC is the SVM classifier using the 26 PCs of ACVs as features. Interestingly, the performance of the hybrid classifier

reaches 93.0%, which is better than the best performing NOXclass MSC. Therefore, for classification problems involving more than two classes of data, better results may be reached by constructing a hybrid multi-stage classifier. At each stage of the classifier, features that are specifically discriminative for the classification problem of the stage should be employed.

When the complete ACVs were used as features, the performance of the SVMs became worse, especially for the classification between obligate and non-obligate interactions. We tested the SVM classifiers using both ACVs and interface features used in NOXclass. The performances of all SVMs using the 26 PCs of ACVs are better than the SVMs using the complete ACVs as features. The reason might be that the SVMs using the complete ACVs are overfitting as the dimensionality of the features used in these SVMs is very big (171). The performances of all the SVMs in the classification of biological interactions and crystal packing contacts are close when the features used in NOXclass are employed. This observation again demonstrates that the six interface features capture the distinction between biological interactions and crystal packing contacts very well.

2.3.6 NOXclass

Our classification program using the six interface properties introduced in Section 2.2 and the SVM algorithms is named *NOXclass*, where *N*, *O*, and *X* stand respectively for the three types of protein interactions non-obligate, obligate, and crystal packing, and *class* means classification. A web server based on the method and the datasets used in this study are available at `http://noxclass.bioinf.mpi-inf.mpg.de/`. Both multi-class and multi-stage SVM classifier are provided. Predictions of interaction types are presented together with the prediction probability values, as well as the values of the interface features used for the prediction. The source code for the program can be downloaded from the same address. The source code of the NOXclass program is distributed under the terms of GNU LGPL[2].

2.4 Discussion

2.4.1 Summary

In this work, we analyzed six interface properties for three types of protein-protein interactions. The interface area was the most important feature in our study for distinguishing biological interactions from crystal packing contacts. The area of a crystal packing interface is typically smaller than that of a biological interface (Figure 2.2). Different cutoffs have been proposed for separating crystal packing contacts from biological interactions (Henrick and Thornton, 1998; Ponstingl *et al.*, 2000). In our analysis we found 650 Å^2 to be a reasonable cutoff for interface area with respect to the binary classification of biological and non-biological interactions.

[2] `http://www.gnu.org/licenses/lgpl.html`

Table 2.12: Performance (LOOCV) of SVM classifiers using ACVs and NOXclass features

Features	Feature Dim	Multi-class SVM	Two-stage SVM		
			Stage 1	Stage 2*	Combined
ACV PC [a,+]	26	86.8%	94.2%	87.8% (90.5%)	89.3%
NOXclass [b,+]	23/22[1]	90.9%	97.9%	86.4% (87.6%)	91.8%
ACV [c,−]	171	79.8%	92.6%	72.6% (79.6%)	80.7%
ACV PC+NOXclass MCC [d,−]	49	88.9%	93.8%	83.3% (89.1%)	87.7%
ACV PC+NOXclass MSC [e,−]	48	89.3%	94.2%	83.7% (87.6%)	87.7%
ACV+NOXclass MCC [f,−]	194	82.7%	93.8%	77.9% (84.7%)	84.3%
ACV+NOXclass MSC [g,−]	193	83.1%	94.2%	79.4% (84.7%)	85.2%
ACV+NOXclass All [h,−]	196	84.4%	94.2%	78.8% (85.4%)	85.2%
Hybrid MSC [i,+]	22/26[2]	-	97.9%	88.6% (90.5%)	93.0%

a) SVM classifiers constructed using the first 26 principal components from ACVs;
b) SVM classifiers as reported in Table 2.6 and 2.8;
c) SVM classifiers constructed using ACVs;
d) SVM classifiers constructed using the first 26 principal components from ACVs and four features used by the multi-class NOXclass as reported in Table 2.6;
e) SVM classifiers constructed using the first 26 principal components from ACVs and three features used by the multi-class NOXclass as reported in Table 2.8;
f) SVM classifiers constructed using ACVs and four features used by the multi-class NOXclass as reported in Table 2.6;
g) SVM classifiers constructed using ACVs and three features used by the multi-class NOXclass as reported in Table 2.8;
h) SVM classifiers constructed using ACVs and all six features from the NOXclass;
i) Hybrid two-stage SVM classifier: stage 1 is taken from the best performing NOXclass two-stage SVM (Table 2.8), and stage 2 is constructed using the first 26 principal components from ACVs;

+) Parameter optimization is performed;
−) Parameter optimization is not performed;

1) Feature dimensionality for MCC is 23, for MSC is 22;
2) Feature dimensionality for stage 1 is 22, for stage 2 is 26;

***)** Accuracy values in parentheses are obtained from LOOCV independent of stage 1.

This threshold separates the BNCP-CS dataset with an accuracy of 93%. Biological interactions where small subunits are involved are better identified using the interface area ratio property in addition.

The 20 amino acids display a variable preference for protein-protein interactions. This was shown by the different contributions of the 20 amino acids to the interface areas of different types of interactions. Obligate and non-obligate interactions show noticeable differences regarding the features based on the amino acid composition.

Residues involved in biological interactions were shown to be more strongly conserved than residues involved in crystal packing contacts (Figure 2.8). With the

increase of the conservation scores of the interface residues, the difference between the three types of interactions are more obvious in terms of their ΔSASA per residue. In particular, conserved residues involved in crystal packing contacts tend to have lower ΔSASA values than biological interactions (Figure 2.9). However, the classification of interaction types using the SVM algorithm did not benefit from including conservation scores. We have included confidence measures for conservation scores and no significant improvement was observed. The effect of confidence measures and conservation scores in the SVM performance should be further investigated (see also related discussion in Section 2.4.3).

The first stage of the two-stage SVM classifier distinguishes crystal packing contacts from biological interactions with an accuracy of 97.9% (see the Two-stage SVM Section). Valdar and Thornton obtained an accuracy of 98.3% on a similar problem (Valdar and Thornton, 2001). Nevertheless, the performances of the two methods are not directly comparable because the datasets are different and, in particular, the biological interactions were restricted to homodimers in the latter method.

In addition to SVMs, there are a variety of other classification methods available, such as decision tree and random forests. In our preliminary test, the decision tree algorithm C4.5 exhibited the worst performance and was abandoned at very early stage. The random forests algorithm has also been tested and the performance of the method is not as good as that of the SVM classifiers.

The nested cross-validation results indicate that there is no considerable difference between the performances of the multi-class and two-stage SVMs. The small variances of these results along with the minor difference between the performances of the SVM implementations indicate that the approach is quite robust.

NOXclass allows the interpretation and analysis of protein quaternary structures. In particular, it generates testable hypotheses regarding the nature of protein-protein interactions, when experimental results are not available. We believe that the NOXclass program will benefit the users of protein structure models, as well as protein crystallographers and NMR spectroscopists.

2.4.2 Related Work after NOXclass

There has been more work published at the same time with or after our NOXclass work. In 2005, Ansari and Helms (2005) carried out a comprehensive statistical analysis of a set of 170 transient protein-protein interactions. This study revealed many new insights about the properties of transient interactions, as well as confirmed several previous findings. Charged residues were found to be dominating in the amino acid composition. The hydrophobicity of interfaces was shown to decrease with the interface size, while polar and charged residues were more frequently discovered at smaller interfaces, implying hydrophobicity is more accountable for binding affinity than for specificity. The authors reported that small interfaces are more often involved in quick and highly specific interactions, in which longtime complexation of subunits is not required. Therefore, the proportion of hydrophobic residues declines when the interface size decreases. Conversely, large interfaces contain more

hydrophobic residues and are mainly stabilized by the hydrophobic effect.

De *et al.* (2005) published the results of a statistical analysis of the interface properties for obligate and non-obligate interactions. The average interface area of obligate interactions was shown to be approximately twice the size of non-obligate interactions. The interface area ratio of non-obligate interactions was discovered to be relatively smaller than obligate interactions. The authors also analyzed the secondary structure composition at different interfaces and noticed that β-sheets across subunits are only observed in obligate interfaces. In addition, non-obligate interactions were observed to involve more irregular secondary structure elements.

The Klebe group presented a systematic study about the classification of permanent and transient interactions using different machine learning methods (Block *et al.*, 2006). The work by Block *et al.* (2006) investigated four different classification methods SVM, C4.5 Decision Tree, K Nearest Neighbors, and Naïve Bayes algorithms for selecting physicochemical properties to distinguish permanent and transient interactions. Three feature selection strategies (filter, wrapper, and genetic algorithms) were applied for extracting discriminating interface features from a set of physicochemical properties. The physicochemical properties were represented in four different ways: two different atomic contact vectors (Mintseris and Weng, 2003), DrugScore potential vectors (Gohlke *et al.*, 2000), and SFCscore descriptor vectors (Sotriffer *et al.*, 2008). The best results were achieved by using the ACV descriptor of interfaces and the decision tree algorithm C4.5 optimized using the genetic algorithm for feature selection. The prediction accuracy is 94.8% for the classification between crystal packing contacts and homodimeric interactions, and 93.6% for the discrimination between permanent and transient interactions.

Bai et al. used the dynamic properties of protein structure to distinguish biological interactions from crystal packings (Bai *et al.*, 2008). The authors employed a Gaussian Network Model (GNM) (Bahar *et al.*, 1997; Haliloglu *et al.*, 1997) to analyze the global and local motions of residues belonging to the different subunits of protein complexes. They discovered that the slow mode fluctuations, which reflect the global motion of subunits, are distinct for biological and non-biological interactions. The authors concluded that this is because global motions like hinge-bending or stretch-contact motion are weak at biological interfaces, because these interfaces are large and specific and thus the subunits are static relative to each other. For non-biological interactions, the interfaces are smaller and the relative motions between subunits are stronger. An accuracy of 89.4% was obtained for identifying crystal packings from biological protein-protein interactions.

Recently, Bernauer *et al.* (2008) published the DoMoVo method, which also uses a SVM algorithm. The authors explored as many as 87 parameters derived from a Voronoi tessellation of protein structures (Bernauer *et al.*, 2005) and a previous similar work (Bahadur *et al.*, 2004). The DoMoVo method obtained favorable results using 21 selected parameters on several different datasets compared to a few other methods, including NOXclass (this work), PISA (Krissinel and Henrick, 2005), and PITA (Ponstingl *et al.*, 2003).

In general, the results presented in these researches regarding the physicochemical properties of interfaces are in agreement with our results. The performance of the classification or prediction approaches have been improved via the introduction of more sophisticated interface features or machine learning techniques.

2.4.3 Outlook

In a recent study carried out by Choi *et al.* (2009), the authors emphasized that as proteins sometimes participate in multiple interactions with different partners, all the known binding regions on the surface of proteins should be taken into account when studying the conservation of protein-protein interfaces. It was shown that amino acids located at interfaces are more conserved than those at the surface regions that do not participate in any known interactions. The distinction in the conservation level is more prominent when multiple binding regions are considered. In our work, we did not take into account the factor that the non-binding region on the surface of proteins in our dataset BNCP-CS might participate in other protein-protein interactions. This could have led to the close distribution of CORn and CORa values for the three types of protein-protein interactions (Figure 2.6). Furthermore, the residues at the crystal packing interfaces may be involved in some other biologically relevant interactions. This may be one of the reasons for us to obtain similar conservation scores for the three types of interactions (Figure 2.10). Further work is desired to verify these hypotheses.

So far, the best performing NOXclass classifier is the two-stage SVM based on the three interface properties IA, IAR, AACa. In the current construction, we used always the same set of features for both stages of classifications. However, we demonstrated in Section 2.3.5 that the classification accuracy may be improved by exploiting different features for the two different classification problems at the two stages. This is one of the directions of further developments for the NOXclass classifier.

In the NOXclass project, only dimeric interactions were considered, while interactions with more than two subunits were ignored. The investigation of such multimeric interfaces is strongly restricted because there is few annotation of interaction types available for multimeric oligomers. Nevertheless, the definitions of the physicochemical properties for dimeric interfaces may be easily extended to multimeric interfaces. Consequently, the comparison of interface properties and classification of interaction types for multimeric oligomers may be carried out in a similar manner to the NOXclass work.

Although the oligomeric states of many proteins may be inferred during the process of protein purification for crystallization, this is not always the case. In addition, this information is not easily available in the literature or well annotated in structure databases like the PDB. Furthermore, there is still a current lack of a well-defined criterion for defining interaction types based on experimental results, although there has been some progress in this area (Nooren and Thornton, 2003b). A number of incompatible definitions have been proposed for the classifications of protein-protein interactions, which also led to various terms describing interaction types. As de-

scribed in Section 1.1.5, Tsai *et al.* (1997a) studied two-state and three-state interactions, Gunasekaran *et al.* (2004) analyzed ordered and disordered interactions, and Mintseris and Weng (2003) suggested terms of folding complexes and recognition complexes. We focused on obligate and non-obligate interactions following the comprehensive and systematic description of the classifications by Nooren and Thornton (2003b), In each of these works, a dataset has been collected and thoroughly analyzed. The essential rules for these classifications are similar to a large extent, and the interactions defined under these terms overlap largely. However, due to the inconsistency of the definitions, these published datasets cannot be easily merged for an analysis on a larger scale. It is desirable to progress further in the development of criteria for the definition of oligomeric states, as well as the enrichment of protein-protein interaction data with known oligomeric states.

Chapter 3

Alignment of Non-covalent Interactions at Protein-Protein Interfaces

In this chapter, we present a method for aligning non-covalent interactions between different protein-protein interfaces and to estimate the statistical significance of their similarity (Zhu *et al.*, 2008). We first discuss the background of the work in Section 3.1. We introduce the alignment methodology in Section 3.2. In Section 3.2.2, we validate the method by applying it on a published dataset that comprises a variety of protein-protein interfaces. The results are compared to two relevant methods. Section 3.2.3 presents four detailed case studies of protein mimicry using the proposed approach. A scoring strategy for the alignment is described and tested in Section 3.3. In the end of the chapter, we discuss possible improvements and applications of the alignment method.

3.1 Introduction

3.1.1 Background

The characterization of protein interfaces provides insights into protein interaction mechanisms. Such analysis is expected to have an impact on the prediction of interaction partners, as well as to assist in the design and engineering of protein interactions and interaction inhibitors. The physicochemical properties of protein-protein interfaces, such as size, geometric shape, residue composition, have been previously investigated extensively (Jones and Thornton, 1996; Lo Conte *et al.*, 1999; Sheinerman *et al.*, 2000; Rodier *et al.*, 2005). Interactions between proteins are classified according to different criteria (see Section 1.1.5). Methods including NOXclass have been developed for distinguishing different interaction types based on interface properties (Bahadur *et al.*, 2004; Mintseris and Weng, 2005; Zhu *et al.*, 2006). Protein-protein interfaces are also compared for identifying common binding modes. The similarity between protein-protein interactions can be investigated on different levels, for instance, the orientation of interaction partners (Aloy *et al.*, 2003),

the location of the binding region relative to the fold of the proteins (Kim *et al.*, 2006; Teyra *et al.*, 2008), or the local structure similarity of interfaces (Shulman-Peleg *et al.*, 2004; Keskin and Nussinov, 2005). Besides, non-covalent interactions at protein-protein or protein-ligand interfaces are often compared in order to characterize binding modes and to identify detailed structural differences. Such work is normally carried out manually because there have been no methods available for comparing non-covalent interactions across interfaces automatically.

3.1.2 Related Work

A detailed comparison of protein-protein interfaces is fundamental for their better characterization and for structure-based classification of protein complexes. With an increasing amount of structural models for protein complexes available in the PDB, protein complexes can now be compared systematically on the structural level. Furthermore, protein-protein interfaces may be compared with respect to the non-covalent atomic interactions across the interfaces. In this section, we discuss these comparisons of interfaces and several interface databases. In addition, we introduce protein mimicry, a phenomenon that leads to similar interactions formed between dissimilar protein subunits.

Structure Simiarity between Interfaces

The structural similarity of protein complexes may be assessed on three levels: i) the similarity of the orientation of the folds of the subunits relative to the folds of their partners, ii) the similarity of the orientation of the binding sites relative to the folds of the subunits, and iii) the local structural similarity of interfaces. They are detailed in the next three paragraphs.

- **Comparison of Interaction Orientations.** In a comprehensive study, Aloy *et al.* (2003) analyzed the relationship between protein sequence similarity and protein interaction orientation. The geometric difference between domain orientations was computed to measure the similarity of two interactions. To calculate the difference in the orientations, first, one protein complex was chosen as the reference and the two domains in the other complex were superposed to the two domains in the reference complexes, respectively. Then the RMSD between a standard set of pseudopoints in the second complex after the two superpositions was calculated and used to evaluate the orientation difference. No binding regions or interface atoms/residues were considered for the computation of the geometric difference. Aloy *et al.* discovered that proteins with high sequence similarities tend to interact in a similar orientation.

- **Comparison of Binding Site Orientations.** Kim and colleagues put forward a method for objectively comparing the orientations of the binding regions relative to the folds of subunits in two complexes (Kim *et al.*, 2006). For two domain-domain interactions under comparison, domains on at least one side

of the two interfaces were required to exhibit similar folds. The method first superimposed the backbones of domains sharing common folds. Then the angle between the two centroids of the binding regions on the surface of the two domains and the common centroid of the superimposed domains was calculated as the measure of orientation difference. In addition, the spatial overlap of atoms at the two binding regions was also considered as part of the measure for orientation difference. The authors divided protein domain-domain interfaces into different groups (face types), resulting in SCOPPI, a structural classification of protein-protein interfaces (Winter *et al.*, 2006). They showed that similar protein domains may interact with distinct partners (non-homologous structures) using similar face types, but similar domains might also interact via different face types. Recently, using a similar method, Henschel *et al.* (2006) identified cases of protein interaction mimicry, where homologous subunits interact with non-homologous partners in the same relative orientation. Similarly, Teyra *et al.* (2008) assessed similarity of protein binding regions according to the overlap of interacting residues after a structure alignment of the backbones of protein domains sharing common fold. Using this method, a classification for protein binding regions in all domain-domain interactions derived from the PDB was integrated into the SCOWLP database, a web-based database for characterization and visualization of protein-protein interfaces (Teyra *et al.*, 2006).

- **Comparison of Interface Local Structures** Local structure comparison of interfaces has been the focus of several other studies. Nussinov and colleagues clustered all known protein-protein interfaces in the PDB by comparing the binding site C_α atoms using a geometric hashing procedure (Tsai *et al.*, 1996; Keskin *et al.*, 2004). Based on the analysis of the resulting clusters, they observed that proteins with different folds and functions may associate to yield interfaces of similar local structures (Keskin and Nussinov, 2005). Shulman-Peleg *et al.* developed I2I-SiteEngine and MAPPIS, programs that compare and align the functional groups at a pair or set of interacting binding sites using a geometric hashing algorithm (Shulman-Peleg *et al.*, 2004; Mintz *et al.*, 2005; Shulman-Peleg *et al.*, 2005). Similar methods have been developed for comparing protein binding sites for small molecules (Schmitt *et al.*, 2002; Najmanovich *et al.*, 2007), and they have been recently reviewed by Domingues and Lengauer (2007).

Comparison of Non-Covalent Interactions at Interfaces

Protein complexes are stabilized by non-covalent interactions formed across interfaces. Non-covalent interactions at protein-protein or protein-ligand interfaces are often compared in order to characterize binding modes and to identify detailed structural differences. As early as in 1990, Yamamoto *et al.* manually compared the binding mode of a papain–substrate complex with that of a papain–inhibitor complex on

the atomic level based on the crystal structures of the complexes for elucidating the inhibitory mechanism of the papain inhibitor. Biswal and colleagues manually examined van der Waals (vdW) interactions and hydrogen bonds at two interfaces corresponding to a polymerase binding to two different inhibitors (Biswal *et al.*, 2005). Deng *et al.* represented interactions at protein-ligand interfaces as a one-dimensional fingerprint descriptor for studying different docking results on the same protein (Deng *et al.*, 2004). Swint-Kruse compared the interfaces of dimeric LacI complexes in distinct functional states (Swint-Kruse, 2004). The differences in fine structures of the interfaces were identified by representing the set of non-covalent interactions as two-dimensional networks formed between interface residues (Swint-Kruse and Brown, 2005). Recently, Keskin and Nussinov (2007) showed that proteins may interact with variable partners via structurally conserved non-covalent interactions. All of the above approaches require precomputed sequence alignments or structure-based alignments of backbone atoms, and do not directly align the non-covalent interactions according to their conserved geometry.

Interface Databases

Databases of protein-protein interfaces are highly desirable for the exploration of protein-structure relationships, as essentially all protein mediated biological processes are based on protein-protein or protein-small molecule interactions. There have been some efforts devoted to implementing such databases previously. For instance, PRISM is a database composed of all two-chain interfaces derived from the PDB (Ogmen *et al.*, 2005). All inter-chain interfaces were compared and clustered using geometric hashing technique (Nussinov and Wolfson, 1991). An interface was defined to be the contacting residues between the interacting chains as well as the neighboring residue in the vicinity of the contacting residue. Only C_α atoms were considered in the comparison of interface structures. In addition, the conservation of interface residues was also investigated based on the multiple structure alignment (Shatsky *et al.*, 2002) of the interfaces. A structurally conserved hotspot was identified if a residue at a certain interface position is conserved in more than half of the interfaces in a non-redundant interface cluster (Keskin *et al.*, 2005). PRISM also provides prediction of putative interactions based on the assumption that if two proteins contain binding sites that resemble those in a known interaction, then the two proteins may also interact via the same regions (Aloy *et al.*, 2003). SCOPPI is a structural classification database of domain-domain interactions derived from the structures in the PQS and based on the SCOP domain definition (Winter *et al.*, 2006). The underlying method compared the orientations of interacting domains in two complexes using geometric measures (Kim *et al.*, 2006). Protein domain-domain interfaces were then divided into different groups (or face types). Comprehensive sequential and structural information is provided for each domain-domain interface. SNAPPI-DB provides a database of domain-domain interactions as well as the application programming interface (API) to the database (Jefferson *et al.*, 2007). A variety of derived data about protein sequences and structures have been integrated

into the SNAPPI-DB, including three widely used domain definitions (SCOP, CATH, Pfam), PQS, GO terms, Interpro, SWISSPROT. PQS was chosen by SNAPPI-DB as the source of protein interaction data in order to not only avoid non-biological contacts contained in the asymmetric units of the PDB models, but also to take into account those interactions that do not appear in the PDB models. Domain-domain interactions were clustered based on the structural similarity and the orientation of the interacting domains. In SNAPPI-DB, a pair of multiple structure alignments, one for the domains on each side of the interfaces in each cluster are generated using STAMP (Russell and Barton, 1992). Recently, Günther *et al.* (2009) presented JAIL, a library of both protein-protein interfaces and protein-nucleic acid interfaces. The clustering of interfaces was again inferred from the structural similarity of the component subunits. All these database assess the similarity between interfaces based on either the structural similarity or the relative orientation of the interacting subunits. The physicochemical nature of the interfaces is not captured, even though it is important to the understanding of the interactions.

Protein Mimicry

The similarity of protein-protein interactions does not necessarily rely on the similarity of the subunit backbone structures. Protein mimicry is a phenomenon resulting in similar protein-protein interactions involving dissimilar subunits.

Molecular mimicry or *protein mimicry* is a term for describing the phenomenon that a protein domain has evolved during evolution such that it mimics the shape of another biological molecule in order to fulfill a similar biological function (Berg *et al.*, 2002). Oldstone (2005) defines molecular mimicry as "similar structures shared by molecules from dissimilar genes or by their protein products". Either several amino acids continuous in sequence or their conformational fit may be shared between molecules of different origins.

One example of protein mimicry is the resemblance between the structure of elongation factor G (EF-G) and the structure of the complex between elongation factor Tu (EF-Tu) and tRNA. The N-terminal region of EF-G is homologous to EF-Tu, and the C-terminal region of EF-G exhibits a structure similar to tRNA. Due to the structural similarity, the EF-G and the complex between the EF-Tu and tRNA interact with the ribosome in a similar way during protein synthesis at the end of each round of polypeptide elongation. After a new amino acid is added to the growing peptide chain, EF-G binds through its EF-Tu-like domain to the EF-Tu binding site and its tRNA-like domain to the tRNA binding site on the ribosome. The binding results in the translocation of the tRNA and moves the mRNA through the ribosome, thereby creating a vacant site for the next cycle of elongation (Green, 2000). Clearly, this protein mimicry is essential for the normal translocation process of mRNA and tRNA within the ribosome during protein synthesis (the EF-G is thus also called *translocase*).

Another well investigated example is the inhibition of serine proteases by a diverse group of inhibitors. Serine proteases are a large family of enzymes responsible for

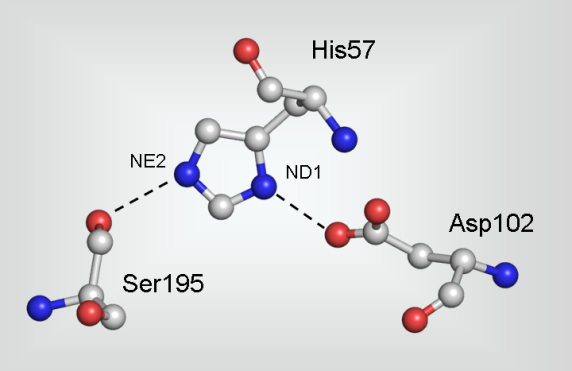


Figure 3.1: Catalytic triad of α–chymotrypsin (PDB ID: 4CHA). Dashed lines stand for hydrogen bonds.

proteolysis. In the catalytic mechanism of serine proteases, the so-called *catalytic triad* plays an essential role. The term catalytic triad refers to three residues inside or close to the active sites of serine proteases. The component residues in the catalytic triad exhibit very similar spatial arrangement, although the overall structures of the proteases may be completely different (see Figure 3.18 and 3.19 for examples). The classic catalytic triad Ser–His–Asp in serine proteases is composed of serine (Ser), histidine (His), and aspartic acid (Asp) residues[1]. The three residues are far apart in the primary structures but are brought close to function together in the tertiary structures of the proteins (see Figure 3.1). Serine proteases are inhibited by a diverse group of inhibitors, which exhibit very low similarity in their primary and tertiary structures. However, it has been discovered that these inhibitors possess a canonical loop structure interacting with the enzyme active sites (Laskowski and Kato, 1980). Although the amino acid sequences of these loops in the inhibitors display a very low similarity, the backbone conformation of the loops is highly conserved. This characteristic canonical loop is the mimicry of the normal substrates. The inhibitors exhibiting the substrate-like canonical loop have been found in four *classes*, including nine *folds* and 12 *superfamilies* in the SCOP database (Jackson and Russell, 2000). Furthermore, some inhibitors demonstrate related function by mimicking each other in the inhibition of the same type of proteases (Radisky and Koshland, 2002).

Molecular mimicry plays an important roles in the development of certain diseases. For example, it is regarded as one of the mechanisms responsible for autoimmune diseases. During an infection by a pathogen, if the infectious agent shares cross-reactive epitopes for B or T cells with the host, the immune response to the pathogen will attack the host as well. The occurrence of molecular mimicry between proteins encoded by infectious agents and self-proteins of hosts was discovered to be very common (Cunningham and Fujinami, 2000; Oldstone, 2005). Such kind of autoimmune assault is believed to contribute to the development of autoimmune diseases (Wucherpfennig, 2001; Levin *et al.*, 2002).

[1]With the discovery of new serine proteases, the component residues of the catalytic triads may differ, but the nucleophile–base–acid pattern of the triads is conserved (Polgár, 2005).

Molecular mimicry is a central topic in the design of new drugs. Peptide mimetics of proteins in the binding to the functional sites of other proteins are within the scope of many rational drug designs. Such peptides are important for the controlled interference of protein-protein or protein-ligand interactions (Eichler, 2008).

3.1.3 Detection of Structural Similarity

Traditionally, it is common to treat protein structures as rigid bodies. Structural alignments are then often visualized and assessed by using least-squares superposition. There are other representations of protein structures, such as distance difference matrices, which contain detailed information about internal motions (Holm and Sander, 1993). In order to compare two or more structures, a measure of structural similarity needs to be defined first. There is no universally accepted definition for structural similarity between proteins. Together with the various representations of protein structures, many measures for assessing the structural similarity between proteins have been proposed (Hasegawa and Holm, 2009). Once a similarity measure is defined, an algorithm for aligning the structures can be developed to optimize the alignment with respect to the similarity measure. A variety of protein structure representations have been used. SSAP generates a set of vectors from C_β atoms (dummy C_β is used for glycine) of each protein structure to be aligned (Taylor and Orengo, 1989). In DALI, proteins are represented as 2D matrices of distances between their C_α atoms (Holm and Sander, 1993). SARF2 considers proteins as a set of SSEs (secondary structure elements) (Alexandrov, 1996). CE (Combinatorial Extension) represents proteins as a set of distances between C_α atoms of octameric fragments in the protein structures (Shindyalov and Bourne, 1998). MAMMOTH (MAtching Molecular Models Obtained from THeory) considers all heptamers of protein structures and computes unit vectors from the consecutive C_α atoms in the heptameric fragments (Ortiz *et al.*, 2002). Subsequently, different comparison algorithms have been proposed to perform structural alignments. SSAP uses a double dynamic programming algorithm: a first dynamic programming step is applied to select matching positions, and a second run of dynamic programming is used to optimize the final alignment. DALI produces from the original distance matrices a set of submatrices, which are joined based on the overlap between corresponding fragments. A branch and bound algorithm is then used to find the optimal alignment. In CE, a combinatorial extension algorithm is used to identify and combine aligned fragment pairs. MAMMOTH also uses a dynamic programming algorithm in order to build the optimal alignment. The general problem of finding the optimal alignment between two proteins is NP-complete (Poleksic, 2009) and all the available solutions are thus heuristic (Sierk and Kleywegt, 2004).

When representing two protein-protein interfaces as 3D structures, the comparison of the two interfaces is transformed into the detection of similar parts between the two structures. In computational geometry, this problem can be formulated as follows. Given two point sets A and B, find a subset in A that is similar to some subset in B. The problem contains two sub-problems to be solved: the first is to establish

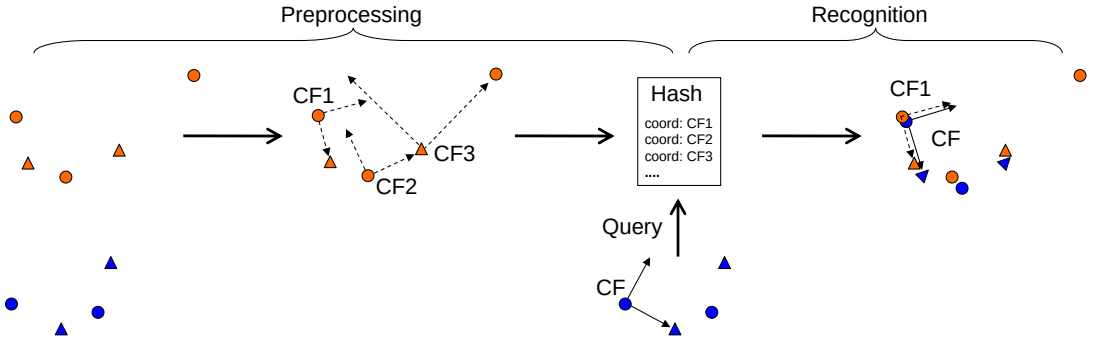


Figure 3.2: Geometric hashing. Typically, geometric hashing comprises two phases: a preprocessing phase and a recognition phrase. (CF: Coordinate Frame)

the equivalence between two subsets of points, and the second is to find a transformation that superposes the two structures. One of the common measures to define the similarity of two subsets of points is the *bottleneck matching metric* (Efrat *et al.*, 2001), restricting the maximum distance between the matched points after superposition to be less than ϵ ($\epsilon \geq 0$). Such two subsets of points are called *ϵ-congruent*. Usually, we also want to maximize the cardinality of the two similar subsets. The optimization problem of finding the ϵ-congruent subsets with maximum cardinality between A and B is called the *largest common point set* (LCP) problem. This problem can be solved in 3D space with a time complexity of $O(m^{16}n^{16}\sqrt{m+n})$, where m and n are the sizes of sets A and B (Ambuhl *et al.*, 2000). Apparently, this time complexity is impractical even for solving LCP problem for small point sets. Therefore, more efficient methods are desirable, if necessary at the price of solution accuracy.

The similar substructures of two 3D objects may be detected by considering the geometric constraints between points and the "labeling" constraint of individual points (e.g., atom types) in the objects. There are various approaches that can be applied for inferring maximum common substructures between two structures. Two of the most widely used algorithms are geometric hashing and clique detection.

Geometric Hashing

Geometric hashing was first introduced in the work of Kalvin *et al.* (1986) and Schwartz and Sharir (1987). It has been originally developed in the computer vision field for recognizing geometric features in a database. Typically, geometric hashing method is composed of two phases: a preprocessing phase and a recognition phase (Wolfson and Rigoutsos, 1997). In the preprocessing phase, the model information is encoded and indexed in a hash table. During the recognition phase, the method accesses the hash table, searching for similar features to a query (Figure 3.2).

We explain in detail how the method works for pairwise comparison of protein structures. In such scenarios, two models to be compared are represented as two sets of spatial points, encoding the atoms, residues, or other physicochemical rep-

resentations of proteins. First, one of the two models is chosen as the reference model. During the *preprocessing* phase, every possible triplet of points in the reference model is selected as the basis for defining a coordinate frame. The coordinates of the remaining points of the reference model in this coordinate frame are computed and all such values for all coordinate frames are stored in a hash table. The keys of the hash table are the coordinates of the points, and the values are the coordinate frames. Then, in the following *recognition* phase, the method selects three points in the other model and builds a coordinate frame accordingly. The coordinates of the remaining points are computed and used to query the previously constructed hash table. For each query coordinate, matches are detected in the hash table. Then the associated coordinate frames to the matches are tallied for votes based on the number of times they are detected. The coordinate frames whose number of votes exceeds a predefined threshold define transformations that realize potential superpositions of the two proteins.

From the above description it is obvious that a big advantage of geometric hashing is that the preprocessing phase for the reference model is independent of the recognition phase, thus can be carried out offline. In addition, in the recognition stage, the lookups of the keys in the hash tables are also independent of each other. Hence, the recognition phase may be performed in parallel on different reference models. These features are favorable, particularly when there are multiple reference models to compare, or a large scale database scan is needed (Wallace *et al.*, 1997; Gold and Jackson, 2006b). Nevertheless, there is a major disadvantage of the geometric hashing method in large-scale application. The geometric hashing requires a large amount of space for holding the hash table. The space complexity of the algorithm is $O(n^3)$, where n is the number of points in the reference model (Fischer *et al.*, 1992).

In computational biology, this approach has long been adopted for comparing the structures or local regions of proteins. Nussinov and Wolfson pioneered in the application of geometric hashing for the comparison of macromolecule structures (Nussinov and Wolfson, 1991; Fischer *et al.*, 1992; Bachar *et al.*, 1993). In the following years, Nussinov and Wolfson groups extended the application of geometric hashing to flexible structure comparison (Verbitsky *et al.*, 1999) and multiple structure alignment (Leibowitz *et al.*, 2001a,b). Meanwhile, Wallace *et al.* (1997) developed TESS, an approach using geometric hashing for searching functional sites on protein surface. Gold and Jackson constructed a searchable protein-ligand binding site database, SitesBase, based on geometric hashing (Gold and Jackson, 2006a,b,c).

Clique Detection

The basic idea of clique detection is to represent the protein structures under comparison as graphs and then to infer common substructures between proteins by detecting common subgraphs in their graph representations. Typically, vertices in the graph represent atoms, residues, or pseudopoints like functional groups in the protein. Pairs of vertices are connected by edges, which are labeled with properties

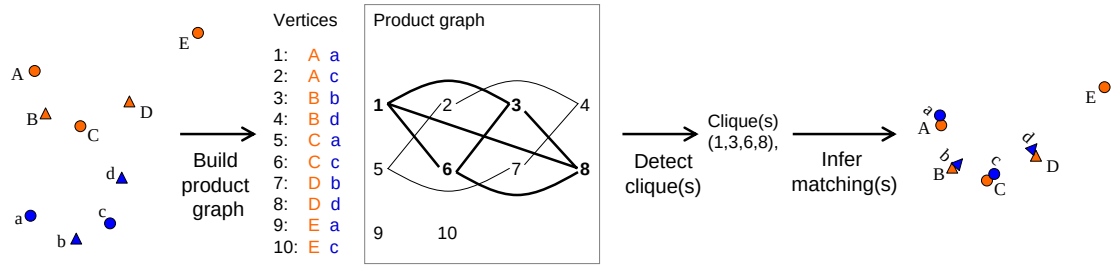


Figure 3.3: Clique detection. In clique detection, first a product graph is built based on input graphs. Then, cliques in the product graph are identified, which correspond to isomorphic subgraphs in the input graphs.

like the distance between vertices. Given two graphs G_1 and G_2, the problem of finding the *maximum common subgraphs* G'_1 and G'_2 ($G'_1 \subseteq G_1$, $G'_2 \subseteq G_2$) between them is known as *maximum common subgraph* (MCS) problem. In graph theory, G'_1 is considered to be *isomorphic* to G'_2 (denoted by $G'_1 \equiv G'_2$) if there is a *bijection* $f : V(G'_1) \rightarrow V(G'_2)$ between the vertices of G'_1 and G'_2, such that any two vertices u and v in G'_1 are adjacent if and only if $f(u)$ and $f(v)$ are adjacent in G'_2. The MCS problem in two graphs can be represented as the *maximum clique* problem in a single *product graph* constructed from the two graphs (Levi, 1972). In the product graph (sometimes called *compatibility graph*), every vertex represents two compatible vertices, each from one input graph. Two vertices of the product graph are defined to be adjacent if the two pairs of vertices that they represent are adjacent in the respective input graphs, and if the two edges are compatible (Figure 3.3). The compatibility of vertices or edges is typically defined according to the features assigned to the vertices or edges in the input graphs. After the product graph is built, cliques are detected in it. A *clique* of a graph is defined as a largest complete subgraph that is not contained in any other complete subgraph[2] (Luce and Perry, 1949; Bron and Kerbosch, 1973; Harary, 1994; Gross and Yellen, 2006). In a *complete graph*, every pair of vertices are adjacent. It has been shown that each clique in the product graph corresponds to a pair of isomorphic subgraphs of the two input graphs (Levi, 1972). Therefore, the maximum common subgraph isomorphism can be determined by detecting the maximum clique in the product graph. Both the MCS problem and the maximum clique problem are *NP-hard* in complexity theory (Garey and Johnson., 1979). Nevertheless, many approaches have been developed to solve the maximum clique problem in reasonable time for practical problems. One of the commonly used algorithm for clique search was presented by Bron and Kerbosch (1973).

As a matter of fact, product graph can encode more general types of local struc-

[2]There exist different definitions of clique. Beside the definition given here, some other authors define a clique of a graph as any of its complete subgraphs and then refer to "maximum clique" (Pemmaraju and Skiena, 2003) or "maximal clique" (Koch *et al.*, 1996). This usage disagrees with the original definition put forward by Luce and Perry (1949), where the concept "clique" was used to model an *exclusive* group of people in social network. We follow the definition in Luce and Perry (1949) in this dissertation.

ture compatibility, such as inter-residue distances. Therefore, clique detection algorithm can be applied not only to MCS problem, but also to problems involving more features for graph vertices or edges.

Clique detection in product graph has a long history of application in *cheminformatics* (Raymond and Willett, 2002; Willett, 2008). Note that because of the inherent complexity of the associated problems (maximum clique problem is NP-hard), most applications of clique detection are limited to objects of relatively small sizes such as small molecules (e.g., ligands) or the local structures of proteins (e.g., binding sites). In structural bioinformatics, the technique is often used for comparing protein local regions. Schmitt *et al.* (2002) developed an encoding scheme for simplifying the description of protein binding cavities and compared binding pockets using a clique detection algorithm. Weskamp *et al.* (2004) put forward a *k-clique hashing* approach for searching similar substructures in protein structure databases. The input graph representations for protein structures are split into a large number of complete subgraphs of the same size k, termed k-cliques. The alignment method contains two steps. In the first step, local matches of these k-cliques are generated by simply examining the labels of nodes and the weights of edges in the k-cliques. In the second step, the final overall matches are assembled from these local matches of k-cliques. This is realized by representing each k-clique by a vertex and building a modified product graph based on the local matches between the k-cliques. A clique search process is followed to identify overall matches. In a recent work, Najmanovich *et al.* (2007) exploited clique detection method together with experimental data for studying the correlation between binding site structural similarity and small molecule structural similarity.

3.2 Alignment of Non-Covalent Interactions

In this section, we describe the methodology for aligning non-covalent interactions at protein-protein interfaces. Specifically, we present in detail the vector representation of non-covalent interactions at protein-protein interfaces, and the approach for the geometrical comparison of these vector representations.

3.2.1 Alignment Algorithm

In our work, two types of non-covalent interactions were considered: van der Waals interactions and hydrogen bonds. These non-covalent interactions were represented as vectors (NCIVs) connecting the centers of two interacting atoms. The goal of the alignment method was to find the largest set of NCIVs in similar geometric orientations. Two NCIVs (each from one interface) were matched in the alignment if they represented the same type of non-covalent interactions, and have similar distances and relative orientations to the other matched NCIVs within the respective interfaces. A graph-based method was applied for aligning NCIVs. The complete procedure was implemented in *Galinter* (Graph-based alignment of protein-protein

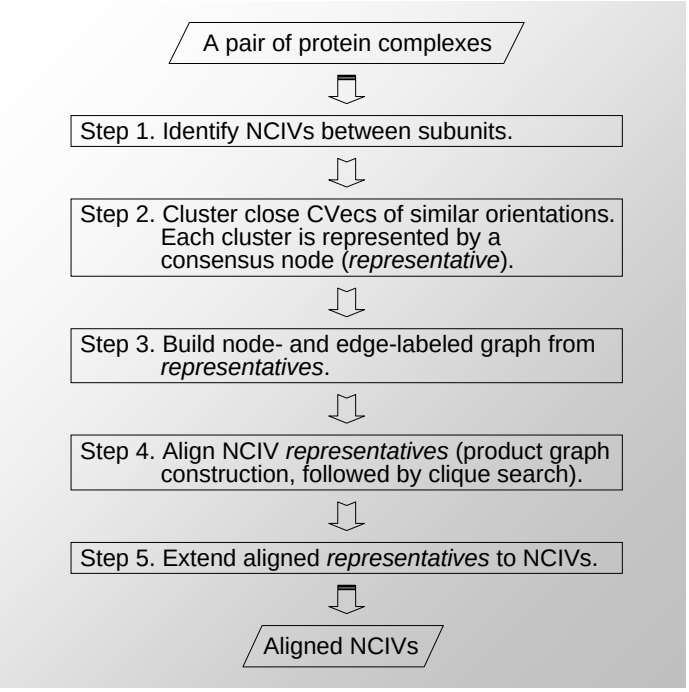


Figure 3.4: Flow chart of the Galinter program. (NCIV: non-covalent interaction vector; CVec: contact vector)

interfaces).

The workflow of the Galinter method is composed of the following five steps:

1. Identifying non-covalent interactions and representing them as vectors at interfaces;
2. Clustering the vector representations of non-covalent interactions;
3. Generating graph representation of interfaces based on the vectors;
4. Aligning clustered vectors between interfaces;
5. Extending aligned clustered vectors to original vectors.

Figure 3.4 provides a schematic overview. We now explain each of these five steps in detail.

Identifying NCIVs

For two protein complexes with known 3D structures, two types of NCIVs between the interacting proteins were distinguished. They were *contact vectors* and *hydrogen bond vectors*, representing van der Waals interactions and hydrogen bonds at interfaces, respectively.

Contact vectors (CVecs) were detected based on a distance criterion and represent van der Waals interactions. A CVec connected two heavy atoms, one in each of the interaction partners, if the distance between them was less than the sum of their respective van der Waals radii plus 1.0 Å. Van der Waals radii values used in this

Table 3.1: Atom radius values used in Galinter (Chothia, 1975).

Atom	Chemical Formula	Radius (Å)
Oxygen	=O or –O–	1.40
Trigonal Nitrogen	>N–	1.65
Tetrahedral Nitrogen	$-NH_3^+$	1.50
Trigonal Carbon	>C=	1.76
Tetrahedral Carbon	>C<	1.87
Sulfur	–S–	1.85

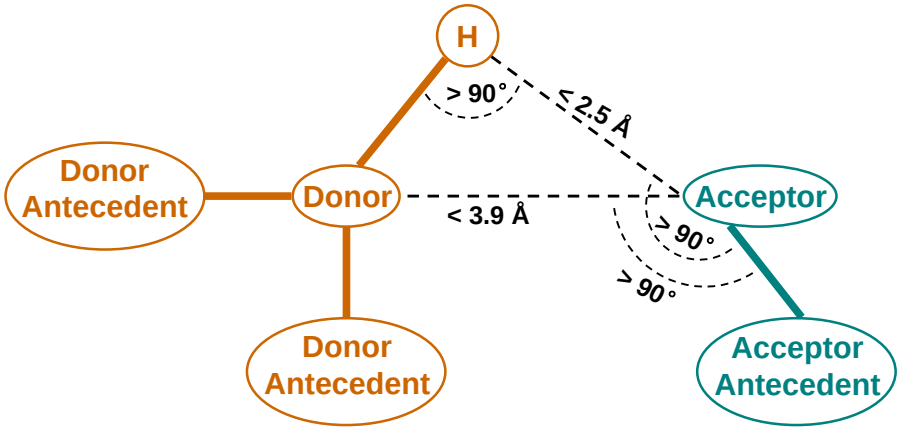


Figure 3.5: Geometric criteria for identifying hydrogen bonds (McDonald and Thornton, 1994).

study were taken from Chothia (1975). The main radii values are listed in Table 3.1. These values were derived from the intermolecular distances of a set of accurate crystal structures (Chothia, 1975). The user specifies one of the two binding sites as the *head* site and the other as the *tail* site. All CVecs point from the tail to the head site.

Hydrogen bond vectors (HVecs) are the second type of NCIV. These were determined by first adding hydrogen atoms to the protein structures with the REDUCE program Word *et al.* (1999a) and then applying a set of geometric criteria (McDonald and Thornton, 1994, see Figure 3.5). The directions of the HVecs encode the hydrogen bonding donor–acceptor direction.

The distance between a pair of NCIVs was defined as the Euclidean distance between their two midpoints. The midpoint of a vector is the mean of its head and tail.

Clustering NCIVs

In this step, two CVecs were grouped into the same cluster if they were closer than 2.0 Å and if the angle between them was less than or equal to 45°. Subsequently, a consensus vector was computed and then used as a *representative* for each cluster. The consensus vector points from the centroid of all the tails of the vectors in the cluster, to the centroid of all the heads. A complete linkage hierarchical clustering algorithm was employed to cluster the NCIVs. The distance between *representatives* was defined in the same way as the distance between NCIVs.

HVecs were not clustered and each HVecs itself was taken as a HVec representative. This is because the number of HVecs at protein-protein interface is generally small and they are usually dispersed at the interface. A statistical analysis carried out by Xu *et al.* (1997) shows that on average only one hydrogen bond is expected per 100 Å^2 at protein-protein interfaces.

This clustering step is based on the observation that often there are small groups (size 2–4) of CVecs with similar orientations (angle difference $\leq 45°$). In the distribution of inter-CVec interaction distances, we observed that there are three distance ranges below 2 Å within which the inter-vdW interaction distance occurs more frequently than the other ranges (see the inset in Figure 3.6a). These three ranges are (0.6, 0.8 Å), (1.05, 1.45 Å), (1.75, 2.00 Å). We inspected a number of vdW interactions at different protein-protein interfaces and discovered that there are a few common patterns of interatomic interactions (see Figure 3.7). Such interactions between neighboring atoms are very close to each other and exhibit similar orientations. It is reasonable to group them as one single consensus interaction. Clustering NCIVs also reduces the size of the alignment problem and enables Galinter to obtain results in reasonable run time (within minutes).

Generating a Graph Representation for Protein-Protein Interfaces

In this step, each protein-protein interface is modeled as an undirected node- and edge-labeled graph $G(V, E)$. Node set V consists of all the NCIV *representatives* obtained in the previous step. Each node is labeled as either a CVec *representative*, or a HVec *representative*. Two nodes u, v are connected by an edge if the distance between the corresponding NCIVs is in the range from 2.0 to 40.0 Å. Each edge is labeled with a 5-tuple *EdgeLabel*. In every *EdgeLabel*, the first value is the distance between the corresponding NCIVs, and the other four values are the distances between each pair of endpoints of these two NCIVs. That is, suppose the two NCIVs are $\overrightarrow{AB}$ and $\overrightarrow{CD}$, and the midpoints of $\overrightarrow{AB}$ and $\overrightarrow{CD}$ are M and N, then $EdgeLabel(u, v) = (d_{MN}, d_{AC}, d_{AD}, d_{BC}, d_{BD})$.

We chose 2.0 Å as lower bound because in the previous clustering step the cluster radius was also 2.0 Å. The upper bound was chosen as 40.0 Å based on the analysis of inter-vdW interaction distances of interface atoms. We obtained a distribution of the inter-vdW interaction distances between heavy atoms at protein-protein interfaces based on a structurally non-redundant two-chain interface dataset (Keskin *et al.*,

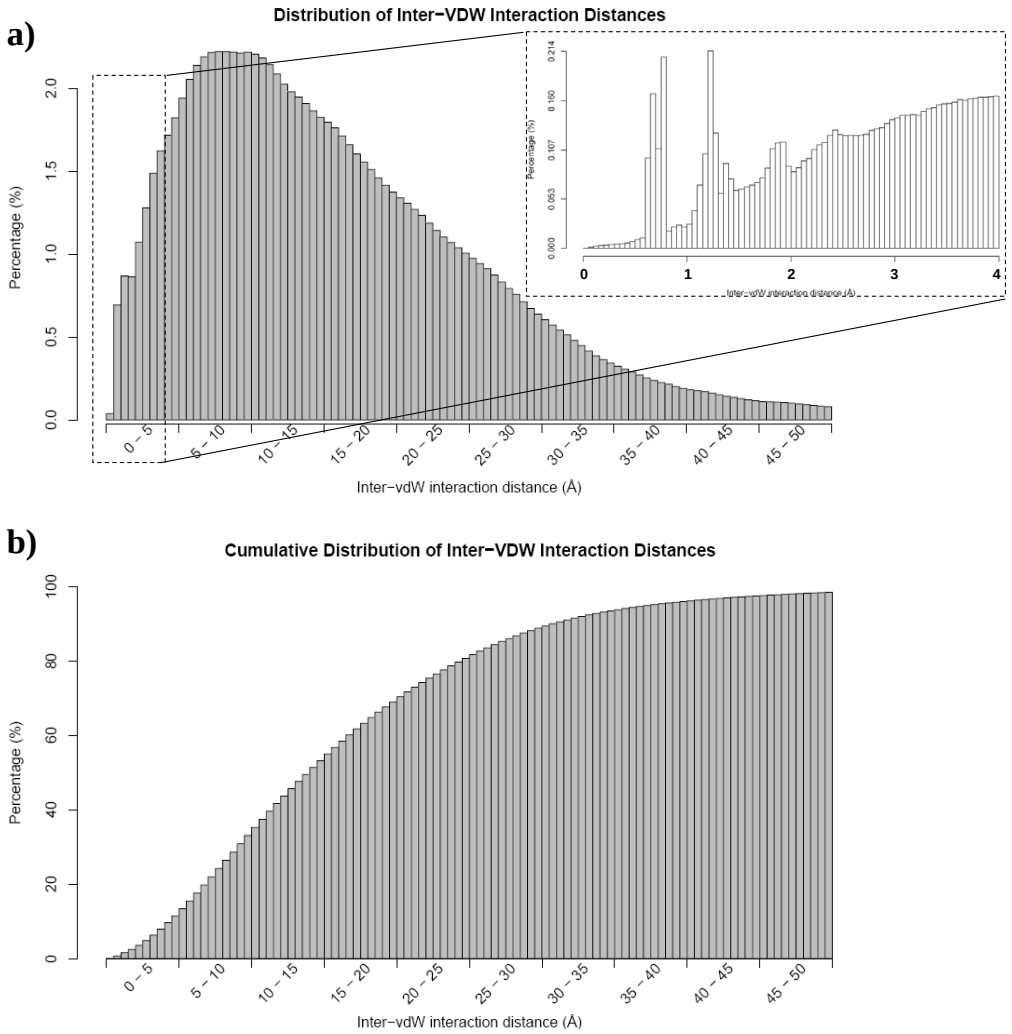


Figure 3.6: Distribution of inter-vdW interaction distances between van der Waals interactions at interfaces in a non-redundant two-chain interface dataset (Keskin *et al.*, 2004). a) the distribution of inter-vdW interaction distances; b) the cumulative distribution of inter-vdW interaction distances. The distance between two vdW interactions is computed as the distance between the midpoints of the vector representations of the vdW interactions. Only inter-vdW interaction distances up to 50 Å are shown.

2004). Of the 969 interfaces in the original dataset (Keskin *et al.*, 2004), a total of 937 were used in the analysis. Some interfaces were excluded because the chain names are unknown or because the two chains do not have any atomic contacts according to our distance criteria. The cumulative distribution of inter-vdW interaction distances is shown in Figure 3.6b. About 95.6% of inter-vdW interaction distances are below 40 Å. Thus the choice of 40 Å as the upper bound excluded less than 5% of the NCIVs. The purpose of introducing these cutoff values was to decrease the product graph size in terms of edge number and thus reduce run time of the following clique search step.

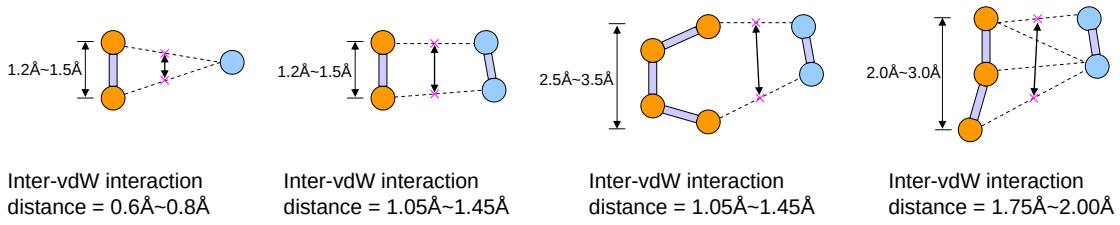


Figure 3.7: Common patterns of close interatomic interactions at protein-protein interfaces. Each sphere represents an atom. Atoms from different subunits are colored differently. Thick sticks represent covalent bonds. Dashed lines denote van der Waals interactions between atoms. The distance between two vdW interactions is computed as the distance between the midpoints (pink crosses) of the vector representations of the vdW interactions.

Aligning Representatives

Given two graphs $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$ representing two protein-protein interfaces, our goal was to find all *maximum common subgraphs* H_1 and H_2 such that

i) $H_1 \subseteq G_1$, $H_2 \subseteq G_2$, H_1 and H_2 are isomorphic $H_1 \equiv H_2$, and

ii) there is no pair (H_1', H_2') such that $H_1 \subseteq H_1' \subseteq G_1$, $H_2 \subseteq H_2' \subseteq G_2$, $H_1' \equiv H_2'$, and H_1', H_2' have more nodes than H_1 and H_2, respectively.

The maximum common subgraph problem was transformed to the maximum clique problem in the traditional fashion (Grindley *et al.*, 1993; Koch *et al.*, 1996). Maximal common subgraphs in G_1 and G_2 were identified by searching for maximal cliques in a product graph of G_1 and G_2 (Levi, 1972; Koch *et al.*, 1996). In our method, the alignment of NCIV representatives comprises two stages: building product graph and detecting cliques in the product graph.

Building Product Graph The product graph $P(V_P, E_P)$ has a node set

$$V_P = \{\, (u_1, u_2) \,|\, V_1 \times V_2 \text{ and } \mathit{label}(u_1) = \mathit{label}(u_2) \,\}$$

In P, two nodes (u_1, u_2) and (v_1, v_2) are connected if and only if (u_1, u_2) and (v_1, v_2) are different, u_1, v_1 are connected in G_1 and u_2, v_2 are connected in G_2 and for each $i \in (1, ..., 5)$:

$$\begin{aligned} |\, \mathit{EdgeLabel}(u_1, v_1)[i] - \mathit{EdgeLabel}(u_2, v_2)[i] \,| \\ \leq TOL_{rep}(\mathit{EdgeLabel}(u_1, v_1)[i], \mathit{EdgeLabel}(u_2, v_2)[i]) \end{aligned}$$

where TOL_{rep} is a tolerance function defined as:

$$TOL_{rep}(a, b) = \begin{cases} 1.0 + (\frac{a+b}{2}) \,/\, 20 & \frac{a+b}{2} < 20.0\,\text{Å} \\ 2.0 & \frac{a+b}{2} \geq 20.0\,\text{Å} \end{cases} \qquad (3.1)$$

The function enforces an upper limit on the difference of two distances.

Compatibility of Distances between NCIVs The tolerance function defined in Equation 3.1 was derived from the analysis of a set of protein-protein interfaces in which the subunits exhibit significant sequence and structural similarity. The idea was to examine the variance of inter-NCIV distances among a group of homologous interfaces, which allows to determine the upper limit for defining the compatibility of inter-NCIV distances. For two interfaces $I_1(A_1, B_1)$ and $I_2(A_2, B_2)$, where A_1 and B_1 are the two interacting subunits involved in I_1 and A_2 and B_2 are the two interacting subunits involved in I_2, we considered them to be homologous if A_1 and A_2 are homologous, and B_1 and B_2 are homologous. In such cases, the two protein complexes A_1B_1 and A_2B_2 are defined to be double-sided homologous. We obtained a set of homologous interfaces from the *pilot dataset* of interfaces used by Shulman-Peleg *et al.* (2004)[3]. There are 22 groups in the dataset, in which only six groups contain double-sided homologous complexes (group 4, 7, 15, 16, 17 and 18. See also Figure 3.13). Three of the six groups contain only two interfaces (group 4, 7 and 15) and are not considered. In the end, we used groups 16, 17, and 18 for the inference of the tolerance function.

The backbone structures of the protein subunits in these three groups are homologous based on SCOP classification. In group 16, all subunits are from SCOP family a.39.1.2. In group 17, all subunits are from SCOP families a.45.1.1 or c.47.1.5. In group 18, all subunits are from SCOP family d.153.1.4. Non-redundancy is assured by sequence comparison using the bl2seq program (Tatusova and Madden, 1999). The results reveal that the subunit proteins have relatively low sequence identity (see Table 3.2).

In order to examine the variance of inter-NCIV distances at different interfaces, the correspondence between NCIV representatives at different interfaces needed to be determined. We implemented this in two steps. First, we determined the correspondence between interface residues. The correspondence was inferred from the multiple structural alignment of the subunit proteins using the MultiProt program (Shatsky *et al.*, 2002). Since all complexes are homo-dimers in the three groups, we aligned only one side of these homo-dimers in each group. Next, the correspondence between NCIV representatives among interfaces were determined by the correspondence between the interface residues that are closest to the NCIV representatives.

The distances between corresponding interface NCIV representatives were computed and compared for the 12 interfaces from groups 16, 17, and 18 (Figure 3.8). The relationship between the average distances and the differences in the distances of NCIV representatives is shown in Figure 3.9. The tolerance function (3.1) was then designed in a way that it is close to the 70-percentile. The choice was made mainly based on the inspection of the protein stuctures. In addition, this design of the tolerance function also led to the best alignment results in our test.

[3]Available at `http://bioinfo3d.cs.tau.ac.il/I2I-SiteEngine/about/clusters_table.html`

Table 3.2: Sequence alignment results for subunit proteins in groups 16, 17, and 18 of the pilot dataset.

Group 16

	1irjAB	1dt7AB	1e8aAB
1bt6AB	28/87 (32%) 6e-12[a]	33/88 (37%) 4e-17	27/78 (34%) 9e-13
1irjAB	-	34/88 (38%) 4e-18	42/91 (46%) 3e-22
1dt7AB	-	-	32/85 (37%) 2e-19

Group 17

	10gsAB	1pd212	1axdAB
1b48AB	65/192 (33%) 5e-26	58/202 (28%) 1e-21	38/157 (24%) 1e-07
10gsAB	-	54/202 (26%) 2e-16	42/158 (26%) 1e-09
1pd212	-	-	23/80 (28%) 9e-05

Group 18

	1iruFG	1g0uOP	1pmaAC
1iruOP	62/239 (25%) 2e-27	82/245 (33%) 3e-34	84/239 (35%) 2e-37
1iruFG	-	70/219 (31%) 1e-28	77/230 (33%) 1e-36
1g0uOP	-	-	83/197 (42%) 2e-40

[a]The three fields in each cell give the values for 1) the number of identical residues/alignment length; 2) (percentage of sequence identity); 3) E-value of the alignment computed using bl2seq.

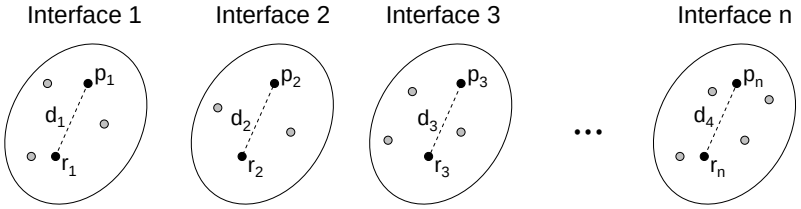


Figure 3.8: Comparison of NCIV representative distances at different interfaces. The nodes p_i and $r_i, i = 1, 2, ..., n$ are equivalent NCIV representatives across interfaces based on the multiple structure alignment of the subunit structures. Each pair of the distances $d_i, i = 1, 2, ..., n$ are compared.

Detecting Cliques in Product Graph After obtaining the product graph, cliques were detected using the Bron-Kerbosch algorithm (Bron and Kerbosch, 1973). The cliques in the product graph correspond to aligned *representatives* between interfaces. Only the largest alignments of *representatives* were considered in the following step.

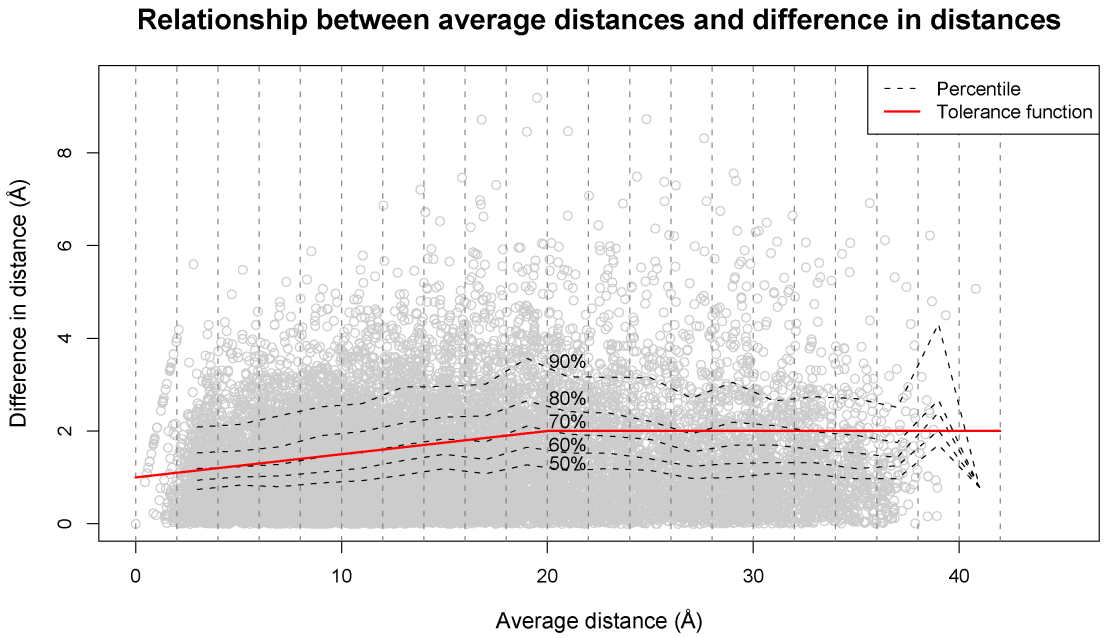


Figure 3.9: Relationship between average distance and difference in distances for interface NCIV representatives. The data are based on the interfaces in group 16, 17, and 18 of the pilot dataset (Shulman-Peleg *et al.*, 2004)

Extending Aligned Representatives to NCIVs

Up to this stage, the alignment consists of aligned *representatives* of NCIV clusters. In this step, these aligned *representatives* were used as "anchors" for deriving the alignment between the original sets of NCIVs. The extension from the aligned representatives to the original NCIVs were performed in an iterative manner. An *expanding* step and a *filtering* step was repeated until no new pairs of NCIVs were able to be aligned.

First, in an expanding procedure, two NCIVs were matched if they fulfilled the following *expanding criteria*:

i) they are of the same type,

ii) they have similar orientations (the angle between them $\leq 45°$) after the transformation based on the superposition of the anchors, and

iii) they have similar distances to the anchors.

In general, the positions of NCIVs at interfaces are close to their NCIV representatives. Thus, we defined a tolerance function TOL_{vec} for distance differences between NCIVs in a similar way to TOL_{rep}:

$$TOL_{vec}(a,b) = \begin{cases} 1.0 + (\frac{a+b}{2}) \,/\, 40 & \frac{a+b}{2} < 20.0\,\text{Å} \\ 1.5 & \frac{a+b}{2} \geq 20.0\,\text{Å} \end{cases} \qquad (3.2)$$

where a and b are the distances to be compared. TOL_{vec} is more restrictive than

TOL_{rep}, as it is applied to actual NCIVs instead of *representatives*.

Our goal was to find a maximum matching between the NCIVs at the two interfaces. However, following the expanding criteria in the expanding step, we detected more than one match for a NCIV. If we consider the two sets of NCIVs at the two interfaces as two disjoint sets of vertices, and the pairwise matches between the NCIVs as edges between corresponding vertices, the problem can then be described as a *maximum bipartite matching* problem (Cormen *et al.*, 2001). A *bipartite* graph $G = (V, E)$ is a graph with vertex partition $V = L \cup R$, where L and R are disjoint and all edges in E are between L and R. A *matching* in a bipartite graph is a subset of edges $M \subseteq E$, in which no two edges share a common vertex. A *maximum matching* in a bipartite graph is a matching with the maximum cardinality. To solve the maximum bipartite matching problem, a commonly employed algorithm is the Hopcroft-Karp algorithm (Hopcroft and Karp, 1973). This algorithm was implemented in the Galinter program for identifying the maximum matching between NCIVs in the expanding step.

After finding all the potential alignments of NCIVs, a filtering procedure was performed in order to remove incompatible matches of NCIVs. In the expanding step, matches between anchors were expanded in a greedy manner. According to the expanding criteria, all matches between NCIVs that satisfy the tolerance defined in Equation 3.2 with respect to their anchors were added to the alignment. But the newly discovered matches did not necessarily fulfill the tolerance criterion with respect to other matched NCIVs in the alignment. Therefore, we refined the expanded set of aligned NCIVs in this filtering step. A pair of aligned NCIVs found in the expanding procedure was discarded if the difference of their distances to any other pair of aligned NCIVs exceeded the tolerance defined by TOL_{vec} in Equation 3.2. For a pair of aligned NCIVs, those other pairs of NCIVs with incompatible distances to the aligned pair were defined as the *incompatible pairs*. The removal of aligned NCIVs with incompatible distances to other aligned NCIVs was implemented using a greedy algorithm. First, for each pair of aligned NCIVs, all incompatible pairs were identified and counted. Then the pair of NCIVs with highest number of incompatible pairs was removed. The numbers of incompatible pairs for the rest pairs of aligned NCIVs were updated accordingly. These two steps were repeated until there is no more incompatible pairs in the alignment.

The resulting matched NCIVs replaced the aligned *representatives* as new anchors, and the expanding and filtering procedures were repeated. Newly found matches of NCIVs were added to the anchors, until no more NCIVs could be matched in the expanding procedure. All resulting alignments of NCIVs were sorted according to their sizes, and the largest alignments and corresponding transformations were reported.

3.2.2 Validation of Alignment Algorithm

The comparison of Galinter to other methods was difficult, since no other interface comparison tool produces alignments of non-covalent interactions. At the same time,

there was a lack of datasets of interfaces that had been thoroughly compared based on the patterns of non-covalent interactions at the interfaces, to which we could benchmark the Galinter method.

Despite of these restrictions, we decided to apply the Galinter program on a published dataset of interfaces, which was compared and clustered by a similar interface comparison program I2I-SiteEngine (Shulman-Peleg *et al.*, 2004). The results obtained by using the Galinter program were compared to the results of I2I-SiteEngine. I2I-SiteEngine compares interfaces by aligning the functional groups at binding sites, instead of aligning molecular interactions within the interface like Galinter. Galinter and I2I-SiteEngine can be regarded as complementary approaches as they use different properties to compare interfaces.

In addition, backbone structure comparison methods like DaliLite (Holm and Park, 2000) can be used to generate interface alignments indirectly. These alignments are indirect in the sense that they do not take the structural similarities of the interfaces into account explicitly. When the interaction orientations of subunits are conserved between complexes, the indirect alignments derived from backbone structure comparison of homologous subunits between complexes provide a coarse way of validating alignments from direct methods like Galinter and I2I-SiteEngine. The alignments based on backbone structures are expected to agree with explicit alignments of non-covalent interactions within the interfaces to some extent but not necessarily to match them.

In this section, we describe the application of the Galinter program on the published dataset of protein-protein interfaces used by I2I-SiteEngine. The alignments were compared to those obtained by using two relevant programs, I2I-SiteEngine and DaliLite. A measure for assessing the agreement between the alignments of interfaces (*interface residue RMSD*, or *irRMSD*) was proposed. We also explain the disagreements in the alignments from different programs.

Pilot Dataset

We applied Galinter to the pilot dataset that was used for testing I2I-SiteEngine. This dataset consists of 64 protein-protein interfaces clustered into 22 groups according to I2I-SiteEngine alignment results. It is composed of a variety of protein complexes, including antigen-antibody, protease-inhibitor, protein-peptide, and protein-protein dimers. There are both homo- and hetero-dimers in the dataset. We excluded eight singleton groups from the dataset. The following analysis is restricted to the remaining 14 non-singleton groups.

We defined the homology of interfaces based on the homology of the interacting subunits. For any pair of complexes to be compared, if at least one subunit of one complex was homologous to at least one subunit of the other complex, then the two complexes were labeled as *S/D-homologous* (single- or double-sided homologous). Otherwise the two complexes were labeled as *non-homologous*. Two subunit structures were considered to be homologous if they belonged to the same *superfamily* in SCOP (Murzin *et al.*, 1995). In nine of the 14 groups, all complexes are S/D-

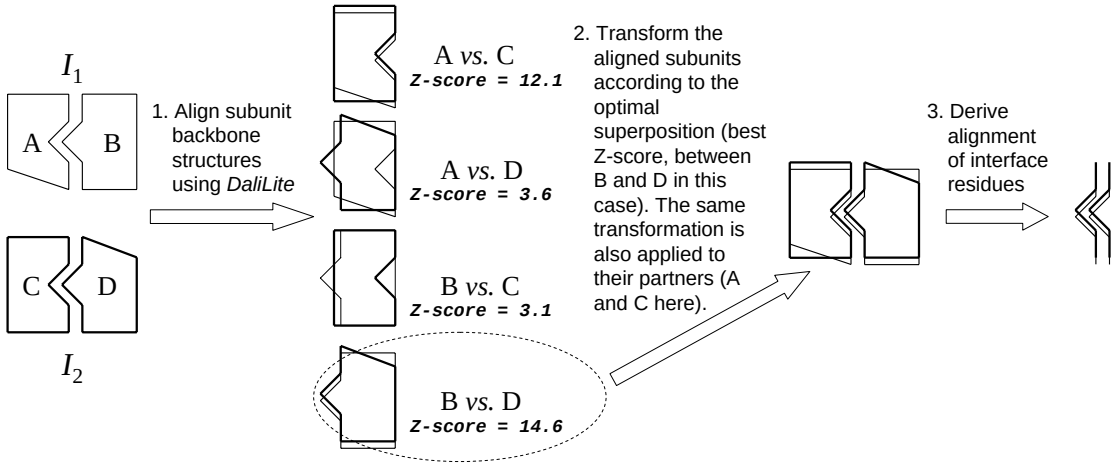


Figure 3.10: Interface alignment based on backbone superposition. First, subunit structures are compared individually at both sides of interfaces using DaliLite. Based on the optimal structure alignment of subunits (subunits B and D in this case as highlighted by dashed circle), their partners are transformed accordingly. A subsequent alignment of interface residues can be derived from the superposed complexes.

homologous to each other within the group. The remaining groups also contain some complexes not related by homology. See the Figure 3.13 for more details.

Comparing Galinter to I2I-SiteEngine and DaliLite

On the pilot dataset, Galinter alignments were compared to the alignments generated by the I2I-SiteEngine interface comparison method. I2I-SiteEngine matches chemical functional groups and associated residues at the binding sites of different interfaces. In addition, we compared the results of both Galinter and I2I-SiteEngine to alignments based on backbone structure, generated with DaliLite (Holm and Park, 2000). Using DaliLite, subunit structures were compared individually at both sides of interfaces. A subsequent alignment of interface residues was derived based on the most significant DaliLite alignment of subunit structures as detailed in Figure 3.10.

Assessing the Agreement of the results

In this work, we defined interface residues as those which contain at least one interface atom, where interface atoms are the atoms involved in interface NCIVs. We compared the alignment of interfaces from the different methods (Galinter, I2I-SiteEngine, and DaliLite) by examining the deviation of C_α atom coordinates of interface residues after corresponding transformations. Given two interfaces I_1 and I_2 and two alignment methods M_a and M_b, we considered I_1 as the reference, and let I_{2a} correspond to the transformed interface I_2 according to the optimal superposition between I_1 and I_2 based on the alignment from method M_a. Analogously, I_2 was transformed to I_{2b} based on the alignment from method M_b. Then, the root-mean-square devia-

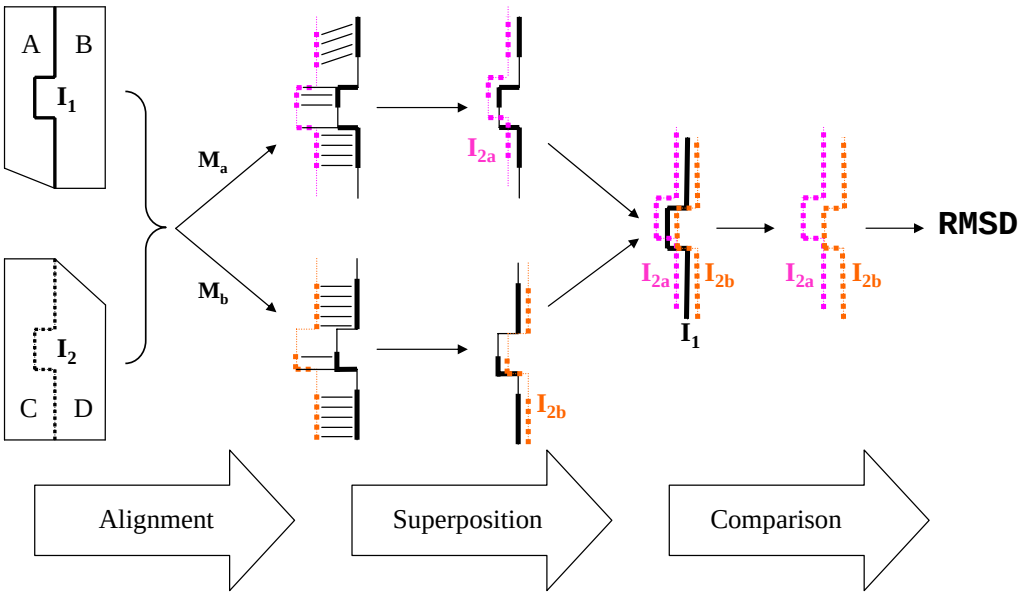


Figure 3.11: Comparison of interface alignments using irRMSD. Two interfaces $I_1(A, B)$ and $I_2(C, D)$ are aligned using two methods M_a and M_b. I_1 is considered the reference. I_{2a} and I_{2b} correspond to the transformed I_2 according to the optimal superpositions between I_1 and I_2 based on the alignments from methods M_a, and M_b, respectively. The interface residue RMSD (irRMSD) is calculated as the RMSD for all C_α atoms of interface residues in I_{2a} and I_{2b} to assess the agreement between the two methods. (NCIV: non-covalent interaction vector)

tion (RMSD) for all C_α atoms of interface residues in I_{2a} and I_{2b} was calculated to assess the agreement between the two methods M_a and M_b. We defined this measure the *irRMSD*, which stands for *interface residue RMSD*. See Figure 3.11 for an illustration of the calculation of irRMSD.

Validation Results

To assess whether Galinter produces valid interface alignments, we compared the results of Galinter to the alignments generated by I2I-SiteEngine and DaliLite.

Validation Results on the Pilot Dataset We applied Galinter to every pair of interfaces within each of the 14 groups from the pilot dataset. There are 240 comparisons in total. The mean run time is 138.5 seconds (median run time 71.5 seconds) on a normal desktop (3.0GHz CPU, 1GB memory) for these comparisons. The alignment results were compared to those of I2I-SiteEngine and DaliLite. The extent of agreement is measured using irRMSD as described in Section 3.2.2.

Figure 3.12 provides a summary of the irRMSD values obtained in the analysis. All pairwise comparisons of interfaces were separated into two groups according to whether the corresponding complexes are S/D-homologous or non-homologous. Of

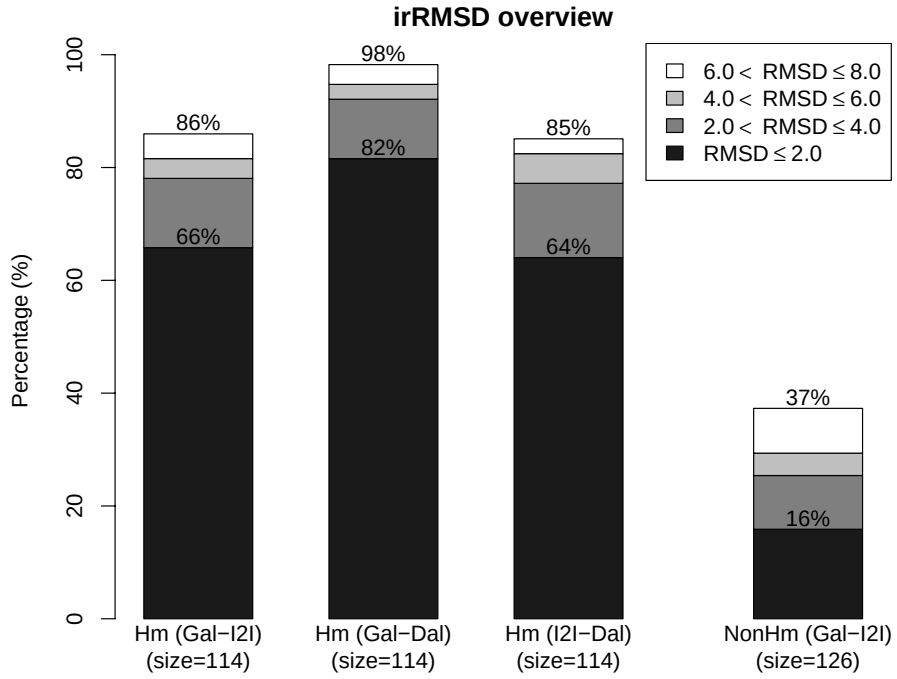


Figure 3.12: Overview of irRMSD values for pairwise comparison of protein-protein interfaces. Most interfaces for non-homologous complexes cannot be compared using backbone alignment method. Thus, for the alignments of non-homologous complex interfaces, only an overview of irRMSD values for the comparison between Galinter and I2I-SiteEngine are shown. (Hm: S/D-homologous; NonHm: non-homologous; Gal: Galinter; I2I: I2I-SiteEngine; Dal: DaliLite)

the 240 pairs of interfaces compared, 114 are S/D-homologous and the remaining 126 pairs are non-homologous. For the alignments of non-homologous interfaces, only irRMSD values for the comparison between Galinter and I2I-SiteEngine are shown, because most non-homologous interfaces could not be aligned using DaliLite as there is no backbone structural similarity between the respective protein complexes.

Figure 3.12 shows that for S/D-homologous interfaces, Galinter alignments usually agree with I2I-SiteEngine alignments. The alignments are similar (irRMSD $\leq$ 2 Å) for 66% of the cases. Galinter and I2I-SiteEngine both produce similar alignments to DaliLite if the interfaces are S/D-homologous. But the agreement between Galinter and DaliLite is generally higher than that between I2I-SiteEngine and DaliLite. When the alignments generated by Galinter were compared to those by DaliLite, 82% of the comparisons have an irRMSD less than or equal to 2 Å. This value is only 64% for the comparison between I2I-SiteEngine and DaliLite. We believe this is mainly because Galinter alignments are derived directly from individual non-covalent interactions, while I2I-SiteEngine aligns pseudopoints, each of which represents the average of a group of atoms. Therefore Galinter provides more precise alignments which agree with those based on DaliLite for S/D-homologous interfaces. For non-homologous interfaces, Galinter and I2I-SiteEngine generate very different

alignments. Less than 40% of the 126 comparisons have irRMSD values below 8 Å. The results for each comparison is given in Figures 3.13, 3.14, and 3.15.

Disagreements between Alignment Results We explored possible causes for the disagreements between the alignments of different methods. For non-homologous interfaces, most of the disagreements are observed in groups 19 and 5. Group 19 consists of coiled-coil interfaces. More than a single solution is expected for the alignment of these repetitive structures. Therefore, it is not surprising that the alignments from different methods disagree. In general, the alignments of both methods result in reasonable superimposition of the helix backbones. Nevertheless, visual inspection reveals that for some of these pairs one of the methods generates better alignments with more matched residues with a comparable RMSD after optimal superposition of the interacting helices. Galinter produces better alignments for five pairs (1ic2CD vs. 1gl2BC, 1ic2CD vs. 1gk4AB, 1gl2AB vs. 1gk4AB, 1gl2BC vs. 1gk4AB, 1gk4AB vs. 1if3AB), and I2I-SiteEngine in three cases (1ic2CD vs. 1if3AB, 1gl2AB vs. 1if3AB, 1gl2BC vs. 1if3AB). For example, in the comparison of 1gl2AB and 1gk4AB, chain B of 1gl2 has 16 helix turns and they are all superposed based on the Galinter alignment, while only 8 helix turns are superposed based on the I2I-SiteEngine alignment (Figure 3.16).

In group 5, there are relatively few similarities between the subunits from different complexes. There seems to be no obvious alignment solution in terms of either structure or evolution. The only evident common feature in these interfaces is that they include two interacting β-strands. The assessment of the results in this group is thus challenging. Bearing this in mind, we investigated the quality of the results by visual inspection of the superposition of the two strands at the interfaces. We found that for 15 pairs Galinter provides better local structure superposition of the interface β-strands, and for five pairs I2I-SiteEngine leads to better superposition of these strands.

The disagreements between Galinter and I2I-SiteEngine for S/D-homologous interfaces arise mainly from group 10, and also to a lesser extent, from the smaller group 4. Interestingly, for these two groups, the Galinter alignments agree with those based on DaliLite.

Conclusion In general, the three methods agree to a large extent, especially when the interfaces are related by homology. Nevertheless, it is not surprising to observe disagreements in the non-homologous groups, considering both that Galinter and I2I-SiteEngine are based on different interface properties and that there are no unique solutions in these groups.

3.2.3 Case Studies

In this section, we present the application of Galinter to four mimicry cases, for which the interfaces have been manually compared before:

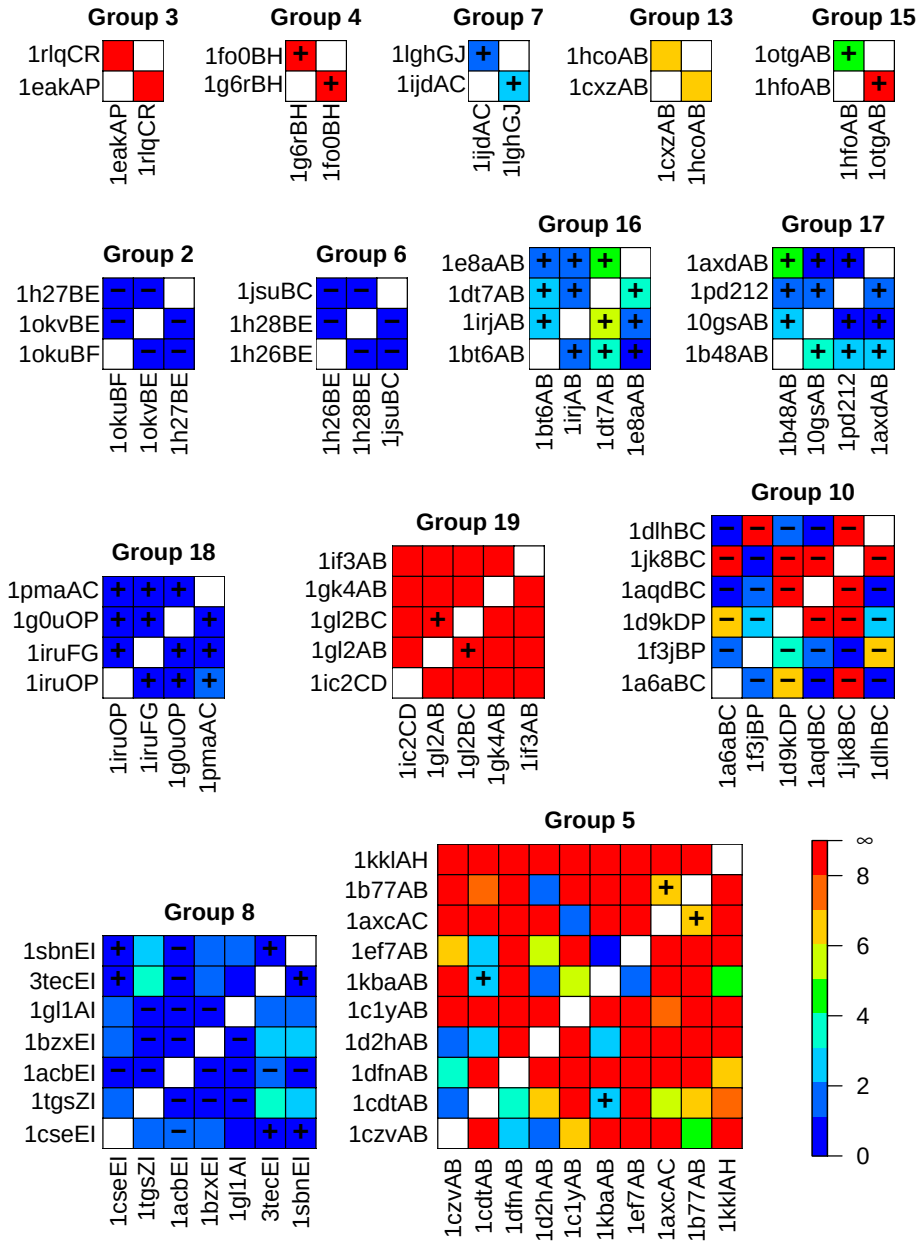


Figure 3.13: Detailed comparison between Galinter and I2I-SiteEngine results. Heat maps for irRMSD values of interface residues. Only the 14 non-singleton groups in the pilot dataset are shown. The heat maps are sorted by size. The columns and rows for each heat map represent interfaces identified by their PDB code and chain names constituting the interfaces. The diagonal grids of all heat maps have been left blank. For S/D-homologous complexes, S/D-homology is indicated in corresponding grids by either a plus sign (+) for double-sided homology, or a minus sign (-) for single-sided homology. The heat maps have been produced using R (R Development Core Team, 2005).

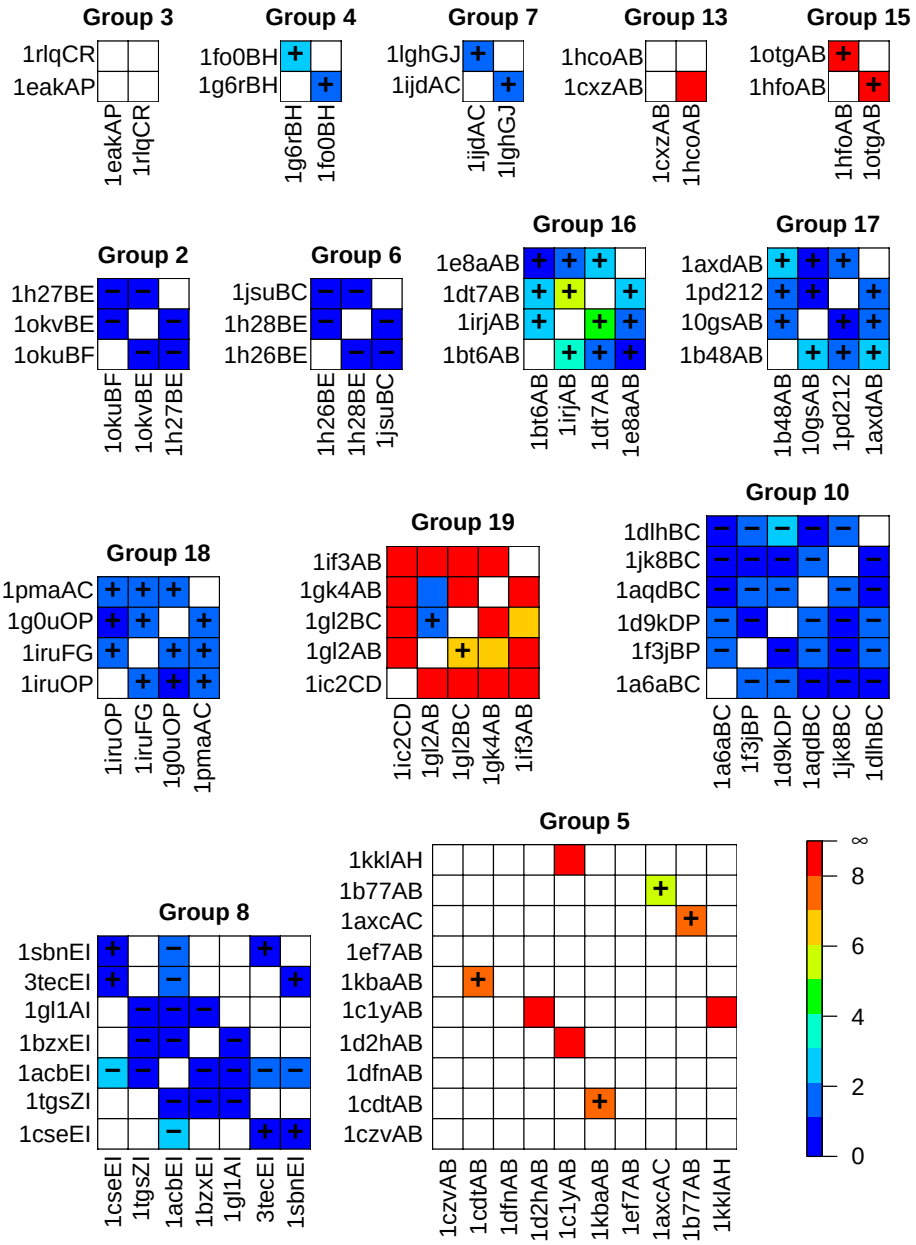


Figure 3.14: Detailed comparison of Galinter and DaliLite results. Heat maps for irRMSD values of interface residues. The same rules as in Figure 3.13 are applied here to generate the heat maps.

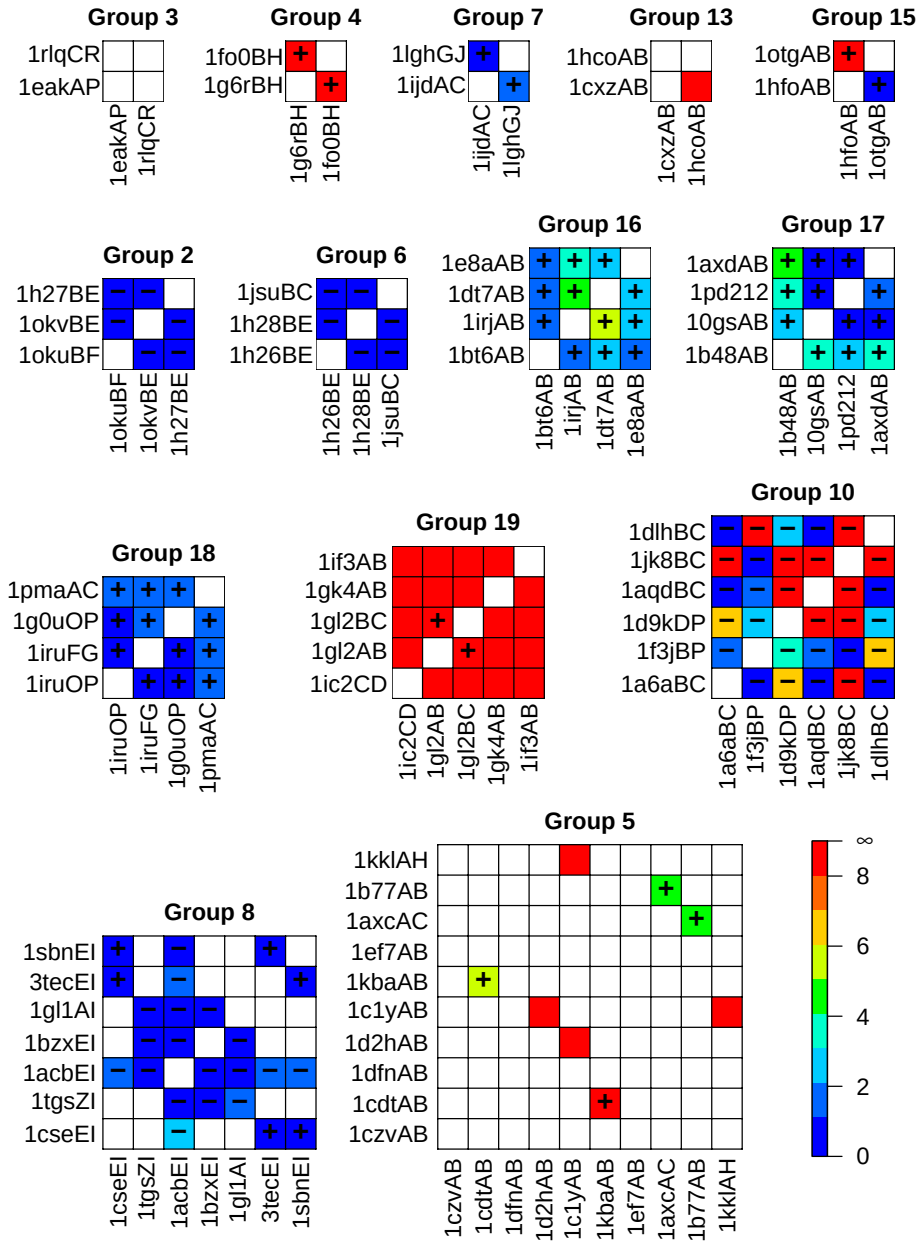


Figure 3.15: Detailed comparison of I2I-SiteEngine and DaliLite results. Heat maps for irRMSD values of interface residues. The same rules as in Figure 3.13 are applied here to generate the heat maps.

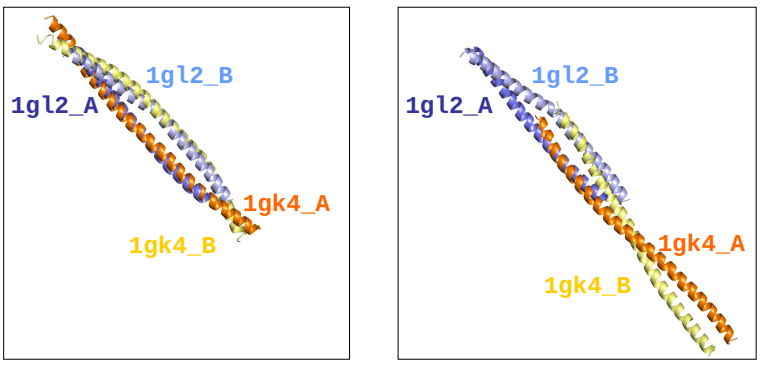


Figure 3.16: Alignment of two coiled coil interfaces (1gl2AB and 1gk4AB) using Galinter and I2I-SiteEngine. Color code: 1gl2A: dark blue; 1gl2B: light blue; 1gk4A: orange; 1gk4B: light yellow.

i) Chymotrypsin and subtilisin interact with the same type of inhibitors, an example of convergent evolution (Wallace *et al.*, 1996; Jackson and Russell, 2000);

ii) Subtilisin and trypsin interact with non-homologous inhibitors, an example of convergent evolution with no structure homology (Wallace *et al.*, 1996; Jackson and Russell, 2000);

iii) A non-peptidic compound SP4206 mimics IL-2Rα in binding to IL-2 (Thanos *et al.*, 2006).

iv) A scorpion-toxin derived compound (CD4M33-F23) mimics CD4 in complex with gp120, a mimicry case relevant to HIV therapy (Huang *et al.*, 2005);

In each of these four cases, the subunits are either homologous only on one side of the interface, or non-homologous on both sides. In the third case, one of the interacting partners is not even a protein.

In addition, we applied I2I-SiteEngine to align three of the four pairs of mimicry interfaces (I2I-SiteEngine is not applicable for the third case). The DaliLite program was also executed for obtaining indirect alignments of interfaces if applicable. The alignment results were compared to those obtained with Galinter.

Two Proteases with Common Inhibitor

The Ser–His–Asp catalytic triad present in many proteases has been intensively analyzed (Berg *et al.*, 2002; Polgár, 2005) (see also Section 3.1.2). This catalytic triad occurs in several protein families that are non-homologous, and therefore have no significant backbone structural similarity (Branden and Tooze, 1999). Specifically, the trypsin-like serine proteases chymotrypsin and subtilisin belong to different SCOP superfamilies (*sccs* codes: b.47.1.2 and c.41.1.1, respectively). Although they lack obvious sequence or structural similarity, they have been found to share as many as three inhibitors (Henschel *et al.*, 2006).

We have analyzed the interactions formed between chymotrypsin and leech pro-

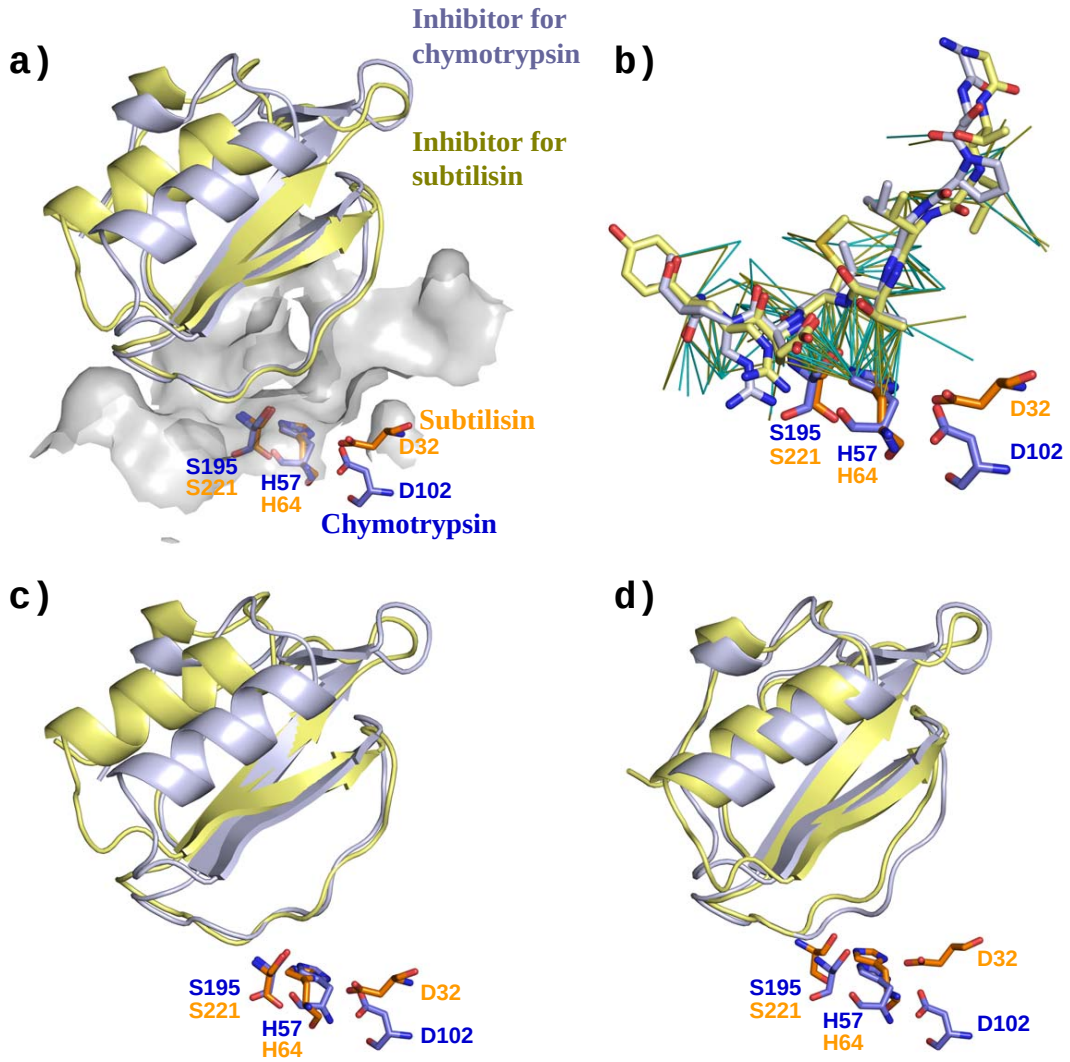


Figure 3.17: Comparison of two protease-inhibitor interfaces (single-sided homologous). **a)** Superposed inhibitors and catalytic triads for chymotrypsin (1acb) and subtilisin (1lw6) according to the Galinter alignment. The inhibitor for Chymotrypsin is shown in light blue and the inhibitor for subtilisin is shown in light yellow. The catalytic triads of chymotrypsin and subtilisin are shown as sticks in dark blue and orange, respectively. The chymotrypsin binding site is shown as a gray surface. **b)** Superposed NCIVs for chymotrypsin/inhibitor interface (1acbEI) and subtilisin/inhibitor interface (1lw6EI) according to the Galinter alignment. NCIVs are depicted as thin lines. Only aligned NCIVs are shown. Chymotrypsin/inhibitor NCIVs are shown in cyan, and subtilisin/inhibitor NCIVs are shown in yellow. **c)** Superposed inhibitors and catalytic triads according to I2I-SiteEngine alignment. **d)** Superposed inhibitors and catalytic triads according to DaliLite alignment.

Table 3.3: Comparison of Alignment Results for 1acbEI and 1lw6EI.

	RMSD (C_α atoms of inhibitor)	RMSD (Functional template atoms of catalytic triads)	irRMSD (Compared to Galinter alignment)
Galinter	2.9 Å	0.5 Å	-
I2I-SiteEngine	4.2 Å	1.1 Å	1.0 Å
DaliLite	1.5 Å	2.2 Å	2.7 Å

teinase inhibitor eglin *c* (PDB code: 1acb, chains E and I), and subtilisin with chymotrypsin inhibitor 2 (PDB code: 1lw6, chains E and I). The two protease inhibitors have similar backbone structures and belong to the same SCOP family (b.40.1.1). The two interfaces contain 299 and 332 NCIVs, respectively. The longest Galinter alignment consists of 117 aligned NCIVs, and the results are visualized in Figure 3.17a and 3.17b. According to this alignment, the two catalytic triads are superposed with an RMSD of 0.5 Å (Figure 3.17a). The RMSD is computed for the overall functional template atoms of the catalytic triads as defined in Wallace *et al.* (1996). Figure 3.17b displays superposed NCIVs according to Galinter alignment. It is noticeable that the NCIVs involving the catalytic serine and histidine residues are well conserved.

For these two protease-inhibitor interfaces, I2I-SiteEngine generates a similar alignment to Galinter with an irRMSD of 1.0 Å (Figure 3.17c). The RMSD for the overall functional template atoms of the two catalytic triads is worse than that calculated based on Galinter alignment (1.1 Å vs. 0.5 Å). In addition, the RMSD for the two inhibitors is 4.2 Å, which is higher than that obtained based on Galinter result (2.9 Å) (Table 3.3).

We also compared the two interfaces based on inhibitor backbone alignment. First the inhibitor structures of the two complexes were aligned using DaliLite. Then the two proteases were superposed accordingly. This way, an alignment of the interfaces was obtained indirectly (see Figure 3.17d). This indirect alignment agrees with the Galinter alignment to a considerable extent (irRMSD = 2.7 Å). Based on this indirect alignment, the RMSD for the overall functional template atoms of the catalytic triads is much larger than the one obtained based on the Galinter alignment (2.2 Å *vs.* 0.5 Å). This is not surprising given that these catalytic residues are not used by DaliLite when computing the alignment. Given that the DaliLite method optimizes the alignment of the backbone structures for the inhibitors, it is natural to observe that the RMSD for the inhibitors is the lowest among the three methods (Table 3.3). Meanwhile, these results also indicate that to compare protein-protein interfaces, an explicit interface alignment approach is more adequate than an approach based on backbone structure.

Two Proteases with Non-Homologous Inhibitors

In this example, the interfaces of two serine proteases with their respective inhibitors are compared. It is different from the interface comparison described in the previous example, as here not only the two proteases are non-homologous, but also their inhibitors. Both the proteases possess the catalytic triad Ser–His–Asp. They were used in the construction of 3D coordinate template for searching the Ser–His–Asp catalytic triad in structural databases (Wallace *et al.*, 1996). The two inhibitors have both been discovered to exhibit the substrate-like canonical loop (see Section 3.1.2) that is responsible for the inhibitory function (Jackson and Russell, 2000).

One of the two interfaces under investigation is formed between subtilisin leech proteinase and inhibitor eglin *c* (PDB code: 1sbn, chains E and I). The other interface is between trypsin and pancreatic secretory trypsin inhibitor (PDB code: 1tgs, chains Z and I). The backbone structures of the two proteases belong to different SCOP classes. Subtilisin 1sbnE is from the SCOP family c.41.1.1, while trypsin 1tgsZ is from the SCOP family b.47.1.2. Their respective inhibitors 1sbnI and 1tgsI are from different SCOP classes d.40.1.1 and g.68.1.1 (see Figure 3.18a). There are 254 and 328 NCIVs at the two interfaces 1sbnEI and 1tgsZI, respectively. Galinter produces a largest alignment containing 113 aligned NCIVs. The subsequent superpositions of the two protease-inhibitor complexes and the two catalytic triads are visualized in Figure 3.18a, b and c. Based on this alignment, the two catalytic triads are superposed with an RMSD of 0.4 Å for their functional template atoms. Figure 3.18c shows the superposition of the catalytic triads as well as the inhibitors.

For the two interfaces involving proteases with non-homologous inhibitors, the I2I-SiteEngine program was also applied. The corresponding superposition of the two catalytic triads and inhibitors are displayed in Figure 3.18d. The RMSD for the functional template atoms of the catalytic triads after superposition is 1.8 Å.

Since there is no homology between either side of the two interfaces, we could not apply the DaliLite method to align the subunit backbone structures and obtain an alignment of the interfaces indirectly. This again reflects the need for methods for the direct comparison of interfaces.

SP4206 Mimic of IL-2Rα in Binding to IL-2

Thanos *et al.* (2003) published the structure of the small compound SP4206 binding to an IL-2 cytokine, which in turn blocks the natural interaction of IL-2 and its receptor IL-2Rα. Interestingly, although the interface size of SP4206 and IL-2 is only half as large as that between IL-2Rα and IL-2, SP4206 and IL-2Rα bind to IL-2 with similar affinities. Thanos and colleagues have discovered that this is mainly because SP4206 utilizes the same hot spot residues as IL-2Rα when interacting with IL-2 (Thanos *et al.*, 2006).

We compared the interface of IL-2Rα and IL-2 (PDB code: 1z92, chains B and A), with the interface formed between SP4206 and IL-2 (PDB code: 1py2, FRH and chain A) using Galinter. The protocol was slightly modified in order to identify hydrogen

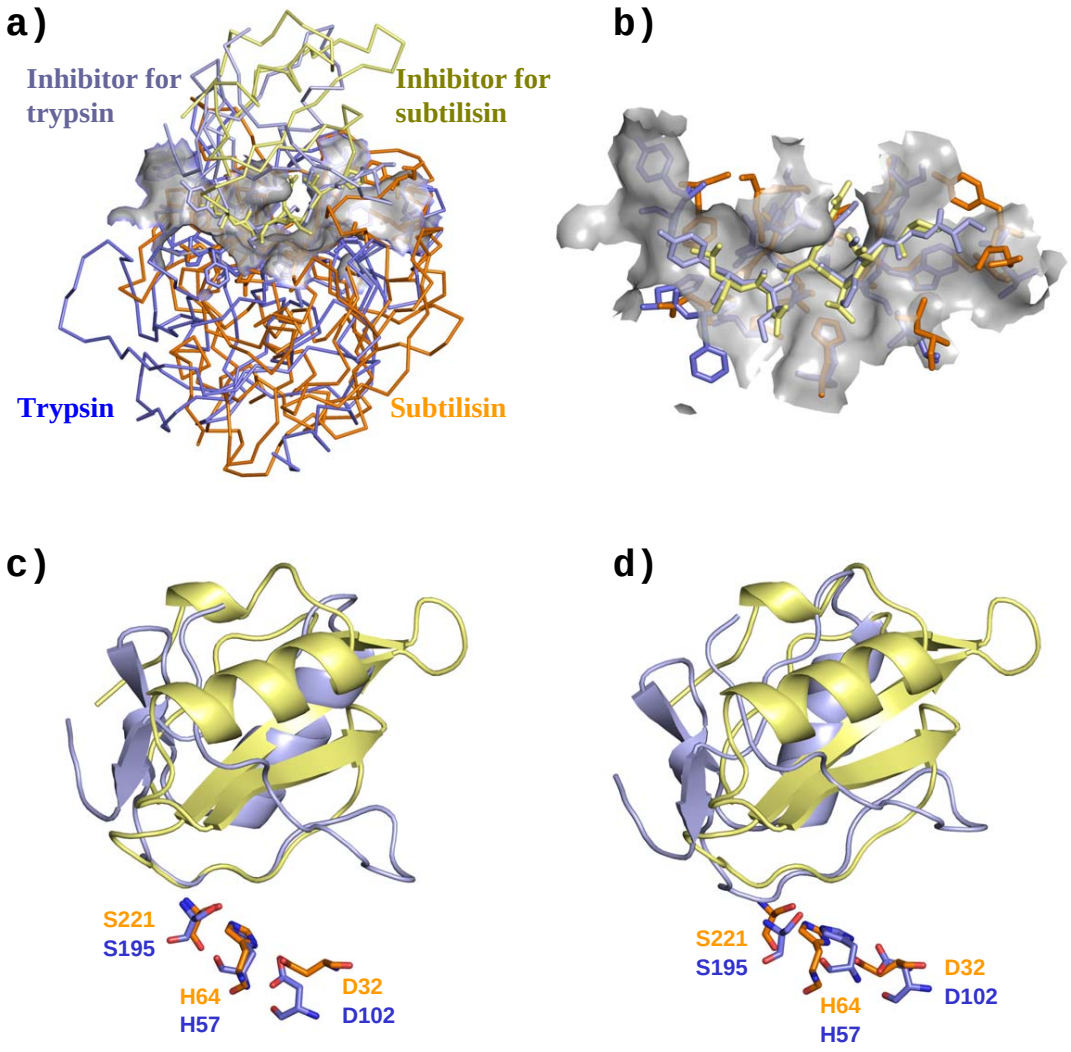


Figure 3.18: Comparison of two protease-inhibitor interfaces (non-homologous). **a)** Superposed subtilisin (1sbnE) and trypsin (1tgsZ) and their inhibitors according to the Galinter alignment of the two interfaces 1sbnEI and 1tgsZI. The inhibitor for subtilisin is shown in light yellow and the inhibitor for trypsin is shown in light blue. The backbones of subtilisin and trypsin are shown orange and in dark blue, respectively. The trypsin binding site is shown as a gray surface. **b)** Superposed canonical loops of the inhibitors and binding site residues according to Galinter alignment. The trypsin binding site is shown as a gray surface. **c)** Superposed inhibitors and catalytic triads according to the Galinter alignment. **d)** Superposed inhibitors and catalytic triads according to the I2I-SiteEngine alignment.

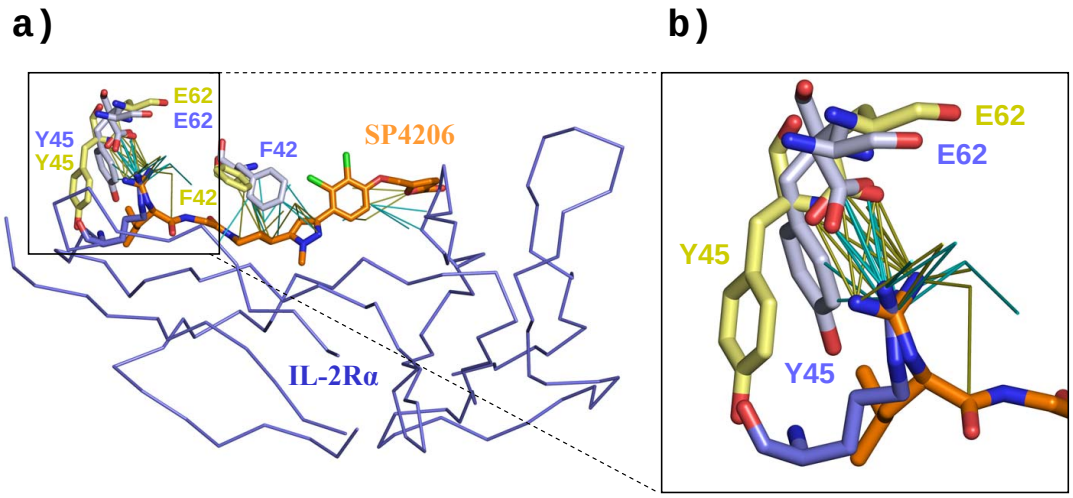


Figure 3.19: SP4206 mimic of IL-2Rα in Binding to IL-2. **a)** Superposed NCIVs according to the Galinter alignment of IL-2Rα/IL-2 interface (1z92BA) in dark and light blue, and of SP4206/IL-2 interface (1py2_A) in orange and light yellow. Only aligned NCIVs are shown. IL-2Rα/IL-2 NCIVs are shown in cyan, SP4206/IL-2 NCIVs are in yellow. The hot spot residues Phe42, Tyr45, and Glu62 in IL-2 are shown as sticks. **b)** An enlarged view of the mimic spot around residue Glu62 in IL-2.

bonds between a non-peptidic molecule and a protein. HBPLUS (McDonald and Thornton, 1994) was used to infer hydrogen bonds within the interface between SP4206 and IL-2. We identified 330 NCIVs for IL-2Rα/IL-2 interface, and 176 NCIVs for SP4206/IL-2 interface. The alignment results are shown in Figure 3.19a. Only a small amount (35 NCIVs) of the interface NCIVs are aligned by Galinter. We found that the main reason for this relatively short alignment is that the IL-2 binding sites adopt different conformations when binding the two partners. Particularly, two of the three hot spot residues on IL-2 binding sites (Phe42 and Tyr45) adopt different side chain formations in the interfaces. Only Glu62 is structurally conserved. In IL-2Rα/IL-2, this residue forms salt bridges with the guanido group of residue Arg36 in IL-2Rα. In SP4206/IL-2, we observe similar interactions between the carboxyl group of IL-2 Glu62 and the guanido group in SP4206 (Thanos *et al.*, 2006). Galinter correctly identifies these conserved interactions (see Figure 3.19b). Apparently the similarities are not uniformly distributed along the interfaces. It is noticeable that in proximity of residue Glu62 the NCIVs are conserved, while NCIVs are only sparsely aligned in the rest of the two interfaces. We label this conserved interface region a *mimic spot*, in analogy to the concept of *hot spot*, which refers to residues contributing to a large fraction of the binding energy (Bogan and Thorn, 1998).

In this mimicry case, one of the subunits participating in the interaction is a non-peptidic molecule (SP4206) and we could not obtain I2I-SiteEngine alignment. I2I-SiteEngine is only applicable to interfaces consisting of interacting proteins as it

relies on the definition of functional groups of amino acids. In this respect, Galinter is more general than I2I-SiteEngine as it can also be applied to interfaces involving non-peptidic molecules. The alignment derived from the backbone alignment of the IL-2 molecules differs from the Galinter alignment with an irRMSD of 3.6 Å. We defined six atoms in the mimic spot to be mimic spot atoms. At the IL-2Rα/IL-2 interface, they are the atoms in the carboxyl group of residue Glu62 (CD, OE1, and OE2) in IL-2, and atoms in the guanido group of residue Arg36 (CZ, NH1, and NH2) in IL-2α. At the SP4206/IL-2 interface, they are the atoms in the carboxyl group of residue Glu62 in IL-2, and the atoms in the guanido group in SP4206 (C17, N4, N1). The RMSD between the mimic spot atoms at the two interfaces based on Galinter alignment is 0.7 Å. This value is 1.0 Å for DaliLite. This suggests that the mimic spot atoms are marginally better aligned by using Galinter than using DaliLite.

A Scorpion-Toxin Derived Mimic of CD4 in Complex with gp120

In order for HIV to infect host cells, the HIV envelope glycoprotein gp120 binds CD4 receptors located on the target cell surfaces. The CD4 binding site for gp120 has been engineered onto a scorpion-toxin protein, resulting in CD4M33-F23. The mimic interaction of CD4M33-F23 in complex with gp120 has been investigated in detail and compared to the native complex structure of CD4 and gp120 (Huang *et al.*, 2005). In particular, Huang and colleagues analyzed the difference distance matrix between the two complexes for gp120 residues surrounding the hot spot residue Phe43 of CD4. This provides a localized measure of the structural mimicry of CD4M33-F23 to CD4. The results show that the structural changes induced by CD4 in gp120 are very closely mimicked by CD4M33-F23.

We compared the natural complex interfaces (PDB code: 1rzj, chains C and G) and mimicry interface (PDB code: 1yym, chains M and G) using Galinter. The numbers of NCIVs are 364 for 1rzjCG and 166 for 1yymMG. In spite of the lack of similarities between the overall folds of CD4 and CD4M33-F23, about 80% (133 NCIVs) of the NCIVs at the CD4M33-F23/gp120 interface were aligned to those at the CD4/gp120 interface. In addition, three of the four interface hydrogen bonds aligned as described in Huang *et al.* (2005) are also aligned in the same way by Galinter (Figure 3.20a).

We also observed that the hot spot residue Phe43 in CD4 (or equivalent residue Phe23 in CD4M33-F23) is in contact with eight residues of gp120 (Asp368, Ile371, Glu370, Asn425, Met426, Trp427, Gly473, and Met475) via 46 vdW interactions of the 133 total aligned NCIVs in both interfaces. All these NCIVs were aligned by Galinter successfully (Figure 3.20b). For this mimicry case, the I2I-SiteEngine alignment agrees with the Galinter result, with an irRMSD of only 0.4 Å. The interface alignment based upon the backbone alignment of the two gp120 molecules using DaliLite also agrees with Galinter and I2I-SiteEngine alignments. The irRMSD values are 0.9 Å (Galinter vs. DaliLite) and 0.6 Å (I2I-SiteEngine vs. DaliLite).

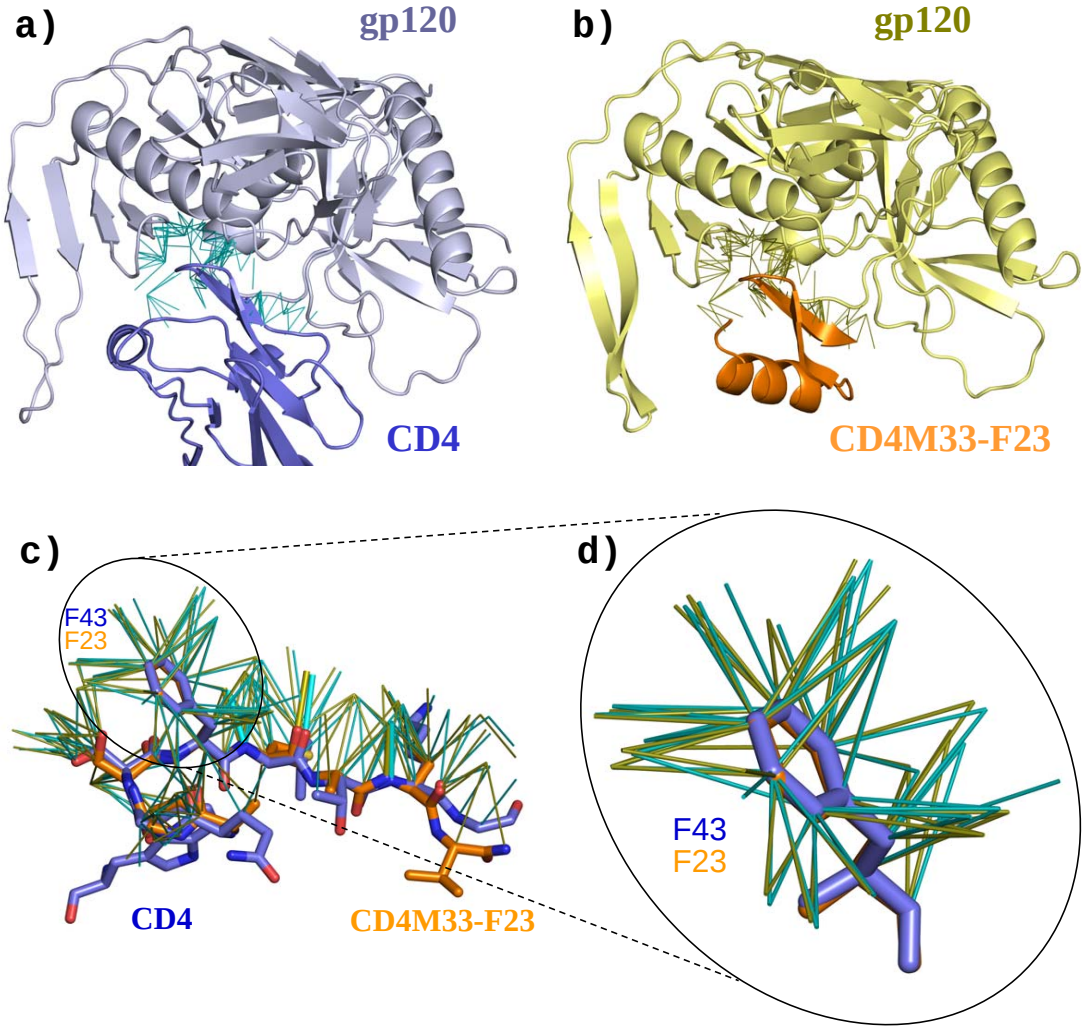


Figure 3.20: A scorpion-toxin derived mimic of CD4 in complex with gp120. **a)** Interaction between gp120 and CD4. **b)** Interaction between gp120 and CD4M33-F23, with gp120 at the same orientation as in **a)**. **c)** Superposed NCIVs for CD4/gp120 interface (1rzjCG) and CD4M33-F23/gp120 interface (1yymMG) according to the Galinter alignment. CD4 is shown in dark blue and CD4M33-F23 is in orange. Only aligned NCIVs are shown. CD4/gp120 NCIVs are shown in cyan, and CD4M33-F23/gp120 NCIVs are in yellow. Hydrogen bonds are shown as thick lines. **d)** An enlarged view of the aligned NCIVs involving the hot spot phenylalanines.

3.3 Scoring of Alignments

In this section, we introduce a statistical scoring method for assessing the significance of interface alignment results.

Galinter produces alignments of the vector representations of non-covalent interactions across protein-protein interfaces. The alignment results are ranked by the number of aligned NCIVs. This number reflects not only the similarity between the interfaces under comparison, but also the total number of NCIVs at the two interfaces. Apparently, a comparison involving an interface composed of a small number of NCIVs never produces long alignments.

When two objects under comparison are of different sizes, the *Tanimoto coefficient* is commonly used as a measure of similarity (Tanimoto, 1958; Willett and Winterman, 1996; Tan *et al.*, 2005). For two objects to be compared A and B, the Tanimoto coefficient $T(A, B)$ is defined as:

$$T(A, B) = \frac{N_{aln}}{N_A + N_B - N_{aln}}$$

where N_{aln} is the size of the matching between A and B, N_A and N_B are the sizes of A and B, respectively.

The Tanimoto coefficient is very simple as it uses only the sizes of the two objects and the size of the matching between them. Therefore, the computation of this measure is very fast, which is a desirable feature for large-scale applications. However, the Tanimoto coefficient has a well-known defect, that is, there is no critical value for it, which provides a measure of statistical significance for the similarity. Consequently, it is hard to identify significant matching based on the value of Tanimoto coefficient.

Recently, Davies *et al.* (2007) presented a probabilistic model for assessing the similarity between protein-ligand binding sites. Using this model, the significance of the binding site similarity can be measured. We developed a scheme for scoring the alignment of protein-protein interfaces based on the same model. In the following sections, we first describe the underlying probabilistic model of the scoring scheme. Then, we estimate the essential parameters in the scoring scheme based on a large scale application of the Galinter program on a dataset of dissimilar interfaces. Finally, the test results of the scoring method are reported.

3.3.1 The Poisson Index

Davies and coworkers proposed a probabilistic model for measuring the significance of the similarity between protein-ligand binding sites (Davies *et al.*, 2007). In their proposed model, the similarity was measured by the *Poisson Index* (PI) based on a statistical model for the matching of binding site atoms. The same information is used for the computation of the PI as for the Tanimoto coefficient. This makes the PI applicable for large scale comparison of binding sites. As a matter of fact, the PI is defined as the probability of finding a match as good as or better than

the observed result. Therefore, it is essentially the p-value for comparison results, suggesting whether the matches between two objects are significant.

Given that the PI measure is simple to compute and capable of measuring statistical significance, we decided to adopt this statistical model and applied it to estimate the significance of the comparison results of NCIVs at protein-protein interfaces. Here we introduce the derivation of the PI.

Poisson Process and Poisson Distribution

Many physical processes in real life can be described as *Poisson processes*. For example, the telephone calls arriving at a switchboard during a fixed period of time, the defects on a specified length of magnetic tape. The occurrences of such events satisfy three conditions:

1. the numbers of occurrences in any two disjoint time intervals or regions are independent;
2. the probability of an occurrence of the event during any time interval or within any region is approximately proportional to the length of the interval or the size of the region;
3. the probability of two or more occurrences during a very short time interval or within a very small regions is of a smaller order of magnitude than the probability of one occurrence, or $p(X \geq 2) = o(p(X = 1))$.

If these three conditions are satisfied, the number of occurrences in the processes in any fixed time interval t has a *Poisson distribution* with mean λt ($\lambda > 0$) (DeGroot and Schervish, 2001). This is why the processes are named Poisson processes. Let random variable X denote the number of occurrences of such an event, the Poisson distribution describing the probability of X taking value x:

$$p(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, ..., \tag{3.3}$$

where $\lambda\,(\lambda > 0)$ is the mean number of occurrences of events during per unit time or within per unit region.

Derivation of the Poisson Index

The derivation of the Poisson Index presented here is in accordance with that introduced in Davies *et al.* (2007). Let A and B be two interfaces to be compared, which are represented as two sets of NCIVs at the two protein-protein interfaces. We use m and n to denote the numbers of NCIVs at the two interfaces A and B. Without loss of generality, we assume $m \leq n$. With a certain transformation in 3D space, the two interfaces are superposed such that the maximum number of NCIVs match. Suppose the two matched subsets of NCIVs are $a \subseteq A$ and $b \subseteq B$, and the number of matched NCIVs is L.

We consider a superpopulation of a set of N vectors in a volume v generated by a homogeneous Poisson process with rate λ. Each vector in this volume may belong to A alone, B alone, both A and B, or none, with respective probabilities p_a, p_b, $\rho p_a p_b$, and $1 - p_a - p_b - \rho p_a p_b$, where ρ is the *a priori* tendency for two vectors to match in the volume v. Under this model, the numbers of vectors belonging to the categories A, B or both are three independent Poisson random variables when N is very large ($N \to \infty$). The means for the three variables are $\lambda v p_a$, $\lambda v p_b$, and $\lambda v \rho p_a p_b$. The total number of vectors for the three variables are $m - L$, $n - L$, L.

The distribution of a Poisson variable x with mean $\lambda\,(\lambda > 0)$ is given in Equation 3.3. For two interfaces consisting of m and n NCIVs ($m \leq n$), the probability of observing L matching NCIVs between two interfaces consisting of m and n NCIVs is

$$\begin{aligned} p(L|m,n) &\propto \frac{(\lambda v p_a)^{m-L}\, e^{-\lambda v p_a}}{(m-L)!} \times \frac{(\lambda v p_b)^{n-L}\, e^{-\lambda v p_b}}{(n-L)!} \times \frac{(\lambda v \rho p_a p_b)^{L}\, e^{-\lambda v \rho p_a p_b}}{L!} \\ &\propto \frac{(\lambda v)^{m+n}\, {p_a}^m\, {p_b}^n\, (\rho/\lambda v)^L\, e^{-\lambda v (p_a + p_b + \rho p_a p_b)}}{(m-L)!\,(n-L)!\,L!} \\ &= \frac{K d^L}{(m-L)!\,(n-L)!\,L!} \end{aligned} \tag{3.4}$$

where $d = \rho/(\lambda v)$ is the propensity of two vectors to match, and K is a normalization constant. K shall be calculated such that

$$\sum_{L=0}^{m} p(L|m,n) = 1,\ (m \leq n) \tag{3.5}$$

that is,

$$K = 1 \Big/ \sum_{L=0}^{m} \frac{d^L}{(m-L)!\,(n-L)!\,L!} \tag{3.6}$$

The probability distribution of L values as described in Equation 3.4 is derived in Green and Mardia (2006).

If we can obtain a background distribution of L values for random matches, the p-value indicating the significance of a non-random match L_{obs} can then be computed as the tail probability in the background distribution of finding a match $L'\,(L' \geq L_{obs})$. This p-value is defined in Davies *et al.* (2007) as the *Poisson Index* (PI)

$$PI = \sum_{L'=L_{obs}}^{m} p(L'|m,n,d). \tag{3.7}$$

3.3.2 Parameter Estimation for Poisson Index

In order to apply the PI to assess the significance of Galinter alignment results using Equation 3.4 and 3.7, we needed to estimate the values for K and d. Hence, we

first collected a dataset of pairs of dissimilar interfaces and ran Galinter to align them. The alignments were considered to be random matches as the interfaces were chosen specifically to make certain that the subunits at both sides of interfaces are non-homologous. Then the parameters in the Equation 3.4 and 3.7 were estimated based on the alignment results.

Dissimilar Interface Pairs

We derived a dataset of pairs of dissimilar domain-domain interfaces using the following protocol, which is similar to the one introduced in Zhu *et al.* (2006), in principle:

1. *Collect domain-domain interfaces.* We collected domain-domain interfaces from SCOPPI, a structural classification of protein domain-domain interfaces (Winter *et al.*, 2006). Only domains belonging to the first seven classes in the SCOP classification hierarchy were considered, including all alpha proteins, all beta proteins, alpha and beta proteins (a/b), alpha and beta proteins (a+b), multi-domain proteins (alpha and beta), membrane and cell surface proteins and peptides, and small proteins. The remaining four classes were excluded because they are not considered to be "true classes" in SCOP (Murzin *et al.*, 1995), including coiled-coil proteins, low resolution protein structures, peptides, and designed proteins.
2. *Remove redundancy in the domain-domain interfaces on the sequence level.* Only domains with sequence identity less or equal to 90% were retained. Using this cutoff, very similar protein domains were removed while sufficient amount of interfaces were still retained.
3. *Remove redundancy in the domain-domain interfaces on the structure level.* All domain-domain interfaces were first divided into groups. Each group was defined by a pair of SCOP folds to which the two interacting domains belong. Then, from each of these fold-fold groups we selected one domain-domain interface whose corresponding PDB structure model has the highest AEROSPACI score (Chandonia *et al.*, 2004). The AEROSPACI score is a measure of the quality of the structural models deposited in the PDB. The higher the score, the better the quality.
4. *Select dissimilar interface pairs.* For each pair of these selected domain-domain interfaces, they were considered dissimilar if the interacting domains on both sides of the interfaces belong to different SCOP classes. For example, two interfaces $I_1(A_1, B_1)$ and $I_2(A_2, B_2)$ are considered to be dissimilar if $class(A_1) \neq class(A_2)$ and $class(B_1) \neq class(B_2)$ and $class(A_1) \neq class(B_2)$ and $class(B_1) \neq class(A_2)$.

We chose to use SCOP fold level in step 3 mainly for the reason that domains belonging to different SCOP folds are expected to have different 3D structures and are not related by homology (Murzin *et al.*, 1995). In step 4, the dissimilarity between interfaces was guaranteed by restricting the subunits on both sides of the

interfaces to be from different SCOP classes and thus possess totally different backbone structures. We relied on these two steps to remove pairs of interfaces that were potentially evolutionary related. Certainly, the similarity between the local structures at protein-protein interfaces is not completely determined by the overall structures of the component proteins. This has been illustrated by examples analyzed in Section 3.2.3. However, after applying these selection criteria, relatively few pairs with interface similarity are expected in the dataset.

We started with SCOPPI 1.69 (based on SCOP 1.69), which consists of 102,083 domain-domain interfaces. Only 13,743 interfaces were retained after removing redundancy using 90% sequence identity. We obtained 1,731 *fold-fold* groups in step 3. In the end, 883,445 pairs of dissimilar interfaces were generated after step 4. We named this dataset DDI_90_Fold. We applied Galinter to align each pair of interfaces in the dataset for generating the background distribution of L values.

In the following computation, we considered only interface sizes larger or equal to 50, i.e., $50 \leq m \leq n$. According to a statistics based on the data collected from the non-redundant protein-protein interaction dataset Keskin *et al.* (2004), there are about 6.5 NCIVs related to each interface residue, or

$$\frac{\text{total number of NCIVs}}{\text{total number of interface residues}} \approx 6.5 \ .$$

Therefore, an interface consisting of 50 NCIVs has less than eight interface residues in total. This implies that there are less than four residues on each binding site on average. Such interfaces with too few interface residues are likely to have very small interface areas and thus are unlikely to be biologically relevant (Zhu *et al.*, 2006). Therefore, we did not consider interfaces containing less than 50 NCIVs. In addition, a statistics on the sizes of the domain-domain interfaces in the dataset DDI_90_Fold shows there are 264 non-covalent interactions at each interface on average. There are only approximately 14.0% of the interfaces containing less than 50 non-covalent interactions (Figure 3.21).

In the end, we used 400,462 alignment results for the estimation of parameters in the PI, resulting in 117,077 different (m, n) pairs.

Parameter Estimation

For each pair of interfaces compared by Galinter, we have the sizes of the two interfaces m and $n\,(m \leq n)$, and the size of the alignment L. As proposed in Davies *et al.* (2007), the values of d can then be estimated using a maximum likelihood estimation (MLE) procedure. The advantage of the MLE method is that no prior distributions or loss functions are required for constructing estimators of parameters. Furthermore, the MLE method generally yields precise estimators when the sample

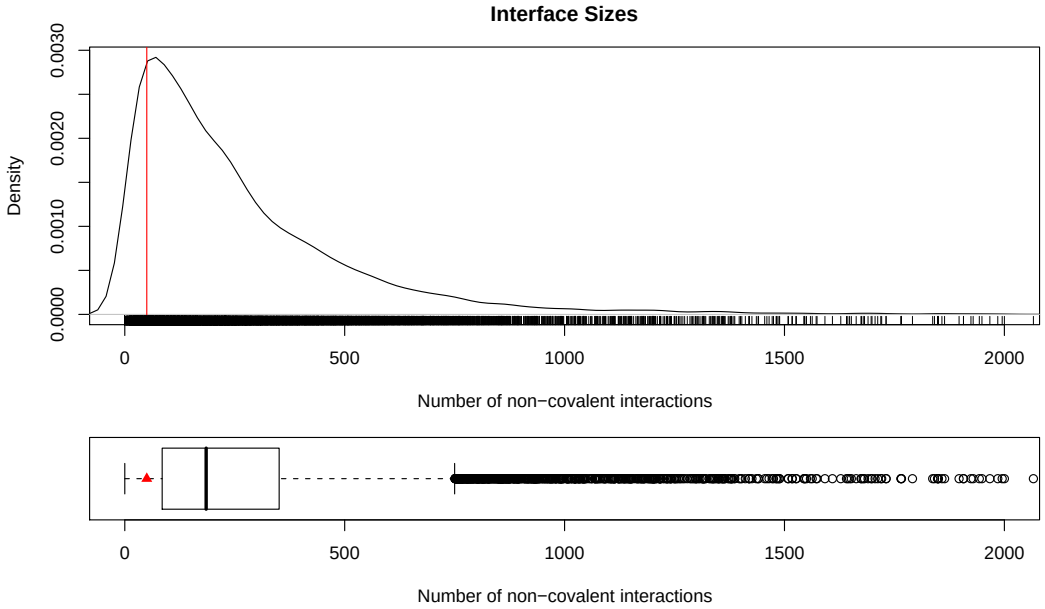


Figure 3.21: The distribution of domain-domain interface size. The density curve has been estimated using Gaussian kernel with R function `density()`. In the boxplot, the lower and upper "hinges" of the box are (basically) the first and third quartiles. The line in the middle of the box is the median. The "whiskers" represent the minimum and maximum values, or 1.5 times the box length if the minimum or maximum value exceeds 1.5 times the box length. Sizes larger than $2,000$ are not shown. The size 50 is marked in red in both plots. There are $1,925$ out of $13,743$ domain-domain interfaces containing less than 50 non-covalent interactions.

is large (DeGroot and Schervish, 2001). The maximum likelihood estimator for d is

$$\begin{aligned}
&\arg\max_{d\in[0,1]} \left[\prod_{i=1}^{q} p(L_i|m,n)\right] \\
\Longleftrightarrow &\arg\max_{d\in[0,1]} \left[\sum_{i=1}^{q} \log p(L_i|m,n)\right] \\
\Longleftrightarrow &\arg\max_{d\in[0,1]} \left[\sum_{i=1}^{q} \log K + L_i \log d - \log(m-L_i)! - \log(n-L_i)! - \log L_i!\right] \qquad (3.8)
\end{aligned}$$

where q is the number of observed interface pairs for the given m and n values. As $d = \rho/(\lambda v)$ is the propensity of two vectors to match, we shall search the value of d in the range of 0 to 1. After obtaining d, the value of K can then be easily calculated using Equation 3.6.

Solving d from its maximum likelihood estimator is a non-linear optimization problem. We employed the L-BFGS-B (Nocedal, 1980) algorithm implemented in the open-source software SciPy (Jones *et al.*, 2001) to find the solution of d for all (m,n) pairs.

We first investigated the d values for a few representative examples of (m,n) values

and examined the consequent theoretical distribution of L values (Equation 3.4) by comparing it to the distribution of the observed L values. This step can be regarded as a validation of the probabilistic model for protein-protein interface similarity. The choice of the (m,n) examples was based on two criteria: a) the occurrence of the (m,n) pair is high in the DDI_90_Fold dataset; b) the m and n values cover different interface sizes in the DDI_90_Fold dataset. The first criterion was enforced such that enough data were available for the estimation of the corresponding d values. After inspecting the number of occurrences of m and n values in the DDI_90_Fold dataset, we chose six different values: 58, 112, 151, 206, 241, 292. These are the interface sizes that appear most frequently and cover different ranges of the interface size (~ 50, ~ 100, ~ 150, ~ 200, ~ 250, ~ 300). Observed L values are rare for any specific m and n pair where $300 < m \leq n$, thus these cases were not analyzed here.

The d values were estimated for all the $\binom{6}{2} + 6 = 21$ combinations of (m, n) where $m, n \in (58, 112, 151, 206, 241, 292)$. We could compute $p(L|m, n)$ and PI values based on the estimated d value using Equation 3.4 and 3.7. We then compared the $p(L|m, n)$ values to the distribution of observed L values, and the PI values to the cumulative distribution of observed L values for each (m, n) combination. The results are shown in Figure 3.22.

To assess the agreement between the empirical distribution and the theoretical cumulative distribution of various L values in Figure 3.22, we computed the statistic used in the *Kolmogorov-Smirnov* (K-S) test (DeGroot and Schervish, 2001) for each pair of m, n values. For a combination of m and n ($m \leq n$) values, let L_0, L_1, $\cdots$, L_k denote observed L values. The empirical distribution function $G(L)$ for the occurrences of various L values is a step function defined as

$$G(L) = \sum_{i=1}^{m} I_{L_i \geq L},$$

where $I_{L_i \geq L}$ represents the indicator function, which equals 1 if $L_i \geq L$ and equals to 0 otherwise. The theoretical cumulative distribution of $p(L|m, n)$ is defined as in Equation 3.7. Here we scaled it by a factor of k, the number of observed L values and denoted the scaled theoretical cumulative distribution function by $F(L)$. We then calculated the Kolmogorov-Smirnov statistic

$$D = sup|F(L) - G(L)|$$

where *sup* denotes the supremum of a set. See Figure 3.23 for an illustration of the K-S statistic.

To evaluate the significance of the obtained Kolmogorov-Smirnov statistic values (D_{obs}), we compared these values to simulated data. For each (m, n) pair, we did a simulation of 10^6 trials. In each trial, sample data were simulated based on the probability density function defined in Equation 3.4, and the sample size was equal to that of the observed data. The Kolmogorov-Smirnov statistic for the simulated data was computed and denoted as D_{sim}. For each (m, n) pair, we report the ratio

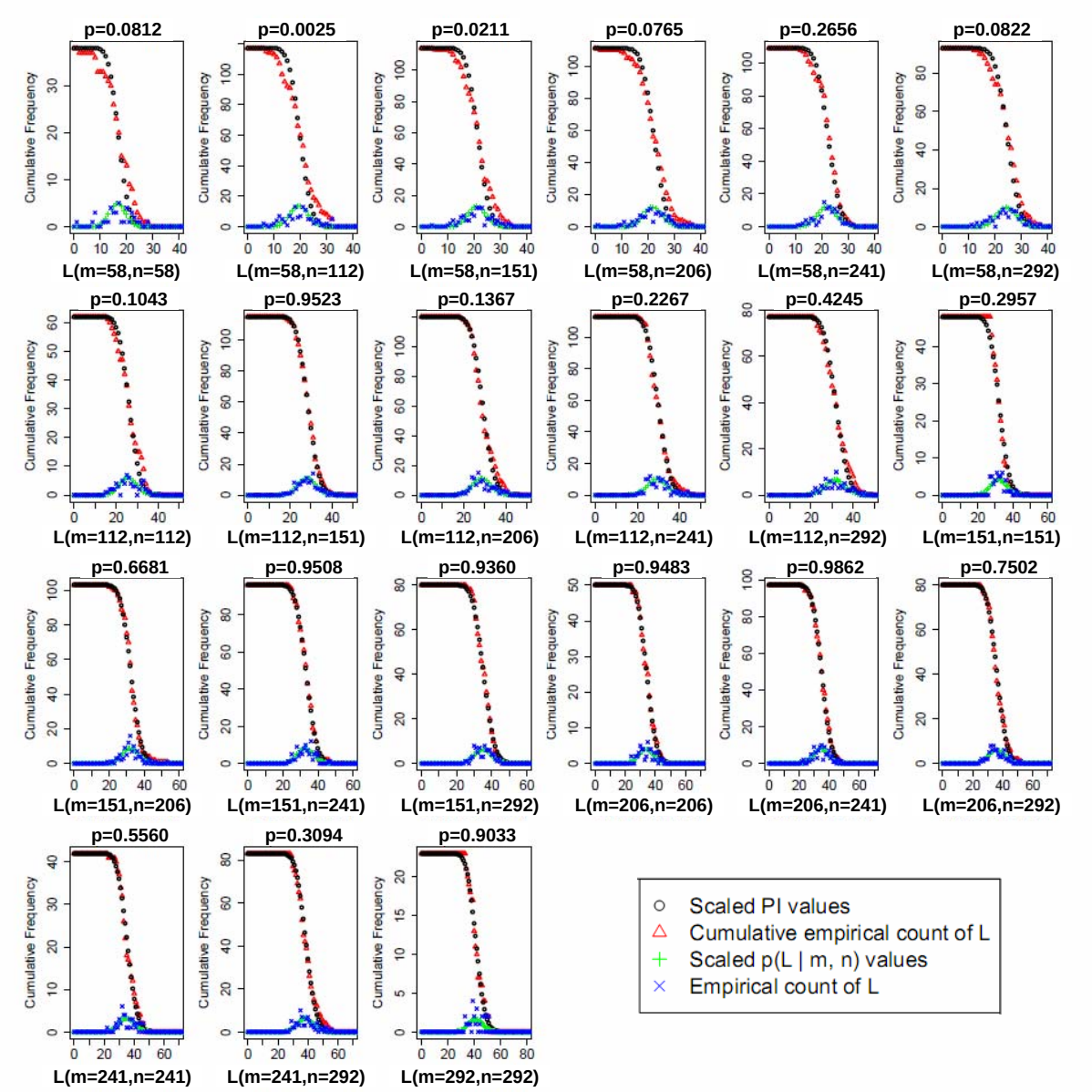


Figure 3.22: Comparison of scaled PI values to the cumulative empirical counts of L, and the $p(L|m,n)$ values to the empirical count of L for each combination (m,n) where $m,n \in (58,112,151,206,241,292)$. The K-S test has been employed to assess the similarity between the distributions of the scaled PI values and of the cumulative empirical counts of L. The p-values for the K-S test results computed from simulations of 10^6 trials are shown on the top of each plot.

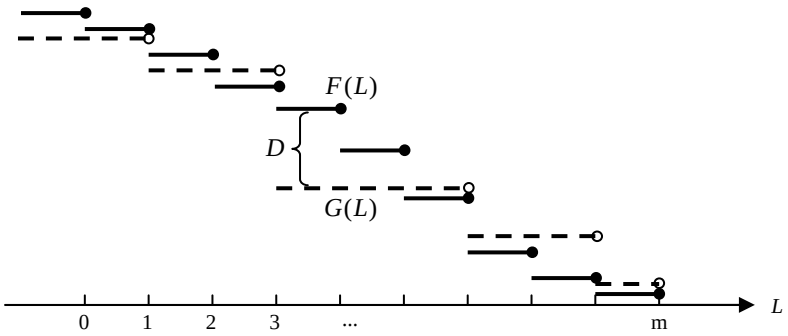


Figure 3.23: Kolmogorov-Smirnov statistic for assessing the agreement between empirical distribution and theoretical cumulative distribution of L. $G(x)$ represents the empirical distribution function for observed L values (dashed line). $F(x)$ is the theoretical cumulative distribution function of L (solid line). The K-S statistic D is computed as the maximum difference between the two distributions.

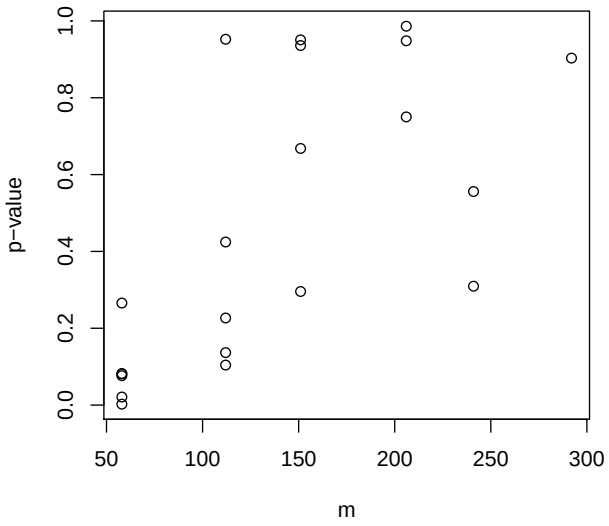


Figure 3.24: Relationship between interface sizes (m) and p-values of the K-S test.

of D_{sim} values that are larger or equal to the D_{obs} value as the p-value for the D_{obs} (see Figure 3.22).

Except for two (m, n) pairs, all the p-values of the K-S test statistic are larger than 0.05, indicating that the distributions of observed L values show no sign of significant discrepancy to the theoretical probability density distribution. The result suggests that the empirical distribution of L values agrees with the theoretical distribution described in Equation 3.4. Therefore, the probabilistic model proposed in the work of Davies *et al.* (2007) seems to be appropriate for the protein-protein interface similarity.

Meanwhile, we noticed that when interface sizes are small, the p-values of the

K-S test statistic are generally smaller (see Figure 3.24). Specifically, the p-values are below the 0.05 significance level for $m = 58$, $n = 112$ and for $m = 58$, $n = 151$ and they are 0.0025 and 0.0211, respectively. This suggests that the Poisson Index is not applicable for the alignment of small protein-protein interfaces (size=$\sim$50 in terms of the number of NCIV). Actually, interfaces containing only 50 or less NCIVs are most probably protein-ligand interfaces, or non-biologically relevant given the small interface area they have. Obviously, an independent estimation of parameters is necessary for adapting the Poisson Index specially for the alignment of protein-peptide or protein-ligand interfaces.

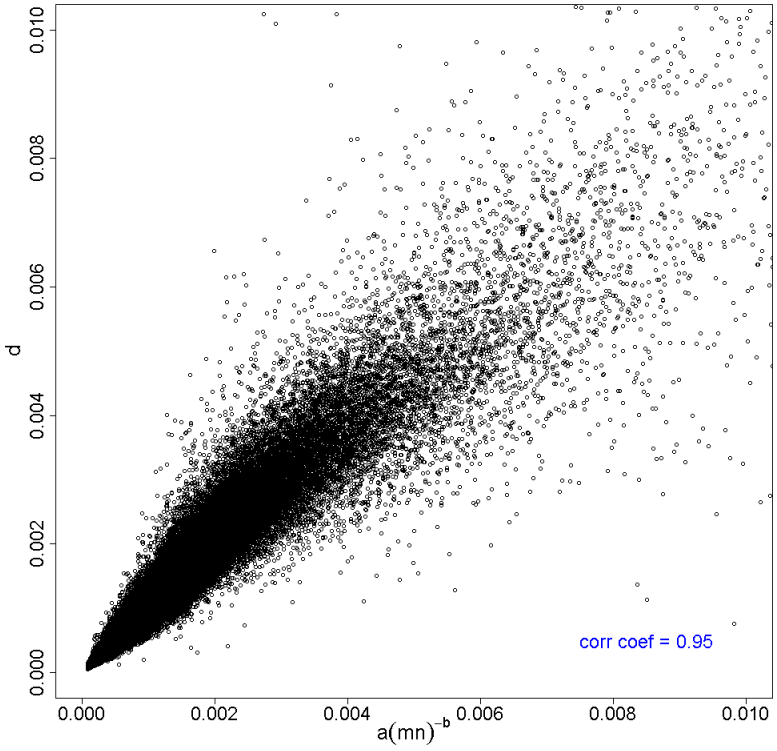


Figure 3.25: Scatter plot of d and fitted $\hat{d} = a(mn)^{-b}$.

General Model for Parameters

There was a d value estimated for each pair of interface sizes (m, n). The next step is to estimate the relationship between d, the propensity of two NCIVs to match, and m and n, the sizes of two interfaces. Davies *et al.* (2007) suggest a model of the form $d = a(mn)^{-b}$ based on the definition of d ($d = \rho/\lambda v$), where v is the volume occupied by the superpopulation, thus proportional to m and n. Using this model, they obtained a very good fit with $R^2 = 0.99$ for their data. Here we have tested a few different models, including $d = a(mn)^{-b}$, $d = a(m)^{-b}$, $d = a(n)^{-b}$, $d = a(m+n)^{-b}$. We transformed these non-linear models to linear models by taking

logarithms on both sides of the equations and used a linear least squares method to fit the models (Hastie *et al.*, 2003). The model proposed by Davies *et al.* (2007) yielded the best R^2 values for our data. Thus we decided to use the same model ($d = a(mn)^{-b}$) in this study. Note that d values were estimated in the form of $\log d$ in the maximum likelihood estimator given in Equation 3.8. Thus, the experimental error or the estimation error for $\log d$ is normally distributed, rather than for d. Therefore, we first transferred the model $d = a(mn)^{-b}$ to $\log d = \log a - b\log(mn)$ and estimated the values of a and b using a least squares method. Then we used these values as the starting points for finding the optimal values of a and b using a maximum likelihood estimation procedure. This is because the least squares corresponds to the maximum likelihood criterion when the experimental errors have a normal distribution.

We employed the `R` function `lm()` to perform the linear regression (R Development Core Team, 2005). It generated results of $a = 18.130$ and $b = -0.898$. The Pearson's correlation coefficient is 0.95 (Spearman's rank correlation coefficient = 0.98) for the original d values and the fitted $\hat{d}$ values using the formula $\hat{d} = 18.130(mn)^{-0.898}$ (see Figure 3.25). It can be seen that the correlation between the d values and the $\hat{d}$ values is relatively low in the top-right quarter of Figure 3.25. This quarter is mainly occupied by small interfaces of small m or n values. If we restrict $m \times n > 10000$, the Pearson's correlation coefficient increases slightly to 0.96 (Spearman's rank correlation coefficient remains at 0.98).

The optimal values were then estimated using an MLE procedure (DeGroot and Schervish, 2001). The maximum likelihood estimator for a and b is

$$
\begin{aligned}
& \underset{a\in[1,\infty),b\in[0,1]}{\arg\max} \left[\prod_{j=1}^{T}\prod_{i=1}^{q_j} p(L_i|m_j,n_j)\right] \\
\Longleftrightarrow\ & \underset{a\in[1,\infty),b\in[0,1]}{\arg\max} \left[\prod_{j=1}^{T}\prod_{i=1}^{q_j} \frac{K_j\, {d_j}^{L_i}}{(m_j-L_i)!\,(n_j-L_i)\,!L_i!}\right] \\
\Longleftrightarrow\ & \underset{a\in[1,\infty),b\in[0,1]}{\arg\max} \left[\prod_{j=1}^{T}\prod_{i=1}^{q_j} K_j\, {d_j}^{L_i}\right] \\
\Longleftrightarrow\ & \underset{a\in[1,\infty),b\in[0,1]}{\arg\max} \left[\prod_{j=1}^{T} {K_j}^{q_j}\, {d_j}^{\sum_{i=1}^{q_j} L_i}\right] \\
\Longleftrightarrow\ & \underset{a\in[1,\infty),b\in[0,1]}{\arg\min} -\log\left[\prod_{j=1}^{T} {K_j}^{q_j}\, {d_j}^{\sum_{i=1}^{q_j} L_i}\right] \\
\Longleftrightarrow\ & \underset{a\in[1,\infty),b\in[0,1]}{\arg\min} -\sum_{j=1}^{T}\left[q_j\log K_j + (\sum_{i=1}^{q_j} L_i)\,\log d_j\right] \\
\Longleftrightarrow\ & \underset{a\in[1,\infty),b\in[0,1]}{\arg\min} \sum_{j=1}^{T}\left[-q_j\log K_j - (\sum_{i=1}^{q_j} L_i)\,(\log a - b\log m_j n_j)\right] \qquad (3.9)
\end{aligned}
$$

where T is the total number of (m_j,n_j) pairs of any interface sizes, q_j is the number

of observed interface pairs for interface size pair (m_j,n_j). Note that the original estimator given in Davies *et al.* (2007) is incorrect. Again, the L-BFGS-B (Nocedal, 1980) algorithm provided in SciPy (Jones *et al.*, 2001) was used to solve the non-linear optimization problem. An MLE process yielded $a = 17.207$, $b = 0.892$ as the optimal values. The final model for the relationship between d and (m, n) integrated in the Galinter program for estimating significance values of alignments is $d = 17.207(mn)^{-0.892}$.

3.3.3 Database Scans using Poisson Index

The Poisson Index was designed to measure the significance of protein-protein interface alignment results. To test its usability, we applied the Poisson Index measure with the estimated parameters in a database scan using Galinter. The purpose of this test was to examine whether similar patterns of non-covalent interactions can be captured by using the Poisson Index. In this test of database scan, we reused the cases study examples reported in Section 3.2.3. For each of the four pairs of interfaces reported in Section 3.2.3, we used one of the two interfaces as a query. We then prepared a database of interfaces that are considered to be dissimilar to the query (see below for details). For each of the query interfaces, one database is made, which contains only interfaces that are dissimilar to the query. Then, the Galinter program was applied to scan the database by aligning all the interfaces in the database to the query. The results from the database scan were compared to the PI value of the alignment between the query and its partner interface in the case study examples (Section 3.2.3). Our goal was to investigate whether the Poisson Index will produce distinct PI values for the alignment between the query and its partner interface from the alignments between the query and the dissimilar interfaces in the database.

Database Scans

From the detailed analysis of interfaces in Section 3.2.3, we used the first mimicry example involving the catalytic triad and the fourth example involving a scorpion-toxin mimicking CD4 in complex with gp120 (1acbEI vs. 1lw6EI, 1rzjCG vs. 1yymMG) as positive test cases. For the third case (SP4206 mimic of IL-2Rα in binding to IL-2 (1z92BA vs. 1py2_A)), the two interfaces differ in many aspects. Only one of hot spot residues at the two interfaces keeps its orientation. Thus, the mimicry is considered imperfect and we used it as a negative example. We did not consider the second case study example because it is related to the first example.

To construct a database containing negative examples against which the query is compared, we followed the rules used to build the DDI_90_Fold dataset of dissimilar interfaces in Section 3.3.2. Specifically, for a given query interface $I(s1, s2)$, where $s1$ and $s2$ are the interacting subunits in the protein complex, we examined all the interfaces in the SCOPPI database. An interface $I'(s1', s2')$ is considered to be dissimilar to $I(s1, s2)$ if their respective subunits belong to different SCOP classes. That is, class($s1$) $\neq$ class($s1'$) and class($s1$) $\neq$ class($s2'$) and class($s2$) $\neq$ class($s1'$)

Table 3.4: Query interfaces and their database sizes.

Query Proteins	PDB	Database Size	Partner Interface
Chymotrypsin/Inhibitor	1acbEI	1643	1lw6EI
Subtilisin/Inhibitor	1lw6EI	1661	1acbEI
CD4/gp120	1rzjCG	1625	1yymMG
IL-2/IL-2Rα	1z92BA	1666	1py2_A

and class($s2$) $\neq$ class($s2'$). There are six interfaces in the three case studies, two of which (1yymMG and 1py2_A) are not available in SCOP 1.69, from which SCOPPI 1.69 has been derived. Therefore, in the end we used four query interfaces. For each query interface, a database of dissimilar interfaces was built to be scanned against, in which the partner interface of the query was added. The PDB ID and the chain names of the four queries and the sizes of their corresponding databases are summarized in Table 3.4.

Scan Results

We performed four database scans in total. A Poisson Index value was obtained for each Galinter alignment of interfaces. The distribution of all the Poisson Index values obtained from the four database scans are shown in Figure 3.26 as density and boxplots.

In three of the four database scans, in which chymotrypsin/inhibitor, subtilisin/inhibitor and CD4/gp120 were used as queries, extreme PI values were observed for the alignments with the mimicry related complexes (described in Section 3.2.3). There are few or no additional observations with lower PI values (Figure 3.26a, b, and c). Such distinct PI values strongly indicate that the corresponding alignments are significant. These results agree with the analysis presented in Section 3.2.3. While for the IL-2/IL-2Rα case depicted in Figure 3.26d, the PI value for the alignment between IL-2/IL-2Rα and IL-2/SP4206 is quite large, at the upper quartile and therefore it is not an extremely low value. This result indicates that the interface similarity between the IL-2/IL-2Rα complex and the IL-2/SP4206 complex is not significant, which agrees with our previous conclusions that this is a case of imperfect mimicry.

After carefully examining the extreme PI values in the database scan results, we concluded that PI $= 2.5 \times 10^{-6}$ is a reasonable significance level. Using this significance level, we discovered a few interesting cases with significant PI values for queries chymotrypsin/inhibitor and subtilisin/inhibitor. These hits and their corresponding PI values are listed in the Table 3.5. As a comparison, the partner interfaces to the queries and the PI values for their alignments are also shown in the same table.

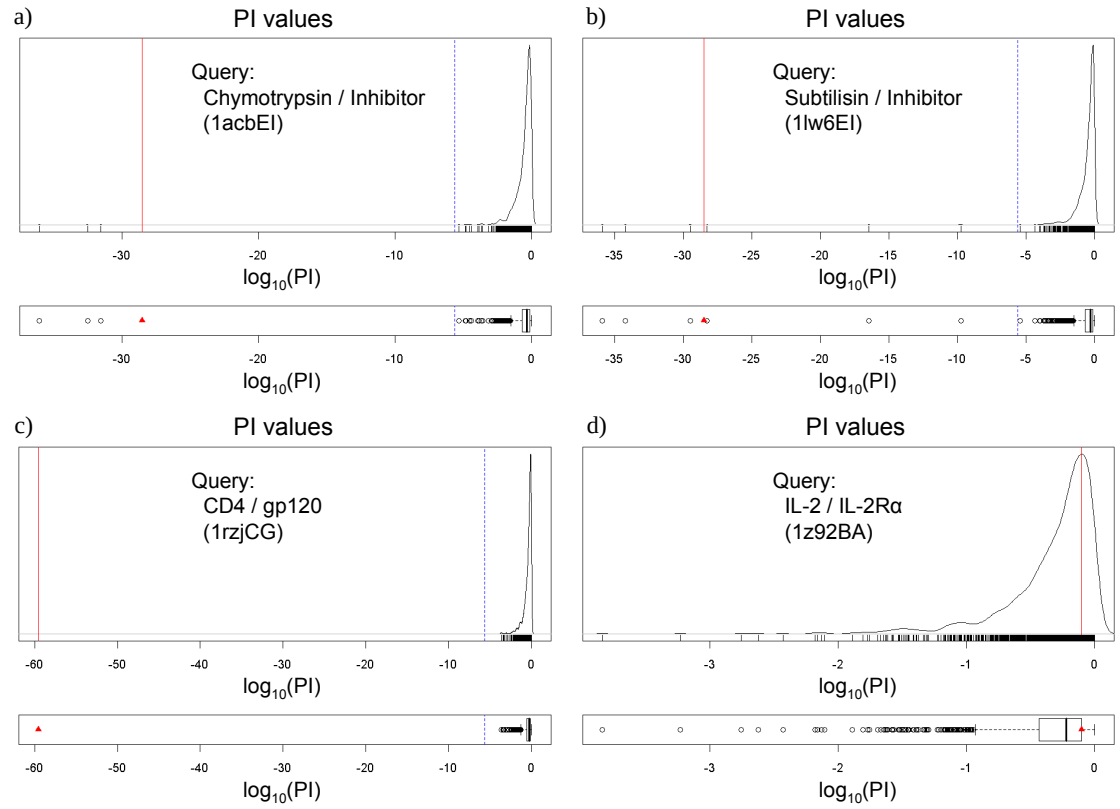


Figure 3.26: Distribution of the Poisson Index values for database scans. For each of the four database scans, both a density estimation and a box-and-whisker plot of the PI values are plotted. In each of the four subgraphs, the red line in the density plot and red triangle in the boxplot indicate the PI value for the alignment between the query interface and its partner interface added to the database. The dashed blue lines in the density plots mark the empirical significance level of the Poisson Index (2.5×10^{-6}).

We investigated all the alignments with PI values lower than the cutoff. In general, they are interfaces formed between proteases and their inhibitors. In other words, the interfaces with extreme low PI values are indeed from protein-protein interactions responsible for the same biological function as the query interfaces, though the backbone structures of the query complexes and the hit complexes are completely distinct. Therefore, the cases detected by the Galinter program from the databases with low PI values are similar to the examples introduces in Section 3.2.3.

For the query chymotrypsin/inhibitor, we detected three interfaces with outstanding PI values when aligned to the query (see Figure 3.26a). It can be seen from Table 3.5 that all the three hits are complexes of a subtilisin and an inhibitor. All the three subtilisins are homologous to the partner interface of the query subtilisin/inhibitor. The four complexes all contain a subtilisin domain from the same SCOP family c.41.1.1. The chymotrypsin domain in the query interface chymotrypsin/inhibitor is from SCOP family b.47.1.2. The inhibitors of these sub-

tilisins present totally different backbone structures. They all belong to different SCOP folds, or even different SCOP classes. By examining the Galinter alignment results, we found that the serine and histidine residues of the catalytic triads are all correctly aligned between interfaces. Here it is demonstrated again that the Galinter program is capable of capturing similar binding patterns at interfaces independent of the backbone structures of interacting subunits.

For the query subtilisin/inhibitor, seven interfaces were discovered where the alignments have significant PI values. They are all formed between serine protease and inhibitor except for an intra-chain interface in the PDB entry 1cu1. The serine proteases are all from the same SCOP superfamily b.41.1. The inhibitors are all from the same SCOP class g (small proteins) except for 1acbEI and 1ophBA. We will look at these two examples, namely 1lw6EI vs. 1ophBA and 1lw6EI vs. 1cu1 in detail in the following paragraphs. The inhibitor backbone structures are all from different SCOP folds or classes. We examined the alignment of interface residues by the Galinter program, the catalytic triads are all well matched.

Table 3.5: Database scan results. Hit interfaces are listed in ascending order of the PI value. Partner interfaces of queries are shown in bold.

Query	PDB	PI value	Interacting subunits	SCOP families
Chymotrypsin/ Inhibitor[a]	1r0r	8.93e-37	Subtilisin/Inhibitor	c.41.1.1 & g.68.1.1
	1oyv	2.98e-33	Subtilisin/Inhibitor	c.41.1.1 & g.69.1.1
	2sic	2.68e-32	Subtilisin/Inhibitor	c.41.1.1 & d.84.1.1
	1lw6EI	**3.09e-29**	Subtilisin/Inhibitor	c.41.1.1 & d.40.1.1
Subtilisin/ Inhibitor[b]	1sgr	1.26e-36	Proteinase/Inhibitor	b.47.1.1 & g.68.1.1
	4sgb	6.00e-35	Proteinase/Inhibitor	b.47.1.1 & g.69.1.1
	1mct	3.39e-30	Trypsin/Inhibitor	b.47.1.2 & g.3.2.1
	1acbEI	**3.09e-29**	Chymotrypsin/Inhibitor	b.47.1.2 & d.40.1.1
	1oph	5.49e-29	Trypsin/Inhibitor	b.47.1.2 & e.1.1.1
	1gl1	3.53e-17	Chymotrypsin/Inhibitor	b.47.1.2 & g.4.1.1
	1eai	1.79e-10	Chymotrypsin/Inhibitor	b.47.1.2 & g.22.1.1
	1cu1	1.85e-10	Protease/Helicase	b.47.1.3 & c.37.1.14

[a]Interacting chymotrypsin and inhibitor (1acbEI) belong to SCOP families b.47.1.2 and d.40.1.1, respectively.

[b]Interacting subtilisin and inhibitor (1lw6EI) belong to SCOP families c.41.1.1 and d.40.1.1, respectively.

The interface formed between chain B and A of the PDB entry 1oph is aligned to the query subtilisin/inhibitor with PI = $5.49e^{-29}$. This is an interface between a S195A trypsin and a α1-protease inhibitor (α1PI) Pittsburgh (Dementiev *et al.*, 2003). In the S195A trypsin the active site serine residue (Ser195) has been mutated to alanine (S195A). The α1PI Pittsburgh is a variant of the serpin α1PI with a P1

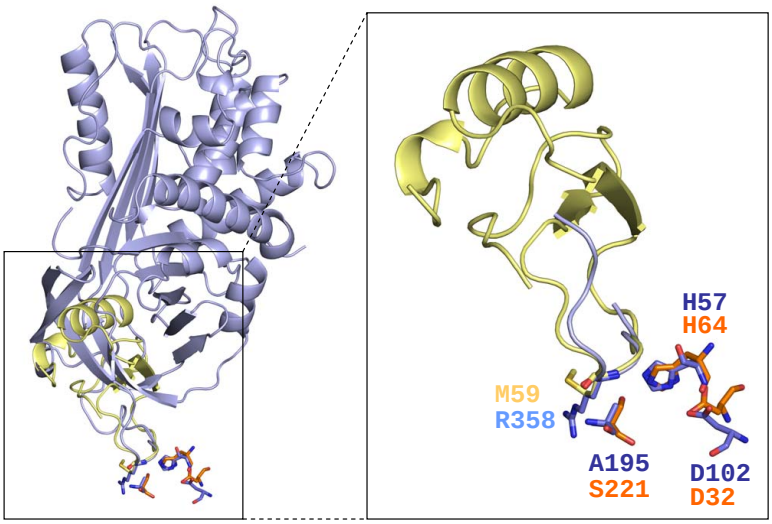


Figure 3.27: Alignment of subtilisin-inhibitor interface (PDB ID: 1lw6) and subtilisin-inhibitor interface (PDB ID: 1oph). Only the inhibitor of the subtilisin-inhibitor complex in 1lw6 (chain I, in light yellow) and the α1PI Pittsburgh in 1oph (chain A, in light blue) are shown as cartoons. The two sets of catalytic triad residues are shown as sticks in colors of orange and blue for 1lw6 chain E and 1oph chain B, respectively. The mutated arginine (R358) residue at the reactive center loop of the α1PI Pittsburgh and the wild-type methionine residue (M59) at the subtilisin inhibitor are also shown as sticks in light blue and light yellow.

mutation of methionine to arginine at the reactive center loop (RCL). In the X-ray structure, the RCL of the α1PI Pittsburgh exhibits a canonical conformation similar to that of non-complexed wild-type serpin. Given all these important mutations, the non-covalent interaction between the S195A trypsin and the α1PI Pittsburgh has been discovered to be very similar to that of classical serine protease and inhibitors (Bode and Huber, 1992). Based on the Galinter alignment, we obtained the matching between the classic catalytic triad Ser–His–Asp of subtilisin (1lw6E) and the mutated triad Ala–His–Asp in the S195A trypsin (1ophB) (see Figure 3.27). This example illustrates that even if the physicochemical properties of the side chains of interface residues are different, the Galinter program is still able to recognize conserved interaction patterns at interfaces.

The hit interface in 1cu1 identified by query subtilisin/inhibitor is an intra-chain interface between a viral protease domain (SCOP family: b.47.1.3, SCOP sid: d1cu1a1) and an RNA helicase domain (SCOP family: c.37.1.14, SCOP sid: d1cu1a3). The protein deposited in PDB 1cu1 is a nonstructural protein of hepatitis C virus (HCV) (Yao *et al.*, 1999). This protein is a bifunctional enzyme with both protease and helicase activities termed nonstructural protein 3 (NS3). The polypeptide chain of the molecule in PDB 1cu1 has been engineered such that it contains also the sequence of a protease activation domain of the nonstructural protein 4A

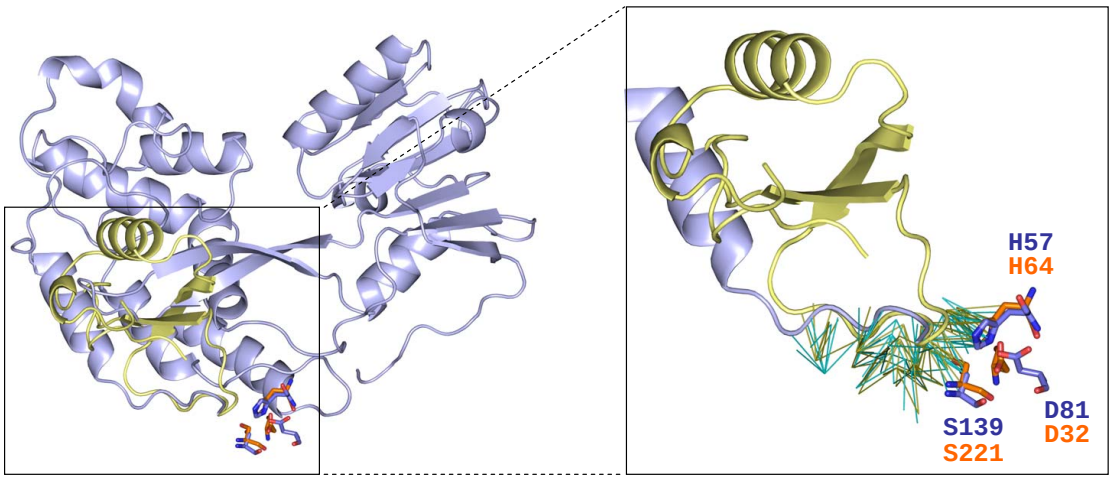


Figure 3.28: Alignment of subtilisin-inhibitor interface (PDB ID: 1lw6) and protease-helicase interface (PDB ID: 1cu1). Only the inhibitor of the subtilisin-inhibitor complex in 1lw6 (in light yellow) and the helicase domain of the chain A of 1cu1 (in light blue) are shown as cartoons. The two sets of catalytic triad residues are shown as sticks in colors of orange and blue for 1lw6 and 1cu1, respectively. In the enlarged view of the aligned interfaces, aligned NCIVs are depicted as lines in yellow for 1lw6 and in cyan for 1cu1.

(NS4A), linked to the N-terminus of the NS3 sequence. The whole molecule is thus named scNS3-NS4A. In scNS3-NS4A, the active site of the protease domain, where the catalytic triad locates is occupied by the C terminal of the scNS3-NS4A, which is part of the helicase domain. This leads to the *autoinhibition* of NS3 protease activity. When the interface between the protease domain and the helicase domain in 1cu1 is aligned to the interface between subtilisin and its inhibitor in 1lw6, the Galinter program successfully recognize the conserved interaction pattern of protease inhibition. The catalytic triads at the two interfaces are also superposed according to the Galinter alignment (see Figure 3.28). The catalytic triad from chain E of 1lw6 consists of Ser221, His64, Asp32 and from chain A of 1cu1 inlucdes Ser139, His57, and Asp81[4]. The RMSD for the functional template atoms after superposition is 0.7 Å.

Conclusion

Based on the probabilistic model proposed by Davies *et al.* (2007), we constructed a statistical scoring scheme for the protein-protein interface similarity. The parameters

[4]Note in the chain A of 1cu1, Asp81 is considered one of the catalytic triad residues, instead of Asp99 as stated in Yao *et al.* (1999), which is the primary reference to the structure model 1cu1 in the PDB. This has been confirmed in Trozzi *et al.* (2003) and by examining 1cu1 in the Catalytic Site Atlas (CSA) (Porter *et al.*, 2004). Asp99 is spatially too distant from the other two residues and is incapable for form a catalytic triad with them.

of the model were estimated based on a large-scale comparison of dissimilar interfaces. The statistical score Poisson Index is capable of assessing the significance of an interface alignment. The Poisson Index was tested in four database scans. Using an empirical significance level (2.5×10^{-6}), we discovered more cases of conserved interaction patterns between interfaces whose component subunits exhibit no structure homology at all. The results also demonstrate that the Galinter program has the ability to capture similar interaction patterns at interfaces independent of either the backbone structures of proteins or the physicochemical properties of interface residue side chains. In the mean time, the database scan results also suggest that non-homology of backbone structures cannot ensure that the binding patterns at the interfaces are different. In other words, protein binding patterns may still be conserved even the protein complexes do not exhibit structure homology. We collected our data in the four databases by forcing the subunits on both sides of the interfaces to be from different SCOP classes than the query interfaces. In spite of this strict criterion, we still discovered several interesting similar interfaces to the queries, two of which have been described in detail.

3.4 Galinter

The Galinter program is accessible at `http://galinter.bioinf.mpi-inf.mpg.de/`. As input for Galinter, users can either specify the PDB ID and chain names of two interfaces, or upload their own structure files in PDB format. Interface alignment results are visualized to end users using Jmol[5]. Protein complexes are superposed based on the alignment of non-covalent interactions at the interfaces. Non-covalent interactions, including van der Waals interactions and hydrogen bonds, are also depicted in the visualization. The details of the alignment results are provided, including

- PI score,
- lists of matched interface atoms and interface residues,
- transformed structure of protein complex in PDB format.

The source code of the Galinter program can also be downloaded as a standalone software. It was implemented using Python and C. The standalone Galinter program provides more useful output. In addition to the output listed for the web user interface, it provides

- Jmol scripts for visualizing superposed protein complexes and NCIVs,
- PyMol scripts for visualizing superposed protein complexes and NCIVs.

The Galinter program is distributed under the GNU General Public License (GPL)[6].

[5]Jmol: an open-source Java viewer for chemical structures in 3D. `http://www.jmol.org/`
[6]http://www.gnu.org/licenses/gpl.html

3.5 Discussion

Comparison to I2I-SiteEngine

In this chapter, we presented Galinter, a novel method for comparing interfaces based on the geometry and type of non-covalent interactions. The proposed method is complementary to existing approaches for the analysis of protein-protein interfaces. The method was applied to the pilot dataset (Shulman-Peleg *et al.*, 2004) in comparison to I2I-SiteEngine and DaliLite. It is reassuring that for S/D-homologous complexes we obtained consistent results with the three methods. In addition, Galinter was applied to comparing mimicry examples, and the results agree with previous human-curated analyses. The results also suggest that Galinter has the potential of assisting in the design of interaction inhibitors. Galinter not only produces an alignment of interface residues, but also an alignment of interface atoms based on the alignment of NCIVs at interfaces. While using I2I-SiteEngine, only matchings of interface residues are generated. Hence the Galinter alignments can be regarded to have higher resolutions than I2I-SiteEngine alignments, and thus provide more information to the user. This is a useful feature for obtaining precise insights on interaction details at protein-protein interfaces. When matched residues are of different amino acid types, Galinter is still able to reveal the correspondence between the atoms in the residues, which discloses the conserved physicochemical features at protein binding sites.

Application to Protein-Ligand Interactions

Galinter was applied not only to the comparison of protein-protein interfaces, but also to the comparison of a protein-protein interface to a protein-ligand interface, i.e., the example of SP4206 mimic of IL-2Rα in binding to IL-2. In principle, the Galinter method may be easily applied to comparing protein-ligand interfaces. In addition, the method is also applicable to interfaces involving non-peptidic molecules. For that usage, the application just needs a module for identifying hydrogen bonds at interfaces involving non-peptidic molecules. Alternatively, external programs can be employed for this task, such as the Probe program (Word *et al.*, 1999b). Furthermore, the parameters needed for the calculation of the Poisson Index should be reestimated for assessing the similarity between protein-ligand interfaces.

Alignment of Molecular Interactions

The Galinter program can align both interfaces formed between different polypeptide chains, or between different protein domains, including domains that are from the same polypeptide chain (see the text about the intrachain protease-helicase interfaces, and Figure 3.27). Essentially, the identification and the alignment of non-covalent interactions at interfaces are two independent steps and may be separated. In theory, the NCIVs aligned by Galinter may be any kind of molecular interactions, including non-covalent interactions formed inside a protein domain. Therefore, the Galinter methodology is theoretically also applicable to the comparison of

non-covalent interactions within proteins or domains. This might be useful for the comparison of structural features of functional relevance.

Mimic Spot

In the comparison of SP4206/IL-2 and IL-2Rα/IL-2, we have observed a non-uniform distribution of conserved NCIVs throughout the two interfaces. The NCIVs involving residue Arg36 on IL-2Rα and its counterpart guanido group on SP4206 are highly conserved. Similar results have be observed in the remaining three case studies in Section 3.2.3. In the two cases of the protease/inhibitor interfaces, a large fraction of aligned NCIVs involve the two catalytic residues serine and histidine. At CD4/gp120 and CD4M33-F23/gp120 interfaces, Phe43 in CD4 and Phe23 in CD4M33-F23, respectively, form 46 NCIVs with eight surrounding residues (see Figure 3.20b). All these NCIVs are aligned and account for 35% of the final alignment. We named these conserved interface regions *similar spots*, or more specifically, *mimic spots*. One possible extension to the functionality of Galinter is the automatic detection of conserved interface regions, as in the case of mimic spots. The relationship between conserved interface regions, mimic spots and hot spots is another interesting topic for further research. Recent results indicate that conserved regions and hot spots overlap to a considerable extent (Shulman-Peleg *et al.*, 2007).

Direction of Hydrogen Bonds

In the current implementation of the Galinter method, the hydrogen bond vectors (HVecs) point from hydrogen bond donors to acceptors. This is different from the assignment of CVec directions, which uniformly point from one binding site to the other. This choice was made because in this way the HVec direction encodes the information about the locations of hydrogen bond donors and acceptors. Therefore, the matching of HVecs between interfaces suggests not only the conservation of hydrogen bonds at the interfaces, but also the conserved pattern of distributions of hydrogen bond donors and acceptors on corresponding binding sites. Nevertheless, we have tested the Galinter program by using the same assignment of directions for HVecs as for CVecs on a few alignment examples. The revised method was applied to align the interfaces presented in Section 3.2.3. The differences in the Galinter alignments are marginal, and the alignments do not seem to improve. For the protease-inhibitor interfaces, the RMSD values for the functional atoms in the catalytic triads became slightly worse ($\sim$0.1 Å) after superimposing the complexes according to the alignments. For the SP4206 mimic of IL-2Rα in binding to IL-2, the result remains the same. For the scorpion-toxin mimic of CD4 in complex with gp120, one less hydrogen bond is aligned. It would be certainly preferable if users are able to determine whether the directions of HVecs should be dependent on donor-acceptor directions, or correspond to the CVec directions.

Contribution of Different Types of Non-Covalent Interactions to the Alignment

In the current implementation, Galinter aligns vdW interactions and hydrogen bonds at interfaces. However, there are other types of non-covalent atomic interactions, especially electrostatic interactions between positively and negatively charged atoms. Thus, we have explored the contribution of short-range electrostatic interactions to the alignment of protein-protein interfaces. Using a definition by Xu *et al.* (1997), we have identified less than three short-range electrostatic interactions on average for each of the 64 interfaces in the pilot dataset used in the manuscript. This is only 1% of the number of vdW interactions. In addition, we have re-ranked the alignment results by assigning a larger weight of 3 to short-range electrostatic interactions (versus a weight of 1 to vdW interactions and hydrogen bonds). Except for four cases (1okvBE vs. 1okuBF, 10gsAB vs. 1axdAB, 1axdAB vs. 10gsAB, 1g0uOP vs. 1iruFG), the top-ranking alignments for the pilot dataset remain the same. Even for these four cases, the new results exhibit considerable similarity to the original alignments (half or more of the aligned NCIVs are the same).

These results indicate that the current method seems to be robust with respect to different weighting of the various types of interactions. Nevertheless, a thorough investigation is required on how to weight different types of non-covalent interactions for interface alignments.

Geometric Hashing Approach

Essentially, the interface alignment problem discussed in this chapter is the largest common point set (LCP) problem (see 3.1.3). In addition to clique detection, another widely used method for solving this problem is geometric hashing. We have not compared the performance of these two techniques. As we mentioned in Section 3.1.3, in the geometric hashing technique, the preprocessing step can be performed independent of the recognition step. Furthermore, the recognition step can be performed in parallel. These are attractive features for large-scale comparisons of binding modes, which are possible applications in drug discovery and protein engineering. Therefore, it is worth to investigate thoroughly the geometric hashing method in the alignment of protein interfaces.

Multiple Alignment

The Galinter program is only capable of performing pairwise alignments of protein-protein interfaces. It would be very useful to develop a multiple alignment tool for identifying common interaction patterns among a group of protein-protein interfaces.

MultiBind is a computational method comparing multiple protein binding sites simultaneously (Maxim Shatsky and Wolfson, 2005; Shatsky *et al.*, 2006). Approximate solutions to the underlying LCP problem are found in two major stages. The first stage is the selection of pairwise transformations for the construction of multiple superposition of the binding sites. The second step is the detection of common physicochemical properties from the multiple superposition. Shulman-Peleg *et al.*

(2005) propose an approach named MAPPIS for the multiple structural alignment of protein-protein interfaces. The approach is an extension to that for the multiple structure alignment of protein binding sites described in Maxim Shatsky and Wolfson (2005). In MAPPIS, the objects under consideration are interfaces, each of which is a pair of interacting binding sites. Like in I2I-SiteEngine, protein-protein interfaces are represented by two sets of pseudopoints, each set from one of the two interacting binding sites. MAPPIS identifies common physicochemical properties and the their interactions shared by a set of protein-protein interfaces. Geometric hashing is the basic technique used by both MultiBind and MAPPIS.

Weskamp *et al.* (2007) introduced a multiple alignment method for protein substructures based on pairwise alignments obtained using clique detection technique. One of the protein substructures is chosen as the pivot and all the rest substructures are aligned to it. Then the multiple alignment is constructed by incrementally merging the precomputed pairwise alignments. This is a star-like alignment scheme commonly used in multiple sequence alignment methods (Setubal *et al.*, 1997). Different pivots are tested in order to detect the multiple alignment with the best overall score.

Since the representations of protein substructures or interfaces in all these methods are pseudopoints denoting the functional groups of the 20 amino acids (Schmitt *et al.*, 2002), the non-covalent interactions that stabilize protein-protein interactions are not compared exactly. Due to the complexity of the related computational problem, the reduction of interface atoms to functional groups is necessary for finding approximate solutions with reasonable time complexity. However, the comparison of pseudopoints representing functional groups at interfaces provides only an indirect estimation of possible molecular interactions. The Galinter approach compares molecular interactions directly and with a better resolution. A multiple alignment version of the Galinter method would allow for the investigation of common binding patterns within sets of protein complexes.

Top-Down vs. Bottom-Up & Galinter 2

The current methodology of Galinter can be viewed as a "top-down" approach. The NCIVs at interfaces are clustered into NCIV representatives, each of which stands for a set of NCIVs covering a patch of interface area. It is the NCIV representatives that are directly aligned by using a clique detection technique. The matching of NCIV representatives is then decomposed to obtain the matching between individual NCIVs. Altogether, the strategy can be described as: first decompose the whole interface into small patches (NCIV representatives), after acquiring the alignment of the patches, decompose patches into individuals (NCIVs) and obtain the alignment of individuals (NCIVs). The central idea is *decomposition*.

Alternatively, one could implement a "bottom-up" approach for the alignment of NCIVs at interfaces. In this strategy, the whole interface is also divided into small patches. But first, the NCIVs in each patch are aligned to the NCIVs in the patches of the other interface. That is, the alignment at the most elementary

layer is constructed first. Then, the alignment of the whole interface is able to be constructed based on the alignments of NCIVs between interface patches. Unlike the "top-down" strategy, no clustering of NCIVs into consensus representatives is required in this approach. Using this strategy, many alignments of small interface patches are synthesized for form the alignment of the whole interfaces. In fact, a very similar approach has been proposed as *k-clique hashing* by Weskamp *et al.* (2004) for searching similar substructure in a protein structure database.

We name the "bottom-up" approach *Galinter 2* consisting of the following steps:

Step 1: Divide the whole interface into small patches. Clustering of NCIVs is needed here for the division of the interface. Usually, interface NCIVs are not evenly distributed across the whole interface. By clustering them into patches, we expect the structure flexibility can be better handled. On the one hand, within each patch where NCIVs are populated, the interface is considered to be relatively rigid. Consequently, the tolerance to the difference between inter-NCIV distances will be set restrictively. On the other hand, each patch is regarded as an independent part and the relative movement between patches will be more tolerated. Interface patches are not necessarily exclusive of each other. The size of each patch in terms of the number of NCIVs contained in the patch should be balanced such that the alignment between patches may be finished in a reasonable run time.

Step 2: Align each pair of patches between two interfaces. Either clique detection or geometric hashing technique may be used to align NCIVs at two interface patches. A matching between two patches is considered to be valid only if the ratio of aligned NCIVs is above a predefined threshold. Alternatively, the Poisson Index may be applied with re-estimated parameters. For each pair of aligned patches P_1, P_2, a transformation $T_{P_1P_2}$ is obtained based on the superposition of the aligned NCIVs at the two patches. This transformation T is important in the next step.

Step 3: Combine the alignments of the patches to the alignment of the whole interfaces. Two interfaces under comparison are represented as complete graphs, with each patch considered a vertex of the graph. All the vertices are connected and edges are labeled by inter-patch distances. Then the product graph is constructed as follows. Every pair of matched patches (P_1, P_2) in Step 2 is defined as a vertex in the production graph. Two vertices (P_1, P_2) and (P_1', P_2') are connected if a) the distances between the respective patches are compatible, i.e. $\text{dist}(P_1, P_1') \approx \text{dist}(P_2, P_2')$; b) the transformation matrices are compatible, i.e. $T_{P_1P_2} \approx T_{P_1'P_2'}$. A clique search step is then carried out in the product graph. The alignment of NCIVs at patches in the maximum cliques are synthesized to form the final alignment of NCIVs between interfaces.

In Galinter, the main reason for the clustering step of NCIVs into consensus representatives is to reduce the sizes of interfaces in terms of NCIVs numbers. This heuristic step introduces the artificial vectors of consensus NCIV representatives. Without this step, the run time of the product graph construction and the clique

detection is too long for practical use. Galinter 2 does not include the clustering of NCIVs into artificial consensus representatives. It avoids the heuristic step by employing two alignment steps at different levels.

Chapter 4

Summary and Outlook

In this section, we first summarize our research about the characterization, classification, and alignment of protein-protein interfaces. Then, we give our outlook on possible future developments and applications of the NOXclass and Galinter tools.

4.1 Summary

In this dissertation, two computational methods on the study of protein-protein interactions and interfaces have been developed and validated. In Chapter 2, we presented the project for characterizing and discriminating different types of protein-protein interactions. We focused on three types of protein-protein interactions, namely, two biologically relevant interactions (obligate and non-obligate) and one that is biologically irrelevant (crystal packing). First, to carry out the analysis on the characterization of the interactions, we curated a balanced dataset of the three types of interactions. Then, we dissected the three types of interactions by using six physicochemical interface features. These features were compared side-by-side for the interactions in the curated dataset. Through these comparisons we illustrated that in the three types of data, obligate interfaces exhibit features that indicate the most "fit" interactions, e.g., the binding regions are the most complementary and the binding areas are the largest, the interfaces are the most hydrophobic, the amino acids at the interfaces have the most different composition from the rest of the protein surfaces and are most conserved. On the contrary, crystal packing cases were demonstrated to be the least "fit" category with respect to the features we have analyzed. The features of non-obligate interactions display intermediate values. Based upon the analysis of the interface features, we constructed a classifier named *NOXclass* for distinguishing the three types of interactions automatically. The classifier uses the six physicochemical features of interfaces and employs a support vector machine algorithm to predict interaction types. According to the inherent relationship of the three interaction types, i.e., two types are biological and one is non-biological, the classifier was implemented in a two-stage scheme. In the first stage, the protein-protein interaction is predicted to be either biological or non-biological. If it is considered biological, it

is processed by the second stage of the classifier and predicted to be either obligate or non-obligate. The classifier was trained and validated using the same dataset curated for the characterization of the three types of interactions. We achieved an accuracy of 91.8% for the classification of three types of data. NOXclass allows for the interpretation and analysis of protein quaternary structures. In particular, it generates testable hypotheses regarding the nature of protein-protein interactions, when experimental results are unavailable. The NOXclass program may be beneficial to the users of protein structure models, as well as protein crystallographers and NMR spectroscopists, e.g., in the inference of protein functions.

In Chapter 3, we introduced the development of a novel interface alignment method and a scoring scheme using a statistical model to measure the significance of the alignment results. In this work, we represented interfaces as vectors denoting the non-covalent interactions across the interfaces and compared the geometry of these vectors using a graph-based approach. The method was named *Galinter* and was compared to a complementary interface comparison program I2I-SiteEngine, as well as a protein backbone comparison program DaliLite. It was shown that the results of the three methods agree to a large extend. We also applied the Galinter methods to four case studies. The applications demonstrated that the Galinter program is capable of identifying similar interaction patterns between different interfaces, independent of the backbone structures of the proteins. To measure the significance of the alignment results, we developed a scoring function using a statistical model based on Poisson process. The parameters used in the scoring function were derived upon a large scale comparison of protein-protein interfaces. It has been validated that the scoring function is helpful for assessing the significance of similar interface patterns identified by the Galinter program. To our knowledge, Galinter is the first program that directly and explicitly aligns different non-covalent interactions at interfaces. The program is capable of aligning not only protein-protein interfaces, but also interfaces in which ligand or non-peptidic molecules are involved. The Galinter program may be utilized to detect conserved patterns of non-covalent interactions, or identify local dissimilar binding modes at interfaces. It also provides an intuitive method for the comparative analysis and visualization of binding modes. Furthermore, it is possible to discover novel functional similarities between proteins by using Galinter and thereby detect similar interfaces between complexes of dissimilar structures.

4.2 Outlook

The total number of current drug targets is of the order of 10^2 (Drews and Ryser (1997) estimated the number to be 482). However, the number of potential drug targets has been estimated to be one order of magnitude higher (Hopkins and Groom, 2002; Imming *et al.*, 2006). Protein-protein interactions regulate a wide variety of important cellular pathways in many biological systems. Therefore, they are considered to be a highly populated class of targets for drug discovery.

The structural characteristics of protein binding sites may deliver common principles that are of general usefulness for future drug design efforts. The characterization effort described in Chapter 2 may be extended to the analysis of other types of interactions and protein-ligand interactions to explore the druggability of proteins with known structures. Machine learning algorithms may be exploited based on the structural features for distinguishing probable binding site residues from non-binding site residues and for deriving putative druggable regions.

One of the fundamental goals of computer-assisted drug design is to predict whether a molecule will bind to a target, and if so, how strong the association will be. Meanwhile, as many drugs actually interact with multiple proteins rather than single targets, it is also important to reduce the probability for the drugs to form unfavorable interactions such that their toxicity is minimized (Hopkins, 2008). Our Galinter method may assist in both tasks. The models of interactions between ligands and targets may be validated by using Galinter to compare them to the interactions involving the targets discovered in known 3D structures. Unfavorable binding may also be detected by searching the binding modes of the modeled complexes in a database where all known binding modes have been deposited.

Protein engineering is another field in which our programs may get involved. The aim of protein engineering is to improve the properties of existing proteins, especially enzymes, by altering their structures (Lutz and Bornscheuer, 2009; Kazlauskas and Bornscheuer, 2009). The understanding of physicochemical properties at interfaces provides guiding information for the modification of existing proteins or the design of new proteins. The comparison of protein binding modes is helpful in the screening of protein variants for improved properties.

In conclusion, we believe that our tools NOXclass and Galinter may play important roles in drug design and protein engineering by giving aid to the understanding of common characteristics of protein interfaces, and by accelerating the search and validation of new drug targets or protein variants of better properties.

Appendices

Appendix A

List of Publications

Journal Papers

Hongbo Zhu, Francisco S Domingues, Ingolf Sommer and Thomas Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics* 2006, 7(27):1–15.

Hongbo Zhu, Ingolf Sommer, Thomas Lengauer and Francisco S Domingues. Alignment of Non-Covalent Interactions at Protein-Protein Interfaces. *PLoS ONE* 2008, 3(4): e1926.

Posters

Hongbo Zhu, Francisco S Domingues, Ingolf Sommer and Thomas Lengauer. Analysis and prediction of protein-protein interaction types. *4th European Conference on Computational Biology (ECCB05)*, Madrid, Spain, Sep. 2005.

Hongbo Zhu, Ingolf Sommer, Thomas Lengauer, and Francisco S Domingues. Alignment of Non-Covalent Interactions at Protein-Protein Interfaces. *16th Annual International Conference Intelligent Systems for Molecular Biology (ISMB2008)*, Toronto, Canada, Aug. 2008.

Bibliography

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**(3), 291–294.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Publishing, New York, USA, 4th edition.

Alexandrov, N. N. (1996). SARFing the PDB. *Protein Eng*, **9**(9), 727–732.

Aloy, P. and Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, **99**(9), 5896–5901.

Aloy, P. and Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nat Biotechnol*, **22**(10), 1317–1321.

Aloy, P., Ceulemans, H., Stark, A., and Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *J Mol Biol*, **332**(5), 989–998.

Ambuhl, C., Chakraborty, S., and Gartner, B. (2000). Common point sets under approximate congruence. In *Proceedings of the 8th Annual European Symposium on Algorithms*, volume 1879/2000, pages 52–64. Springer Berlin / Heidelberg.

Ansari, S. and Helms, V. (2005). Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, **61**(2), 344–355.

Armon, A., Graur, D., and Ben-Tal, N. (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol*, **307**(1), 447–463.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1), 25–29.

Bachar, O., Fischer, D., Nussinov, R., and Wolfson, H. (1993). A computer vision based technique for 3-d sequence-independent structural comparison of proteins. *Protein Eng*, **6**(3), 279–288.

Bader, G. D. and Hogue, C. W. V. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, **20**(10), 991–997.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. F., Pawson, T., and Hogue, C. W. V. (2001). BIND–the biomolecular interaction network database. *Nucleic Acids Res*, **29**(1), 242–245.

Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, **22**(1), 78–85.

Bahadur, R. P., Chakrabarti, P., Rodier, F., and Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, **336**(4), 943–955.

Bahar, I., Atilgan, A. R., and Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, **2**(3), 173–181.

Bai, H., Ma, W., Liu, S., and Lai, L. (2008). Dynamic property is a key determinant for protein-protein interactions. *Proteins*, **70**(4), 1323–1331.

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

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21 Suppl 1**, i38–i46.

Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry*. W. H. Freeman and Company, New York, USA, 5th edition.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**(1), 235–242.

Bernauer, J., Poupon, A., Azé, J., and Janin, J. (2005). A docking analysis of the statistical physics of protein-protein recognition. *Phys Biol*, **2**(1-2), S17–S23.

Bernauer, J., Bahadur, R. P., Rodier, F., Janin, J., and Poupon, A. (2008). DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, **24**(5), 652–658.

Betel, D., Breitkreuz, K. E., Isserlin, R., Dewar-Darch, D., Tyers, M., and Hogue, C. W. V. (2007). Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput Biol*, **3**(9), 1783–1789.

Biswal, B. K., Cherney, M. M., Wang, M., Chan, L., Yannopoulos, C. G., Bilimoria, D., Nicolas, O., Bedard, J., and James, M. N. G. (2005). Crystal structures of the RNA-dependent RNA polymerase genotype 2a of hepatitis C virus reveal two conformations and suggest mechanisms of inhibition by non-nucleoside inhibitors. *J Biol Chem*, **280**(18), 18202–18210.

Blakely, B. T., Rossi, F. M., Tillotson, B., Palmer, M., Estelles, A., and Blau, H. M. (2000). Epidermal growth factor receptor dimerization monitored in live cells. *Nat Biotechnol*, **18**(2), 218–222.

Block, P., Paern, J., Hüllermeier, E., Sanschagrin, P., Sotriffer, C. A., and Klebe, G. (2006). Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins*, **65**(3), 607–622.

Bode, W. and Huber, R. (1992). Natural protein proteinase inhibitors and their interaction with proteinases. *Eur J Biochem*, **204**(2), 433–451.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, **31**(1), 365–370.

Bogan, A. A. and Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J Mol Biol*, **280**(1), 1–9.

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

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

Branden, C. and Tooze, J. (1999). *Introduction to Protein Structure*. Garland Publishing, Inc., New York, USA, 2nd edition.

Breg, J. N., van Opheusden, J. H., Burgering, M. J., Boelens, R., and Kaptein, R. (1990). Structure of arc repressor in solution: evidence for a family of beta-sheet DNA-binding proteins. *Nature*, **346**(6284), 586–589.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, **16**(9), 575–577.

Carugo, O. and Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Sci*, **6**(10), 2261–2263.

Chakrabarti, P. and Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins*, **47**(3), 334–343.

Chandonia, J.-M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S. E. (2004). The ASTRAL compendium in 2004. *Nucleic Acids Res*, **32**(Database issue), D189–D192.

Chang, C.-C. and Lin, C.-J. (2005). *LIBSVM: a Library for Support Vector Machines*.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., and Cesareni, G. (2007). MINT: the molecular interaction database. *Nucleic Acids Res*, **35**(Database issue), D572–D574.

Chen, X.-W. and Liu, M. (2005). Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**(24), 4394–4400.

Chen, Y.-C., Lo, Y.-S., Hsu, W.-C., and Yang, J.-M. (2007). 3D-partner: a web server to infer interacting partners and binding models. *Nucleic Acids Res*, **35**(Web Server issue), W561–W567.

Chiu, W., Baker, M. L., Jiang, W., Dougherty, M., and Schmid, M. F. (2005). Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure*, **13**(3), 363–372.

Choi, Y. S., Yang, J.-S., Choi, Y., Ryu, S. H., and Kim, S. (2009). Evolutionary conservation in multiple faces of protein interaction. *Proteins*, **77**(1), 14–25.

Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**(5498), 304–308.

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

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. MIT Press and McGraw-Hill.

Cunningham, M. W. and Fujinami, R. S., editors (2000). *Molecular Mimicry, Microbes, And Autoimmunity*. ASM Press, Washington DC, USA.

Dafas, P., Bolser, D., Gomoluch, J., Park, J., and Schroeder, M. (2004). Using convex hulls to extract interaction interfaces from known structures. *Bioinformatics*, **20**(10), 1486–1490.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, **23**(9), 324–328.

Dasgupta, S., Iyer, G., Bryant, S., Lawrence, C., and Bell, J. (1997). Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, **28**(4), 494–514.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, **1**, 131–156.

Davies, J. R., Jackson, R. M., Mardia, K. V., and Taylor, C. C. (2007). The Poisson Index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics*, **23**(22), 3001–3008.

Davis, F. P. and Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**(9), 1901–1907.

Davis, F. P., Braberg, H., Shen, M.-Y., Pieper, U., Sali, A., and Madhusudhan, M. S. (2006). Protein complex compositions predicted by structural similarity. *Nucleic Acids Res*, **34**(10), 2943–2952.

Davis, L. I. (1995). The nuclear pore complex. *Annu Rev Biochem*, **64**, 865–896.

De, S., Krishnadev, O., Srinivasan, N., and Rekha, N. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct Biol*, **5**(15), 1–16.

de Berg, M., van Kreveld, M., Overmars, M., and Schwarzkopf, O. (2000). *Computational Geometry: Algorithms and Applications*. Springer, Berlin, Germany, 2nd edition.

Deane, C. M., Salwiński, Ł., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, **1**(5), 349–356.

DeGroot, M. H. and Schervish, M. J. (2001). *Probability and Statistics*. Addison Wesley, Boston, USA, 3rd edition.

DeLano, W. L. (2002). The PyMOL molecular graphics system, on world wide web `http://www.pymol.org`.

Dementiev, A., Simonovic, M., Volz, K., and Gettins, P. G. W. (2003). Canonical inhibitor-like interactions explain reactivity of alpha1-proteinase inhibitor Pittsburgh and antithrombin with proteinases. *J Biol Chem*, **278**(39), 37881–37887.

Deng, Z., Chuaqui, C., and Singh, J. (2004). Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J Med Chem*, **47**(2), 337–344.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**(31), 7133–7155.

Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., , and Weingessel, A. (2005). *e1071: Misc functions of the department of statistics (e1071)*, 1.5-8 edition.

Domingues, F. S. and Lengauer, T. (2007). Inferring protein function from protein structure. In *Bioinformatics - From Genomes to Therapies*, volume 3, pages 1211–1252. Wiley-VCH.

Draper, D. E. and Reynaldo, L. P. (1999). RNA binding strategies of ribosomal proteins. *Nucleic Acids Res*, **27**(2), 381–388.

Drews, J. and Ryser, S. (1997). The role of innovation in drug development. *Nat Biotechnol*, **15**(13), 1318–1319.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. Wiley, New York, USA, 2nd edition.

Dvir, A., Conaway, J. W., and Conaway, R. C. (2001). Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Curr Opin Genet Dev*, **11**(2), 209–214.

Dyson, H. J. and Wright, P. E. (2005). Elucidation of the protein folding landscape by NMR. *Methods Enzymol*, **394**, 299–321.

Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., and Gerstein, M. (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, **18**(10), 529–536.

Efrat, A., Itai, A., and Katz, M. J. (2001). Geometry helps in bottleneck matching and related problems. *Algorithmica*, **31**(1), 1–28.

Eichler, J. (2008). Peptides as protein binding site mimetics. *Curr Opin Chem Biol*, **12**(6), 707–713.

Emsley, J. (1980). Very strong hydrogen bonding. *Chem. Soc. Rev.,*, **9**, 91–124.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**(6757), 86–90.

Espina, V., Woodhouse, E. C., Wulfkuhle, J., Asmussen, H. D., Petricoin, E. F., and Liotta, L. A. (2004). Protein microarray detection strategies: focus on direct detection technologies. *J Immunol Methods*, **290**(1-2), 121–133.

Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., and Tress, M. L. (2009). Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform*, **10**(3), 233–246.

Fass, D., Bogden, C. E., and Berger, J. M. (1999). Crystal structure of the N-terminal domain of the DnaB hexameric helicase. *Structure*, **7**(6), 691–698.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, **27**(8), 861–874.

Fiaux, J., Bertelsen, E. B., Horwich, A. L., and Wuthrich, K. (2002). NMR analysis of a 900K GroEL GroES complex. *Nature*, **418**(6894), 207–211.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, **340**(6230), 245–246.

Finn, R. D., Marshall, M., and Bateman, A. (2005). iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**(3), 410–412.

Fischer, D., Bachar, O., Nussinov, R., and Wolfson, H. (1992). An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. *J Biomol Struct Dyn*, **9**(4), 769–789.

Frigerio, F., Coda, A., Pugliese, L., Lionetti, C., Menegatti, E., Amiconi, G., Schnebli, H. P., Ascenzi, P., and Bolognesi, M. (1992). Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c complex at 2.0 a resolution. *J Mol Biol*, **225**(1), 107–123.

Fryxell, K. J. (1996). The coevolution of gene family trees. *Trends Genet*, **12**(9), 364–369.

Fukuhara, N. and Kawabata, T. (2008). HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res*, **36**(Web Server issue), W185–W189.

Gallet, X., Charloteaux, B., Thomas, A., and Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *J Mol Biol*, **302**(4), 917–926.

Garcia, K. C. and Teyton, L. (1998). T-cell receptor peptide-mhc interactions: biological lessons from structural studies. *Curr Opin Biotechnol*, **9**(4), 338–343.

Garey, M. R. and Johnson., D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, San Francisco, USA.

Gilman, A. G. (1987). G proteins: transducers of receptor-generated signals. *Annu Rev Biochem*, **56**, 615–649.

Glaser, F., Steinberg, D. M., Vakser, I. A., and Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, **43**(2), 89–102.

Glaser, F., Pupko, T., Paz, I., Bell, R. E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, **19**(1), 163–164.

Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D., and Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J Mol Biol*, **299**(2), 283–293.

Gohlke, H., Hendlich, M., and Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol*, **295**(2), 337–356.

Gold, N. D. and Jackson, R. M. (2006a). Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *J Mol Biol*, **355**(5), 1112–1124.

Gold, N. D. and Jackson, R. M. (2006b). A searchable database for comparing protein-ligand binding sites for the analysis of structure-function relationships. *J Chem Inf Model*, **46**(2), 736–742.

Gold, N. D. and Jackson, R. M. (2006c). SitesBase: a database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res*, **34**(Database issue), D231–D234.

Goodsell, D. S. and Olson, A. J. (2000). Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct*, **29**, 105–153.

Green, P. J. and Mardia, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, **93**(2), 235–254.

Green, R. (2000). Ribosomal translocation: EF-G turns the crank. *Curr Biol*, **10**(10), R369–R373.

Grindley, H. M., Artymiuk, P. J., Rice, D. W., and Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol*, **229**(3), 707–721.

Gross, J. L. and Yellen, J. (2006). *Graph theory and its applications*. Chapman & Hall/CRC, Boca Raton, USA.

Gunasekaran, K., Tsai, C.-J., and Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol*, **341**(5), 1327–1341.

Günther, S., von Eichborn, J., May, P., and Preissner, R. (2009). JAIL: a structure-based interface library for macromolecules. *Nucleic Acids Res*, **37**(Database issue), D338–D341.

Haliloglu, T., Baharand, I., and Erman, B. (1997). Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, **79**(16), 3090–3093.

Harary, F. (1994). *Graph Theory*. Westview Press, New York, USA.

Hasegawa, H. and Holm, L. (2009). Advances and pitfalls of protein structural alignment. *Curr Opin Struct Biol*, **19**(3), 341–348.

Hastie, T., Tibshirani, R., Tibshirani, R., and Friedman, R. (2003). *The Elements of Statistical Learning*. Springer-Verlag, New York, USA, 1 edition.

Headd, J., Ban, Y., Brown, P., Edelsbrunner, H., Vaidya, M., and Rudolph, J. (2007). Protein-protein interfaces: Properties, preferences, and projections. *J Proteome Res*, **6**(7), 2576–2586.

Henrick, K. and Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci*, **23**(9), 358–361.

Henschel, A., Kim, W. K., and Schroeder, M. (2006). Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics*, **22**(5), 550–555.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Res*, **32**(Database issue), D452–D455.

Ho, T. K. (1995). Random decision forests. In *Third International Conference on Document Analysis and Recognition (ICDAR'95)*, volume 1, pages 278–282. IEEE Computer Society.

Holm, L. and Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**(6), 566–567.

Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**(1), 123–138.

Hopcroft, J. E. and Karp, R. M. (1973). An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J. Comput.*, **2**(4), 225–231.

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol*, **4**(11), 682–690.

Hopkins, A. L. and Groom, C. R. (2002). The druggable genome. *Nat Rev Drug Discov*, **1**(9), 727–730.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Trans. Neural Networks,*, **13**(2), 415–425.

Huang, C., Stricher, F., Martin, L., Decker, J. M., Majeed, S., Barthe, P., Hendrickson, W. A., Robinson, J., Roumestand, C., Sodroski, J., Wyatt, R., Shaw, G. M., Vita, C., and Kwong, P. D. (2005). Scorpion-toxin mimics of CD4 in complex with human immunodeficiency virus gp120 crystal structures, molecular mimicry, and neutralization breadth. *Structure*, **13**(5), 755–768.

Hubbard, S. J. and Thornton, J. M. (1993). NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College London. http://www.bioinf.manchester.ac.uk/naccess/.

Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res*, **10**(8), 1204–1210.

Imming, P., Sinning, C., and Meyer, A. (2006). Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov*, **5**(10), 821–834.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, **98**(8), 4569–4574.

IUPAC (2005). *IUPAC Compendium of Chemical Terminology*. Electronic version, http://goldbook.iupac.org/.

Jackson, R. M. and Russell, R. B. (2000). The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. *J Mol Biol*, **296**(2), 325–334.

Janin, J. (1997). Specific versus non-specific contacts in protein crystals. *Nat Struct Biol*, **4**(12), 973–974.

Janin, J. and Rodier, F. (1995). Protein-protein interaction at crystal contacts. *Proteins*, **23**(4), 580–587.

Janin, J., Bahadur, R. P., and Chakrabarti, P. (2008). Protein-protein interaction and quaternary structure. *Q Rev Biophys*, **41**(2), 133–180.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**(5644), 449–453.

Jefferson, E. R., Walsh, T. P., Roberts, T. J., and Barton, G. J. (2007). SNAPPI-DB: a database and api of structures, interfaces and alignments for protein-protein interactions. *Nucleic Acids Res*, **35**(Database issue), D580–D589.

Jeffrey, G. A. (1997). *An Introduction to Hydrogen Bonding*. Oxford University Press, New York, USA.

Jensen-Smith, H., Currall, B., Rossino, D., Tiede, L., Nichols, M., and Hallworth, R. (2009). Fluorescence microscopy methods in the study of protein structure and function. *Methods Mol Biol*, **493**, 369–379.

Johnsson, N. and Varshavsky, A. (1994). Split ubiquitin as a sensor of protein interactions in vivo. *Proc Natl Acad Sci U S A*, **91**(22), 10340–10344.

Jones, E., Oliphant, T., Peterson, P., *et al.* (2001). SciPy: Open source scientific tools for Python. http://www.scipy.org/.

Jones, S. and Thornton, J. (1995). Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol*, **63**(1), 31–65.

Jones, S. and Thornton, J. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*, **93**(1), 13–20.

Jones, S. and Thornton, J. (1997a). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, **272**(1), 121–132.

Jones, S. and Thornton, J. (1997b). Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*, **272**(1), 133–143.

Kalvin, A., Schonberg, E., Schwartz, J. T., and Sharir, M. (1986). Two-dimensional, model-based, boundary matching using footprints. *The International Journal of Robotics Research*, **5**(4), 38–55.

Karlsson, R. (2004). Spr for molecular interaction analysis: a review of emerging application areas. *J Mol Recognit*, **17**(3), 151–161.

Kazlauskas, R. J. and Bornscheuer, U. T. (2009). Finding better protein engineering strategies. *Nat Chem Biol*, **5**(8), 526–529.

Kenworthy, A. K. (2001). Imaging protein-protein interactions using fluorescence resonance energy transfer microscopy. *Methods*, **24**(3), 289–296.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., and Hermjakob, H. (2007). IntAct–open source resource for molecular interaction data. *Nucleic Acids Res*, **35**(Database issue), D561–D565.

Keskin, O. and Nussinov, R. (2005). Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Eng Des Sel*, **18**(1), 11–24.

Keskin, O. and Nussinov, R. (2007). Similar binding sites and different partners: implications to shared proteins in cellular pathways. *Structure*, **15**(3), 341–354.

Keskin, O., Tsai, C. J., Wolfson, H., and Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, **13**(4), 1043–1055.

Keskin, O., Ma, B., and Nussinov, R. (2005). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, **345**(5), 1281–1294.

Kim, W. K., Henschel, A., Winter, C., and Schroeder, M. (2006). The many faces of protein-protein interactions: A compendium of interface geometry. *PLoS Comput Biol*, **2**(9), e124.

Koch, I., Lengauer, T., and Wanke, E. (1996). An algorithm for finding maximal common subtopologies in a set of protein structures. *J Comput Biol*, **3**(2), 289–306.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artif Intell*, **97**, 273–324.

Krissinel, E. and Henrick, K. (2005). Detection of protein assemblies in crystals. In *Computational Life Sciences: First International Symposium, CompLife 2005, Konstanz, Germany*, volume 3695, pages 163–174. Springer-Verlag GmbH.

Krissinel, E. and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, **372**(3), 774–797.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., and Robles, V. (2006). Machine learning in bioinformatics. *Brief Bioinform*, **7**(1), 86–112.

Laskowski, M. and Kato, I. (1980). Protein inhibitors of proteinases. *Annu Rev Biochem*, **49**, 593–626.

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph*, **13**(5), 323–330.

Lee, T. I. and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, **34**, 77–137.

Leibowitz, N., Fligelman, Z. Y., Nussinov, R., and Wolfson, H. J. (2001a). Automated multiple structure alignment and detection of a common substructural motif. *Proteins*, **43**(3), 235–245.

Leibowitz, N., Nussinov, R., and Wolfson, H. J. (2001b). MUSTA–a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins. *J Comput Biol*, **8**(2), 93–121.

Levi, G. (1972). A note on the derivation of maximal common subgraphs of two directed or undirected graphs. *Calcolo*, **9**(4), 341–352.

Levin, M. C., Lee, S. M., Kalume, F., Morcos, Y., Dohan, F. C., Hasty, K. A., Callaway, J. C., Zunt, J., Desiderio, D., and Stuart, J. M. (2002). Autoimmunity due to molecular mimicry as a cause of neurological disease. *Nat Med*, **8**(5), 509–513.

Levy, E. D. (2007). Piqsi: protein quaternary structure investigation. *Structure*, **15**(11), 1364–1367.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan c. elegans. *Science*, **303**(5657), 540–543.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, **2**(3), 18–22.

Lichtarge, O., Bourne, H., and Cohen, F. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*, **257**(2), 342–358.

Lin, X., Liu, M., and wen Chen, X. (2009). Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms. *BMC Bioinformatics*, **10 Suppl 4**(S5), 1–14.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J Mol Biol*, **285**(5), 2177–2198.

Lockless, S. W. and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**(5438), 295–299.

Lodish, H., Berk, A., Zipursky, L. S., Matsudaira, P., Baltimore, D., and Darnell., J. E. (1999). *Molecular Cell Biology*. W H Freeman & Co, New York, USA, 4th edition.

Luce, R. D. and Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika*, **14**(2), 95–116.

Lutz, S. and Bornscheuer, U., editors (2009). *Protein Engineering Handbook*. Wiley-VCH, Weinheim, Germany.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A*, **100**(10), 5772–5777.

Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**(5428), 751–753.

Maxim Shatsky, Alexandra Shulman-Peleg, R. N. and Wolfson, H. J. (2005). *Recognition of Binding Patterns Common to a Set of Protein Structures*, volume 3500, pages 440–455. Springer-Verlag GmbH.

McDonald, I. K. and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, **238**(5), 777–793.

Miernyk, J. A. and Thelen, J. J. (2008). Biochemical approaches for discovering protein-protein interactions. *Plant J*, **53**(4), 597–609.

Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *J Mol Biol*, **196**(3), 641–656.

Minarowska, A., Gacko, M., Karwowska, A., and Minarowski, Ł. (2008). Human cathepsin D. *Folia Histochem Cytobiol*, **46**(1), 23–38.

Mintseris, J. and Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins*, **53**(3), 629–639.

Mintseris, J. and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci U S A*, **102**(31), 10930–10935.

Mintz, S., Shulman-Peleg, A., Wolfson, H. J., and Nussinov, R. (2005). Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. *Proteins*, **61**(1), 6–20.

Monod, J., Wyman, J., and Changeus, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J Mol Biol*, **12**, 88–118.

Monti, M., Orr, S., Pagnozzi, D., and Pucci, P. (2005). Interaction proteomics. *Biosci Rep*, **25**(1-2), 45–56.

Moont, G., Gabb, H. A., and Sternberg, M. J. (1999). Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, **35**(3), 364–373.

Moore, P. B. (1998). The three-dimensional structure of the ribosome and its components. *Annu Rev Biophys Biomol Struct*, **27**, 35–58.

Morell, M., Ventura, S., and Avilés, F. X. (2009). Protein complementation assays: approaches for the in vivo analysis of protein interactions. *FEBS Lett*, **583**(11), 1684–1691.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**(4), 536–540.

Najmanovich, R. J., Allali-Hassani, A., Morris, R. J., Dombrovsky, L., Pan, P. W., Vedadi, M., Plotnikov, A. N., Edwards, A., Arrowsmith, C., and Thornton, J. M. (2007). Analysis of binding site similarity, small-molecule similarity and experimental binding profiles in the human cytosolic sulfotransferase family. *Bioinformatics*, **23**(2), e104–e109.

Neuvirth, H., Raz, R., and Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, **338**(1), 181–199.

Neves, S. R., Ram, P. T., and Iyengar, R. (2002). G protein pathways. *Science*, **296**(5573), 1636–1639.

Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, **35**(151), 773–782.

Nogales, E. and Grigorieff, N. (2001). Molecular machines: putting the pieces together. *J Cell Biol*, **152**(1), F1–10.

Nooren, I. M. A. and Thornton, J. M. (2003a). Diversity of protein-protein interactions. *EMBO J*, **22**(14), 3486–3492.

Nooren, I. M. A. and Thornton, J. M. (2003b). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol*, **325**(5), 991–1018.

Nussinov, R. and Wolfson, H. J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A*, **88**(23), 10495–10499.

Ofran, Y. and Rost, B. (2003). Analysing six types of protein-protein interfaces. *J Mol Biol*, **325**(2), 377–387.

Ogmen, U., Keskin, O., Aytuna, A. S., Nussinov, R., and Gursoy, A. (2005). PRISM: protein interactions by structural matching. *Nucleic Acids Res*, **33**(Web Server issue), W331–W336.

Oldstone, M. B., editor (2005). *Molecular Mimicry: Infection Inducing Autoimmune Disease*. Springer-Verlag Berlin Heidelberg.

Olmea, O. and Valencia, A. (1997). Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Fold Des*, **2**(3), S25–S32.

Ortiz, A. R., Strauss, C. E. M., and Olmea, O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci*, **11**(11), 2606–2621.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999a). Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol*, **1**(2), 93–108.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N. (1999b). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, **96**(6), 2896–2901.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H.-W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, **21**(6), 832–834.

Panchenko, A. and Przytycka, T., editors (2008). *Protein-Protein Interactions and Networks: Identification, Computer Analysis, and Prediction*. Springer-Verlag.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, **14**(9), 609–614.

Pazos, F. and Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, **47**(2), 219–227.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**(8), 4285–4288.

Pemmaraju, S. V. and Skiena, S. S. (2003). *Computational discrete mathematics: combinatorics and graph theory with Mathematica*. Cambridge University Press.

Petsko, G. A. and Ringe, D. (2003). *Protein Structure and Function: Primers in Biology*. Wiley-Blackwell.

Piehler, J. (2005). New methodologies for measuring protein interactions in vivo and in vitro. *Curr Opin Struct Biol*, **15**(1), 4–14.

Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., Greenblatt, J., Jessulat, M., Krogan, N., Luo, X., and Golshani, A. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics*, **7**(365), 1–15.

Pitre, S., North, C., Alamgir, M., Jessulat, M., Chan, A., Luo, X., Green, J. R., Dumontier, M., Dehne, F., and Golshani, A. (2008). Global investigation of protein-protein interactions in yeast saccharomyces cerevisiae using re-occurring short polypeptide sequences. *Nucleic Acids Res*, **36**(13), 4286–4294.

Poleksic, A. (2009). Algorithms for optimal protein structure alignment. *Bioinformatics*, **25**(21), 2751–2756.

Polgár, L. (2005). The catalytic triad of serine peptidases. *Cell Mol Life Sci*, **62**(19-20), 2161–2172.

Ponstingl, H., Henrick, K., and Thornton, J. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**(1), 47–57.

Ponstingl, H., Kabir, T., and Thornton, J. M. (2003). Automatic inference of protein quaternary structure from crystals. *J. Appl. Cryst.*, **36**(5), 1116–1122.

Ponstingl, H., Kabir, T., Gorse, D., and Thornton, J. M. (2005). Morphological aspects of oligomeric protein structures. *Prog Biophys Mol Biol*, **89**(1), 9–35.

Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004). The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, **32 Database issue**, D129–D133.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database–2009 update. *Nucleic Acids Res*, **37**(Database issue), D767–D772.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods*, **24**(3), 218–229.

Qi, Y., Klein-Seetharaman, J., and Bar-Joseph, Z. (2005). Random forest similarity for protein-protein interaction prediction from multiple sources. *Pac Symp Biocomput*, pages 531–542.

R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Radisky, E. S. and Koshland, D. E. (2002). A clogged gutter mechanism for protease inhibitors. *Proc Natl Acad Sci U S A*, **99**(16), 10316–10321.

Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T., and Albrecht, M. (2007). Computational analysis of human protein interaction networks. *Proteomics*, **7**(15), 2541–2552.

Raymond, J. and Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, **16**(7), 521–533.

Reese, J. C. (2003). Basal transcription factors. *Curr Opin Genet Dev*, **13**(2), 114–118.

Remy, I. and Michnick, S. W. (2004). A cDNA library functional screening strategy based on fluorescent protein complementation assays to identify novel components of signaling pathways. *Methods*, **32**(4), 381–388.

Remy, I., Wilson, I. A., and Michnick, S. W. (1999). Erythropoietin receptor activation by a ligand-induced conformation change. *Science*, **283**(5404), 990–993.

Rhodes, G. (2000). *Crystallography Made Crystal Clear*. Academic Press, San Diego, USA, 2nd edition.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, **17**(10), 1030–1032.

Rodier, F., Bahadur, R. P., Chakrabarti, P., and Janin, J. (2005). Hydration of protein-protein interfaces. *Proteins*, **60**(1), 36–45.

Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B. T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J Cell Biol*, **148**(4), 635–651.

Ruschhaupt, M., Huber, W., Poustka, A., and Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Stat Appl Genet Mol Biol*, **3**, Article37.

Russell, R. B. and Barton, G. J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**(2), 309–323.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M., and Sali, A. (2004). A structural perspective on protein-protein interactions. *Curr Opin Struct Biol*, **14**(3), 313–324.

Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. *Nature*, **422**(6928), 216–225.

Sánchez, R., Pieper, U., Mirković, N., de Bakker, P. I., Wittenstein, E., and Sali, A. (2000). MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res*, **28**(1), 250–253.

Sander, O., Sing, T., Sommer, I., Low, A. J., Cheung, P. K., Harrigan, P. R., Lengauer, T., and Domingues, F. S. (2007). Structural descriptors of gp120 v3 loop for the prediction of hiv-1 coreceptor usage. *PLoS Comput Biol*, **3**(3), e58.

Sander, O., Domingues, F. S., Zhu, H., Lengauer, T., and Sommer, I. (2008). Structural descriptors of protein-protein binding sites. In A. Brazma, S. Miyano, and T. Akutsu, editors, *Proceedings of 6th Asia-Pacific Bioinformatics Conference*, Advances in Bioinformatics and Computational Biolgoy, pages 79–88, Kyoto, Japan. Imperial College Press, London.

Schalley, C. A., editor (2006). *Analytical Methods in Supramolecular Chemistry*. Wiley-VCH, Weinheim, Germany, 1st edition.

Schmitt, S., Kuhn, D., and Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, **323**(2), 387–406.

Schölkopf, B., Tsuda, K., and Vert, J.-P., editors (2004). *Kernel Methods in Computational Biology*. MIT Press, Massachusetts, USA.

Schwartz, J. T. and Sharir, M. (1987). Identification of partially obscured objects in two and three dimensions by matching noisy characteristic curves. *Int J Rob Res*, **6**(2), 29–44.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol*, **18**(12), 1257–1261.

Setubal, J. C., Setubal, J., and Meidanis, J. (1997). *Introduction to Computational Molecular Biology*. PWS Publishing, Boston, USA.

Shatsky, M., Nussinov, R., and Wolfson, H. (2002). Multiprot - a multiple protein structural alignment algorithm. In *Algorithms in Bioinformatics*, volume 2452 of *Lecture Notes in Computer Science*, pages 235–250, Berlin/Heidelberg, Germany. Springer Verlag.

Shatsky, M., Shulman-Peleg, A., Nussinov, R., and Wolfson, H. J. (2006). The multiple common point set problem and its application to molecule binding pattern detection. *J Comput Biol*, **13**(2), 407–428.

Shaw, A., Fortes, P. A., Stout, C. D., and Vacquier, V. D. (1995). Crystal structure and subunit dynamics of the abalone sperm lysin dimer: egg envelopes dissociate dimers, the monomer is the active species. *J Cell Biol*, **130**(5), 1117–1125.

Sheinerman, F. B., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr Opin Struct Biol*, **10**(2), 153–159.

Shindyalov, I. N. and Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, **11**(9), 739–747.

Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. part II. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, **3**(4), e43.

Shulman-Peleg, A., Mintz, S., Nussinov, R., and Wolfson, H. J. (2004). Protein-protein interfaces: Recognition of similar spatial and chemical organizations. In *Proceedings of the Fourth International Workshop on Algorithms in Bioinformatics*, volume 3240, pages 194–205. Springer.

Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. J. (2005). MAPPIS: Multiple 3D alignment of protein-protein interfaces. In *CompLife*, volume 3695, pages 91–103. Springer.

Shulman-Peleg, A., Shatsky, M., Nussinov, R., and Wolfson, H. J. (2007). Spatial chemical conservation of hot spot interactions in protein-protein complexes. *BMC Biol*, **5**(43), 1–11.

Sierk, M. L. and Kleywegt, G. J. (2004). Déjà vu all over again: finding and analyzing protein structure similarities. *Structure*, **12**(12), 2103–2111.

Skrabanek, L., Saini, H. K., Bader, G. D., and Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Mol Biotechnol*, **38**(1), 1–17.

Smith, T. L. and Sauer, R. T. (1995). P22 Arc repressor: role of cooperativity in repression and binding to operators with altered half-site spacing. *J Mol Biol*, **249**(4), 729–742.

Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*, **28**(18), 3442–3444.

Sotriffer, C. A., Sanschagrin, P., Matter, H., and Klebe, G. (2008). SFCscore: scoring functions for affinity prediction of protein-ligand complexes. *Proteins*, **73**(2), 395–419.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, **311**(4), 681–692.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, **34**(Database issue), D535–D539.

Stein, A., Russell, R. B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, **33 Database Issue**, D413–D417.

Stein, A., Panjkovich, A., and Aloy, P. (2009). 3did update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Res*, **37**(Database issue), D300–D304.

Stryer, L. (1978). Fluorescence energy transfer as a spectroscopic ruler. *Annu Rev Biochem*, **47**, 819–846.

Swint-Kruse, L. (2004). Using networks to identify fine structural differences between functionally distinct protein states. *Biochemistry*, **43**(34), 10886–10895.

Swint-Kruse, L. and Brown, C. S. (2005). Resmap: automated representation of macromolecular interfaces as two-dimensional networks. *Bioinformatics*, **21**(15), 3327–3328.

Tamames, J., Ouzounis, C., Casari, G., Sander, C., and Valencia, A. (1998). EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, **14**(6), 542–543.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley, Boston, USA.

Tanimoto, T. T. (1958). An elementary mathematical theory of classification and prediction. Technical report, IBM Internal Report.

Tatusova, T. A. and Madden, T. L. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*, **174**(2), 247–250.

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *J Mol Biol*, **208**(1), 1–22.

Teyra, J., Doms, A., Schroeder, M., and Pisabarro, M. T. (2006). SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, **7**(104), 1–7.

Teyra, J., Paszkowski-Rogacz, M., Anders, G., and Pisabarro, M. T. (2008). SCOWLP classification: structural comparison and analysis of protein binding regions. *BMC Bioinformatics*, **9**(9), 1–11.

Thanos, C. D., Randal, M., and Wells, J. A. (2003). Potent small-molecule binding to a dynamic hot spot on IL-2. *J Am Chem Soc*, **125**(50), 15280–15281.

Thanos, C. D., DeLano, W. L., and Wells, J. A. (2006). Hot-spot mimicry of a cytokine receptor by a small molecule. *Proc Natl Acad Sci U S A*, **103**(42), 15422–15427.

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. (2000). From structure to function: approaches and limitations. *Nat Struct Biol*, **7 Suppl**, 991–994.

Toby, G. G. and Golemis, E. A. (2001). Using the yeast interaction trap and other two-hybrid-based approaches to study protein-protein interactions. *Methods*, **24**(3), 201–217.

Trozzi, C., Bartholomew, L., Ceccacci, A., Biasiol, G., Pacini, L., Altamura, S., Narjes, F., Muraglia, E., Paonessa, G., Koch, U., De Francesco, R., Steinkuhler, C., and Migliaccio, G. (2003). In vitro selection and characterization of hepatitis c virus serine protease variants resistant to an active-site peptide inhibitor. *Journal of Virology*, **77**(6), 3669–3679.

Tsai, C. J., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1996). A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol*, **260**(4), 604–620.

Tsai, C. J., Xu, D., and Nussinov, R. (1997a). Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci*, **6**(9), 1793–1805.

Tsai, C. J., Lin, S. L., Wolfson, H. J., and Nussinov, R. (1997b). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*, **6**(1), 53–64.

Tsoka, S. and Ouzounis, C. A. (2000). Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nat Genet*, **26**(2), 141–142.

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., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, **403**(6770), 623–627.

Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*, **11**(4), 739–756.

Valdar, W. and Thornton, J. (2001). Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol*, **313**(2), 399–416.

Valdar, W. S. J. (2002). Scoring residue conservation. *Proteins*, **48**(2), 227–41.

Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Curr Opin Struct Biol*, **12**(3), 368–73.

van de Locht, A., Lamba, D., Bauer, M., Huber, R., Friedrich, T., Kröger, B., Höffken, W., and Bode, W. (1995). Two heads are better than one: crystal structure of the insect derived double domain kazal inhibitor rhodniin in complex with thrombin. *EMBO J*, **14**(21), 5149–5157.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag, New York, USA.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York, USA.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Verbitsky, G., Nussinov, R., and Wolfson, H. (1999). Flexible structural comparison allowing hinge-bending, swiveling motions. *Proteins*, **34**(2), 232–254.

Volkmann, N. and Hanein, D. (2003). Electron microscopy. In P. E. Bourne and H. Weissig, editors, *Structural bioinformatics*, pages 115–133. Wiley-Liss.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**(6887), 399–403.

Wall, M. A., Coleman, D. E., Lee, E., Iñiguez-Lluhi, J. A., Posner, B. A., Gilman, A. G., and Sprang, S. R. (1995). The structure of the G protein heterotrimer Gi alpha 1 beta 1 gamma 2. *Cell*, **83**(6), 1047–1058.

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci*, **5**(6), 1001–1013.

Wallace, A. C., Borkakoti, N., and Thornton, J. M. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. application to enzyme active sites. *Protein Sci*, **6**(11), 2308–2323.

Weskamp, N., Kuhn, D., Hüllermeier, E., and Klebe, G. (2004). Efficient similarity search in protein structure databases by k-clique hashing. *Bioinformatics*, **20**(10), 1522–1526.

Weskamp, N., Hüllermeier, E., Kuhn, D., and Klebe, G. (2007). Multiple graph alignment for the structural analysis of protein active sites. *IEEE/ACM Trans Comput Biol Bioinform*, **4**(2), 310–320.

Willett, P. (2008). From chemical documentation to chemoinformatics: 50 years of chemical information science. *J Inform Sci*, **34**(4), 477–499.

Willett, P. and Winterman, V. (1996). A comparison of some measures for the determination of inter-molecular structural similarity measures of inter-molecular structural similarity. *Quant Struct-Act Relat*, **5**(1), 18–25.

Winter, C., Henschel, A., Kim, W. K., and Schroeder, M. (2006). SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res*, **34**(Database issue), D310–D314.

Wolfson, H. J. and Rigoutsos, I. (1997). Geometric hashing: an overview. *IEEE Computational Science and Engineering*, **4**(4), 10–21.

Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (1999a). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*, **285**(4), 1735–1747.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S., and Richardson, D. C. (1999b). Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. *J Mol Biol*, **285**(4), 1711–1733.

Wucherpfennig, K. W. (2001). Structural basis of molecular mimicry. *J Autoimmun*, **16**(3), 293–302.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic Acids Res*, **28**(1), 289–291.

Xu, D., Tsai, C. J., and Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng*, **10**(9), 999–1012.

Xu, Q., Canutescu, A., Obradovic, Z., and Dunbrack, R. L. (2006). ProtBuD: a database of biological unit structures of protein families and superfamilies. *Bioinformatics*, **22**(23), 2876–2882.

Yamamoto, D., Ishida, T., and Inoue, M. (1990). A comparison between the binding modes of a substrate and inhibitor to papain as observed in complex crystal structures. *Biochem Biophys Res Commun*, **171**(2), 711–716.

Yan, Y. and Marriott, G. (2003). Analysis of protein interactions using fluorescence technologies. *Curr Opin Chem Biol*, **7**(5), 635–640.

Yao, N., Reichert, P., Taremi, S. S., Prosise, W. W., and Weber, P. C. (1999). Molecular views of viral polyprotein processing revealed by the crystal structure of the hepatitis c virus bifunctional protease-helicase. *Structure*, **7**(11), 1353–1363.

Yuste, R. (2005). Fluorescence microscopy today. *Nat Methods*, **2**(12), 902–904.

Zhang, C., Vasmatzis, G., Cornette, J. L., and DeLisi, C. (1997). Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol*, **267**(3), 707–726.

Zhang, S.-W., Pan, Q., Zhang, H.-C., Zhang, Y.-L., and Wang, H.-Y. (2003). Classification of protein quaternary structure with support vector machine. *Bioinformatics*, **19**(18), 2390–2396.

Zhou, H.-X. and Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **23**(17), 2203–2209.

Zhou, H. X. and Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, **44**(3), 336–343.

Zhu, H., Domingues, F. S., Sommer, I., and Lengauer, T. (2006). NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, **7**(27), 1–15.

Zhu, H., Sommer, I., Lengauer, T., and Domingues, F. S. (2008). Alignment of non-covalent interactions at protein-protein interfaces. *PLoS ONE*, **3**(4), e1926.