2009

Protein-protein interface: database, analysis and prediction

Feihong Wu
Iowa State University

Protein-protein interface: database, analysis and prediction

by

Feihong Wu

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Bioinformatics and Computational Biology; Computer Science

Program of Study Committee:
Vasant Honavar, Co-major Professor
Robert L. Jernigan, Co-major Professor
Drena Dobbs
Dimitris Margaritis
Guang Song

Iowa State University

Ames, Iowa

2009

DEDICATION

I would like to dedicate this thesis to my wife Lihui Liu without whose support I would not have been able to complete this work.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES ix

ACKNOWLEDGEMENTS xii

CHAPTER 1. General Background 1

1.1 Protein-Protein Interface Analysis and Prediction 2

1.2 Experimental Methods 4

1.2.1 X-ray Crystallography 4

1.2.2 Nuclear Magnetic Resonance (NMR) Spectroscopy 4

1.2.3 Site-Directed Mutagenesis 5

1.2.4 Fluorescence Resonance Energy Transfer (FRET) 5

1.2.5 Chemical Cross-Linking 5

1.3 Computational Methods 6

1.3.1 Protein Docking 6

1.3.2 Evolutionary Methods 6

1.3.3 Patch Based Analysis 7

1.3.4 Machine Learning Methods 8

1.4 Databases 10

1.5 Research Aims 10

1.6 Dissertation Organization 11

CHAPTER 2. Comparing Kernels For Predicting Protein Binding Sites From Amino Acid Sequence 14

2.1 Introduction 14
2.2 Materials 16
2.2.1 42 Peptidase Protein-Protein Interface Data Set 16
2.2.2 56 Protein-DNA Interface Data Set 16
2.2.3 109 Protein-RNA Interface Data Set 16
2.3 Method 17
2.3.1 Support Vector Machines and Kernel Functions 17
2.3.2 Input Representation and Kernel Function Definition 18
2.3.3 Performance Measures 20
2.4 Experimental Results 20
2.5 Related Work 22
2.6 Summary 23
CHAPTER 3. PPIDB – A Database of Protein-Protein Interface 24
3.1 Background 25
3.2 Results and Discussion 30
3.2.1 System Architecture 30
3.2.2 Data Collection Layer 30
3.2.3 Data Publication Layer 34
3.2.4 An Application Case - SHB_PPIS 36
3.3 Future Work 37
3.4 Methods 37
3.4.1 Database Structure 37
3.5 Acknowledgements 38
3.6 Figures 38
3.7 Tables 41
3.8 Appendix 41
CHAPTER 4. Structural Analysis of Protein-Protein Dimeric Interfaces 45
4.1 Introduction 46

4.2 Materials and Methods . . . 47
4.2.1 Dataset . . . 47
4.2.2 Surface versus Non-surface . . . 47
4.2.3 Interface, Exterior and Interior . . . 47
4.2.4 Interface Propensity . . . 48
4.2.5 Residue-Residue Contact Preference . . . 48
4.2.6 Side Chain Orientation . . . 49
4.2.7 Surface Roughness . . . 49
4.2.8 Solid Angle . . . 49
4.2.9 Protrusion-cx Value . . . 49
4.2.10 Surface Micro-Environment: Hydrophobicity and Interface Cluster Size . . . 50
4.3 Analyses Results . . . 50
4.3.1 Residue-Residue Contact Preference . . . 51
4.3.2 Residue Composition and Propensity . . . 53
4.3.3 Variation Entropy . . . 53
4.3.4 Conservation Score . . . 55
4.3.5 Secondary Structure . . . 56
4.3.6 Side Chain Orientation . . . 57
4.3.7 Surface Roughness . . . 58
4.3.8 Solid Angle . . . 59
4.3.9 Protrusion-cx value . . . 60
4.3.10 Surface Micro-Environment: Hydrophobicity and Interface Cluster Size . . . 60
4.4 Discussion . . . 60
4.4.1 A Summary of Protein-Protein Dimeric Interfaces . . . 60
4.4.2 Comparison with Previous Studies . . . 63
4.4.3 The Influence of Different Interface Definitions . . . 63
4.4.4 The Distribution of Interface Size . . . 64
4.4.5 Cutoff of Surface Definition . . . 64

4.4.6 Application: A Case Study . . . 65
4.4.7 Conclusion . . . 67
4.5 Acknowledgements . . . 68

CHAPTER 5. NB_PPIPS - A Naive Bayes Method to Predict Protein-Protein Interaction Sites . . . 69

5.1 Introduction . . . 70
5.2 Materials and Methods . . . 73
5.2.1 Dataset . . . 73
5.2.2 Surface versus Non-surface . . . 74
5.2.3 Interface versus Non-interface . . . 74
5.2.4 Variation Entropy . . . 74
5.2.5 Side Chain Orientation . . . 74
5.2.6 Surface Roughness . . . 75
5.2.7 Solid Angle . . . 75
5.2.8 Protrusion-cx Value . . . 75
5.2.9 Surface Micro-Environment: Hydrophobicity . . . 76
5.2.10 Naive Bayes Classifier . . . 76
5.2.11 Homology-based Structure Modeling . . . 79
5.2.12 Predicting Interfaces on Modeled Structures . . . 79
5.3 Experimental Results . . . 80
5.3.1 Structural and Evolutionary Features Improve Interface Prediction . . . 80
5.3.2 Easy-to-predict and Hard-to-predict Interfaces . . . 80
5.3.3 Predicting Interfaces on Modeled Structures . . . 81
5.4 NB_PPIPS Protein-Protein Interface Prediction Server . . . 86
5.5 Summary . . . 87

CHAPTER 6. Conclusion . . . 89

6.1 Contributions . . . 89
6.2 Future Work . . . 90

BIBLIOGRAPHY 92

LIST OF TABLES

Table 2.1 Comparison of the amino acid identity kernel $\mathbf{K}_i$, the alignment kernel $\mathbf{K}_a$, and several substitution kernels $\mathbf{K}_{sh}$ $\mathbf{K}_{sj}$ and $\mathbf{K}_{sm}$ (derived from HENS920102, JOHM930101, and MCLA720101 substitution matrices respectively). Accuracy (*ac*), recall(*re*), precision (*pr*), and correlation coefficient (*cc*) shown are estimated using leave-one-out cross-validation. 21

Table 4.1 Comparison of interface definitions: ΔASA-based and distance-based 63

Table 4.2 prediction results: chain B of protein 1lj9 67

Table 5.1 Prediction results of different Naive Bayes classifiers with different feature compositions: 1 – sequence, 2 – sequence+side chain orientation, 3 – sequence+variation entropy, 4 – sequence+roughness, 5 – sequence+solid angle, 6 – sequence+hydrophobicity, 7 – sequence+cx, 8 – sequence+cx+hydrophobicity+solid angle+variation entropy 81

Table 5.2 Prediction results of modeled protein structures 84

LIST OF FIGURES

Figure 3.1 PPIDB System Architecture 39

Figure 3.2 Work Flow of PPIDB Data Collection 40

Figure 3.3 Interface Visualization 41

Figure 3.4 Database Schema 42

Figure 4.1 Grid map of residue-residue contact preference 52

Figure 4.2 Percentage frequencies of amino acid residues in the exterior, interface and interior regions 54

Figure 4.3 Propensities of amino acid residues in the exterior, interface and interior regions 54

Figure 4.4 Propensities of amino acid residues in the interface region relative to the surface 54

Figure 4.5 Variation entropy distribution 55

Figure 4.6 Propensities of variation entropy in the interface region relative to the surface 56

Figure 4.7 Conservation score distributions and propensities of interfaces relative to surfaces 56

Figure 4.8 Secondary structure distribution 57

Figure 4.9 Side chain orientation distribution 58

Figure 4.10 Propensities of side chain orientation in the interface region relative to the surface 58

Figure 4.11 Interface propensities of surface roughness 59

Figure 4.12 Interface propensities of solid angle 59

Figure 4.13 Interface propensities of protrusion (cx value) 60

Figure 4.14 Interface propensities of hydrophobicity (average contact energy) and size of interface cluster. 61

Figure 4.15 Interface size distribution 64

Figure 4.16 Relative solvent accessible area distribution 65

Figure 4.17 Interaction Sites Identification of Chain B of Protein *1lj9* Under Two Approaches: voting method (the left) and voting method+refinement strategy (the right). The chain B is shown in grey, with the residues of interest shown in space fill and color coded as follows: black, interface residues identified as such by the classifier (TPs); light grey, interface residues missed by the classifiers (FNs); and dark grey, residues incorrectly classified as interface residues (FPs). For clarity, interface residues for the chain A (gray wireframe) are not shown. The structure diagrams were generated using RasMol (156). 67

Figure 5.1 Interaction Sites Recognition of "easy-to-predict" proteins: 1aih (chain D), 1cdc (chain B), 1igu (chain B), 1joc (chain B), 1lgp (chain A) and 1lj9 (chain B). The predicted chain B is shown in green, with the residues of interest shown in space fill and color coded as follows: red, interface residues identified as such by the classifier (TPs); yellow, interface residues missed by the classifiers (FNs); blue, residues incorrectly classified as interface residues (FPs) and orange, residues correctly classified as non-interface residues (TNs). For clarity, interface residues for partner chains (gray wireframe) are not shown. The structure diagrams were generated using RasMol (156). 82

Figure 5.2 Interaction Sites Recognition of "hard-to-predict" proteins: 1czf (chain A) and 1iqu (chain A) 82

Figure 5.3 Recall_precison curves of the 8 Naive Bayes classifiers with different feature compositions 83

Figure 5.4 Recall_precison curves of the four prediction experiments on protein sequences, modeled protein structures and actual protein structures 85

Figure 5.5 Interaction Site Recognition Using Modeled Structure and Actual Structure (protein 2b5a, chain D): The left displays the prediction of modeled structure (re=0.79 and pr=0.64) and the right shows the prediction of the actual structure (re=0.71 and pr=0.68). RMSD: 1.93 and SE:96.2% 85

Figure 5.6 snapshot of NB_PPIPS running results 86

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis.

First and foremost, Prof. Vasant Honavar, for his constant encouragement and kindly sustained supports. His insightful suggestions help me overcome many difficulties in my Ph.D. research work. I have learned from him not only in scientific knowledge but also scientific attitudes.

I would also like to thank my committee members for their efforts and contributions to this work: Prof. Robert Jernigan and Prof. Drena Dobbs for their kind instructions; Prof. Dimitris Margaritis and Prof. Guang Song for their conversations with me from which I benefit a lot.

Many thanks go to Michael Terribilini, Peter Zaback, Jae-hyung Lee, Fadi Towfic, Yasser El-Manzalawy, Dr. Jyotishman Pathak, Dr. Changhui Yan, Dr. Jie Bao, and many other members in AI lab for their rewarding discussions and helpful suggestions.

My Ph.D research is supported by research assistantships from Bioinformatics and Computational Biology Program and the Artificial Intelligence Research Laboratory funded through grants from the National Institutes of Health (GM 066387) to Vasant Honavar.

CHAPTER 1. General Background

Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters and molecular machines in cells. The diverse roles played by proteins in cells are mediated by interactions between proteins (130), and between proteins and DNA (152), proteins and RNA (30) and between proteins and ligands (33). Protein-protein interactions involve the association of protein molecules. The interactions between proteins are important in virtually every biological process in a living cell. Protein-protein interactions play a pivotal role in energy conversion; ion transportation; DNA replication and transcription; RNA splicing; reaction catalysis and signal transduction. Consequently, understanding the sequence and structural determinants as well as the biophysical and biochemical mechanisms of interactions between proteins is crucial for understanding of cellular function. The study of protein-protein interactions encompasses:

- Experimental detection of pairwise protein-protein interactions (175). Dozens of methods have been developed for this task although only two broad classes of methods have been used on a large scale: Fragment complementation assays such as the yeast two-hybrid (Y2H) system (153) in which split proteins are reconstituted by fusions of interacting proteins; Biophysical methods which include structure determination and mass spectroscopic (MS) identification proteins in complexes. At present, it is unclear whether any particular method should be favored over others; hence, multiple complementary methods are used in practice to cover the interactome (the full complement of interactions among proteins) within a cell or organism of interest.

- Computational analysis of protein-protein interaction networks (47; 17; 178). Such networks provide a global picture of protein-protein interactions that can further be analyzed to identify putative functional modules (21; 171; 174), nodes that play important roles (e.g., hubs) (83); or to determine topological features (degree distribution, hierarchical structure, modularity, etc.

(49; 146; 192; 96). Comparative analysis of two or more networks of the same type from different species can help identify conserved functional modules (159; 168; 135; 196).

- Protein docking (90; 120) methods for modeling the structure of complexes formed by interactions between proteins. Docking methods can be used to address questions such as: can two proteins with known (or predicted) structures bind to each other? If so, what is the strength of the interaction between them? What spatial configuration do they adopt in their bound state?

- Analysis and characterization of protein-protein interfaces in terms of their physico-chemical features (127), topological features (86), geometric features (35; 140; 110; 187) and residue contact preferences (41; 133). Barring a few exceptions (133; 191), much of the published analyses of protein-protein interfaces have been carried out using relatively small datasets. Against this background, comprehensive and systematic analyses of protein-protein interfaces is of interest.

- Computational prediction of protein-protein interfaces or binding sites (169; 164; 165; 195) using amino acid sequence features (134; 58), structural features (when the structure of the target protein is available but the structure(s) of complexe(s) it forms with other protein(s) are unknown (127; 19; 20; 100), evolutionary information (114; 115; 148; 105) as well as a combination of different types of information (158; 112; 113). Many of these methods have been evaluated using relatively small datasets of protein-protein interfaces. Systematic comparison of the alternative methods using large datasets and different classes of interfaces is important for understanding their relative strengths and weaknesses, assessing the reliability of individual methods, and for developing improved methods for predicting protein-protein interaction sites.

1.1 Protein-Protein Interface Analysis and Prediction

The primary focus of this thesis is on the analysis and characterization of protein-protein interfaces and on the development of methods for predicting protein-protein interface sites. Advances in protein methods for identifying the amino acid residues that contribute to the specificity and affinity of the protein-protein interaction can complement experimental methods for construction of protein-protein interaction networks, and enhance the effectiveness of protein docking methods (by helping focus the

search for interfaces to parts of the protein surface that are predicted to be part of an interface), and ultimately, assist the discovery the physicochemical principles of protein macromolecule associations and contribute to advances in rational drug design.

Protein-protein interactions can be classified into several categories based on the sequence identity of constituent proteins, and the strength and the duration of the interactions. The interfaces between subunits of a protein-protein complex are called homo-oligomeric interfaces when the two subunits share a high degree of sequence identity; otherwise, they are called hetero-dimeric interfaces. Based on the strength and duration of interactions, protein-protein interactions can be classified into permanent (obligate) or transient (non-obligate) interactions (133; 129). Permanent or strong interactions result in stable protein-protein complexes, whereas transient or weak interactions result in unstable protein-protein complexes. Transient interactions play a major role in the regulation of several important cellular processes. However, they are much harder to study than obligate interactions due to the lack of the physically stable complexes.

Experimental methods while generally expensive and time-consuming, in many instances, offer the only direct methods available for reliable identification of amino acid residues that participate in protein-protein interactions. Computational methods on the other hand are inexpensive, and can be applied in settings where experimental data are unavailable or too expensive to obtain. However, before computational methods can be applied on a large scale, it is important to determine the reliability of predictions generated by such methods through rigorous statistical cross-validation as well as direct experimental verification of their predictions. We now proceed to briefly review the experimental and computational methods that are available for identification or prediction of protein-protein interfaces.

1.2 Experimental Methods

1.2.1 X-ray Crystallography

X-ray crystallography technique works as follows: shed a beam of X-ray on crystal material; the diffraction caused by the electrons in the crystal produce a three-dimensional picture of the density of electrons; atomic positions and bonds can then be determined. X-ray is regarded as a gold standard of structure determination due to its precision (71) and has been successfully applied on the determination of large protein complexes structures (132; 12; 182; 69). Most resolved structures deposited in the Protein Data Bank (PDB) (14) come from the X-ray method. The challenge with X-ray is the requirement of highly purified protein complexes and favorable conditions for crystallization. Proteins such as membrane proteins, virus envelopes and participants in transient protein-protein interactions rarely form stable crystals (23). Thus, it is impossible to determine the structures of such proteins or protein complexes using X-ray crystallography. Moreover, crystallized protein complexes may not imply biologically relevant conformations. This creates controversy about the reliability of PPIS identified using only structures derived from X-ray crystallography.

1.2.2 Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR imposes external magnetic fields on the nuclei of atoms and aligns them into high or low energy levels. When agitated with an alternating magnetic field, the nuclei shifts between energy levels and produces NMR spectroscopy from which physical, chemical and electronic information is obtained to deduce protein macromolecule structures. Compared to X-ray crystallography, NMR produces lower-resolution protein structures and only works for smaller proteins (about 30-40 kDa). However, the latest development of transverse relaxation-optimized spectroscopy (TROSY) (138) technology has enabled NMR's applications to large proteins (up to 1,000 kDa). NMR with chemical shift, cross-saturation and TROSY has specifically been developed for identifying protein-protein interfaces (138; 163; 54; 170).

1.2.3 Site-Directed Mutagenesis

Site-directed mutagenesis (181) is a molecular biology method to change particular base pairs in a piece of DNA. Gene functions are identified through the comparison of the wild type genotype with the mutated DNA. For instance, alanine scanning mutagenesis alters the codon sequence by substituting specific amino acid residue with alanine and investigates the change of binding affinity of the wild type protein. Thus, specific protein-protein interaction sites can be recognized (37; 38; 7). Evidence shows that interface residues do not contribute equally to the protein-protein binding; the binding affinity comes mainly from a few central residues (dubbed "hot spots"), which are surrounded by less important residues (34). Several research teams (15; 22) have recently begun to explore the relationship between hot spots and protein-protein interfaces.

1.2.4 Fluorescence Resonance Energy Transfer (FRET)

FRET can be applied to recognize transient protein-protein interactions in living cells. When two fluorescent molecules are close enough (within $60\dot{A}$), energy is transferred from one excited fluorescent molecule (the donor) to the other (the acceptor), which can be detected via their spectrum peaks. Using fluorescent molecules as tags, FRET can identify the distance of interactions groups between two proteins (139).

1.2.5 Chemical Cross-Linking

Chemical cross-links are chemical bonds formed in a chemical reaction between two cross-linking reagents. Two proteins tagged with chemical reagents will be covalently cross-linked if they interact with each other. The formation of a cross-link not only identifies the close proximity of the two proteins, but also reveals their contact regions. Combined with subsequent mass spectrometry, chemical cross-linking is well-suited for investigating interacting sites in transient protein complexes (145; 119).

1.3 Computational Methods

1.3.1 Protein Docking

Protein docking refers to the computational modeling of a protein complex formed by two or more unbounded proteins as components. It is helpful to disclose genetic diseases caused by mutated proteins and design rational drugs. Protein docking correlates tightly with PPIS: on the one hand, interaction sites can be easily determined based on the success of the protein docking; on the other hand, unbounded proteins often have been studied before the docking, their interaction sites or "hot-spots" often have already been identified, hence the knowledge of the binding sites significantly reduces the searching space of configurations in the protein docking process (167). Protein docking usually consists of four steps: 1) choose protein surface representation, 2) list a set of possible configurations of protein complexes, 3) evaluate the configurations based on associate energy and select the nearly correct configurations and 4) refine the candidate models by accommodating side-chains. During the process, protein docking utilizes the prior knowledge such as binding sites, NMR conformations and known structures of protein complex homologues to restrict the configuration space (167; 120; 42). Protein docking methods are developed with the Critical Assessment of Predicted Interactions (CAPRI) contest (80). They have been greatly improved using fast Fourier transformation (90; 57; 116; 29), flexible docking (150; 59), analysis of highly-populated low-energy regions (51) and the combination of external biochemical or biophysical data (13), all of which rely on shape complementarities of protein molecule surfaces to select the candidate models. The challenge of docking stems from the unpredictable conformation changes which occur upon binding (63).

1.3.2 Evolutionary Methods

Protein interaction sites are assumed to be more conserved across different protein families than other residues to maintain protein functionality, so attempts have been made to use conservation to identify interaction sites. The Evolutionary Trace (ET) method (114; 115) generates multiple sequence alignment given an input protein sequence and builds up a phylogenetic tree which is subdivided into groups in terms of sequence divergence. Subsequently, each residue position in the multiple sequence

alignment is assigned a score to rank its evolutionary importance by taking into account of its variation within and between groups. High rank residues detected by ET method are clustered spatially (105) or are combined with machine learning methods to identify functional important sites (148). The conservation of residues is also defined in ways other than ET method, such as in consurf_hssp database (61). There is a trend of incorporating the conservation scores of residues as features into machine learning classifiers to improve their prediction (111; 180; 16; 112). Although some works show homo-dimers have more conserved interface residues than surface residues (176), other works reveal that conservation score does not actually improve in the discrimination of interface residues from surface residues (100). The discrepancy might be explained by the recent discovery that protein-protein interfaces are slightly more conserved than surfaces when estimated based on residues rather than surface patches (24).

1.3.3 Patch Based Analysis

Protein-protein interfaces can be regarded as amino acid residues or surface patches. Patch based analysis studies the physicochemical properties of interfaces using surface patches as units. Some properties of the patch depend on aggregate properties of the surface residues it contains (like hydrophobicity and propensity), others depend on the geometry of the molecule surface (like accessible surface area and protrusion). Patch analysis on six properties (solvation potential, hydrophobicity, accessible surface area, residue interface propensity, planarity and protrusion) shows that protein-protein interfaces are more hydrophobic, planner, more global and more protruding than protein surfaces (88; 85; 86). These results have been applied to predict protein-protein interfaces. The general steps are as follows (87; 124): 1) define surface patches, 2) calculate six properties for each surface patch, and 3) score each surface patch in terms of six properties, rank the patches in terms of scores and select out potential interfaces. Patch based analysis emphasizes more on the aggregate role of protein-protein interface residues in contrast to residue based analysis, because residues in protein-protein interfaces do not contribute equally to the protein-protein binding (34). However, patch analysis has a potential drawback: It is necessary to know the structure of the target protein structure. The combinations of patch-based analysis and residue-based analysis have been tried to cultivate advantages from both (19; 100)– pre-

dictions are mainly made on residues, while the contributions by the patches to which each residue belongs are also considered in each prediction.

1.3.4 Machine Learning Methods

The prediction of protein-protein interaction sites can be formulated as a supervised learning problem as follows (25): given a training data set in which each instance has the form $(\mathbf{x}, y)$, learn a function $y = f(\mathbf{x})$ out of it. Here $\mathbf{x} = (x_1, x_2, ..., x_n)$ is a feature vector, y is a 0 vs. 1 two-valued class label in which 0 represents an instance of non-interaction sites and 1 represents an instance of interaction sites. The application of a machine learning method for predicting protein-protein interface residues consists of three steps: 1) Construct a classifier using the training dataset. 2) Assess the performance of the trained classifier on an independent test dataset. 3) If the performance is acceptable, use the classifier to make predictions on new proteins of interest. Typically, steps 1 and 2 are repeated several times (using statistical cross-validation) to get reliable estimates of the performance of the trained classifiers.

Successful application of machine learning methods in practice has to address several challenges:

- No single machine learning method outperforms all other methods on all problems. Hence, a broad class of machine learning algorithms have been applied to the problem of predicting protein-protein interfaces including: Naive Bayes (127), Neural Network (134; 48; 194; 28), Support Vector Machine (98; 148; 180; 16; 189; 19; 43; 141), Linear Regression (111; 100; 112; 74), Baysian Network (20), Hidden Markov Model (55; 128) and Conditional Random Field (123). However, in the absence of direct and systematic comparisons on large datasets, it is unclear as to how the different methods compare against each other and whether their relative strengths can be synergistically combined to generate more reliable predictions.

- The performance of classifiers trained using machine learning methods often depends on the specific attributes or features that are used to encode the inputs. Good feature representations often incorporate domain knowledge in a form that can be exploited by the machine learning algorithm. A variety of features have been used in the prediction of protein-protein interface residues including the identity of amino acid residues in the sequence neighborhood of a target residue (134; 189; 48), physicochemical properties of amino acid residues (127; 194; 28; 98;

16; 19; 43; 141; 74; 20; 55; 128; 123), and evolutionary information (111; 112; 180), as well as their combinations (188; 158; 19; 143). In the absence of direct and systematic comparisons of the different choices of data representations, and different types of features on large datasets, it is unclear as to what the optimal choice of features or the optimal combination of feature representations is for predicting protein-protein interface sites.

- The predictions of a classifier trained on a small training set may not generalize beyond the training data. This often requires feature selection to reduce the dimensionality of the input to the classifier as well as incorporation of penalties for overly complex classifiers.

- Obtaining reliable estimates of the expected performance of a classifier on new data requires steps to ensure that the training and test data used in cross-validation are indeed independent and that the overall distribution of data instances is representative of the scenario in which the trained classifier is intended to be used.

- The proportion of interface residues is often much smaller than that of non-interface residues. Machine learning methods that simply optimize the accuracy (the fraction of correct predictions) tend to favor the majority class at the expense of the minority class. Hence, it is necessary to incorporate steps to cope with class imbalance in training such classifiers and use performance measures that provide a comprehensive picture of the tradeoff between sensitivity and specificity of predictions (25).

- The protein-protein interface data is necessarily incomplete. Whereas the class labels associated interface residues in the dataset can be generally relied on (assuming that the experimental method used to identify the interfaces is accurate), the class labels associated with non-interface residues in the dataset are inherently unreliable: A residue that is labeled as a non-interface residue may in fact participate in an as yet undiscovered protein-protein interaction. Hence, care needs to be exercised in interpreting the predictions generated by classifiers trained on such data.

1.4 Databases

Several databases that focus on different aspects of protein-protein interaction have been developed in recent years. These include: databases of interaction partner or network with data deduced from high-throughput experiments or literature reports, e.g., (186; 8; 27); databases of structurally-defined interfaces between pairs of protein domains with data deduced from Protein Data Bank (PDB) (14), e.g., PIBASE (39), 3DID (4), ProtCom (101), iPfam (149), InterPare (64), amid which iPfam (149) also includes interfaces at amino acid resolutions; databases of co-crystallized complexes, e.g., DOCK-GROUND (44), databases of structural classification of protein-protein interfaces, e.g., SCOPPI (183), tools for characterization and visualization of protein sequence and structure, e.g., STING (125) or SCOWLP (173), and databases of protein-peptide interfaces, e.g., DOMINO (26), electrostatic-surface of protein functional sites, e.g., eF-site (97), amino acid hotspots in protein interfaces, e.g., BID (53), and protein surface regions for functional annotation of proteins, e.g., SURFACE (52) and SPIN-PP (Spi), dataset of non-redundant interface structures (95). Most structural data is derived from Protein Data Bank (PDB) (14) which consists of protein complexes resolved by X-ray or NMR. However, the deposit entries in PDB usually are asymmetric units (ASU), from which the complete crystal macro-molecules can be reconstructed via crystallographic symmetry operation. This imposes difficulties on determining the oligomeric state of proteins and motivates the creation of the Protein Quaternary Structure (PQS) database (72), which separates multiple copies of protein molecules, applies crystal-lographic symmetry operations and removes crystal packing for PDB entries. Published datasets and analyses of protein-protein interfaces have used different definitions of interfaces making it difficult to directly compare the results of analyses or methods for predicting protein-protein interfaces. Hence, there is a need for a comprehensive database of protein-protein interface residues from which large datasets can be extracted based on user-supplied definitions of interfaces.

1.5 Research Aims

The long term goal of our research is to discover the sequence and structural correlates of the protein-protein interfaces. Our guiding hypothesis is that protein-protein interaction rules can be "learned" using machine learning algorithms, which is trained on experimentally well-characterized

data sets to identify protein-protein interaction sites. The specific aims are:

- **Aim1: Develop a database of all known protein-protein complexes.** We will develop a protein-protein interface database (PPIDB), which includes all known protein-protein complexes deposited in the PDB. PPIDB will facilitate protein-protein interface query with customized criteria and extraction of protein-protein interface data sets for statistical analysis and computational learning. PPIDB will update periodically to synchronize with the PDB.
- **Aim2: Analyze protein-protein interface data to discover sequence and structural determinants of protein-protein interfaces.** After establishing the database, we will analyze the data to discover key sequence and structural features of protein-protein interfaces. For example, analysis of amino acid composition of interfaces can reveal whether some amino acids are preferred in protein-protein interfaces.
- **Aim3: Implement and evaluate machine learning algorithms to predict protein-protein interaction sites.** Based on the analysis, we will develop machine learning algorithms to identify protein-protein interaction sites. The implemented classifiers, will assist in experimental work e.g., in the identification of ITK kinase binding sites. The classifiers will be integrated into the online tools of PPIDB system as well.

1.6 Dissertation Organization

The dissertation is organized as follows:

- **Chapter 1** This chapter addresses the general background of the protein-protein interaction sites (PPIS) problem, specific aims of our study and outlines the dissertation.
- **Chapter 2** We have explored kernel methods to determine binding sites on proteins, specifically from protein-protein, protein-DNA, and protein-RNA complexes. We examine three different kernels functions: identity kernel, sequence-alignment kernel, and amino acid sub-

stitution matrix kernel to learn support vector machine classifiers on three data sets: peptidase interface, DNA-binding site and RNA-binding dataset. These results, which have been published in IEEE Joint Conference on Neural Networks, 2006, show that the substitution matrix kernel method improves the predictions. Feihong Wu carried out the computational experiments and drafted the manuscript; Byron Olson contributed to the discussions and draft editing. Drena Dobbs and Vasant Honavar contributed to experimental design, discussions and manuscript preparation.

– **Chapter 3** We have built Protein-Protein Interface Database (PPIDB) , a comprehensive database of protein-protein interfaces extracted from experimentally determined protein complex structures deposited in the current version of Protein Data Bank (PDB). At present, PPIDB consists of $71,486$ binary protein-protein interfaces. PPIDB supports the extraction of well-characterized datasets of protein-protein interface residues for computational analyses. The database is accessible through the Web Interface `http://ppidb.cs.iastate.edu` and a set of Web services `http://ppidb.cs.iastate.edu/axis/services/Version?wsdl`. Feihong Wu designed the system architecture, built up the database, specified the Web service functions and implemented the Web interface, prepared an initial draft of the manuscript and participated in later manuscript revisions. Rafael Jordan designed the Web services, implemented the Web Interface and Web Services, wrote an initial draft of the Web services and Web interface sections of the manuscript and participated in later manuscript revisions. Jyotishman Pathak participated in the implementation of the Web service and the draft editing. Peter Zaback designed the web site and contributed significantly to the draft editing. Changhui Yan participated in discussions on database design, data integration and manuscript reviews. Drena Dobbs and Vasant Honavar participated in database design, discussions, manuscript preparation and revisions. The manuscript is to be submitted to BMC Bioinformatics.

– **Chapter 4** We have analyzed protein-protein dimeric interfaces. We have studied five parameters (amino acid composition, secondary structure, variation entropy, conservation score, side chain orientation) properties to differentiate interfaces from protein exterior

and interior regions and eight parameters (variation entropy, conservation score, side chain orientation, surface roughness, solid angle, cx value, hydrophobicity and interface cluster size) to discriminate interfaces from surfaces. The results of our analysis show that interface residues have side chains pointing inward; interfaces are rougher, tend to be flat, moderately convex or concave and protrude more relative to non-interface surface residues. Interface residues tend to be surrounded by hydrophobic neighbors and form clusters consisting of three or more interfaces residues. Feihong Wu carried out the computational experiments and prepared the draft; Fadi Towfic contributed to the discussions and the draft editing. Drena Dobbs and Vasant Honavar contributed to discussions and manuscript preparation. The results have been submitted to International Journal of Data Mining.

– **Chapter 5** We have built NB_PPIPS, a Naive Bayes classifier to predict protein-protein interaction sites out of protein surfaces. NB_PPIPS improves its prediction by incorporating evolutionary and structural properties. Evaluated on modelled protein structures, NB_PPIPS reveals the importance of protein structures to the prediction. NB_PPIPS implemented as an online server at `http://watson.cs.iastate.edu/nb_ppips`. The results have been submitted to journal BMC Bioinformatics. Feihong Wu carried out the computational experiments, implemented the NB_PPIPS server and web site, and prepared the draft; Fadi Towfic contributed to discussions and the draft editing; Drena Dobbs and Vasant Honavar contributed to discussions and manuscript editing. The manuscript will be submitted to the IEEE Transactions on Bioinformatics and Computational Biology.

– **Chapter 6** This chapter summarizes the study and future work.

CHAPTER 2. Comparing Kernels For Predicting Protein Binding Sites From Amino Acid Sequence

A paper published in the 2006 International Joint Conference on Neural Networks

Feihong Wu, Byron Olson, Drena Dobbs, Vasant Honavar

Abstract The ability to identify protein binding sites and to detect specific amino acid residues that contribute to the specificity and affinity of protein interactions has important implications for problems ranging from rational drug design to analysis of metabolic and signal transduction networks. Support vector machines (SVM) and related kernel methods offer an attractive approach to predicting protein binding sites. An appropriate choice of the kernel function is critical to the performance of SVM. Kernel functions offer a way to incorporate domain-specific knowledge into the classifier.

We compare the performance of three types of kernels functions: identity kernel, sequence-alignment kernel, and amino acid substitution matrix kernel in the case of SVM classifiers for predicting protein-protein, protein-DNA and protein-RNA binding sites. The results show that the identity kernel is quite effective in on all three tasks. The substitution kernel based on amino acid substitution matrices that take into account structural or evolutionary conservation or physicochemical properties of amino acids yields modest improvement.

2.1 Introduction

Proteins are the principal catalytic agents, structural elements, signal transmitters, transporters and molecular machines in cells. Hence, assigning them putative functions from sequences alone remains one of the most challenging problems in functional genomics. Protein-protein, protein-DNA, and protein-RNA interactions play a pivotal role in protein function. Experimental detection of residues in protein-protein interaction surfaces must come from determination of the structure of protein-protein,

protein-DNA and protein-RNA complexes. However, experimental determination of structures of such complexes is a time-consuming and expensive enterprise. Hence, there is a need for reliable computational methods for identifying protein-protein, protein-DNA and protein-RNA binding sites from the amino acid sequence of the protein. Machine learning methods in general, and support vector machines and related kernel methods in particular, offer an attractive approach to construction of sequence-based classifiers for identifying such binding sites (157; 189; 188; 179; 172).

The SVM (18) classifies inputs into two classes using a hyperplane in a high-dimensional space. If the patterns are not separable in the original n-dimensional pattern space, a suitable non-linear kernel function is used to implicitly map the patterns in the n-dimensional input space into a typically higher (finite or even infinite)dimensional feature space in which the patterns become separable. SVM selects the hyperplane that maximizes the margin of separation between the two classes from among all separating hyperplanes. The kernel function measures the similarity between pairs of patterns in the feature space. An appropriate choice of the kernel function is critical to the performance of SVM. An ideal kernel function assigns a higher similarity score to any pair of patterns that belong to the same class label than it does to any pair of patterns that belong to different classes. Kernel functions provide a means of incorporating domain-specific knowledge into an SVM. Hence, there is a large body of work aimed at designing suitable kernels for protein sequence classification (109; 108). Against this background, we investigate the effect of incorporating various types of biological information into SVM kernels for protein-protein, protein-DNA, and protein-RNA binding site prediction.

The rest of this paper is organized as follows: Section 2 describes the three data sets used in the study. Section 3 introduces the kernel methods and describes the design of the three types of kernel functions. Section 4 presents the experimental results comparing the performance of SVM classifiers trained using the different kernel functions considered in this study. Section 5 briefly describes related work on SVM applications in bioinformatics. Section 6 concludes with a summary of the paper and an outline of some promising directions for further research.

2.2 Materials

The data sets used in this study are available for download at `http://www.cild.iastate.edu/GM066387/homepage.htm`.

2.2.1 42 Peptidase Protein-Protein Interface Data Set

Protein-protein interactions play a central role in protein function. A peptidase is an enzyme that digests proteins through the breaking of peptide bonds. The peptidase interface data set consists of 42 peptidase chains (with sequence identity $< 40\%$) from the MEROPS database (147). Interface residues (binding sites) – amino acids in the sequence that bind to another protein, are identified on the basis of observed decrease in solvent accessible surface area (ASA) in the bound complex relative to that of the monomer. The ASA is computed using the Naccess program (76)(http://wolf.bms.umist.ac.uk/naccess/). A residue is defined as a interface residue if the reduction in ASA in the complex is $> 1\dot{A}^2$ (85). Relative solvent accessibility is defined as the ratio of ASA to the nominal maximal ASA of the residue by Rost and Sander (151). A residue is defined as a surface residue when the relative accessibility is greater than 25%. This data set consists of 1694 interface residues out of 5513 total surface residues.

2.2.2 56 Protein-DNA Interface Data Set

Protein-DNA interactions play a pivotal role in DNA replication and transcription. The 56 protein-DNA binding data set, first published by Jones (84), includes 56 non-homologous protein chains. The definition of interface residues is the same as in the 42 peptidase interface data set. This results in 1752 interface residues out of 12665 total residues.

2.2.3 109 Protein-RNA Interface Data Set

Protein-RNA interactions are vitally important in a wide range of biological processes, including regulation of gene expression, protein synthesis, and replication and assembly of many viruses. The 109 protein-RNA binding data set (172) consists of 109 non-homologous protein chains. Interface residues are determined using software ENTANGLE (1). The data set consists of 3518 interface residues out of $25,118$ total residues.

2.3 Method

2.3.1 Support Vector Machines and Kernel Functions

The SVM classifies inputs into two classes using a hyperplane in a high-dimensional space. If the patterns are not separable in the original *n*-dimensional pattern space, a suitable non-linear kernel function is used to implicitly map the patterns in the *n*-dimensional input space into a typically higher (finite or even infinite)dimensional feature space in which the patterns become separable. SVM selects the hyperplane that maximizes the margin of separation between the two classes C_+ and C_- from among all separating hyperplanes. The kernel function measures the similarity between pairs of patterns in the feature space. Given the training data set with *m* labelled examples

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), ..., (\mathbf{x}_m, y_m)$$

$$where \begin{cases} y_k = 1 & \text{if } \mathbf{x}_k \in C_+; \\ y_k = -1 & \text{if } \mathbf{x}_k \in C_-, \end{cases}$$

the SVM produces a decision function:

$$\begin{aligned} D(\mathbf{x}) &= \sum_{k=1}^{m} \alpha_k \mathbf{K}(\mathbf{x}_k, \mathbf{x}) + b \\ &\text{such that} \\ &\text{if D(x)} > 0, \mathbf{x} \in A \\ &\text{otherwise } \mathbf{x} \in B \end{aligned}$$

where the kernel function **K** defines a kernel matrix K whose entries K_{ij} correspond to similarities between pairs of training instances (i.e. $K_{ij} = \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$). A valid kernel function needs to satisfy the Mercer conditions which requires the kernel matrix to be positive semi-definite (104). The optimization procedure used in training a support vector machine coefficients essentially solves a quadratic programming problem. The *weights* α_k $(1 \leq k \leq m)$ and the *bias b* are determined by the SVM algorithm. The training samples with non-zero weights are called the support vectors.

2.3.2 Input Representation and Kernel Function Definition

In this study, the SVM was trained to predict whether or not a residue is in the interaction site. The input to the SVM consists of the identity of amino acids within a window of 11 contiguous residues, corresponding to the target residue flanked by five sequence neighbors residues on each side. The desired output of the classifier is a 1 if the target residue is an interface residue (class C_+) and -1 (class C_-) otherwise. The training set consists of 11-residue subsequences extracted from the protein sequences, with each window labelled with the corresponding class label.

A kernel function defines similarity between two fixed length sequences $S_a = a_1 a_2 ... a_n$ and $S_b = b_1 b_2 ... b_n$ in which $a_i, b_i (1 \leq i \leq n)$ are amino acids and n is the width of the window. We define three kernel functions: the *identity kernel*, the *alignment kernel* (154), and the *substitution kernel* (177).

Definition 1 (identity kernel) *The identity kernel counts the number of matching residues between the two strings S_a, S_b.*

$$\begin{aligned} \boldsymbol{K}_i(S_a, S_b) &= \sum_{k=1}^{n} \boldsymbol{e}(a_k, b_k) \\ where & \begin{cases} \boldsymbol{e}(a_k, b_k) = 1, & \text{if } a_k = b_k; \\ \boldsymbol{e}(a_k, b_k) = 0 & \text{otherwise}. \end{cases} \end{aligned}$$

It is easy to show that the resulting kernel matrix K is a positive semidefinite matrix.

Definition 2 (alignment kernel) *Let $\boldsymbol{A}$ be a matrix of alignment scores obtained by locally aligning each pair of strings S_a, S_b, in the training set.*

$$\boldsymbol{A}(S_a, S_b) = \boldsymbol{align}(S_a, S_b)$$

*where $\boldsymbol{align}(S_a, S_b)$ is the alignment score based on local alignment of S_a and S_b. The **align** function, and hence the matrix $\boldsymbol{A}$ is not guaranteed to be positive definite. To circumvent this problem, we define*

the alignment kernel $\boldsymbol{K}_a$ *as follows:*

$$\boldsymbol{K}_a(S_a, S_b) = \begin{cases} \boldsymbol{A}(S_a, S_b) - \lambda_g & \text{if } S_a = S_b; \\ \boldsymbol{A}(S_a, S_b) & \text{otherwise} \end{cases}$$

where λ_g *is the smallest eigenvalue of the matrix of pairwise alignment scores* $\boldsymbol{A}$*. The resulting kernel matrix* K *is a positive semidefinite matrix.*

Definition 3 (substitution kernel) *Let* $\boldsymbol{M}_s$ *be an amino acid substitution matrix (***?** *). Substitution matrices are not typically positive definite. We can create a positive semidefinite matrix* $\boldsymbol{M}$ *from a substitution matrix* $\boldsymbol{M}_s$ *as follows:*

1. *Find the minimal entry* min *of* $\boldsymbol{M}_s$
2. *Find the maximal entry* max *of* $\boldsymbol{M}_s$
3. $\boldsymbol{M}(i,j) = \frac{\boldsymbol{M}_s(i,j) - \min}{\max - \min}$
4. *Find the least eigenvalue* λ *of* $\boldsymbol{M}$
5. $\boldsymbol{M}(i,i) = \boldsymbol{M}(i,i) - \lambda$

The substitution kernel is defined as follows:

$$\boldsymbol{K}_s(S_a, S_b) \quad = \quad \sum_{k=1}^{n} \boldsymbol{M}(a_k, b_k)$$

The resulting kernel matrix K *is a positive semidefinite matrix.*

Amino acid substitution matrices are symmetric matrices expressing the rate of substitution of one amino acid by another. A variety of substitution matrices based on physical, chemical and biological properties of amino acids as well as evolutionary and structural considerations are available in the AAindex database (91). For example, HENS920102, a well known BLOSUM62 matrix, is based on evolutionary considerations; The substitution matrix JOHM930101 is based on structural considerations, and MCLA720101 is based on chemical properties of amino acids.

2.3.3 Performance Measures

Let *TP* be the number of true positives(residues predicted to be interaction sites that are actually interaction sites); *FP* the number of false positives(residues predicted to be interaction sites that are actually non-interaction sites); *TN* the number of true negatives; *FN* the number of false negatives. the performance measures *ac*(accuracy), *re*(recall), *pr*(precision) and *cc*(correlation coefficient) defined as follows:

$$
\begin{aligned}
ac &= \frac{TP + TN}{TP + FP + TN + FN} \\
re &= \frac{TP}{TP + FN} \\
pr &= \frac{TP}{TP + FP} \\
cc &= \frac{TP * TN - FN * FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}
\end{aligned}
$$

2.4 Experimental Results

We trained SVM classifiers for predicting whether or not a target residue is a (protein-protein, protein-DNA, or protein-RNA) interface residue based on the amino acid identities of its sequence neighbors using the identity kernel $\mathbf{K}_i$, alignment kernel $\mathbf{K}_a$ and substitution kernel $\mathbf{K}_s$. The classifiers were trained and evaluated (using leave-one-out cross-validation) on the 3 data sets: P (42 peptidase protein-protein interface data set), D (56 protein-DNA interface data set) and R (109 protein-RNA interface data set). The alignment kernel was derived using the BLOSUM62 (HENS920102) substitution matrix. The substitution kernel was derived using 3 different substitution matrixes and we got 3 substitution kernels: $\mathbf{K}_{sh}$ with evolution based substitution matrix HENS920102, $\mathbf{K}_{sj}$ with structure based matrix JOHM930101 and $\mathbf{K}_{sm}$ with chemical similarity based matrix MCLA720101. Our SVM classifiers with different kernels were implemented based on WEKA machine learning package (184).

When data sets have unbalanced class representation (as in the case with the data sets used in this study), the traditional performance measure of accuracy can present a misleading picture of the

effectiveness of the classifier. Hence we report multiple performance measures including accuracy, recall, precision, and correlation coefficient. The results are summarized in Table 2.1.

Table 2.1 Comparison of the amino acid identity kernel $\mathbf{K}_i$, the alignment kernel $\mathbf{K}_a$, and several substitution kernels $\mathbf{K}_{sh}$ $\mathbf{K}_{sj}$ and $\mathbf{K}_{sm}$ (derived from HENS920102, JOHM930101, and MCLA720101 substitution matrices respectively). Accuracy (*ac*), recall(*re*), precision (*pr*), and correlation coefficient (*cc*) shown are estimated using leave-one-out cross-validation.

data set	kernel function	*ac*	*re*	*pr*	*cc*
P	$\mathbf{K}_i$	60.3%	54.9%	42.0%	16.6%
	$\mathbf{K}_a$	63.7%	47.6%	43.9%	16.6%
	$\mathbf{K}_{sh}$	63.4%	48.1%	44.0%	17.7%
	$\mathbf{K}_{sj}$	63.6%	49.7%	44.5%	**18.9%**
	$\mathbf{K}_{sm}$	62.0%	51.4%	42.7%	17.0%
D	$\mathbf{K}_i$	64.0%	69.6%	30.0%	25.0%
	$\mathbf{K}_a$	63.9%	66.0%	29.4%	22.7%
	$\mathbf{K}_{sh}$	64.1%	69.3%	29.7%	24.4%
	$\mathbf{K}_{sj}$	64.4%	68.1%	29.8%	24.3%
	$\mathbf{K}_{sm}$	65.1%	69.6%	30.3%	**25.7%**
R	$\mathbf{K}_i$	71.2%	60.3%	34.8%	25.1%
	$\mathbf{K}_a$	69.2%	53.1%	31.9%	18.0%
	$\mathbf{K}_{sh}$	72.1%	58.4%	35.3%	24.9%
	$\mathbf{K}_{sj}$	72.2%	58.9%	35.5%	**25.3%**
	$\mathbf{K}_{sm}$	71.6%	58.6%	34.8%	24.3%

The performance of the identity kernel is competitive with that of other kernels on all three prediction tasks.

The substitution kernel, depending on the data set used, and the specific substitution kernel chosen, sometimes outperforms the identity kernel. In the case of the peptidase protein-protein interface data set, the substitution kernel yields a 13.9% **relative** improvement in correlation coefficient over the identity kernel when the JOHM930101 substitution matrix is used; In the case of the other two data sets, the relative improvement in correlation coefficient offered by the substitution kernel is quite small: 2.8% (using MCLA720101 substitution matrix on the protein-DNA interface data set) and 0.8% (using

JOHM930101 substitution matrix on the protein-RNA interface data set) respectively.

The alignment kernel does not perform as well as the other kernels on these data sets. This might be due to the fact that the substitution matrix used for aligning sequences (BLOSUM62) may be suboptimal for the data sets used. (Note that the results of the substitution kernel varies with on the specific substitution matrix used).

2.5 Related Work

Kernel methods have been widely applied in computational biology, and many kernel functions have been specifically designed for biological data (157; 179). Several authors have explored the use of support vector machines for secondary structure prediction (75) (66). Bram et al. (177) have examined the effects of amino acid substitution matrix on the effectiveness of SVM kernels for secondary structure prediction. Jaakkola et al. (77) have derived a Hidden Markov Model (HMM) profile based SVM-Fisher kernel for remote homology detection. Leslie et al. (109) have explored the p-spectrum kernel and a mismatch kernel (108) for protein function classification. Saigo et al. (154) have proposed a string alignment kernel for protein remote homology detection. Lanckriet et al. (103) have developed a method based on semi-definite programming for optimal linear combination of multiple kernels for protein function prediction.

Several authors have explored the application of machine learning approaches to classification of protein-protein, protein-DNA, and protein-RNA interface sites from amino acid sequences. Yan et al. (189; 188) have used SVM for identifying protein-protein interface residues among surface residues using amino acid sequence information. Sen et al. (158) have proposed an approach to combining several different sources of information (including amino acid sequence, evolutionary conservation, and structure comparison) to improve the accuracy of protein-protein interface residues. Yan et al. (190) have explored the use of several types of information derived from amino acid sequences to train a Naive Bayes classifier on the 56 protein-DNA data set used in this study. The result obtained using amino acid sequence identity alone (correlation coefficient of 24%) is comparable to that of the SVM reported here. However, addition of residue entropy of the target residue (obtained from

multiple sequence alignment) with other sequences in the training data set as an additional input to the classifier improved the correlation coefficient to 28%. Terribilini et al. (172) have used a Naive Bayes classifier to predict protein-RNA interface residues from amino acid sequence. On the data set of 109 protein-RNA interfaces which is the same as the protein-RNA interface data set used in our study, the Naive Bayes classifier yields a correlation coefficient of 35%, which is better than that of SVM trained using sequence kernels. However, the reported performance of Naive Bayes classifier for protein-RNA interface prediction was obtained with a window size of 25 (as opposed to a window size of 11 used in our study).

2.6 Summary

We have compared the performance of 3 types of kernels to predict protein-protein, protein-DNA, and protein-RNA interfaces from amino acid sequence information alone. Our results suggest that the identity kernel is competitive with apparently more sophisticated kernels on all three prediction tasks. Our results also suggest the possibility of improving the performance of the SVM classifiers using kernel functions derived using amino acid substitution matrices. Yan et al. (190) have shown that it is possible to improve the accuracy of protein-DNA interface prediction by using sequence entropy of the target residue as an additional input to the Naive Bayes classifier. Sen et al. (158) have reported improved accuracy of protein-protein interface prediction using multiple types of information. Hence, there is reason to expect that the performance of the SVM classifiers reported in this paper can be further improved by using other types of information such as sequence conservation score (60), predicted or known secondary structure, sequence entropy, sequence disorder, sequence entropy, among others. Work in progress is aimed at exploring these possibilities.

Acknowledgements: This research was supported by a grant from the National Institutes of Health (GM066387) to Vasant Honavar. We thank C. Yan and M. Terribilini for providing data sets and discussions.

CHAPTER 3. PPIDB – A Database of Protein-Protein Interface

A paper submitted to the Journal of BMC Bioinformatics

Feihong Wu, Rafael Jordan, Jyotishman Pathak, Peter Zaback, Changhui Yan, Drena Dobbs and Vasant Honavar

Abstract Protein-protein interactions play a key role in biological processes like signal transduction and metabolism network. Although more and more protein structures are resolved, the experimental work of identifying protein interfaces lags behind. Analyzing and predicting protein-protein interfaces with computational methods, is increasingly used in this area that calls for large benchmark datasets extracted using an unanimous definition.

We have built a Protein-Protein Interface Database (PPIDB) which extracted $71,486$ binary protein interfaces from experimentally determined protein complex structures deposited in the current version of PDB (Protein Data Bank). PPIDB not only identifies protein-protein interaction sites, but also integrates protein functional annotations (Gene Ontology identifier), sequential properties (sequence and residues), structural properties (solvent accessible area and secondary structure) and evolutional properties (variation entropy and conservation score) from other public domain databases. PPIDB is designed to facilitate the construction of well-characterized datasets of protein-protein interface residues for computational analysis. It incorporates three widely used protein-protein interface residue definitions and allows users to specify desired thresholds for each definition. Queries based on the protein name, PDB ID, or Gene Ontology (GO) identifier and batch retrieval of interface residues for single polypeptide chains or for pairs of chains are supported. Tools for filtering datasets on the basis of amino acid sequence similarity and for visualizing interface residues within the primary amino acid sequence mapped onto the 3D structure of proteins are provided. PPIDB is periodically updated and synchronized with the PDB. PPIDB can be accessed through a Web query interface or programmati-

cally through a set of Web services. A sequence homologue based protein-protein interface prediction server (SHB_PPIPS), designates an example of using PPIDB data via Web services.

PPIDB is accessible through a Web Interface `http://ppidb.cs.iastate.edu` and a set of Web services `http://ppidb.cs.iastate.edu/axis/services/Version?wsdl`.

3.1 Background

Protein-protein interactions are essential for virtually all biological processes, including DNA replication and transcription, RNA splicing, signal transduction and metabolism network. Identifying the sequence and structural determinants of specificity and affinity of protein-protein interactions is important not only for macromolecular recognition, but also for practical applications such as rational drug design. High throughput genomic sequencing and structural genomics projects have led to an explosion in the number of available protein sequences and experimentally determined protein structures. In contrast, detailed experimental characterization of protein-protein interfaces has lagged behind because the traditional experimental methods are costly and time-consuming. In this context, computational methods have been employed as a complementary way to analyze and/or predict protein-protein interactions (58; 87; 134; 48; 194; 28; 189; 188; 19; 148; 158).

Protein-protein interaction consists of three problems: First is the "if" problem in which one tries to determine whether two monomeric proteins interact to form a protein complex. This problem emerges in identifying metabolic roles performed by proteins in metabolic networks. Second is the "where" problem in which one tries to identify the interaction sites of a protein interacting with its partners. This problem occurs in annotating protein functions or identifying critical protein function domains. The third is the "how" problem or the protein docking problem in which one tries to build the interaction model given two unbounded proteins. These three kinds of interactions problems are highly correlated with each other: people tend to answer the "if" problem in terms of the functional domains of proteins. After knowing that two proteins interact, it is of great interest to further explore the details of the interaction – the interacting sites and how residues from two proteins coordinate to form protein complexes. In this paper, we primarily focus on investigating the "where" problem, which bridges the

first and the third problems, where protein-protein interface refers to protein-protein binding sites.

Even narrowed down, the study is complicated by the inconsistent definitions of protein-protein interface residues: Jones and Thornton (86) defined a residue as an interface residue if its change of solvent accessible area (ΔASA) is greater than $1\dot{A}^2$ when the protein transits from its monomeric state to its dimeric state; Ofran and Rost (133) suggested that two residues interact if their closest atoms are within a distance cutoff of $6\dot{A}$; Fariselli (48) considered two residues to interact if the distance between their α-carbon atoms is $< 12\dot{A}$. Such inconsistencies make it hard to compare analysis results of researchers since the way in which the definition influences the results remains unknown. Under such circumstances, our first goal is to allow users to choose any of the three definitions freely.

A number of authors (88; 85; 86; 129; 133; 40; 176; 15; 93; 94; 16; 24; 70; 137; 160; 191) have analyzed protein-protein interfaces: Jones and Thornton (88; 85; 86) disclosed that interfaces are hydrophobic, planar and more accessible. Ofran (133) studied amino acid composition and contact preference of six types of interfaces (intra-domain, domain-domain, homo-obligomer, homo-complex, hetero-obligomer and hetero-complex). Nooren (129) and De (40) emphasized spotting the discrepancies between transient interfaces and permanent interfaces. Valdar and Thornton (176) showed homodimeric interface residues are more conserved than surface residues. Their study is complemented by Caffrey's work (24), which showed protein-protein interfaces are slightly more conserved than surfaces when estimated based on residues, but are rarely more conserved than surfaces when estimated based on surface patches. Bogan and Thorn (15) demonstrated that central residues of the protein interface (dubbed "hot spots") impose more influence on the stability of protein complexes. Additionally, Keskin (93; 94) discovered that hot spots clustered in tight-packed regions contribute to the stability of complexes in a modular way. Bordner (16) discovered that interfaces are hydrophobic and less polar, which agrees with Thornton's results. However, he also asserted that interfaces are not more accessible than surfaces, which contradicts with Thornton's results. Headd (70) studied unexpectedly preferred residue-residue pairs in protein-protein interfaces. Pal (137) compared segmentation of interfaces with that of crystal packing contacts. Sheinerman (160) revealed the central role of electrostatic interactions

in protein-protein association. Yan (191) compared the interfaces with protein cores and non-interface surfaces using a large dataset. Protein-protein interfaces are divided into homodimer interfaces vs. heterodimer interfaces in terms of pairwise sequence similarity, transient interfaces vs. permanent interfaces in terms of binding affinity and antibody-antigen vs. protease-inhibitor in terms of functionality. Our second goal, therefore, is to provide queries of protein-protein interfaces and their correlated properties with constraints on pairwise sequence similarity or protein functionality.

Many approaches have been proposed for predicting protein interaction sites from protein sequences and/or structures. Gallet (58) proposed a hydrophobic moment method to predict protein-protein interaction sites. Ofran (134) predicted heterodimer interfaces using neural networks only from protein sequences. Fariselli (48) implemented a neural network method to identify heterodimer protein interfaces using spatial neighboring residues of the target residue as inputs. Likewise, Zhou (194; 28) predicted interaction sites by the neural network method fed with the spatial neighboring residues. Koike (98) employed SVM (support vector machine) to predict interaction sites using spatial neighboring residues. Lichtarge et al. (114; 115; 148) devised a ET (evolution trace) method to connect evolutionary information with protein-protein interface. The ET method was subsequently extended by Landgraf (105) in the analysis of spatial residue clusters. Li (111) and Wang (180) adopted the idea of integrating conservation scores into prediction as well. Bordner (16) predicted protein-protein interface with SVM in terms of sequence profiles and evolutionary rates of spatial neighbors. Yan (189; 188) identified protein interface residues out of surface residues by considering the sequential neighbors of a target residue using a support vector machine method and later refined his work with a two stage approach. Jones (87; 124) predicted four types of protein interfaces (homodimers, small and large protomers from hetero-complexes and antigens) based on their patch analysis method. Neuvirth (127) presented a structure-based program to identify interacting sites by assigning scores to residues in terms of thirteen different properties. Bradford (19; 20) proposed a method to combine the SVM and the patch analysis, followed with an "expert" Bayesian network method. Kufareva (100) recently presented a PIER algorithm which scores surface patches through twelve patch descriptors defined by atom properties and reflects the surface patch scores back to surface residues. Sen et al. (158) re-

ported improved accuracy of predicting hydrolase-inhibitor interfaces by combining several methods. Generative methods, such as Hidden Markov Model (by Friedrich (55), by Nguyen (128)) and Conditional Random Field (by Li (123)) are also applied in addition to the discriminative methods like neural network and SVM. Hoskins (74) disclosed the usage of abnormal exposed secondary structure in the prediction of protein-protein interfaces. Liang (112; 113) combined the side chain energy, conservation score and residue propensity into his PINUP prediction implementation. Dong (43) defined profile-level interface propensity to predict various protein-protein interfaces. Porollo (141) applied the difference of accessible area between monomeric and oligomeric states of a protein in protein-protein interface prediction. Zhou (195; 143) recently built a meta server by combining results from different predictors.

In spite of these advances, most published studies of protein-protein interfaces have relied on analysis of relatively small datasets barring several notable exceptions (185; 134; 133; 191), partly due to the difficulty of extracting well-defined sets of protein-protein interface residues from the PDB. The lack of widely-available large datasets, together with the absence of widely-agreed upon definitions of "interface residues" has made it difficult to compare results of various studies. It is unclear whether conclusions drawn from analysis of relatively small datasets generalize beyond the specific datasets analyzed. To facilitate direct comparison of different analysis and prediction results, e.g., those obtained using automated machine learning approaches, there is an urgent need for a comprehensive database of well-characterized protein-protein interfaces, with support for generation of large benchmark datasets. Our third goal is to supply tools to extract large representative datasets of protein-protein interface for analysis and prediction.

The Protein-Protein Interface Database (PPIDB) is designed to address these needs. PPIDB allows identification of interfaces on a residue-by-residue basis and facilitates extraction of large datasets in a machine-readable format. What then, distinguishes PPIDB from other databases and associated tools that focus on various aspects of protein-protein interactions and protein-protein interfaces? PPIDB focuses on interfaces of protein-protein complexes at the amino acid residue level and enables users

to extract interfaces using a flexible set of parameters. Thus, PPIDB provides unique capabilities that complement existing databases of interacting partners, e.g., DIP (186), MINT (27), databases of structurally-defined interfaces between pairs of protein domains, e.g., PIBASE (39), 3DID (4), ProtCom (101) and InterPare (64), tools for visualization of interactions between pairs of domains, e.g., iPfam (149), databases of co-crystallized complexes, e.g., DOCKGROUND (44), databases of structural classification of protein-protein interfaces, e.g., SCOPPI (183), tools for analysis and visualization of protein sequence and structure, e.g., STING (125) or SCOWLP (173), and databases of protein-peptide interfaces, e.g., DOMINO (26), protein functional sites, e.g., eF-site (97), amino acid hotspots in protein interfaces, e.g., BID (53), and protein surface regions for functional annotation of proteins, e.g., SURFACE (52). Three publicly available databases with goals similar to PPIDB are InterPare (64), DOCKGROUND (44) and SNAPPI-DB (82) all of which seek to provide a comprehensive database of protein-protein interfaces extracted from the PDB. InterPare contains both inter-chain and intra-chain interfaces, but has not been updated since 2004 and consequently, has significantly limited coverage relative to PPIDB. DOCKGROUND is a relational database of co-crystalized protein-protein complexes that allows datasets to be generated based on either sequence or structural similarity. SNAPPI-DB is an object-oriented database of domain-domain interactions observed in structural data at the atomic level. Unlike PPIDB, neither InterPare nor DOCKGROUND provides flexibility in interface definition (e.g., user-specified choices of parameters such as distance thresholds), generation of datasets based on additional user-defined criteria or support of flexible programmatic access to the underlying database. SNAPPI-DB stores domain-domain interactions for multiple domain definitions. It also provides programmatic access to the underlying database through its application programmer's interface (API). This is similar to PPIDB where such access is supported through a Web service interface. Unlike SNAPPI-DB, PPIDB focuses on a database of interface residues for use in large-scale analysis of protein-protein interfaces. In short, PPIDB complements existing databases of protein-protein interfaces without duplicating functionality.

3.2 Results and Discussion

3.2.1 System Architecture

PPIDB consists of two layers: a data collection layer and a data publication layer (See Figure 1). The data collection layer connects to online public domain databases, downloads raw data such as protein-protein complexes, calculates protein-protein interfaces and stores the final data - residue, sequence and structural properties into the database. The data collection layer consists of two components: the data loading program and the MySQL 5.0.4 database server. The data loading program is written in Java and operates the MySQL database through JDBC (Java Database Connectivity). The work flow of the data collection is illustrated in Figure 1. The data publication layer accepts end users' query requests, retrieves data in the database and returns the formalized results. The data publication layer also has two components: Tomcat 5.0.28 Web server and Apache Axis 1.2 Web services (2). End users are able to access PPIDB data either through Web browsers or by running their client programs to invoke Web services.

3.2.2 Data Collection Layer

3.2.2.1 Data Sources

The data collection layer integrates data from the following public domain databases:

- PDB (Protein Data Bank) (14) is the macromolecular structure database from which PPIDB extracts protein-protein interaction chain pairs and interaction sites. PPIDB deducts most attributes of protein chains such as molecule name, sequence, secondary structure etc. from PDB. Moreover, PPIDB updates itself periodically to synchronize with PDB.

- PQS (Protein Quaternary Structure file server) (72) is the complementary data source of PDB for PPIDB to decide on protein-protein interaction chain pairs. It contains quaternary states for macromolecular structures in PDB determined by X-ray crystallography. It attempts to remove duplicate quaternary structures and crystal compacts, which are noise in protein-protein interface data.

- HSSP (Homology-derived Secondary Structure of Proteins) (155), a database that derives the multiple sequence alignment of proteins with known structures, contains the variation entropy of each protein residue. The variation entropy denotes evolutionary conservations of protein residues, while interface residues are believed to be evolutionary-insensitive to keep the protein's functionalities.

- CONSURF-HSSP (61) makes use of multiple sequence alignments of HSSP and estimates the evolutionary rates of residues by an algorithmic tool called Rate4Site (142). It can detect functionally important regions overlooked by HSSP. PPIDB includes the conservation coding, a discrete scale of relative entropy as a complementary to the HSSP's variation entropy.

- GO (Gene Ontology) (6)is the database that provides ontological annotation of genes and gene products. GO organizes biological vocabularies into three branches: molecular functions, biological process and cellular components. Each term in the vocabulary is represented by a GO Identification (GO ID). The GO IDs for a protein chain denote its biological roles. PPIDB integrates the GO IDs of each protein chain to support queries based on GO ID.

3.2.2.2 Data Collection

The primary source of data used by PPIDB is the Protein Data Bank (PDB). The basic approach is: 1) identify the distinct quaternary structures that are components of a given PDB complex; 2) extract interacting chain pairs out of quaternary structures; and 3) identify protein-protein interface residues in each chain pair.

Before deciding on the protein-protein interfaces, it must be determined if two protein chains interact. In general, two protein chains are considered to interact with each other if they are in a biological unit, defined as a macromolecule that has been shown or is believed to be functional. A biological unit can be identified through function analysis or computationally through the quaternary structures of the protein complex. Due to the existence of point-group symmetry in crystallographic studies of protein complex structures, the deposited entries in the PDB are actually Asymmetric Symmetry Units (ASU) which may include a portion of a biological unit or multiple copies of a biological unit. However, quaternary structures can be regenerated through ASUs using the symmetry operation. The Protein Quaternary

Structure Server, PQS (72), generates quaternary structures for those proteins in PDB determined by X-ray crystallography. PQS also tries to distinguish functionally relevant contacts from crystal packing contacts based on physico-chemical parameters of the quaternary structure, since protein-protein interactions are believed to exist inside a quaternary structure with functional relevant contacts. We determine the binary protein interaction chain pairs as follows: for those protein complexes having corresponding entries in PQS, we only consider possible interactions of any two chains in the same quaternary structure; for those having no corresponding entries in PQS, we assume any two chains of the protein complex might interact. Two chains are considered to be interacting if the mean ΔASA is $\geq 200\dot{A}^2$. For example, in PDB, protein complex *1a3a* has 4 chains A, B, C, D. In PQS, they are divided into two quaternary structures: (A, C) and (B, D). So we only need to determine if interactions exist between chains A and C, B and D. On the other hand, PDB protein complex *1ciw* also has 4 chains A, B, C and D, but is regarded as one quaternary structure in PQS, so we need to determine if any combination of two chains will interact. Based on our calculations, we find that there are only four such interaction pairs: (A, B), (A, D), (B, C) and (C, D) out of six candidate chain pair combinations. The ΔASA is calculated by NACCESS (76).

To identify protein-protein interface residues in each chain pair, PPIDB adopts the three different definitions in previous studies: by ΔASA (86), by closest atom distance (133) and by α-carbon atom distance (48). Interface residues (IRs) of a protein chain refer to the set of amino acid residues that belong to protein-protein interfaces formed by that chain with any of the other protein chains in the same PDB structure. Thus, if a protein chain A forms interfaces with chains B and C, the IRs of protein chain A include the residues of chain A located in the interface between chains A and B and in the interface between chains A and C.

To do protein-protein interface analysis and prediction, we need non-redundant representative data sets in which protein sequence homologues are removed from the dataset. In a non-redundant data set, the sequence identity between any two protein chains should be lower than a threshold. PDB builds

clusters of chains[1] under sequence identities of 30%, 50%, 70%, 90% and 95% weekly. PPIDB relies on the PDB clustering to remove redundancy in dataset generation. Table 1 lists the number of protein chains in PPIDB under various sequence identity values.

PPIDB calculates the sequence similarity of any two interacting chains using the ALIGN program[2] thereby allowing users to filter out homodimeric or heterodimeric interfaces in terms of sequence similarity. In addition, PPIDB incorporates the molecule name and GO IDs of each protein chain to support queries with respect to protein functions and biological roles. PPIDB handles the protein-protein interfaces in a residue-based way, so it has properties such as interface status, surface status, secondary structure, variation entropy etc. residue by residue.

PDB updates monthly to add new protein complex structures, remove obsolete ones and make changes to incomplete ones. PPIDB's data loading program follows the PDB's monthly updating logs and periodically updates PPIDB to synchronize with PDB. Initially set up in terms of the PDB version on Oct 20, 2007, PPIDB has been updated to keep pace with the PDB version on June 1, 2008.

3.2.2.3 Contents & Statistics

PPIDB contains a total of $22,225,472$ amino acid residues, of which $4,097,830$ correspond to interface residues in protein-protein complexes using the definition of ΔASA (the numbers of interface residues are 4,722,127 and 6,886,548 respectively using definitions of closest atom distance and α-carbon distance). The current version of PPIDB is derived from a total of $51,458$ protein structures retrieved from the PDB. These structures correspond to a list of protein sequences for the PDB snapshot on June 1st, 2008, which can be downloaded from ftp://ftp.rcsb.org/pub/pdb/derived_data/pdb_seqres.txt/. After removing complexes containing only protein-DNA, protein-RNA and crystal contacts, 38,815 PDB structures remained, from which the PPIDB data were derived. These 38,815 complexes consist of 94,220 polypeptide chains (17,909 of them are monomeric chains and do not participate in the protein-protein interfaces) drawn from 2,724 species with 1,847 unique GO function labels. The 76,311

[1] ftp://ftp.rcsb.org/pub/pdb/derived_data/NR
[2] ftp://ftp.virginia.edu

chains participate in 71,486 inter-chain protein-protein interfaces. Pair-wise sequence identity between pairs of chains is $> 90\%$ for 47,604 and $< 30\%$ for 21,567 of these interfaces.

3.2.3 Data Publication Layer

PPIDB can be accessed through a Web interface or a set of Web services. The Web interface allows users to access the PPIDB data through a Web browser in a traditional way, while Web services (2) enable users to access the PPIDB data in their own client programs. Web services are self-contained, self-describing, modular applications that can be published, accessed and invoked over the Web. The Web service technology aims to connect computer systems and enable data exchange data and other tasks without human interference.

3.2.3.1 Web Interface

The Web Interface[3] supports the following six types of queries:

1. *Identify interacting chain pairs within a protein complex.* Through this query, a user can determine if any two chains in a complex interact. For example, a user might know that the protein complex 1ciw has 4 chains in PDB, namely, A, B, C and D. However, with the help of this query the user can find out which of these 4 chains interact. In this case, there are only 4 interaction pairs: (A, B), (A, D), (B, C) and (A, D).

2. *Identify protein-protein interfaces for a single chain.* By specifying a single protein chain, a user can receive the interacting sites by providing the following additional information: interface definition, threshold, the similarity of two chains or the other chain. For example, the protein complex *1d5x* has 3 chains A, B and C in which there are 2 interacting chain pairs:(A,B), (A,C). The similarity between A and B is 27% and the similarity between A and C is 18.9%. Now consider the chain *1d5xA*, if one only wants to know the interactions, one can query without any constraints and get the combined interface sites of *1d5xA* from the two interaction pairs: AB and AC. If one cares only about interfaces whose similarity of chains are greater than 20%, the

[3]http://einstein.cs.iastate.edu/ppidb/Search

interactions between A and C can be filtered out; or, one can query the interacting residues of chain A with C by specifying chain C directly.

3. *Basic Search*. This is the way to generate non-redundant protein-protein interface datasets for analyses. A user can specify additional options such as the interface definition, the protein complex resolution, etc. to customize the generation of datasets, which are XML format files sent out via email.

4. *Search by GO IDs*. Through this query, a user can generate non-redundant protein-protein interface datasets by specifying gene annotation. For example, one might want to get the interfaces of proteins that have the molecular function "trypsin activity" which corresponds to GO ID 0004295 and interfaces of those that act on the biological process "proteolysis" which corresponds to GO ID 0006508. The user can input the corresponding GO IDs "0004295 0006508" and apply the function of search by GO to obtain what the desired interface datasets.

5. *Search protein chains by molecule name*. Using this query, a user can generate non-redundant protein-protein interface datasets by specifying by specifying the molecule name. For example, to generate the interfaces of all peptidases but not the inhibitors, one can include the key word "peptidase" in the search and exclude the key word "inhibitor" to get the result.

6. *Batch query for a list of PDB IDs*. A user can also generate protein-protein interface dtasets by providing a protein ID list. List items could be a PDB ID (e.g. 1ciw), a protein chain (e.g. 1ciwA), or a protein chain pair (e.g. 1ciwAB).

3.2.3.2 Visualization

Using the Web interface the computed protein-protein interfaces can be displayed with 3-D representation of the protein structure, which PPIDB carries out by embedding the Jmol[4] open source viewer in the dynamically generated Web pages. This visualization provides an intuitive way to study the correlation of protein interfaces and protein structures. Figure 3 demonstrates an example of an interface query of a single protein chain.

[4]http://www.jmol.org

3.2.3.3 Web Services

PPIDB provides seven running services: *infoComplex*, *interactingChains*, *interfacesResiduesOneChain*, *interfaceResiduesTwoChains*, *interfacesBasicSearch*, *interfacesGoId* and *interfacesMoleculeName*, which are the infrastructure for Web interface implementation. The specifications of the Web services, including function name, input and output parameters, are described in a Web Service Definition Language (WSDL) (32) file[5]. Conforming to the specifications, users can invoke Web Services in their own programs ("client program") by connecting to the Web Services, sending out a request with input parameters and getting back a response with the output parameters. Client programs can be written independent of software languages, which greatly facilitates the use of PPIDB data. PPIDB provide examples of client programs written in Java and Python[6]. A sample Web service client program written in Java is shown in Additional File 1.

3.2.4 An Application Case - SHB_PPIS

Sequence Homologue Based Protein-Protein Interface Prediction Server (SHB_PPIS) is an online server used to predict protein-protein interfaces in term of protein sequence. As an application of PPIDB, it demonstrates how easily PPIDB data can be used and seamlessly integrated into other independent applications.

SHB_PPIS aims to identify interface residues in a given protein sequence using the Naive Bayes (45) machine learning method: each residue r_0 is regarded as an instance with class label c_1 (interface residue) or c_0 (otherwise). Each instance has 11 features $\mathbf{x} = r_{-5}, r_{-4}, .., r_{-1}, r_0, r_1, ..., r_5$ corresponding to 11 sequential neighbor residues: five each on both sides of the target residue r_0 and r_0 itself. Let $P(c_k|\mathbf{x})$ be the conditional probability that an instance belongs to class c_k (k=0 or 1) given that it has feature vector **x**, the Naive Bayes classifier works as follows:

$$\frac{P(c_1|\mathbf{x})}{P(c_0|\mathbf{x})} = \frac{P(c_1) \times \prod_{j=-5}^{5} P(x_j|c_0)}{P(c_0) \times \prod_{j=-5}^{5} P(x_j|c_1)}$$

$$\text{If } \frac{P(c_1|\mathbf{x})}{P(c_0|\mathbf{x})} > 1, r_0 \text{ is an interface residue;}$$

[5]http://einstein.cs.iastate.edu/axis/WebServices.jws?wsdl
[6]http://einstein.cs.iastate.edu/ppidb/ZDownload

Otherwise , r_0 is non-interface site.

The conditional probabilities $P(c_k|\mathbf{x})$ are calculated from training data sets with known class labels of residues. SHB_PPIS is expected to boost the prediction by using sequence homologues to generate the training dataset. The working procedure is:

1. Run blastp (3) against the target protein sequence to obtain its sequence homologues.
2. Identify protein-protein interaction residues of the sequence homologues using the PPIDB Web service *interfacesResiduesOneChain*.
3. Form the training dataset, learn the Naive Bayes classifier and predict the interaction residues out of the target protein sequence.

3.3 Future Work

Current work is directed toward the development of several additional Web services for analysis of protein-protein interface datasets and visualization of the results (e.g., relative amino acid propensities, surface roughness, local curvature of interfaces and non-interfaces). We also plan to annotate PPIDB and the associated Web services with metadata conforming to W3C standards to enable other research groups to integrate PPIDB with further data resources and utilize services offered by PPIIDB as part of larger analysis workflows or pipelines. We plan to add several additional tools for the prediction of protein-protein interface residues and for the analysis of protein-protein interfaces based on various physicochemical, structural and geometric properties of interfaces.

3.4 Methods

3.4.1 Database Structure

The PPIDB database is implemented as a relational database. The simplified version of the current database schema is shown in Figure 4. The key tables CHAIN, CHAIN_CHAIN and RESIDUE store information of single protein chains, protein chain-chain interfaces and residues respectively. Tables MOLECULE, SPECIES and COMPLEX store protein molecule names, protein species and protein

complex source respectively. Table CHAIN has many-to-one relationships with tables MOLECULE, SPECIES and SOURCE, while table CHAIN_CHAIN has a many-to-one relationship with table CHAIN, which is often used in schema design of relational databases to remove store redundancy. Table RESIDUE stores properties of individual residues such as amino acid, entropy, etc. Table CHAIN stores properties for individual chains like sequence, secondary structure, surface, etc. Table CHAIN_CHAIN stores properties relating to binary interfaces like similarity and interfaces under default threshold (ΔASA: $1\dot{A}$; α-carbon distance: $12\dot{A}$; closest atom distance: $6.0\dot{A}$). The database also stores intermediate data such as calculated ΔASA, α-carbon distance and minimal atom distance of residues in the table RESIDUE_CHAIN1, RESIDUE_CHAIN2 and RESIDUE_RESIDUE. Consequently, if users query for interfaces under the default thresholds, PPIDB will immediately return interfaces from its storage; otherwise, PPIDB will calculate the interface based on the saved intermediate data, which takes a little more time.

3.5 Acknowledgements

This research was supported in part by a grant from the National Institutes of Health (GM066387) to Vasant Honavar.

3.6 Figures

Figure 1 -PPIDB System Architecture

PPIDB system is composed of two parts: the data collection layer (left) and the data publication layer (right). The data collection layer pumps data into the database. The data publication layer presents the data through a Web Interface and a set of Web services.

Figure 2 - Work Flow of PPIDB Data Collection

The PPIDB data collection procedure depicts how protein-protein interfaces are extracted from public domain data sources (PDB, PQS, etc.), processed and deposited into the database.

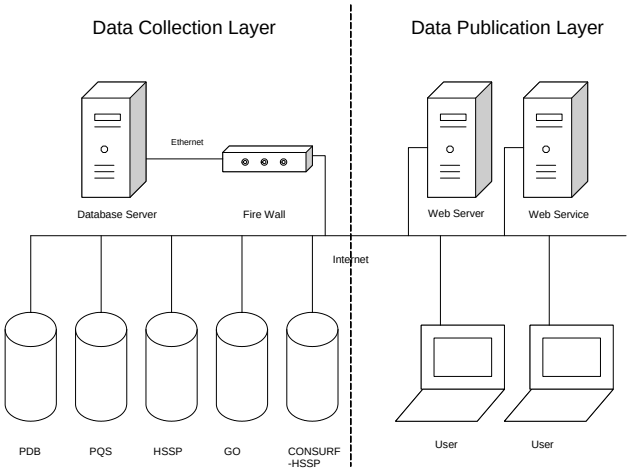

Figure 3.1 PPIDB System Architecture

.

Figure 3 - Interface Visualization

Interaction Sites of Protein Chain *1d5xA*. Chain A is shown as white *spacefill*. Interacting chains B and C, are shown as yellow *ribbons*. Interface residues of chain A, defined by loss $1\dot{A}$ of ASA, are green. The structure diagram was generated by PPIDB Web query interface with Jmol plug-ins.

Figure 4 -Database Schema

The main tables and relationships of the PPIDB database are shown. The arrows represent the foreign key constraints.

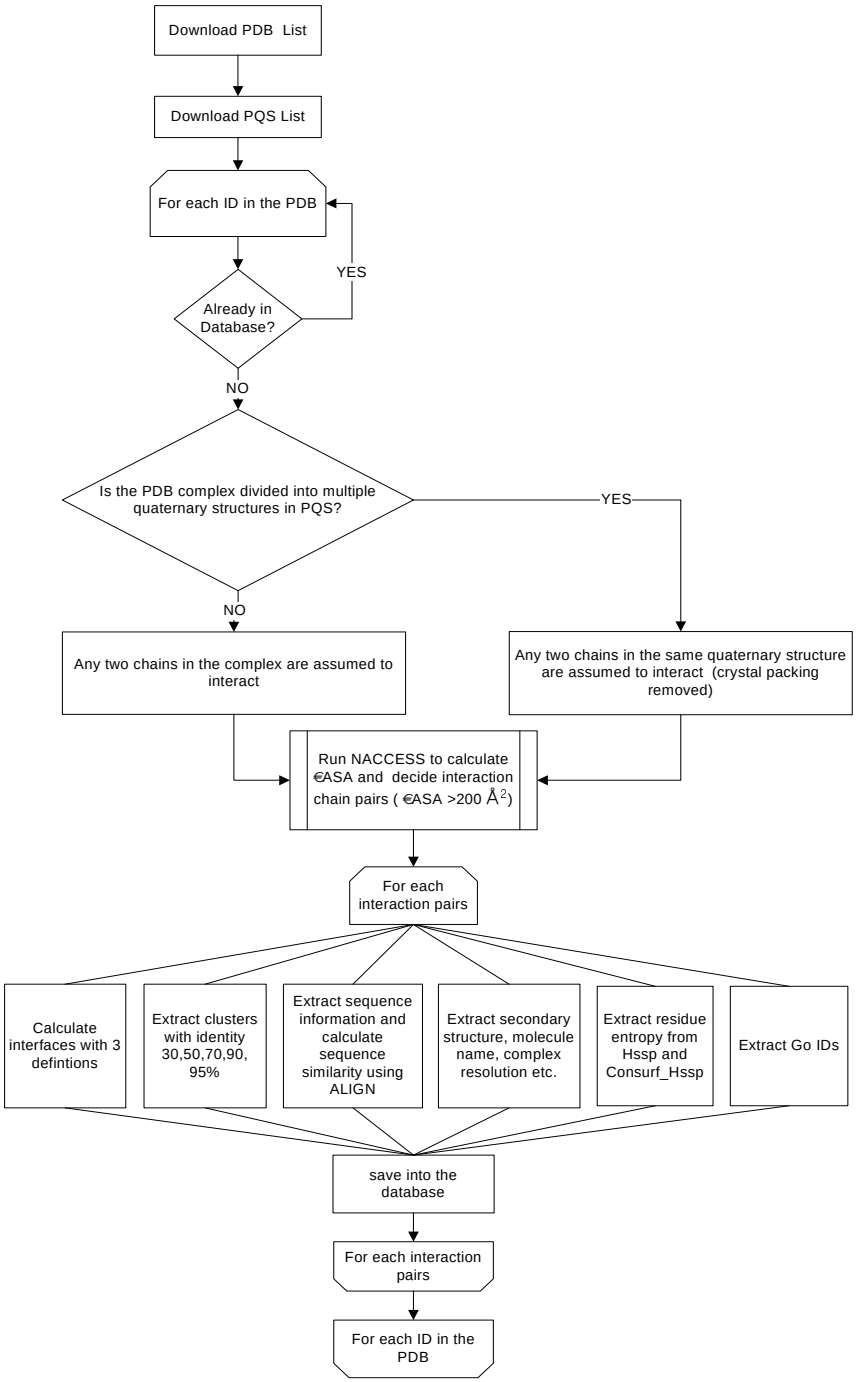

Figure 3.2 Work Flow of PPIDB Data Collection

.

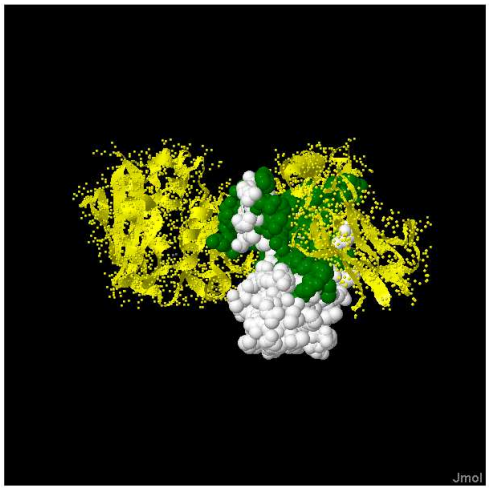

Figure 3.3 Interface Visualization

.

3.7 Tables

Table 1 - The sizes of non-redundant datasets with various sequence identity cutoffs

Sequence Identity Cutoff	Size of Non-Redundant Dataset
30	6492
50	10684
70	12365
90	14022
95	14871

3.8 Appendix

A sample: Web service client program written in java

This program invokes the Web service *interfacesMoleculeName*. It retrieves non-redundant protein chains with 90% similarity, including "peptidase" and excluding "inhibitor" from the molecule name and considering only protein complexes obtained using X-ray crystallography with a resolution $\leq 3\ \dot{A}$. Residues with loss of ASA upon complexation greater than $1\ \dot{A}^2$ are defined as interfaces. More details about the use of PPIDB Web services can be found online in the sections *"Documentation"* and *"Download"* on the PPIDB web page.

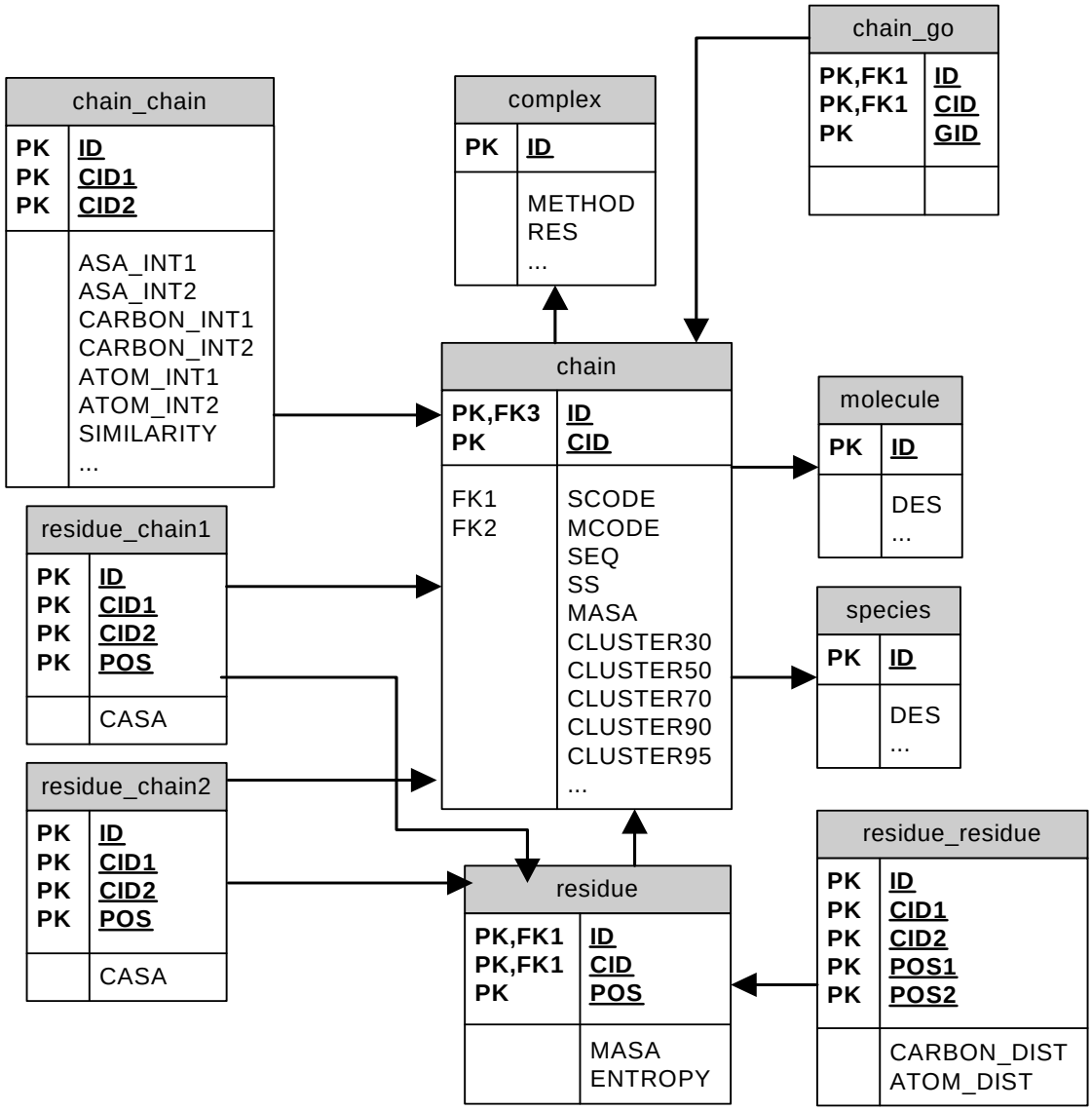

Figure 3.4 Database Schema

.

```
/* Sample invocation of the interfacesMoleculeName search service*/
import java.net.MalformedURLException; import java.net.URL; import
java.rmi.RemoteException; import javax.xml.namespace.QName; import
javax.xml.rpc.ServiceException; import org.apache.axis.client.Call;
import org.apache.axis.client.Service;

public class MoleculeNameServiceClient {
    // Interface definitions
    final static int atomDistance = 1;
    final static int alphaCarbonDistance = 2;
    final static int ASAChange = 3;
    // Experimental methods
    final static int XRayCrystallography = 1;
    final static int NMR = 2;

    public static void main(String[] args)
        throws ServiceException, MalformedURLException {
        //Step 1. Setup for the web service invocation
        Service service = new Service();
        Call call = (Call)service.createCall();
        call.setTargetEndpointAddress
        (new URL("http://einstein.cs.iastate.edu/axis/WebServices.jws?wsdl"));
        call.setOperationName(new QName("", "interfacesMoleculeName"));

        //Step 2: Parameters of the service
        String includedText = "peptidase";
        String notIncludedText = "inhibitor";
        double interfaceDefinitionThreshold = 1.0;
        // interfaces defined by loss of ASA under complexation
        int experimentalMethod = XRayCrystallography;
        double xRayResolutionThreshold = 3.0;
        double chainSimmilarityLowerLimit = 0.0;
        double chainSimmilarityUpperLimit = 100.0;
        double RASAthresholdSurface = 0.125;
        int identityFilter = 90;
        //identityFilter possible values: 30, 50, 70, 90, 95
        int typeOfService = ASAChange;

        //Step 3: Invocation of the service
        HashMap chains = null;
```

```
41          try {
42              chains = (HashMap )call.invoke(
43                      new Object[]{includedText, notIncludedText,
44                      new Double(interfaceDefinitionThreshold),
45                      new Double(RASAthresholdSurface),
46                      new Integer(identityFilter),
47                      new Double(chainSimmilarityLowerLimit),
48                      new Double(chainSimmilarityUpperLimit),
49                      new Integer(experimentalMethod),
50                      new Double(xRayResolutionThreshold),
51                      new Integer(typeOfService)});
52          } catch (RemoteException e) {
53              System.out.println("Remote exception: " + e.toString());
54              e.printStackTrace();
55          }
56
57          //Step 4. Retrieve and print the results
58          String state = (String)chains.get("pstate");
59          if (state.equals("ok")) {
60          // if there was not error within the web services
61              Object [] pdbChains  = (Object []) chains.get("data");
62              for (int i = 0; i < pdbChains.length; i++){
63                  String pdbChain = (String)pdbChains[i];
64                  System.out.println(pdbChain);
65              }
66          } else {
67          //  print the web service errors
68              System.out.println("Problem with interfacesMoleculeName\n Status: " + state);
69              System.out.println("Message error: "+((String)chains.get("pmessage")));
70          }
71      }
72  }
```

CHAPTER 4. Structural Analysis of Protein-Protein Dimeric Interfaces

A paper submitted to the International Journal of Data Mining

Feihong Wu, Fadi Towfic, Drena Dobbs, Robert Jernigan and Vasant Honavar

Abstract Protein-protein interfaces are analyzed on a series of properties to differentiate them from protein exterior and interior regions. The residue-based analysis is carried out on a large, non-redundant dataset of $2,383$ protein chains extracted from dimeric complexes. The residue-residue contact preference map discloses the importance of cysteine-bridge, salt-bridge and aromatic residues to interfaces. Five parameters (amino acid composition, secondary structure, variation entropy, conservation score, side chain orientation) show that interfaces can exist amid exterior and interior regions, but are more likely to be closer to the exterior. Furthermore, protein interfaces can be analyzed according to structural properties when protein's structure is known. We consider eight parameters: variation entropy, conservation score, side chain orientation, surface roughness, solid angle, cx value, hydrophobicity and interface cluster size. The results of our analysis show that interface residues have side chains pointing inward; interfaces are rougher, tend to be flat, moderately convex or concave and protrude more relative to non-interface surface residues. Interface residues tend to be surrounded by hydrophobic neighbors and form clusters of three or more interfaces residues. These findings are consistent with previous published studies using much smaller datasets. We find that none of the sequence or structure derived features carries a strong enough signal to allow reliable prediction of protein-protein interfaces. This underscores the need for developing sophisticated machine learning methods that can discover sequence and structural correlates of protein-protein interfaces.

4.1 Introduction

Protein-protein interactions play a pivotal role in cellular processes such as DNA replication and transcription, RNA splicing, signal transduction and metabolic networks. Therefore, understanding the sequence and structural determinants of protein-protein interactions is crucial for understanding biological processes, including those that play a role in diseases and efforts to design therapeutic drugs. Many studies of protein-protein interface residues have been carried out to identify specific physicochemical characteristics that contribute to protein-molecule recognition. These studies have covered a wide scope and analyzed a variety of interface types (homo versus hetero dimer, transient versus permanent interface, etc.), considered different amino acid characteristics and representations, and used different definitions of interfaces (measured at the level of residues or surface patches). At the residue level, interfaces differ in terms of amino acid composition, inter-residue contact preference and degree of conservation across orthologous proteins relative to non-interfaces. At the surface patch level, interfaces are more hydrophobic, planar and protruding relative to non-interfaces (133; 176; 85; 86; 40; 129; 127; 70; 41; 169; 164; 165; 195). Additionally, various surface descriptors associated with surface residues have been examined to study protein surfaces. However, since most of this work was carried out on relatively small datasets, one might suspect that the small dataset size influenced the conclusions from previous studies. In this paper, the analysis is carried on a large dataset of $2,383$ protein chains. All protein residues are studied in the context of interface, interior or exterior; Interface residues are studied relative to surface residues and a series of parameters are evaluated in order to obtain discriminants to differentiate interfaces from non-interfaces or surfaces. Specifically, we attempt to answer the following questions: Can interfaces be differentiated from non-interfaces? from surfaces? Which properties are the most useful in distinguishing interfaces from non-interface residues?

The rest of the paper is organized as follows: Section II describes the dataset and residue properties examined in this study. Section III presents the results of our analysis: residue-residue contact preference on the interface; comparing interface, interior and exterior residues based on five properties; comparing interface and non-interface surface residues based on eight properties. Section IV summarizes the results, compares the study to related work and discusses future applications.

4.2 Materials and Methods

4.2.1 Dataset

We extracted protein-protein interface residues from complexes in PDB (14) using the following procedure: The protein entries with resolution $\leq 3\dot{A}$ were then checked with the Protein Quaternary structure file Server (PQS) (72) to regenerate quaternary structures, from which protein dimers were kept, while crystal packing and protein multimers were filtered out. Next, protein dimers with one chain of ≤ 20 amino acids were removed. We selected chains out of the protein dimer complexes such that no two chains share sequence identity $\geq 30\%$. The protein sequence identity information was obtained from the PDB (ftp://ftp.rcsb.org/pub/pdb/derived_data/NR/). The final dataset (PPI2383) includes 2383 protein chains derived from 2316 protein dimers. The dataset consists of 452 heterodimeric and 1931 homodimeric interfaces (Interfaces between chains with $\geq 90\%$ sequence identity are defined as homodimeric interfaces. All others are defined as heterodimeric interfaces.)

4.2.2 Surface versus Non-surface

Surface residues are defined by Miller et al. (122) as those residues having relative accessible surface area $\geq 5\%$. The relative accessible surface area is calculated through the Naccess program (76).

4.2.3 Interface, Exterior and Interior

We follow Ofran and Rost's definition of interface residues: Two residues are considered to be in contact if the closest distance between any two atoms, one from each residue, is less than $6\dot{A}$; A residue having at least one contact residue from the interacting partner chain is considered to be an interface residue (133). Additionally, non-interface residues are classified into two types: exterior residues lie on the surface and interior residues are embedded beneath the surface. So the residues can be divided into three disjoint classes: interface, exterior and interior. Our analysis focuses on interface residues extracted from surface residues as well as the comparison of interface residues to the exterior and inte-

rior residues of a protein.

4.2.4 Interface Propensity

Assume there exists a residue-based property (such as amino acid type, variation entropy or residue roughness etc.) with k discrete values: $(v_1, v_2, ..., v_k)$, accordingly, protein residues are divided into k disjoint subsets $S_1, S_2, ..., S_i, ..., S_k$ in terms of the residue property. Let R_i and r_i be the number of residues and interface residues in the set S_i in respect, then:

$$\begin{aligned} p_i &= \frac{r_i}{\sum_{i=1}^{k} r_i} \\ P_i &= \frac{R_i}{\sum_{i=1}^{k} R_i} \\ IP_i &= \log_2 (\frac{p_i}{P_i}) \end{aligned}$$

IP_i is the interface propensity (IP) of the property at value v_i. IP estimates the interface residues' occurring tendency at a specified property value. IP > 0 denotes that the specified property value is more preferred in the interface than in the protein as a whole. Similarly, IP < 0 denotes that it is less preferred in the interface. The concept of interface propensity can also be extended to exterior and interior residues.

4.2.5 Residue-Residue Contact Preference

Let i and j be two types of amino acid residue, $p(i, j)$ is the probability of contacts between residue i and j and $q(i), q(j)$ is the respective probability of occurrence of residue i, j. The residue-residue contact preference $L(i, j)$ is defined as:

$$L(i,j) = \log_2 \frac{p(i,j)}{q(i)q(j)}$$

4.2.6 Side Chain Orientation

The *side chain orientation* of a residue is defined as the angle between two vectors. The first vector connects the geometrical center of the side chain of the residue with its C_α atom. The second vector connects the geometrical center of the protein chain with the C_α atom of the residue. The angle is confined to the range from 0 to π, within which angles $(0, \frac{\pi}{2})$ and $(\frac{\pi}{2}, \pi)$ correspond to side chains pointing directly inward and outward, respectively.

4.2.7 Surface Roughness

Using Richard's (107) method, a *molecular surface* (A_s) is produced by rolling a solvent sphere with radius R against the target protein. Lewis (110) defined *surface roughness* as follows: $\mathbf{D} = 2 - \frac{\partial \log A_s}{\partial \log R}$. It denotes the degree of irregularity of a surface. Here, each surface residue is assumed to have its own molecular surface and roughness. Roughness is calculated by varying the radius R from $0.2\dot{A}$ to $4.0\dot{A}$, in steps of $0.1\dot{A}$. The molecular surface area A_s is calculated using the Molecule Surface Package (MSP) (36).

4.2.8 Solid Angle

Solid angle, first proposed by Connolly (35) as a measure of the shape of local regions of protein surfaces, is calculated as the fraction of a sphere intersecting the protein when the sphere is centered at a point on the protein surface. The range of a solid angle is $(0, 4\pi)$. A point with solid angle $< 2\pi$ lies on a surface that is *locally convex*. A point with $> 2\pi$ lies on surface that is *locally concave*. The MSP software package implemented by Connolly (36) uses discrete dots to represent the molecule surface and generates a solid angle for each dot. The solid angle of a surface residue is calculated as the average of the solid angles of all the surface dots that belong to the residue. The sphere radius is set as $6\dot{A}$ by default in the computation.

4.2.9 Protrusion-cx Value

Pintar (140) devised a metric called *cx value* to estimate the *protrusion* of protein atoms. The basic idea, similar to that of the solid angle, is to center a sphere at an atom and calculate the ratio of volume

occupied by the protein and the volume left free by the protein. The cx value is a real number between 0 and 15. High cx values correspond to protruding atoms. Here, protrusion is defined over surface residues instead of atoms. A surface residue's protrusion is represented by the cx value of its C_α atom. The cx values are computed using the C++ program provided by Pintar with default parameters (140).

4.2.10 Surface Micro-Environment: Hydrophobicity and Interface Cluster Size

Although some interface residues (dubbed "hot spots") contribute more to the binding affinity than other residues (15), most interface residues are not solitary. Interface residues have a tendency to form clusters on the surface. This tendency is the basis of analysis of interfaces using surface patches or spatial clusters (86; 193; 105). Here, we define a *surface micro-environment* for each surface residue to examine whether residue preferences of interfaces are sensitive to the micro-environment or context in which the residue resides. Given a target residue, its surface micro-environment is defined as the set of surface residues whose C_α atom is $< 7\dot{A}$ away from the C_α of the target residue. By this definition, each residue is included in its own surface micro-environment. Two surface micro-environment parameters are of interest here: the hydrophobicity and the interface cluster size. The hydrophobicity of a target residue is defined as the average hydrophobicity of all the residues in its surface micro-environment, while hydrophobicities of each residue type R_i are denoted with an energy value e_i, which is derived from residue contact energies[1] (193). The residue contact energies represent the degree of hydrophobic force between residue pairs. Hence, e_i can be regarded as an estimation of hydrophobicity: the lower the e_i value, the more hydrophobic the residue. As a result, the average energy e_i denotes the hydrophobicity of the surface micro-environment of the target residue. The interface cluster size refers to the number of interface residues within a target residue's micro-environment. We anticipate a larger cluster size for interface residues.

4.3 Analyses Results

We now proceed to describe the results of our analyses of protein-protein interfaces. Our analyses are aimed at exploring the following questions:

[1]The e_i values of 20 residues are: F -5.12, M -4.91, I -4.88, L -4.65, W -4.36, V -4.17, C -4.00, Y -3.24, A -2.82, H -2.75, G -2.34, T -2.30, P -2.22, R -2.18, S -2.07, Q -1.98, E -1.94, N -1.90, D -1.81, K -1.50

- Are some of the 20 amino acids are over-represented in protein-protein interfaces relative to non-interfaces?
- Are some residue-residue contacts over-represented in protein-protein interfaces relative to non-interfaces?
- Are some types of amino acid residues (e.g., hydrophobic ones) over-represented in interface relative to non-interfaces?
- Are interface residues more conserved (in evolutionary terms) than non-interface residues?
- Are some secondary structures over-represented in protein-protein interfaces relative to non-interfaces?
- Are some side-chain orientations more prevalent in protein-protein interfaces relative to non-interfaces?
- Do molecular surfaces of interfaces differ from those of non-interface residues in terms of their roughness?
- Are convex, concave or flat molecular surfaces over-represented in interfaces relative to non-interfaces?
- Do molecular surfaces of interfaces tend to be more or less protruding relative to non-interfaces?

4.3.1 Residue-Residue Contact Preference

Figure 4.1, the grid map, shows the 2D distribution of residue-residue contact preference. Residues are placed in increasing hydrophobicity order along the x- and y-axis. The residue-residue contact preference is calculated as log odds ratio of the observed frequency of the residue pair over its expected frequency. Each grid represents the contact preference value of two residues from the x- and y-axis and is filled with different colors. The colors, from black to white, represent contact preference from low values to high values. In the grid map, Cys-Cys stands out with the highest preference, known for its formation of the cysteine-bridges. The contact preferences between residues with opposite charges

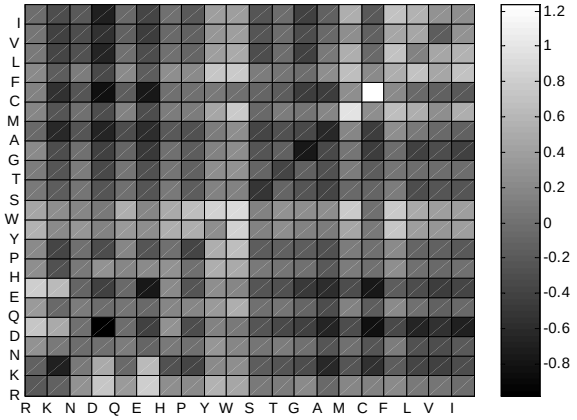

Figure 4.1 Grid map of residue-residue contact preference

Each grid corresponds to a residue-residue contact preference. The filled color measures the extent of preference: white-the most preferred, black-the least preferred. The color bar on the right shows the contact preference values to which the colors correspond. The 20 amino acids are arrayed with increasing hydropathy along the right and up side of the grid map. Cys-Cys outstands as the most preferred contact. A "favorable contact zone" at the top right region and an "unfavorable contact zone" at the top left region are noticeable.

(like Arg-Glu, Arg-Asp, His-Asp and Lys-Asp) are also high, which confirms the earlier findings that interfaces are rich in salt-bridges (118; 161; 73). Contacts consisting of any of Tyr, Trp, Phe and His are generally favored due to aromatic residues' specificity in interfaces. A "favorable zone" exists in the upper-right of the grid map, whose paired components are highly hydrophobic residues like Met, Cys, Phe, Leu, Val and Ile. An "unfavorable zone" is found in the upper-left region and consists primarily of pairs composed of one hydrophobic residue and one hydrophilic residue. The existence of the two regions is consistent with the Glaser's conclusion that contacts between large hydrophobic residues are highly preferred and contacts between pairs of hydrophobic and polar residues are not preferred (62). Keskin investigated the residue contacts at protein-protein interfaces using e^0-"solvent-mediated" potentials and e^r-"residue-mediated" potentials (92). It can be verified that those highly preferred contacts have low values of residue-mediated potential. e.g. e^r(Cys-Cys)=-7.23,e^r(Arg-Glu)=-4.13 and e^r(Tyr-Trp)=-5.12 etc.

4.3.2 Residue Composition and Propensity

The dataset *PPI2383* contains $113,553$ interface residues, $323,270$ exterior residues and $131,632$ interior residues. The occurrence frequencies of each amino acid in the interface, exterior and interior regions are calculated and shown in the y-axis (see Figure 4.2). Along the x-axis, amino acids are placed in order of increasing hydrophobicity based on the Kyte and Doolittle hydropathy index (102). Figure 4.2 shows that more hydrophobic amino acids are over-represented in the protein interiors and in interfaces. This is consistent with the results of earlier analysis by Janet Thornton's group (88). The residue propensities of interface, exterior and interior regions are calculated and compared in Figure 4.3. Note that hydrophobic residues have high propensities in interior regions and low propensities in exterior regions. The hydropathy difference between interior regions and exterior regions is apparent, while the hydropathy of interface regions appears to be neutral between those of interior and exterior regions. Korn et. al (99) and Argos (5) drew similar conclusions concerning the hydropathy of the interface. Furthermore, the calculation of the interface residue propensity out of surface residues (104,789 interface residues and 323,270 exterior residues) in Figure 4.4 reveals that hydrophobic residues like Met, Cys, Phe, Leu, Val and Ile have high affinities for the interface. On the other hand, less hydrophobic residues like Lys, Asn, Asp, Gln and Gly have low affinitities for the interface. Aromatic residues His, Phe, Trp and Tyr all have high propensity values. Argos (5) discovered this rule and postulated that aromatic residues can glue together protein subunits. Charged residues show no consistent affinities: propensities of Arg and His are high, while those of Lys, Asp and Glu are low. Our results corroborate Janet Thorton's (88) analysis, which concluded that interface residues pose as hydrophobic patches of surface residues.

4.3.3 Variation Entropy

The HSSP database (155) provides multiple sequence alignments (MSAs) of all proteins in PDB. The protein homologues in the MSA are selected based on a rigorous sequence identity threshold such that they are also highly likely to be structural homologs. Each residue is assigned a variation entropy, which is calculated based on the occurrence frequency of each amino acid at a given position within the MSA. The variation entropy value of a residue denotes its conservation degree and ranges between 0-

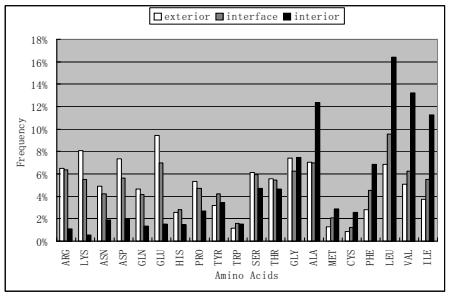

Figure 4.2 Percentage frequencies of amino acid residues in the exterior, interface and interior regions

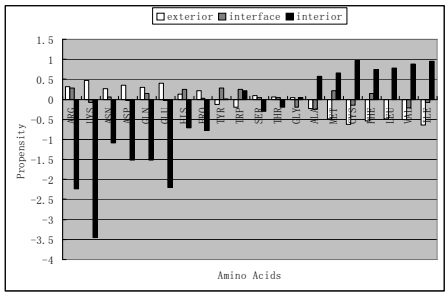

Figure 4.3 Propensities of amino acid residues in the exterior, interface and interior regions

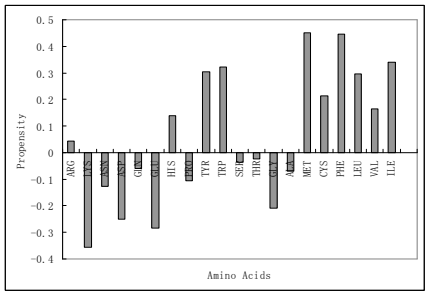

Figure 4.4 Propensities of amino acid residues in the interface region relative to the surface

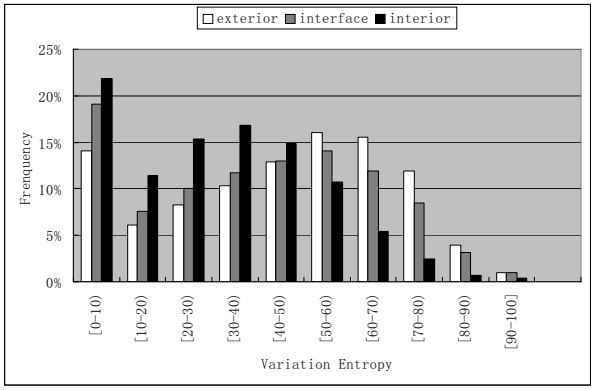

Figure 4.5 Variation entropy distribution

The variation entropy is extracted from the HSSP database (155). Values range at 0-100 and are divided into 10 equal-width zones. The lower variation entropy, the more conserved the residues are.

100. High variation entropy suggests variable residues, while low entropy suggests conserved residues. Figure 4.5 displays the distribution of variation entropy within exterior, interface and interior regions. Variation entropy values between 0 and 100 are placed into 10 equal-sized bins (along the x-axis), The number of residues in each bin, calculated as a percentage of the total number, is plotted along the y-axis. It is evident that interior residues are the most conserved among the three, concentrated in the zones with small variation entropy values. Interface residues are similar to exterior residues. However, as the variation entropy increases along the x-axis, interface residues tend to occupy positions with smaller entropy scores compared to exterior residues. This tendency implies that interface residues are more conserved than exterior residues over evolutionary time. This tendency is also apparent in the propensities of interface residues relative to the surface residues, as depicted in Figure 4.6. This is consistent with Valdar (176)'s amino acid conservation analysis in six homodimer protein families.

4.3.4 Conservation Score

Utilizing the multiple sequence alignments (MSAs) in the HSSP database, Consurf-HSSP database (61) calculates the rate of evolution at each residue position through a phylogenetic tree-based algorithm. The rate of evolution is normalized to a conservation score ranging at 0-9. In contrast to variation entropy of HSSP, high scores correspond to highly conserved residues. In the Figure 4.7, the diagram on the left shows the distribution of conservation scores in exterior, interface and interior

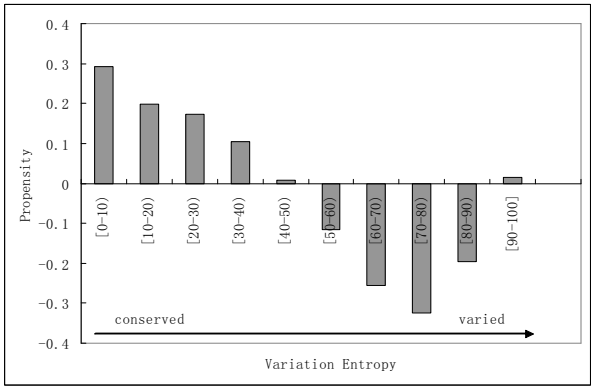

Figure 4.6 Propensities of variation entropy in the interface region relative to the surface

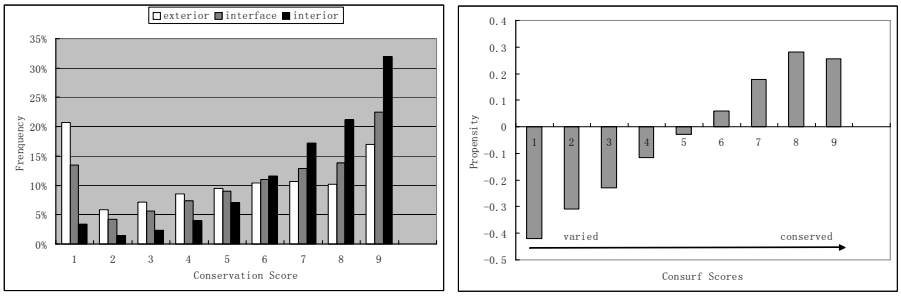

Figure 4.7 Conservation score distributions and propensities of interfaces relative to surfaces

Conservation scores range at 0-9. The higher conservation scores, the more conserved the residues.

regions respectively, and the diagram on the right shows the propensity of interface residues extracted from surface residues. The results, comparable to the variation entropy distribution, lead to the same conclusion that evolutionary conservation can help distinguish interface residues from other surface residues.

4.3.5 Secondary Structure

Protein secondary structure is studied to bridge the gap between the sequence and the structure. Kabsch and Sander (89) defined eight secondary structures: H (alpha helix), B (beta-bridge), E (extended strand), G (3-helix), I (5-helix), T (hydrogen bonded turn), S (bend) and _ (coil). They implemented a program (DSSP) to compute protein secondary structure. Figure 4.8 shows the distribution of

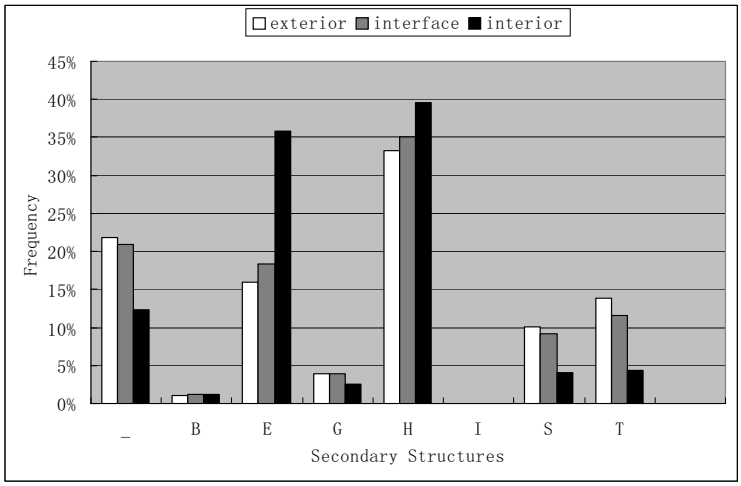

Figure 4.8 Secondary structure distribution

The eight secondary structures: H (alpha helix), B (beta-bridge), E (extended strand), G (3-helix), I (5-helix), T (hydrogen bonded turn), S (bend) and _ (coil) are computed through the program DSSP (89).

the secondary structures in the exterior region, the interface region and the interior region respectively. Interface residues are more similar to exterior residues based on the distribution. H (alpha helix) is the most common configuration in all the three regions. However, the interior region also contains a large amount of E (extended strand) but a small amount of other secondary structures in contrast to exterior and interface regions.

4.3.6 Side Chain Orientation

Rackovsky and Scheraga first proposed side chain orientation as a metric to estimate hydrophobic forces (144). They studied the residue orientations in 13 native proteins and found that polar and non-polar residues have various orientation preferences. Later Yan (187) extended the studies to 48 proteins concerning exposed, interfacial and buried residues. Similar conclusions were reached: side chain orientation has a high correlation with hydrophobicity. Figure 4.9 depicts the distribution of side chain orientations in the exterior, interface and interior regions. Interior residues tend to have small side chain orientations, or point *inward*, which may be a result of the shrinkage of side chains to fit into the protein conformation and minimize the free energy. This conclusion is supported by other metrics such as: amino acid composition, variation entropy, among others. Interface residues are more similar

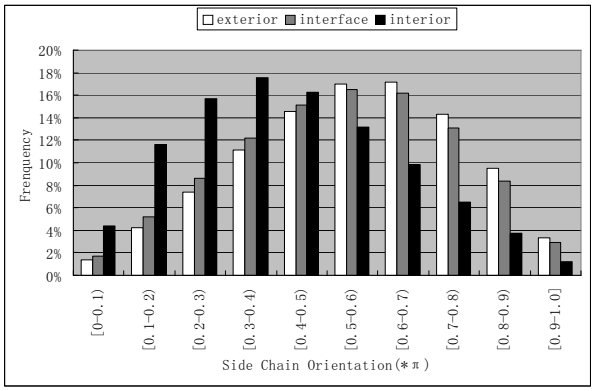

Figure 4.9 Side chain orientation distribution

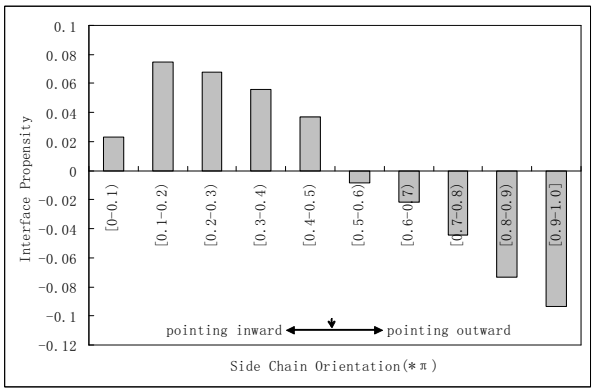

Figure 4.10 Propensities of side chain orientation in the interface region relative to the surface

to exterior residues than interior residues in the side chain orientation distribution. Figure 4.10 shows the side chain orientation propensity of interface residues relative to non-interface surface residues. Interface residues with side chain orientation $< \frac{\pi}{2}$ are overrepresented, implying that interface residues tend to point *inward*. Although the interface residue side chain's tendency to point *inward* is clear, the small propensity values (between -0.1 and 0.1) imply that the tendency is not all that well-pronounced.

4.3.7 Surface Roughness

The difference in surface roughness between interface residues and non-interface surface residues is shown in Figure 4.11. Greater surface roughness values denote a smoother surface of protein residues.

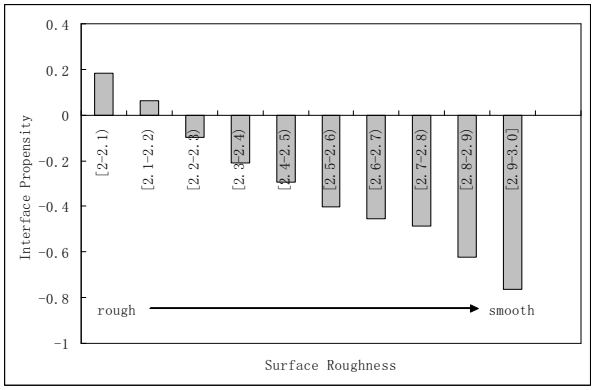

Figure 4.11 Interface propensities of surface roughness

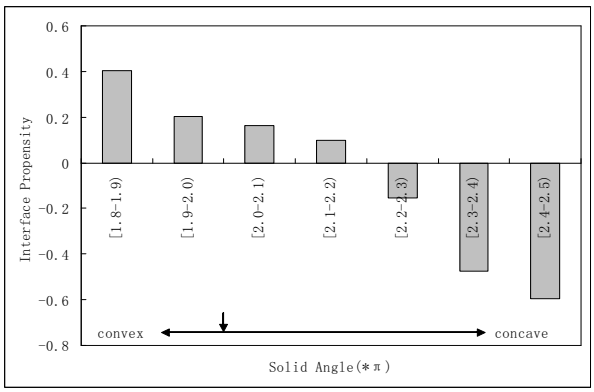

Figure 4.12 Interface propensities of solid angle

The histogram shows that interface residues tend to lie in rougher regions of the surface. The smoother a surface residue, the less likely it is to be an interface residue.

4.3.8 Solid Angle

The difference in solid angle values between the interface and non-interface surface residues is highlighted in Figure 4.12. Solid angles of surface residues mostly lie between 1.8π to 2.5π. Note that the solid angle 2π denotes a "flat" local region, whereas the solid angles $< 2\pi$ and $> 2\pi$ denote "concave" and "convex" local regions respectively. Figure 4.12 shows that interface residues favor moderately concave $(1.8\pi - 2.0\pi)$, flat (2.0π) or moderately convex $(2.0\pi - 2.2\pi)$ local regions but not highly convex regions $(2.2\pi - 2.5\pi)$ or highly concave regions.

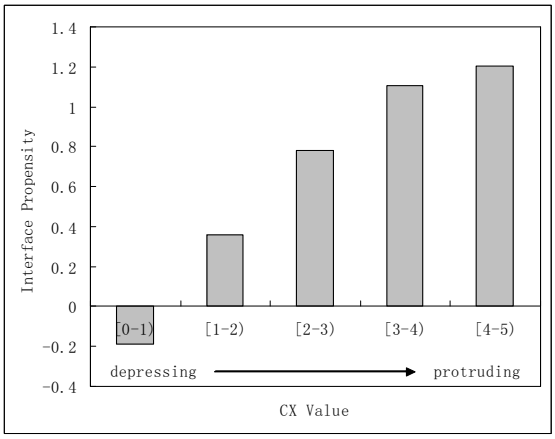

Figure 4.13 Interface propensities of protrusion (cx value)

4.3.9 Protrusion-cx value

Figure 4.13 compares the protrusion in interface and non-interface surface regions. Although the cx values range from 0-15, the cx values of surface residues corresponding to their C_α atoms are concentrated in the range 0-5. Large cx values correspond to protruding atoms. The fact that the propensities increase as the cx values increase suggests that the interface residues prefer to be protruding.

4.3.10 Surface Micro-Environment: Hydrophobicity and Interface Cluster Size

Figure 4.14 shows the propensities of the two parameters related to the surface micro-environment: the hydrophobicity and interface cluster size. The hydrophobicity in the upper figure, estimated through average contact energy, shows that interface residues reside at more hydrophobic environments. The lower figure demonstrates that an interface residue tends to be clustered with three or more interface residues on the protein surface.

4.4 Discussion

4.4.1 A Summary of Protein-Protein Dimeric Interfaces

Several physicochemical properties are studied on a big dataset to characterize the protein-protein dimeric interfaces. Our analysis has shown that:

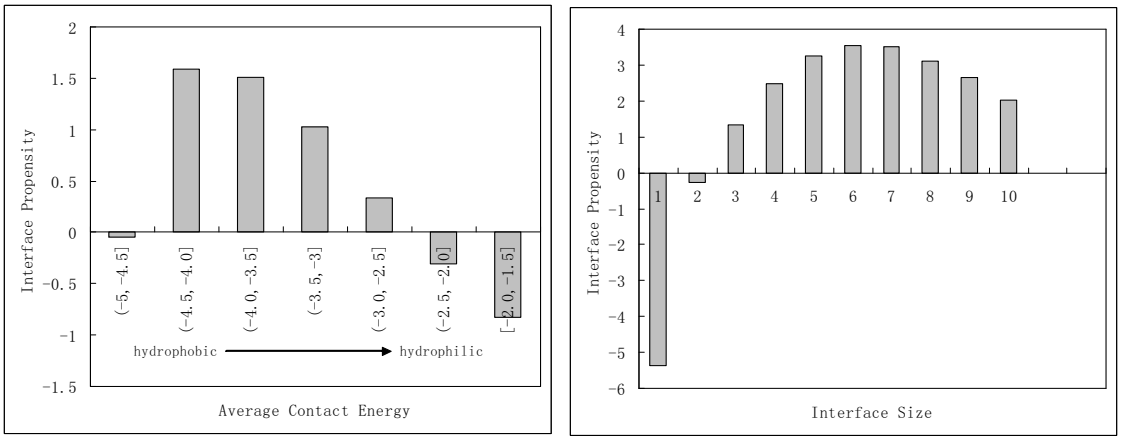

Figure 4.14 Interface propensities of hydrophobicity (average contact energy) and size of interface cluster.

- Interfaces consist of many cysteine-bridges (Cys-Cys) and salt-bridges (Arg-Glu, Arg-Asp, His-Asp and Lys-Asp)
- Aromatic residues (Tyr, Trp, Phe and His) are favored in interfaces.
- Hydrophobic-hydrophobic residue pairs overtake hydrophobic-hydrophilic residue pairs in interfaces.

Our analysis of protein-protein interfaces with respect to amino acid composition, secondary structure, variation entropy, conservation score, side chain orientation has shown that:

- Interior residues tend to be the most hydrophobic, exterior residues tend to be the least hydrophobic and interface residues falling between the two extremes.
- Compared to the exterior and interface residues, interior residues have the most extended strand secondary structures.
- Based on variation entropy and conservation score (derived from multiple sequence alignment and phylogenetic tree construction respectively), interior residues tend to be among the most conserved, the exterior residues among the least conserved, with the interface residues falling between the two extremes.

- Side chains of interior residues tend to point inward, while those of exterior residues tend to point outward, and those of interface residues tend to have orientations between the two extremes, but closer to those of the exterior residues.

Our analysis of interface regions of protein structures with respect to variation entropy, conservation score, side chain orientation, surface roughness, solid angle, cx value, hydrophobicity and interface cluster size has shown that:

- Side chains of interface residues tend to point inward.
- Interfaces tend to be rougher than the rest of the protein surface.
- Interfaces tend to be moderately concave, flat or moderately convex but not highly convex or concave (as measured by the solid angle).
- The C_α atoms of interface residues protrude more in terms of cx value.
- Interface residues often reside in a hydrophobic micro-environment.
- Interface residues are often clustered on the surface.

The results of our analyses of a large dataset of protein-protein interface residues show that the interface residues have distinctive physico-chemical properties in contrast to non-interface residues. Our results confirm that protein-protein interfaces usually have a hydrophobic core with polar residues and water molecules scattered around (106; 136), and are conserved across protein families (114) and that the residues tend to interact through hydrogen bonding and electrostatic forces (78; 85). However, the distinction between interface residues and interior and exterior residues is not very crisp. This presents some challenges in reliable prediction of interface residues from amino acid sequences. In this context, it might be useful to focus on specific sub-categories of interfaces (133), or "hot spots" (15) or on distinguishing interface residues from surface resides (86; 87). However, the focus on surface residues requires the knowledge of the structure of the target protein, or, at the very least, reliable prediction of surface residues from sequence.

4.4.2 Comparison with Previous Studies

Ofran et al. (133) investigated residue-residues contact preferences and amino acid compositions; Thornton et al. (88; 85) studied protein-protein interfaces, interior regions and exterior regions and later focused on surface interfaces. Yan et al. (191) explored physical-chemical properties of protein-protein interfaces on a large dataset extracted from PDB. Connolly et al. (35), Lewis et al. (110), Pintar et al. (140) and Yan et al. (187) focused on one specified structural property of protein-protein interfaces. Most analyses were performed on relatively small datasets (e.g. two datasets of Thornton et al. (88; 85) are 32 and 54 protein-protein dimers respectively.) Among the studies that used large datasets, Ofran et al. (133) limited their studies to sequence residues; Lewis et al. (110) to surface roughness; Yan et al. (191) did not consider structural properties. Against this background, the results presented in this paper are based on comprehensive analyses of protein-protein interfaces using a large dataset of binary protein-protein interfaces associated with $2,383$ non-redundant protein chains.

4.4.3 The Influence of Different Interface Definitions

Our definition of interfaces, adapted to Ofran et al's method (133), is based on the closest distance between atoms. While Thorton's definition of interface, at the residue level, is based on ΔASA: a residue is regarded as an interface residue when the loss of solvent accessible area (ΔASA) is $\geq 1\dot{A}$ during the protein's complexation. The numbers of interface versus non-interface in the dataset, PPI2383 based on the two definitions are listed in table 4.1. The two definitions have 96.97% of the assignments in common. However, 14.63% of the interfaces defined by distance are categorized as non-interfaces on the basis of ΔASA.

Table 4.1 Comparison of interface definitions: ΔASA-based and distance-based

	interface by ΔASA	non-interface by ΔASA
interface by distance	96,933	16,615
non-interface by distance	629	454,288

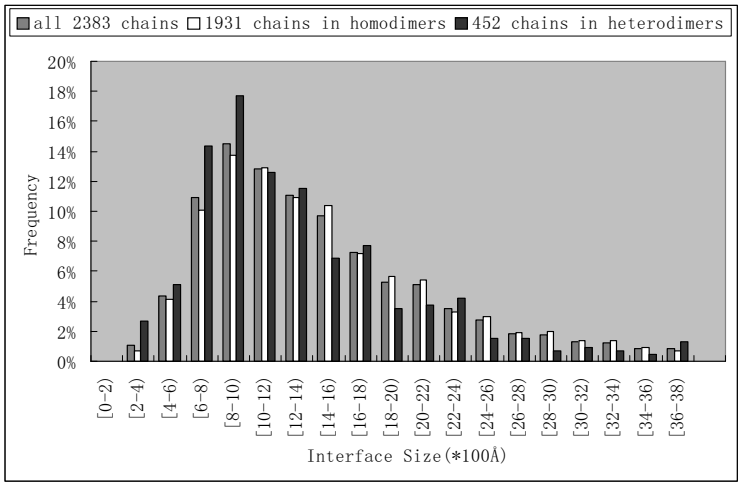

Figure 4.15 Interface size distribution

4.4.4 The Distribution of Interface Size

We repeated Thorton's analysis of the interface size of protein subunits. Interface size is calculated as the loss of solvent accessible area (ΔASA) during the process of protein complexation. Figure 4.15 shows the interface size distribution of the dataset PPI2383, $1,931$ chains of which are in homodimer complexes and 452 chains in heterodimer complexes. Fewer than 1.0% of the interfaces have size less than $400\dot{A}^2$. This is due to the fact that PQS (72) discriminates crystal packing from interfaces using a cutoff of $400\dot{A}^2$ mean accessible area loss. The interface sizes range from $200\dot{A}^2$ up to $4000\dot{A}^2$, 81.0% of which fall into the range $400 - 2200\dot{A}^2$ and the peak is at $800 - 1000\dot{A}^2$. Comparison of the the interface sizes of heterodimers and homodimers shows that heterodimers tend to have interface sizes around $600 - 1400\dot{A}^2$ with a sharp peak in the $800 - 1000\dot{A}^2$ range, whereas the interface sizes of homodimers tend to be more evenly distributed.

4.4.5 Cutoff of Surface Definition

Protein-protein interface residues are defined in terms of atom-atom distance. Protein surface residues are defined in terms of relative solvent accessible area (RASA). The RASA distribution is shown in Figure 4.16. As can be seen from the figure, a large fraction of non-interface residues (up to

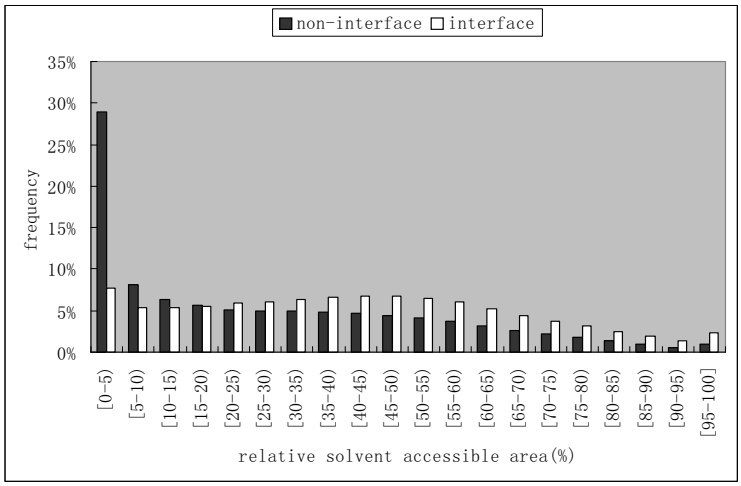

Figure 4.16 Relative solvent accessible area distribution

28.94%) have RASA $\leq 5\%$; whereas a relatively small fraction (7.72%) of the interface residue with RASA $\leq 5\%$. It appears that more interface residues are included among the surface residues when $5\% - 20\%$ range for RASA is used to identify the surface residues. The optimal choice of RASA cutoffs is unclear; and the interface residues that are classified as non-surface residues deserve careful examination.

4.4.6 Application: A Case Study

The distinct characteristics of residues with high versus low interface propensities, with respect to the structural characteristics analyzed above, suggests the possibility that these characteristics might be useful in improving the performance of classifiers trained to predict protein-protein interface residues. Although physicochemical properties of amino acids have been widely used, only a few studies have attempted to exploit geometric features of protein interfaces in building classifiers for predicting protein-protein interface residues (20; 126; 124; 165; 195). To explore the potential utility of such an approach, we examined the transcriptional regulatory protein SlyA (pdb entry 1lj9) by combining the five structural properties using a simple voting method to identify the interface residues of chain B: For each surface residue, we calculated its side chain orientation, surface roughness, solid angle, cx-value and hydrophobicity. If the value of a property lies in the region where the value is preferred in the interface

(based on propensity estimates), the surface residue is "voted" to be an interface residue based on that property. If a surface residue is voted to be as an interface residue based on at least 3 of the 5 properties, it is predicted to be an interface residue; otherwise, it is predicted to be a non-interface residue. We then use the tendency of the interface residues to be clustered together on the protein surface to refine the predictions of the voting method as follows: If a surface residue that is predicted to be an interface residue by the method described above has ≤ 2 neighbors in its surface micro-environment that are also predicted to be interface residues, it is reclassified as a non-interface residue; If a residue predicted to be non-interface residue has ≥ 4 neighbors in its surface micro-environment that are predicted to be interface residues, it is reclassified as an interface residue.

Let *TP* be the number of true positives (residues predicted to be interface residues that are actually interface residues); *FP* the number of false positives (residues predicted to be interface residues that are actually non-interface residues); *TN* the number of true negatives; *FN* the number of false negatives. The numerical performance measures *ac* (accuracy), *re* (recall), *pr* (precision) and *cc* (correlation coefficient) are defined as follows:

$$
\begin{aligned}
ac &= \frac{TP + TN}{TP + FP + TN + FN} \\
re &= \frac{TP}{TP + FN} \\
pr &= \frac{TP}{TP + FP} \\
cc &= \frac{TP * TN - FN * FP}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}
\end{aligned}
$$

The results of prediction using the voting method followed by the refinement strategy described above are summarized in Table 5.1 and Figure 4.17. We see that the use of the five structural properties results in fairly accurate prediction of the interface residues. The results also suggest that refining the predictions based on the clustering tendency of the interface residues further improves the quality of the predictions in terms of precision and recall. It is worth noting that the results are significantly better than those obtained based on analysis of sequence neighbors of the target residues (precision=55%, recall=53%). These results suggest the possibility of using structural properties of interfaces to reliably

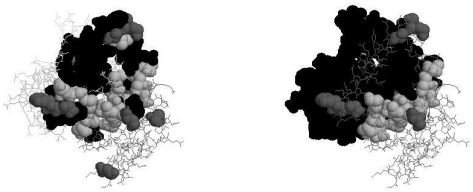

Figure 4.17 Interaction Sites Identification of Chain B of Protein *1lj9* Under Two Approaches: voting method (the left) and voting method+refinement strategy (the right). The chain B is shown in grey, with the residues of interest shown in space fill and color coded as follows: black, interface residues identified as such by the classifier (TPs); light grey, interface residues missed by the classifiers (FNs); and dark grey, residues incorrectly classified as interface residues (FPs). For clarity, interface residues for the chain A (gray wireframe) are not shown. The structure diagrams were generated using RasMol (156).

identify protein-protein interface residues when only the structure of a protein (but not that of protein-protein complex(es) in which it participates) is available.

Table 4.2 prediction results: chain B of protein 1lj9

classifiers	*ac*	*re*	*pr*	*cc*
voting	76%	66%	82%	53%
voting+refinement	82%	77%	86%	64%

4.4.7 Conclusion

Our analyses of dimeric protein-protein interfaces shows that protein-protein interfaces differ from protein interiors and protein surfaces with respect to several sequence and structure derived features. However, none of the sequence-or structure-derived features carries a strong enough signal to allow reliable prediction of protein-protein interface residues. This underscores the need for developing sophisticated machine learning methods that can discover sequence and structural correlates of protein-protein interfaces. Our analyses suggest that the use of structural features (when structure of the target protein is available) can improve the reliability of predicted interfaces.

4.5 Acknowledgements

This research was supported in part by a grant from the National Institutes of Health (GM066387) to Vasant Honavar.

CHAPTER 5. NB_PPIPS - A Naive Bayes Method to Predict Protein-Protein Interaction Sites

A paper submitted to IEEE Transactions on Bioinformatics and Computational Biology

Feihong Wu, Fadi Towfic, Drena Dobbs, Robert Jernigan and Vasant Honavar

Abstract Reliable identification of protein-protein interaction sites in proteins is one of the major challenges in computational biology. We describe a machine learning approach to predicting protein-protein interface residues on the surface of a protein. The proposed method is useful in settings where the structure of the target protein is available although the structures of its binding partners or the complexe(s) formed by it with other protein(s) are unavailable. We use a large non-redundant dataset of $2,383$ proteins. We explore several alternative representations of the input to the classifier starting with a sequence window consisting of the target residue and its sequence neighbors. We compare the performance of the resulting classifier estimated using 10-fold cross-validation with that of classifiers that utilize evolutionary and structure-derived features. Our results show that the Naive Bayes classifier trained using a combination of sequence, structure, and evolutionary information substantially outperforms its counterpart that is trained using sequence information alone. We find that some interfaces are easy-to-predict and others hard-to-predict. In the case of the former, interface predictions with high precision and recall can be achieved using sequence information alone whereas in the case of the latter, even addition of structural and evolutionary features yields only marginal improvement in predictions. We explore the feasibility of adapting this approach to settings where the structure of the target protein is unavailable. We model the structure of the target protein using its known structural homologues and use the modeled structure instead of the actual structure to determine the surface residues and to calculate the structural features to be used as inputs to the classifier. Our results show that the closer the modeled structure is to the actual structure, the more reliable are the interface residue predictions gen-

erated by the classifier. The methods have been implemented in an online server NB_PPIPS, accessible at `http://watson.cs.iastate.edu/nb_ppips`.

5.1 Introduction

Protein-protein interactions play a pivotal role in cellular processes such as DNA replication and transcription, RNA splicing, signal transduction and metabolic networks. Therefore, understanding the sequence and structural determinants of protein-protein interactions is crutical for the determination of protein functions, understanding biological processes and designing therapeutic drugs. Protein-protein interactions play a pivotal role in protein function. Completion of many genomes is being followed rapidly by large-scale efforts to identify interacting protein pairs experimentally, in order to decipher the networks of interacting proteins. Experimental proteomics projects have already resulted in complete 'interactomes' (56). While such efforts yield a catalog of interacting proteins, experimental detection of residues in protein-protein interaction surfaces must come from determination of the structure of protein-protein complexes. However, determination of protein-complex structures using X-ray and NMR methods lags far behind the number of known protein sequences. Hence, there is a need for the development of reliable computational methods for identifying protein-protein interface residues (169; 164; 165).

When the structure of a target protein and its putative binding partner are both available, docking methods (131; 166; 67; 121) can be used to identify interaction sites. However, such an approach is impractical when the structure of the putative binding partner is unknown. Even in cases where the structure of the binding partner is available, the computational effort involved in docking can be significantly reduced if reliable predictions of binding sites can be used to guide docking. Hence, there is an urgent need for reliable computational methods for predicting protein-protein interface residues from amino acid sequence of the target protein, and when available, its structure, but not the structure of its binding partners.

Studies of "hot spot" regions (93; 94; 15; 189; 188) suggest that interaction sites are pre-organized

in the unbound states. This raises the possibility of identifying the interface residues of a protein without knowledge of the structures of its binding partners, or for that matter, the identities of its binding partners. Analyses of protein-protein complexes from the Protein DataBank (14) (31; 79; 81; 5; 133; 176; 88; 85; 86; 40; 129; 127; 193) have shown that protein-protein interface residues differ from non-interface residues with respect to a number of characteristics. These studies covered a wide scope over a variety of interface types (homo-interface or hetereo-interface, transient or permanent interfaces, etc.) and properties (sequential properties, evolutionary properties and structural properties). Despite the differences among different types of interfaces (133; 86), interfaces tend to be more hydrophobic, more conserved and more protruding as compared to non-interface residues (31; 79; 81; 5; 86).

Several groups have attacked the problem of predicting protein-protein interface residues using only the sequence features of the target protein (134; 58). However, there is much room for improvement in the sensitivity and specificity of such methods. Consequently, recent studies have focused on incorporating additional types of information e.g., the degree of conservation of residues in sequence and/or structure, structural features of the target protein, among others: Fariselli and colleagues (48) and Zhou (194; 28) have developed neural network classifiers to identify heterodimer protein interfaces using spatial neighbors of the target residue as inputs to the classifier. Koike et al. (98) have developed a support vector machine classifier using an input representation similar to that used by Zhou et al. (194; 28). Besides discriminative models like neural network and SVM, generative models such as hidden Markov model(by Friedrich (55), by Nguyen (128)) and conditional random field (by Li (123)) are also applied. Lichtarge et al. (114; 115; 148) devised an evolution trace (ET) method that assigns a score to each sequence residue based on phylogenetic analysis and uses the resulting score to predict protein-protein interface residues. Landgraf (105) used ET scores of clusters of residues in 3-dimensional structure of a target protein to predict protein-protein interface residues. Li (111) and Wang (180) have also used conservation scores of residues in predicting interface residues. Bordner et al. (16) have developed SVM classifiers for predicting protein-protein interface using sequence profiles and evolutionary rates of spatial neighbors of a target residue. Jones et al. (87; 124) have developed a method for predicting four types of protein interfaces (homodimers, small and large protomers from

hetero-complexes and antigens) using *patch analysis* of surface residues. Neuvirth et al. (127) have described a structure-based method for identifying interacting sites by assigning scores to residues in terms of thirteen different properties. Bradford et al. (19; 20) have proposed a method to combine the SVM and patch analysis, followed by a Bayes network for combining the predictions. Kufareva et al. (100) have recently presented a PIER algorithm that scored surface patches using twelve patch descriptors defined by atom properties and reflected the surface patch scores back to surface residues. Sen et al. (158) reported improvements in predicting hydrolase-inhibitor interfaces by combining several methods. Hoskins and colleagues (74) have described the use of abnormal exposed secondary structure in the prediction of protein-protein interfaces. Liang et al. (112; 113) have explored a combination of side chain energy, conservation score and residue propensity in predicting protein-protein interfaces. Dong et al. (43) have introduced profile-level interface propensity to predict protein-protein interfaces. Porollo and colleagues (141) have applied the difference of accessible area between monomeric and oligmeric states of a protein in protein-protein interface prediction. Zhou et al. (195; 143) have recently constructed a meta server by combining results from various predictors.

Our previous work of protein-protein interface analysis (185) studied six physicochemical residue properties (variation entropy, side chain orientation, surface roughness, solid angle, cx value, hydrophobicity) in a large dataset consisting of $2,383$ protein chains. The results of our analysis show that the interface residues have side chains pointing inwards; interfaces are rougher, tend to be flat or moderately convex or concave (but not highly convex or concave) and protrude more relative to non-interface surface residues. Interface residues tend to be surrounded by hydrophobic neighbors.

Against this background, this paper explores whether it is possible to exploit such properties to improve the sensitivity and specificity of protein-protein interface predictions. We describe a machine learning approach to predicting protein-protein interface residues on the surface of a protein. The proposed method is useful in settings where the structure of the target protein is available although the structures of its binding partners or the complexe(s) formed by it with other protein(s) are unavailable. We use a large non-redundant dataset of $2,383$ proteins. We explore several alternative representations

of the input to the classifier starting with a sequence window consisting of the target residue and its sequence neighbors. We compare the performance of the resulting classifier estimated using 10-fold cross-validation with that of classifiers that utilize evolutionary and structure-derived features. Our results show that the Naive Bayes classifier trained using a combination of sequence, structure, and evolutionary information substantially outperforms its counterpart that is trained using sequence information alone. We explore the feasibility of adapting this approach to settings where the structure of the target protein is unavailable. We model the structure of the target protein using its known structural homologues and use the modeled structure instead of the actual structure to determine the surface residues and to calculate the structural features to be used as inputs to the classifier. Our results show that the closer the modeled structure is to the actual structure, the more reliable are the interface residue predictions generated by the classifier. The methods have been implemented in an online server NB_PPIPS accessible at `http://watson.cs.iastate.edu/nb_ppips`.

5.2 Materials and Methods

5.2.1 Dataset

We extracted protein-protein interface residues from complexes in PDB (14) using the following procedure: The protein entries with resolution $\leq 3\dot{A}$ were then checked with the Protein Quaternary structure file Server (PQS) (72) to regenerate quaternary structures from which protein dimers are kept, while crystal packing and protein multimers are filtered out. Next, protein dimers with one chain of ≤ 20 amino acids were removed. We selected chains out of the protein dimer complexes such that any two chains would have sequence identity $\leq 30\%$. The protein sequence identity information is obtained from PDB (ftp://ftp.rcsb.org/pub/pdb/derived data/NR/). The final dataset includes $2,383$ protein chains coming from $2,316$ protein dimers. The dataset consists of 452 heterodimer interfaces and 1931 homodimer interfaces. (Homodimer interfaces are distinguished from heterodimer interfaces through sequence identity. Interfaces between chains with $\geq 90\%$ sequence identity are defined as homodimer interfaces. All others are defined as heterodimer interfaces.)

5.2.2 Surface versus Non-surface

Surface residues are defined by Miller et al (122) as those residues having a relatively accessible surface area of $\geq 5\%$. The accessible surface area is calculated using the Naccess program (76).

5.2.3 Interface versus Non-interface

We follow Ofran and Rost's (133) definition of interface residues: Two residues are considered to be in contact if the closest distance between any two atoms, one from each residue, is $\leq 6\dot{A}$. A surface residue having at least one contact residue from the interacting partner chain is considered to be an interface residue, otherwise it is a non-interface residue. Hence, the dataset consists of $104,789$ interface residues and $323,270$ non-interface residues.

5.2.4 Variation Entropy

The HSSP database (155) provides multiple sequence alignments(MSAs) of all proteins in PDB. The protein homologues in the MSA are selected based on a rigorous sequence identity threshold so that they are also structure homologues. Each residue is assigned a variation entropy, which is calculated based on the amino acid occurring frequencies at its position within the MSA. The variation entropy value of a residue denotes its conservation degree and ranges between 0-100. High variation entropy suggests variable residues, while low entropy suggests conserved residues. Our previous study (185) showed that interface residues are more conserved than surface residues: they are over-represented at 0-50 of variation entropy.

5.2.5 Side Chain Orientation

The *side chain orientation* of a residue is defined as the angle between two vectors: The first vector connects the geometrical center of a side chain of the residue with its C^{α} atom. The second vector connects the geometrical center of the protein chain with the C^{α} atom of the residue. The angle is confined in the range of 0 to π. Our previous study (185) shows that the side chains of interface residues prefer to point inward, having side chain orientation $< \frac{\pi}{2}$.

5.2.6 Surface Roughness

Using Richard's (107) method, a *molecule surface* is produced by rolling a solvent sphere with radius R against the target protein. The area of the resulting surface, A_s, depends on R. Lewis (110) defined *surface roughness* as follows $\mathbf{D} = 2 - \frac{\partial \log A_s}{\partial \log R}$. It denotes the degree of irregularity of a surface. Here, each surface residue is assumed to have its own molecule surface and roughness. Roughness is calculated by varying the radius R from $0.2\dot{A}$ to $4.0\dot{A}$, in steps of $0.1\dot{A}$. The molecule surface area A_s is calculated using the Molecule Surface Package (MSP) (36). Our previous study (185) shows that interface residues have rougher molecular surfaces with surface roughness values within $2.0 - 2.2$.

5.2.7 Solid Angle

Solid angle, first proposed by Connolly (35) as a measure of gross shape of local regions of protein surfaces, is calculated as the fraction of a sphere intersecting the protein by centering a sphere at a point on the protein surface. The range of a solid angle is $(0, 4\pi)$.The MSP software package (36) implemented by Connolly uses discrete dots to represent the molecule surface and generates a solid angle for each dot. The solid angle of a surface residue is calculated as the average of the solid angles of all the surface dots that belong to the residue. The sphere radius is set as $6\dot{A}$ by default in the computation. Our previous study (185) shows that interface residues favor moderately concave $(1.8\pi - 2.0\pi)$, flat (2.0π) or moderately convex $(2.0\pi - 2.2\pi)$ local regions.

5.2.8 Protrusion-cx Value

Pintar (140) devised a metric called *cx value* to estimate the *protrusion* of protein atoms. The basic idea, similar to that of the solid angle, is to center a sphere at an atom and calculate the ratio of volume occupied by the protein and the volume left free by the protein. The cx value is a real number between 0 and 15. High cx values correspond to protruding atoms. Here, protrusion is defined over surface residues instead of atoms. A surface residue's protrusion is represented by the cx value of its C^{α} atom. The cx values are computed by the C++ program provided by Pintar with default settings of parameters. Our previous study (185) shows that interface residues have protruding C_{α} atoms because cx values concentrate in the range $1 - 5$.

5.2.9 Surface Micro-Environment: Hydrophobicity

We define a *surface micro-environment* for each surface residue to examine whether the residue preferences of interfaces are sensitive to the micro-environment or the context in which the residue appears. Given a target residue, its surface micro-environment is defined as the set of surface residues whose C^{α} atom is $< 7\dot{A}$ away from the C^{α} of the target residue. By this definition, each residue is included in its own surface micro-environment. The hydrophobicity of a target residue is defined as the average hydrophobicity of all the residues in its surface micro-environment, while the hydrophobicity of each residue type R_i is denoted with an energy value e_i derived from residue contact energies[1] (193). The residue contact energies designate the degree of hydrophobic force between residue pairs. Hence, e_i can be regarded as an estimation of hydrophobicity: the smaller e_i, the more hydrophobic the residue. As a result, the average energy e_i denotes the hydrophobicity of the surface micro-environment of the target residue. Our previous study (185) shows that interface residues primarily reside in more hydrophobic environments (-4.5 -2.5).

5.2.10 Naive Bayes Classifier

A framework of classification is to classify each instance with d-dimension feature vector $\mathbf{x} = (x_1, x_2, ..., x_d)$ into one of e classes $(c_1, c_2, ..., c_k, ...c_e)$. Let $P(c_k|\mathbf{x})$ be the conditional probability that an instance belongs to class c_k given we know it has feature vector $\mathbf{x}$. According to *Bayes' rule* (45), the conditional probability $P(c_k|\mathbf{x})$ can be computed from the conditional probabilities of occurrence of particular vectors of feature values given each class $P(\mathbf{x}|c_k)$ and unconditional probability of occurrence of each class $P(c_k)$ as follows:

$$P(c_k|\mathbf{x}) \quad = \quad P(c_k) \times \frac{P(\mathbf{x}|c_k)}{P(\mathbf{x})}$$

[1]The e_i values of 20 residues are: F -5.12, M -4.91, I -4.88, L -4.65, W -4.36, V -4.17, C -4.00, Y -3.24, A -2.82, H -2.75, G -2.34, T -2.30, P -2.22, R -2.18, S -2.07, Q -1.98, E -1.94, N -1.90, D -1.81, K -1.50

If we can estimate $P(c_k|\mathbf{x})$ for a classification problem, an instance with feature vector **x** can be assigned with the class c_k which $P(c_k|\mathbf{x})$ is highest to minimize the classification errors.

Naive Bayes classifier is derived from *Bayes' rule* and assumes that occurrence of all features are independent. Hence we have:

$$P(\mathbf{x}|c_k) = \prod_{j=1}^{d} P(x_j|c_k)$$

Thus, our goal in interface prediction is to classify a residue r_0 as an interaction site (c_1) or non-interaction site (c_0) based on various features. For example, considering five sequence neighbors $r_{-5}, r_{-4}, .., r_{-1}, r_1, .., r_4, r_5$ on either side of the target residues r_0, we get a 11-dimensional feature vector $\mathbf{x} = (x_{-5}, x_{-4}, ..., x_4, x_5)$ where each x_i denotes the amino acid identity of the corresponding residue. Assuming that the 11 residues in the sequence window are independent given the class, we have a Naive Bayes classifier that uses only the sequence information:

$$\frac{P(c_1|\mathbf{x})}{P(c_0|\mathbf{x})} = \frac{P(c_1) \times \prod_{j=-5}^{5} P(x_j|c_0)}{P(c_0) \times \prod_{j=-5}^{5} P(x_j|c_1)}$$

If $\dfrac{P(c_1|\mathbf{x})}{P(c_0|\mathbf{x})} > 1, r_0$ is an interaction site;

Otherwise , r_0 is non-interaction site.

To improve the baseline classifier, we add evolutionary and structural features to the feature vector **x**. We consider seven features (variation entropy, side chain orientation, surface roughness, solid angle, cx value and spatial neighborhood hydrophobicity) of the target residue r_0. Features are discretized into bins of equal width. For example, values of spatial neighborhood hydrophobicity fall in the range (-5, -1.5). They are discretized into seven bins of equal width (-5,-4.5], (-4.5,-4], ..., (-2.0,-1.5) and numeric value for the feature is assigned to the appropriate bin. The Naive Bayes classifier that combines multiple types of information is constructed using a greedy strategy as follows: Structural and evolutionary features are considered one at a time, in the order of decreasing impact, keeping the feature if the classifier performance improves and discarding the feature otherwise.

We trained the Naive Bayes classifiers on the dataset using ten-fold sequence-based cross validation (25): the $2,383$ protein chains were partitioned into 10 disjoint subsets. In each round, 9 subsets were used to generate the training dataset and the remaining subset was used to generate the test dataset. The Naive Bayes classifier was trained using the training dataset and evaluated on the test dataset. The performance estimates (accuracy, precision, recall, etc.) are averaged over the 10 runs. To assess the performance of our Naive Bayes classifiers, we use the following measures (11):

Let **TP** be the number of true positives (residues predicted to be interaction residues that are in fact interaction residues); **FP** the number of false positives (residues predicted to be interaction residues that are actually non-interaction residues); **TN** the number of true negatives; **FN** the number of false negatives. We then have performance measures *ac*(accuracy), *re*(recall), *pr*(precision), *tpr* (true positive rate) and *fpr* (false positive rate)and *cc*(correlation coefficient) defined as follows:

$$
\begin{aligned}
ac &= \frac{\mathbf{TP}+\mathbf{TN}}{\mathbf{TP}+\mathbf{FP}+\mathbf{TN}+\mathbf{FN}} \\
re &= \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}} \\
pr &= \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FP}} \\
tpr &= \frac{\mathbf{TP}}{\mathbf{TP}+\mathbf{FN}} \\
fpr &= \frac{\mathbf{FP}}{\mathbf{TN}+\mathbf{FP}} \\
cc &= \frac{\mathbf{TP}*\mathbf{TN}-\mathbf{FN}*\mathbf{FP}}{\sqrt{(\mathbf{TP}+\mathbf{FN})(\mathbf{TN}+\mathbf{FP})(\mathbf{TP}+\mathbf{FP})(\mathbf{TN}+\mathbf{FN})}}
\end{aligned}
$$

None of these measures individually gives a comprehensive picture of the performance of the classifier. This is especially true in settings where the class distribution is unbalanced (as is the case in interface prediction: the proportion of interface residues is much smaller than that of non-interface residues). It is possible to trade off true positive rate against false positive rate or alternatively, precision against recall by varying the classification threshold. The ROC curve (50) and the precision-recall curve show the tradeoff between the false positive rate and true positive rate and between precision and recall (respectively) of classifiers.

5.2.11 Homology-based Structure Modeling

We use the FUGUE (162) program to identify the distant homologues of a target protein to use as structural templates. We selected 10% (209) of the proteins in the dataset randomly as target proteins. We discard structure templates with sequence identity $> \%95$ to ensure that FUGUE does not include the target protein itself with its distant homologues and the condition that normalized z-score $>= 6.0$ is enforced to retain only structure templates with high confidence. We then provide the multiple sequence alignment of the target sequence with its homologues to the MODELLER (117) program. MODELLER automatically calculates a modeled protein structure based on the constraints imposed by the sequence alignment and the structure templates. Two parameters are calculated to determine how well the modeled structures approximate the actual structures. Surface equivalence (SE) is defined as the ratio of "equivalent" residues of the modeled structure to the actual structure (two corresponding residues are equivalent if they are both embedded or on the surface). SE is a coarse metric to estimate the surface similarity of two proteins. RMSD, calculated with BioShell package (65), is the root mean square deviation of the modeled structure and the actual structure, denoting the structure similarity of two proteins.

5.2.12 Predicting Interfaces on Modeled Structures

As noted above, a random sample of 209 proteins was selected and set aside from the original dataset of 2383 proteins. We trained a Naive Bayes classifier using the best performing combination of sequence, structural features on the remaining set of 2174 proteins. The performance of the trained classifier was evaluated on the 209 modeled proteins, by using the modeled structures instead of the actual structures to generate the input to the classifier.

5.3 Experimental Results

5.3.1 Structural and Evolutionary Features Improve Interface Prediction

The results of our experiments are shown in Table 5.1 and Figure 5.3. The classifier that uses the target residue and its sequence neighbors provides the baseline. Various structural features are added to explore the impact each feature on the performance of the resulting classifier. The results of our experiments show that almost all the structural and evolutionary features considered improve the accuracy of the classifier over that of the baseline classifier, with cx value yielding the largest improvement, followed by hydrophobicity, solid angle, roughness and variation entropy. Side chain orientation did not yield an improvement in performance over the baseline. The Naive Bayes classifier that was constructed using a combination of sequence, structural and evolutionary features shown in the last line of Table 5.1, incorporates cx value, hydrophobicity, solid angle, variation entropy. It has 60.7% recall and 34.6% precision, as compared to 56.2% and 29.3% respectively for the baseline classifier. The observed difference is statistically significant based on paired t-tests which yield p values less than 0.0001. The precision-recall curve (see Figure 5.3) of the classifier that exploits sequence and structural as well as evolutionary features dominates the precision-recall curve for the baseline classifier.

5.3.2 Easy-to-predict and Hard-to-predict Interfaces

Analysis of the interface residue predictions on individual proteins indicates that the proteins whose interfaces are predicted with high precision and recall using structural features are also well predicted from sequence alone (although the latter are often inferior to the former). This can be accounted for by the fact that we only predict interface residues on the surface, which means that even sequence-based predictor indirectly utilizes some structural information in the form of knowledge of the surface.

We also find that that interfaces in some proteins are "easy-to-predict" (with recall $> 60\%$ and precision $> 50\%$), whereas interfaces in other proteins are "hard-to-predict" (with recall $< 30\%$ and precision $< 20\%$). Interestingly, we find that "easy-to-predict" proteins interfaces are predicted with

relatively high precision and recall using only sequence-derived features. "Hard-to-predict" proteins interfaces tend to be difficult to predict with high precision and recall even using structural features. Upon closer examination, we find that the "easy-to-predict" interfaces tend to have large protein surfaces (with 80% or more of residues being surface residues) and protein interfaces consisting of residues that are clustered along the sequence (with more than > 10 interface residues nearly contiguous in sequence). Examples of "easy-to-predict" interface is shown in 5.1. In contrast, "hard-to-predict" interfaces have large protein surfaces but interface residues that are form small isolated clusters along the sequences (< 6 nearly contiguous residues in sequence) (See figure 5.2). One possible explanation for this observation might be the fact that all of the classifiers considered in this study use a primarily sequence-based representation, albeit augmented by structural and evolutionary features. It would be interesting to explore alternative representations based on surface neighborhood or structural neighborhood of target residues.

Table 5.1 Prediction results of different Naive Bayes classifiers with different feature compositions: 1 – sequence, 2 – sequence+side chain orientation, 3 – sequence+variation entropy, 4 – sequence+roughness, 5 – sequence+solid angle, 6 – sequence+hydrophobicity, 7 – sequence+cx, 8 – sequence+cx+hydrophobicity+solid angle+variation entropy

Naive Bayes Classifier	*ac*	*re*	*pr*	*cc*	*auc*
1	56.1%	56.2%	29.3%	10.6%	0.589
2	56.3%	55.6%	29.4%	10.5%	0.589
3	57.0%	56.8%	30.2%	12.0%	0.610
4	57.9%	57.8%	31.0%	13.6%	0.612
5	57.0%	58.1%	31.0%	13.8%	0.613
6	58.9%	57.8%	31.5%	14.8%	0.625
7	60.2%	56.1%	32.2%	15.3%	0.629
8	62.3%	60.7%	34.6%	20.3%	0.675

5.3.3 Predicting Interfaces on Modeled Structures

Our use of structural features in predicting protein-protein interfaces relies on the availability of the structures of target proteins. In light of the fact that the use of structural features improves the precision

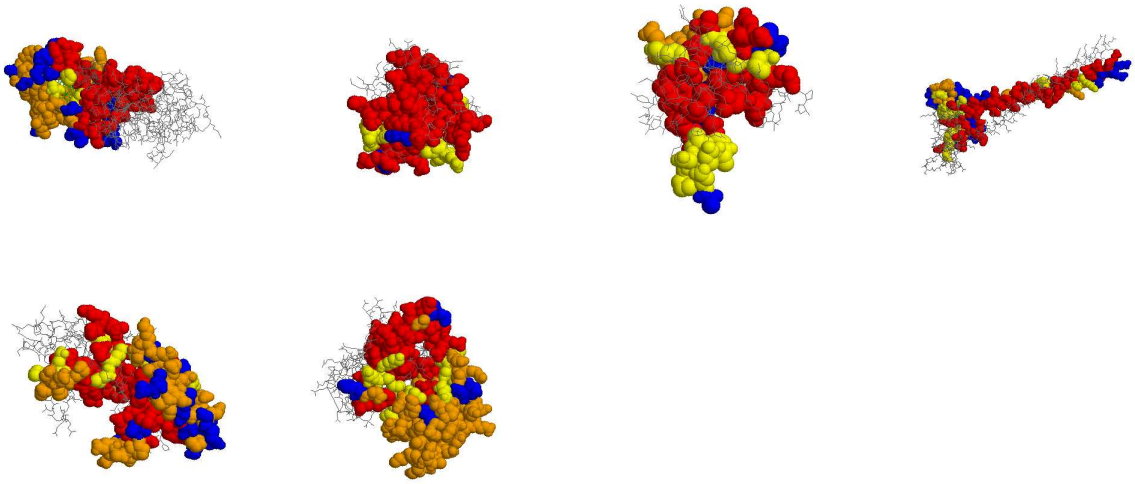

Figure 5.1 Interaction Sites Recognition of "easy-to-predict" proteins: 1aih (chain D), 1cdc (chain B), 1igu (chain B), 1joc (chain B), 1lgp (chain A) and 1lj9 (chain B). The predicted chain B is shown in green, with the residues of interest shown in space fill and color coded as follows: red, interface residues identified as such by the classifier (TPs); yellow, interface residues missed by the classifiers (FNs); blue, residues incorrectly classified as interface residues (FPs) and orange, residues correctly classified as non-interface residues (TNs). For clarity, interface residues for partner chains (gray wireframe) are not shown. The structure diagrams were generated using RasMol (156).

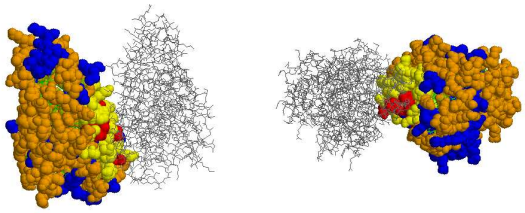

Figure 5.2 Interaction Sites Recognition of "hard-to-predict" proteins: 1czf (chain A) and 1iqu (chain A)

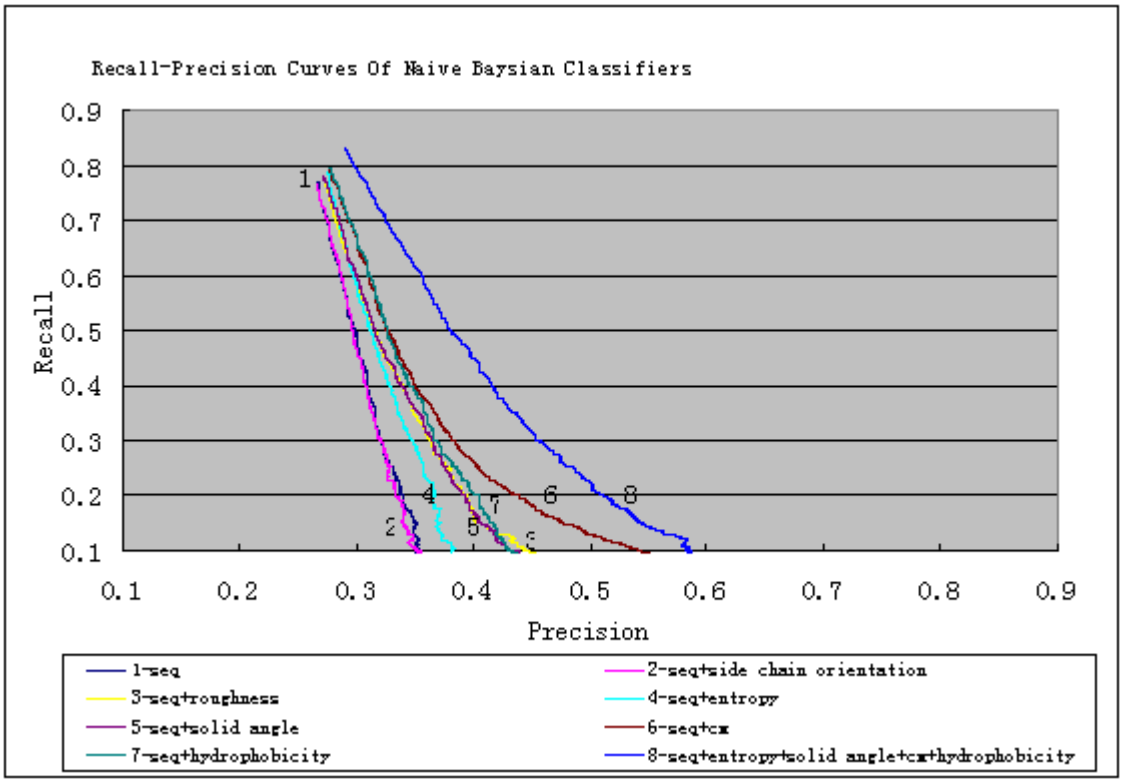

Figure 5.3 Recall_precison curves of the 8 Naive Bayes classifiers with different feature compositions

and recall of interface predictions, it is natural to ask whether similar gains can be obtained if we use modeled structures in place of actual structures of the target proteins. We used the Naive Bayes classifier using the best performing combination of sequence, structure, and evolutionary features described above, trained on a dataset extracted from 2174 of the original 2383 proteins, to predict the protein-protein interfaces in the 209 modeled protein structures. The SE and RMSD between the 209 modeled proteins and their corresponding actual structures are $81.6\% \pm 10.1\%$ and 16.46 ± 36.82 respectively. Three predictions are performed, p1 – using only protein sequence, p2 – using modeled structure and p3 – using actual structures. The prediction results are shown in Table 5.2 as prediction experiments **p1**, **p2** and **p3** and in Figure 5.4 as curves 1, 2 and 3.

Comparison among these three predictions shows that interface prediction the modeled structures is inferior to the predictions on the actual structures, but marginally better than prediction using only sequence information. A possible explanation for the disparity between the predictions on actual versus modeled structures might be the poor accuracy of the modeled structures. To explore this possibility, we conducted an additional experiment – p4: we selected well-modeled proteins with RMSD < 2.0 from the larger set of 209 modeled proteins. The performance of the classifier on this subset of 24 well-modeled proteins (experiment **p4**) are in Table 5.2 and curve 4 in Figure 5.4. The results of this experiment show that the more accurate the structure models, the greater the improvement obtained by using structural features in predicting protein-protein interfaces.

Table 5.2 Prediction results of modeled protein structures

prediction experiment	*ac*	*re*	*pr*	*cc*	*auc*
p1 : $sequencealone$	56.8%	55.8%	32.7%	12.0%	0.589
p2 : $modeledstructures$	55.8%	66.0%	31.8%	14.8%	0.624
p3 : $actualstructures$	63.2%	61.2%	36.2%	19.7%	0.676
p4 : $well - modeledstructures$	57.4%	65.5%	33.6%	17.0%	0.645

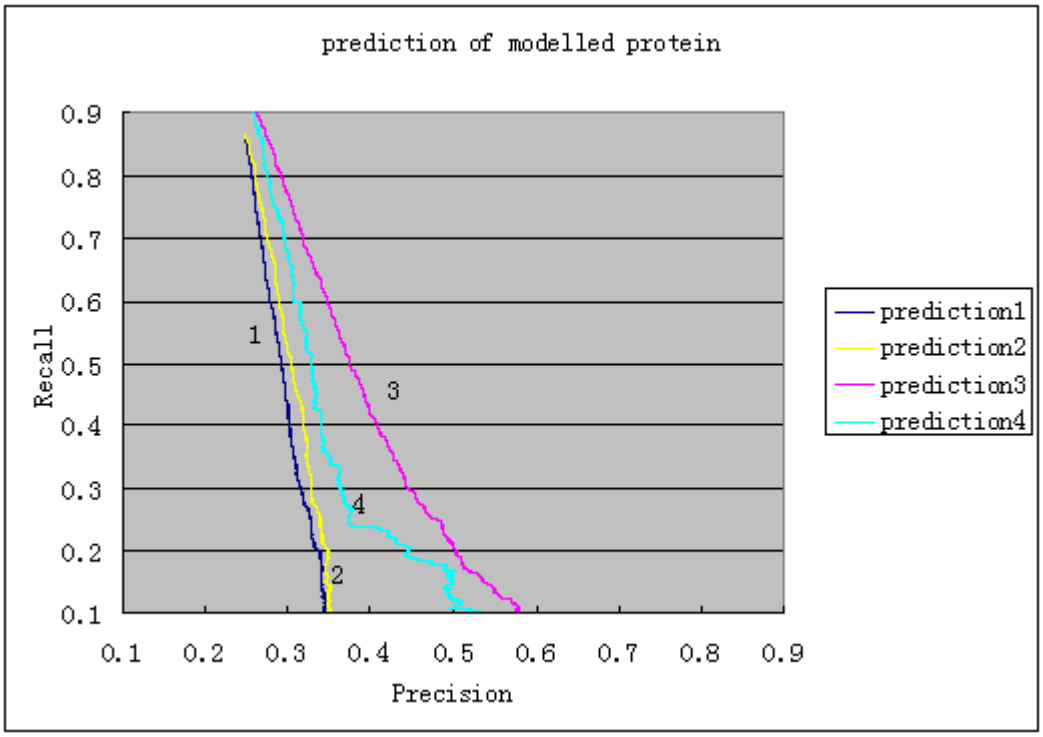

Figure 5.4 Recall_precison curves of the four prediction experiments on protein sequences, modeled protein structures and actual protein structures

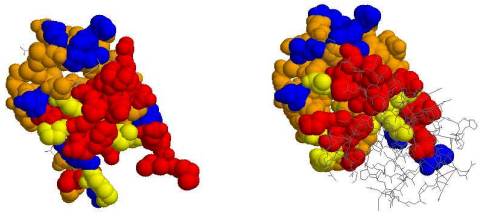

Figure 5.5 Interaction Site Recognition Using Modeled Structure and Actual Structure (protein 2b5a, chain D): The left displays the prediction of modeled structure (re=0.79 and pr=0.64) and the right shows the prediction of the actual structure (re=0.71 and pr=0.68). RMSD: 1.93 and SE:96.2%

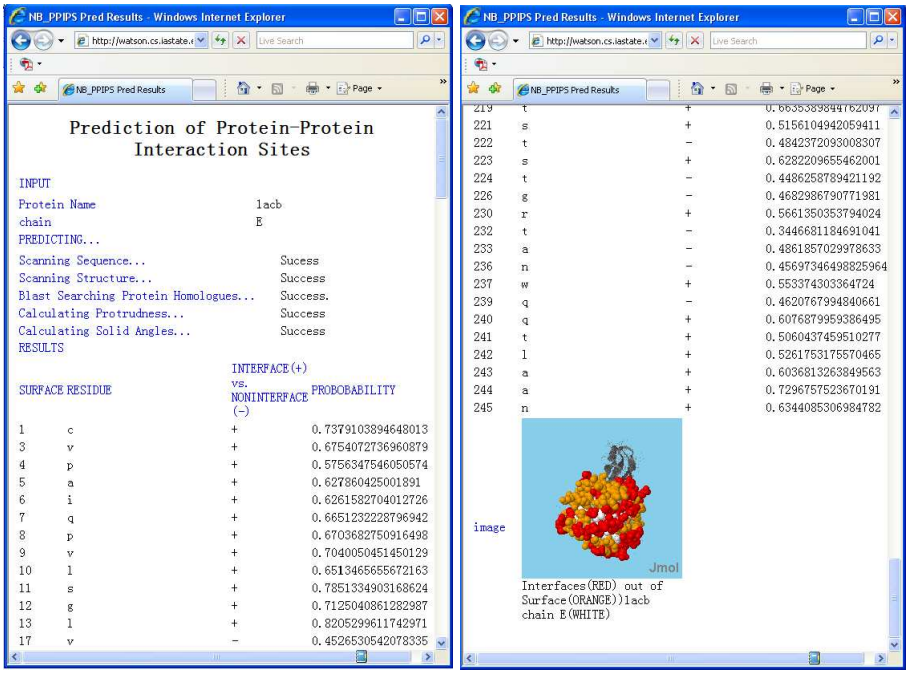

Figure 5.6 snapshot of NB_PPIPS running results

.

5.4 NB_PPIPS Protein-Protein Interface Prediction Server

We have made available NB_PPIPS, an on-line server for predicting protein-protein interfaces given the actual or modeled structure of the target protein. NB_PPIPS implements a Naive Bayes classifier trained on the $2,383$ protein dataset, with 16 input features: eleven sequence neighboring residues, variation entropy, cx value, solid angle and hydrophobicity. NB_PPIPS can be accessed via `http://watson.cs.iastate.edu/nb_ppips`. NB_PPIPS integrates several pieces of software into a web server: WEKA machine learning package (184) to build the classifier, blastp (3) to calculate variation entropy, the cx program (140) to calculate cx value and the msp program (36) to calculate solid angle. NB_PPIPS parses an input protein structure to identify its surface residues, calculates the structural features of each residue and outputs the probability that a residue is an interaction residue (Only surface residues are assigned non-zero probability of being an interface residue). The process of parsing, calculating and predicting can be monitored on the webpage. The predicted interface residues can be visualized on the structure using Jmol (Jmo) plug-in. Figure 5.6 shows a screen shot of the the result page produced by the server.

5.5 Summary

Reliable identification of protein-protein interaction sites in proteins is one of the major challenges in computational biology. We have described a machine learning approach to predicting protein-protein interface residues in settings where the structure of the target protein is available although the structures of the its binding partners or that of complexe(s) formed by it with other protein(s) are unavailable. We have explored several alternative representations of the input to the classifier starting with a sequence window consisting of the target residue and its sequence neighbors. The results of our experiments using 10-fold cross-validation on a large non-redundant dataset of $2,383$ proteins show that the Naive Bayes classifier trained using a combination of sequence, structure, and evolutionary information substantially outperforms its counterpart that is trained using sequence information alone.

We found that that interfaces in some proteins are "easy-to-predict" with relatively high precision and recall using only sequence-derived features whereas interfaces in other proteins are "hard-to-predict" despite the use of structural features, although the use of structural features yields improvements in both cases. We observed that the "easy-to-predict" interfaces tend to have large protein surfaces and protein interfaces consisting of many residues that are clustered along the sequence. In contrast, "hard-to-predict" interfaces have large protein surfaces but interface residues that form small isolated clusters along the sequence.

We also explored the use of a modeled protein structure in place of the actual structure of the target protein as input to the classifier. Our results show that the feasibility of this approach depends on the accuracy of the modeled structure. We have made the structure-based Naive Bayes classifier for predicting protein-protein interfaces as an online server NB_PPIPS accessible at `http://watson.cs.iastate.edu/nb_ppips`.

The identification of protein-protein interface residues is complicated in part because of the great diversity of proteins as well as their interactions. Protein-protein interfaces can be divided into six categories in terms of inter versus intra-molecule, homodimer versus heterodimer and permanent ver-

sus transient. Each of these categories differs from others in terms of amino acid composition (133; 40; 129). Protein-protein interfaces are usually more conserved across protein families in terms of the residues that participate in the interactions although such conservation is surprisingly less pronounced (176; 24). Protein-protein interfaces usually form a hydrophobic core with polar residue margins, but all interface residues do not appear contribute equally to the protein-protein binding affinity (106; 136; 15).

Although a variety of machine learning methods for predicting protein-protein interfaces have been explored, there is limited understanding of the relative strengths and limitations of the different methods. Different studies sometimes use different definitions of interface and surface (48; 133; 86), and report different measures of performance (and in some instances, use different definitions for the same performance measures). The difficulty of comparing different methods is further compounded by the unavailability of the datasets used in the study, and the algorithms used for classification. Therefore, there is a need for systematic comparisons of a broad class of methods, using a variety of data representations and types of sequence and structure-derived features, to understand their relative strengths and weaknesses and to develop new approaches that synergistically combine multiple methods. The dataset used in this study is one of the largest non-redundant datasets of protein-protein interfaces and hence can serve as a basis for comparison of multiple methods.

Work in progress is aimed at:

- Systematic and rigorous comparison of protein-protein interface predictions using a large dataset and a broad class of prediction methods and data representations
- Further exploration of what makes some interfaces easy and others hard to predict
- Development of customized predictors for specific target proteins (as opposed to a single predictor trained on the entire dataset)

CHAPTER 6. Conclusion

6.1 Contributions

Our study focuses on using computational methods to discover sequence and structural correlates of protein-protein interfaces, an important problem in the study of protein-protein interactions. Advances in methods predicting protein-protein interfaces can lead to better methods for identifying functionally important sites of proteins, extending the range of docking problems (by helping localize promising regions on the protein surface for docking), and focusing experimental work (e.g., site-specific mutation studies aimed at uncovering the sequence and structural correlates of protein-protein interactions) and better methods for rational drug design.

The work described in this thesis is organized around three main components:

- **PPIDB: A comprehensive Database of protein-protein interfaces**: We have assembled a comprehensive protein-protein interface database – PPIDB, extracted from the PDB. PPIDB design allows its contents to be updated periodically as PDB is updated. PPIDB provides a well-characterized dataset of protein-protein interfaces. It provides tools for generating large benchmark datasets for analyses and prediction of protein-protein interfaces, and for comparison of alternative prediction algorithms and data representations. PPIDB currently contains $71,486$ inter-chain protein-protein interfaces and keeps updating. PPIDB provides programmatic access to the database through web services. These features make PPIDB distinctive and complementary to other databases such as InterPare (64), DOCKGROUND (44), and SNAPPI-DB (82), etc.

- **Analyses of Protein-Protein Interfaces**: We have carried out analyses of physicochemical and

structural properties of protein-protein interfaces and non-interfaces. Most of the previous analyses of protein-protein interfaces were limited to small datasets (88; 85) or are confined to a subset of the properties of interest (133; 191; 110). Our analyses utilized a large, non-redundant dataset of $2,383$ proteins. We investigated a number of sequence, evolutionary and structural properties. The results of these analyses confirm many of the results of previous analyses using small datasets while providing additional insights regarding the sequence and structure-derived properties that carry information that can be used as a basis for predicting protein-protein interfaces.

- **Improved methods for predicting protein-protein interfaces**: We have developed classifiers to predict protein-protein interfaces from protein sequence, and when available, the structure of the target protein. We have shown that the use of structural features greatly improves the precision and recall of predicted protein-protein interface residues. The resulting classifier has been implemented as an online server.

6.2 Future Work

Identification of protein-protein interfaces is an important, challenging and fast-evolving area of research in computational biology. Some promising directions for further work include:

- **Enhancing functionality of PPIDB to support analyses and prediction of protein-protein interfaces:** It would be useful to develop additional web services to support different types of analyses of protein-protein interface datasets and for visualization of the analyses results (e.g., relative amino acid propensities, surface roughness, local curvature of interfaces and non-interfaces). It would be useful to annotate PPIDB and the associated Web services with metadata to enable other research groups to integrate PPIDB with other data resources and utilize the services offered by PPIDB in largera workflows.

- **Extending the analyses of interfaces to multimeric, transient and protein-ligand complexes:** Our work has largely focused on protein-protein dimeric interfaces (mostly permanent, physical interactions). However, many proteins form multimeric complexes consisting of multiple inter-

acting subunits; transient protein-protein interactions are very important in signal transduction; protein-ligand interactions are extremely important in the context of rational drug design.

- **Analysis and prediction of protein-RNA and protein-DNA interfaces:** Protein-protein interactions may share some of the characteristics of other macromolecular interactions. Hence, it would be interesting to compare protein-protein interfaces with protein-RNA and protein-DNA interfaces.

- **Using interface predictions to focus on experimental investigations:** It would be interesting to use interface predictions on specific experimental targets (such as ITK kinase binding sites) to focus experimental investigations and to use the experiments in turn to verify and help refine the computational predictions.

- **Applying protein-protein interface predictions to improve the reliability of protein-protein interaction networks constructed from high-throughput experiments:** The presence of false positives in protein-protein interaction datasets poses a challenge in analysis and interpretation of such networks: it has been shown that some reported protein interactions cannot be reconciled with known protein complexes (46). The lack of complete experimental data on the interaction networks presents additional challenges (9; 10). Vidal's group (68), for example, observed that the incompleteness of protein-protein interaction networks can lead to misleading conclusions from topological analysis of the networks. Protein-protein interface predictions can be used, in conjunction with docking studies, to improve the reliability of protein-protein interaction networks.

BIBLIOGRAPHY

[Jmo] Jmol: an open-source java viewer for chemical structures in 3d. http://www.jmol.org/.

[Spi] Spin-pp server.

[1] Allers, J. and Shamoo, Y. (2001). Structure-based analysis of protein-RNA interactions using the program entangle. *J Mol Biol*, 75-86:311(1).

[2] Alonso, G., Casati, F., Kuna, H., and Machiraju, V. (2004). *Web Services: Concepts, Architectures and Applications*. Springer-Verlag.

[3] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402.

[4] Amelie Stein, R. B. R. and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res.*, 33:D413–D417.

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

[6] Ashburner, M., Ball, C., Blake, J., and et al. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9.

[7] Ashkenazi, A., Presta, L., Marsters, S., Camerato, T., Rosenthal, K., Fendly, B., and Capon, D. (1990). Mapping the cd4 binding site for human immunodeficiency virus by alanine- scanning mutagenesis. *Proc Natl Acad Sci*, 87:7150–7154.

[8] 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):242–5.

[9] Bader, J. S. (2003). Greedily building protein networks with confidence. *Bioinformatics*, 19(15):1869–1874.

[10] Bader, J. S., Chaudhuri, A., Rothberg, J. M., and Chant, J. (2003). Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22:78–85.

[11] Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

[12] Ben-Shem, A., Frolow, F., and Nelson, N. (2003). Crystal structure of plant photosystem i. *Nature*, 426:630–635.

[13] Ben-Zeev, E. and Eisenstein, M. (2003). Weighted geometric docking: incorporating external information in the rotation-translation scan. *Proteins*, 52(1):24–7.

[14] Berman, H., Westbrook, J., Feng, Z., and et al. (2000). The protein data bank. *Nucleic Acids Res*, 28:235–242.

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

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

[17] Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K., Lee, I., and Marcotte, E. M. (2004). Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292–299.

[18] Boser, B. E., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Computational Learing Theory*, pages 144–152.

[19] Bradford, J. and Westhead, D. (2005). Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–94.

[20] Bradford, J. R., Needham, C. J., Bulpitt, A. J., and Westhead, D. R. (2006). Insights into protein-protein interfaces using a bayesian network prediction method. *J Mol Biol.*, 362(2):365–86.

[21] Brembeck, F. H., Abraham, C., Kietzmann, S., Droege, A., Koeppen, S., Timm, J., Goedde, A., Stelzl, U., Haenig, C., Korn, B., Goehler, H., Zenkner, M., Mintzlaff, S., Krobitsch, S., Worm, U., Lalowski, M., Stroedicke, M., Schoenherr, A., Bock, N., and Toksoz, E. (2005). A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, 122:957–968.

[22] Burgoyne, N. J. and Jackson, R. M. (2006). Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces. *Bioinformatics*, 22:1335–1342.

[23] C, O. and H, M. (1997). Crystallization of membrane proteins. *Curr. Opin. Struct. Biol.*, 7:699701.

[24] Caffrey, D., Somaroo, S., Hughes, J., and et al. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 13(1):190–202.

[25] Caragea, C., Sinapov, J., Dobbs, D., and Honavar, V. (2007). Assessing the performance of macromolecular sequence classifiers. *Proceedings of the IEEE Conference on Bioinformatics and Bioengineering*, pages 320–326.

[26] Ceol, A., Chatr-aryamontri, A., Santonico, E., Sacco, R., Castagnoli, L., and Cesareni, G. (2006). Domino: a database of domain-peptide interactions. *Nucleic Acids Res.*, 35:D557–D560.

[27] 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:D572–D574.

[28] Chen, H. and Zhou, H. (2005). Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against nmr data. *Proteins*.

[29] Chen, R., Li, L., and Weng, Z. (2003). Zdock: an initial-stage protein-docking algorithm. *Proteins*, 52:80–87.

[30] Chen Y, V. G. (2005). Protein families and rna recognition. *FEBS J.*, 272(9):2088–97.

[31] Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 285(4):256(5520).

[32] Christensen, E., Curbera, F., and et al. (2001). Web Services Description Language, Version 1.1. In *http://www.w3.org/TR/wsdl*.

[33] CJ, M., S, L.-M., and JD, H. (2001). The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov Today*, 6(21):1101–1110.

[34] Clackson, T. and Wells, J. (1995). A hot spot of binding energy in a hormone-receptor interface. *science*, 267:383–386.

[35] Connolly, M. L. (1986). Measurement of protein surface shape by solid angles. *Journal of Molecular Graphics*, 4(1):3–6.

[36] Connolly, M. L. (1993). The molecular surface package. *J Mol Graph*, 11(2):139–41.

[37] Cunningham, B. and Wells, J. (1980). High-resolution epitope mapping of hgh-receptor interactions by alanine-scanning mutagenesis. *science*, 244:1081–1085.

[38] Cunningham, B. and Wells, J. (1991). Rational design of receptor-specific variants of human growth hormone. *Proc Natl Acad Sci*, 88:3407–3411.

[39] Davis, F. P. and Sali, A. (2005). Pibase: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21(9):1901–7.

[40] De, S., Krishnadev, O., and et al. (2005). Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct Biol*, pages 5–15.

[41] de Vries, S. J. and Bonvin, A. M. J. J. (2006). Intramolecular surface contacts contain information about protein-protein interface regions. *Bioinformatics*, 22(17):2094–8.

[42] Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. (2003). Haddock: a protein-protein docking approach based on biochemical or biophysical data. *J. Am. Chem. Soc.*, 125:1731–1737.

[43] Dong, Q., Wang, X., Lin, L., and Guan, Y. (2007). Exploiting residue-level and profile-level interface propensities for usage in binding sites prediction of proteins. *BMC Bioinformatics*, 8:147.

[44] Douguet, D., Chen, H.-C., Tovchigrechko, A., and Vakser, I. A. (2006). Dockground resource for studying protein-protein interfaces. *Bioinformatics*, 22(21):2612–8.

[45] Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* Wiley-Interscience, New York.

[46] Edwardsa, A. M., Kusa, B., Jansenb, R., Greenbaumb, D., Greenblatta, J., and Gerstein, M. (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics*, 8(10):529–536.

[47] EV, K., YI, W., and GP, K. (2002). The structure of the protein universe and genome evolution. *Nature*, 420(6912):218–23.

[48] Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur J Biochem*, 269(5):1356–61.

[49] Farkas, I., Jeong, H., Vicsek, T., Barabasi, A. L., and Oltvai, Z. N. (2003). The topology of the transcription regulatory network in the yeast, saccharomyces cerevisiae. *Physica A*, 318:601–612.

[50] Fawcett, T. (2004). Roc graphs: Notes and practical considerations for researchers.

[51] Fernandez-Recio, J., Totrov, M., and Abagyan, R. (2004). Identification of proteincprotein interaction sites from docking energy landscapes. *J. Mol. Biol.*, 335:843–865.

[52] Ferre, F., Ausiello, G., Zanzoni, A., and Helmer-Citterich, M. (2004). Surface: a database of protein surface regions for functional annotation. *Nucleic Acids Res*, 32(Database issue):D240–4.

[53] Fischer, T., Arunachalam, K., Bailey, D., and et al. (2003). The binding interface database (bid): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, 19(11):1453–4.

[54] Frickel, E.-M., Riek, R., Jelesarov, I., Helenius, A., Wthrich, K., and Ellgaard, L. (2002). Trosy-nmr reveals interaction between erp57 and the tip of the calreticulin p-domain. *Proc Natl Acad Sci*, 99(4):1954–1959.

[55] Friedrich, T., Pils, B., Dandekar, T., Schultz, J., and Muller, T. (2006). Modelling interaction sites in protein domains with interaction profile hidden markov models. *Bioinformatics*, 22:2851–2857.

[56] Futschik, M. E., Chaurasia, G., and Herzel, H. (2007). Comparison of human proteincprotein interaction maps. *Bioinformatics*, 23(5):605–611.

[57] Gabb, H. A., Jackson, R. M., and Sternberg, M. J. E. (1997). Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.*, 272:106–120.

[58] Gallet, X., Charloteaux, B., Thomas, A., and Brasseur, R. (2000). A fast method to predict protein interaction sites from sequences. *J Mol Biol*, 302:917–26.

[59] Gervasio, F. L., Laio, A., and Parrinello, M. (2005). Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.*, 127:2600–2607.

[60] Glaser, F., Pupko, T., Paz, I., and et al. (2003). Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19(1):163–4.

[61] Glaser, F., Rosenberg, Y., Kessel, A., and et al. (2005). The consurf-hssp database: the mapping of evolutionary conservation among homologs onto pdb structures. *Proteins*, 58(3):610–7.

[62] 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.

[63] Goh, C.-S., Milburn, D., and Gerstein, M. (2004). Conformational changes associated with protein-protein interactions. *Curr. Opin. Struct. Biol.*, 14:104–109.

[64] Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J., Bolser, D. M., Oh, D., Kim, D.-S., and Bhak, J. (2005). A protein domain interaction interface database: Interpare. *BMC Bioinformatics*, 6:207.

[65] Gront, D. and Kolinski, A. (2006). Bioshellła package of tools for structural biology computations. *Bioinformatics*, 22(5):621–622.

[66] Guo, J., Chen, H., Sun, Z., and Lin, Y. (2004). A novel method for protein secondary structure prediction using dual-layer svm and profiles. *Proteins*, 54(4):738–43.

[67] Halperin, I., Ma, B., HaimWolfson, and Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–43.

[68] Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast proteincprotein interaction network. *Nature*, 430:88–93.

[69] Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F., and Yonath, A. (2001). High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, 107(5):679–88.

[70] Headd, J. J., Ban, Y. E. A., Brown, P., Edelsbrunner, H., Vaidya, M., and Rudolph, J. (2007). Protein-protein interfaces: properties, preferences, and projections. *J Proteome Res.*, 6(7):2576–86.

[71] Hendrickson, W. A. (2000). Synchrotron crystallography. *Trends in Biochemical Sciences*, 25(12):637–643.

[72] Henrick, K. and Thornton, J. (1998). Pqs: a protein quaternary structure file server. *Trends Biochem Sci*, 23(9):358–61.

[73] Honig, B. and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–9.

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

[75] Hua, S. and Sun, Z. (2001). A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol*, 308(2):397–407.

[76] Hubbard, S. and Thornton, J. (1996). Naccess - atomic solvent accessible area calculations. *http://wolf.bi.umist.ac.uk/naccess*.

[77] Jaakkola, T., Diekhans, M., and Haussler, D. (2000). A discriminative framework for detecting remote protein homologies. *J Comput Biol*, 7(1-2):95–114.

[78] Janin, J. (1996). Protein-protein recognition. *Prog. Biophys. Mol. Biol.*, 64:145–166.

[79] Janin, J. and Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem*, 265:16027–16030.

[80] Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J. E., Vajda, S., Vakser, I., and Wodak, S. J. (2003). Capri: a critical assessment of predicted interactions. *Proteins*, 52:2–9.

[81] Janin, J., Miller, S., and Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204(1):155–64.

[82] Jefferson, E. R., Walsh, T. P., Timothy J. Roberts, ., and Barton, G. J. (2007). Snappi-db: a database and api of structures, interfaces and alignments for protein-protein interactions. *Nucleic Acids Res.*, 35:D580–D589.

[83] Jeong, H., Mason, S. P., Barabsi, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411:41–42.

[84] Jones, S., Shanahan, H., Berman, H., and Thornton, J. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res*, 31(24):7189–98.

[85] Jones, S. and Thornton, J. (1996). Principles of protein-protein interactions. *Proc Natl Acad Sci*, 93(1):13–20.

[86] Jones, S. and Thornton, J. (1997a). Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, 272(1):121–32.

[87] Jones, S. and Thornton, J. (1997b). Prediction of protein-protein interaction sites using patch analysis. *J Mol Biol*, 272(1):133–43.

[88] Jones, S. and Thornton, J. M. (1995). Protein-protein interactions: A review of protein dimer structures. *Progress in Biophysics and Molecular Biology*, 63(1):31–59.

[89] Kabsch, W. and C, C. S. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–637.

[90] Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc Natl Acad Sci*, 89:2195–2199.

[91] Kawashima, S. and Kanehisa, M. (2000). Aaindex: amino acid index database. *Nucleic Acids Res*, 28(1):374.

[92] Keskin, O., Bahar, I., Badretdinov, A., Ptitsyn, O., and Jernigan, R. (1998). Empirical solvent-mediated potentials hold for both intra-molecular and intermolecular inter-residue interactions. *Protein Sci*, 7:2578–2586.

[93] Keskin, O., Ma, B., and Nussinov, R. (2005a). Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, 345(5):1281–94.

[94] Keskin, O., Ma, B., Rogale, K., Gunasekaran, K., and Nussinov, R. (2005b). Protein-protein interactions: organization, cooperativity and mapping in a bottom-up systems biology approach. *Phys. Biol.*, 2(2):S24–S35.

[95] 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):1043–55.

[96] Khanin, R. and Wit, E. (2006). How scale-free are biological networks. *J Comput Biol*, 13:810–818.

[97] Kinoshita, K. and Nakamura, H. (2004). ef-site and pdbjviewer: database and viewer for protein functional sites. *Bioinformatics*, 20(8):1329–30.

[98] Koike, A. and Takagi, T. (2004). Prediction of protein-protein interaction sites using support vector machines. *Protein Engineering Design and Selection*, 17:165–173.

[99] Korn, A. P. and Burnett, R. M. (1991). Distribution and complementarity of hydropathy in mutisunit proteins. *Proteins*, 9(1):37–55.

[100] Kufareva, I., Budagyan, L., Raush, E., Totrov, M., and Abagyan, R. (2007). Pier: Protein interface recognition for structural proteomics. *Proteins*, 67(2):400–417.

[101] Kundrotas, P. J. and Alexov, E. (2007). Protcom: searchable database of protein complexes enhanced with domain-domain structures. *Nucleic Acids Res.*, 35:D575–D579.

[102] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–32.

[103] Lanckriet, G., Deng, M., Cristianini, N., Jordan, M., and Noble, W. (2004). Kernel-based data fusion and its application to protein function prediction in yeast. *Pac Symp Biocomput*, pages 300–11.

[104] Lanckriet, G. R. G., Cristianini, N., Bartlett, P. L., Ghaoui, L. E., and Jordan, M. I. (2002). Learning the kernel matrix with semi-definite programming. In *ICML*, pages 323–330.

[105] Landgraf, R., Xenarios, I., and Eisenberg, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.*, 307(5):1487–502.

[106] Larsen, T. A., Olson, A. J., and Goodsell, D. S. (1998). Morphology of proteincprotein interfaces. *Structure*, 6:421–7.

[107] Lee, B. and Richards, F. (1971). The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol*, 55:379–400.

[108] Leslie, C., Eskin, E., Cohen, A., Weston, J., and Noble, W. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–76.

[109] Leslie, C., Eskin, E., and Noble, W. S. (2002). The spectrum kernel: A string kernel for svm protein classification. In *Pacific Symposium on Biocomputing*, pages 564–575.

[110] Lewis, M. and Rees, D. (1985). Fractal surfaces of proteins. *Science*, 230(4730):1163–1165.

[111] Li, J.-J., Huang, D.-S., Wang, B., and Chen, P. (2006). Identifying protein-protein interfacial residues in heterocomplexes using residue conservation scores. *Int J Biol Macromol.*, 38(3-5):241–7.

[112] Liang, S., Zhang, C., Liu, S., and Zhou, Y. (2006). Protein binding site prediction using an empirical scoring function. *Nucleic Acids Research*, 32(13):3698–3707.

[113] Liang, S., Zhang, J., Zhang, S., and Guo, H. (2004). Prediction of the interaction site on the surface of an isolated protein structure by analysis of side chain energy scores. *Proteins*, 57(3):548–57.

[114] Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257:342–358.

[115] Lichtarge, O. and Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struct Biol.*, 12((1)):21–7.

[116] Mandell, J. G., Roberts, V. A., Pique, M. E., Kotlovyi, V., Mitchell, J. C., Nelson, E., Tsigelny, I., and Ten, L. F. (2001). Protein docking using continuum electrostatics and geometric fit. *Protein Eng*, 14:105–113.

[117] Mart-Renom, M. A., Stuart, A. C., Fiser, A., Snchez, R., Melo, F., and ali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29:291–325.

[118] McCoya, A. J., Epaa, V., and Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. *J Mol Biol*, 268(2):570–84.

[119] Melcher, K. (2004). New chemical crosslinking methods for the identification of transient protein-protein interactions with multiprotein complexes. *Current Protein and Peptide Science*, 5(4):287–296.

[120] Mendez, R., Leplae, R., Maria, L. D., and Wodak, S. J. (2003a). Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins*, 52(1):51–67.

[121] Mendez, R., Leplae, R., Maria, L. D., and Wodak, S. J. (2003b). Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins*, 52(1):51–67.

[122] Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *J Mol Biol*, 196(3):641–656.

[123] Ming-Hui Li, Lei Lin, X.-L. W. and Liu, T. (2007). Protein-protein interaction site prediction based on conditional random fields. *Bioinformatics*, 23(5):597–604.

[124] Murakami, Y. and Jones, S. (2006). Sharp2: protein-protein interaction predictions using patch analysis. *Bioinformatics*, 22(14):1794–5.

[125] Neshich, G., Togawa, R., Mancini, A., and et al. (2003). Sting millennium: A web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res*, 31(13):3386–92.

[126] Neuvirth, H., Heinemann, U., Birnbaum, D., Tishby, N., and Schreiber, G. (2007). Promateusan open research approach to protein-binding sites analysis. *Nucleic Acids Res.*, 35:W543W548.

[127] 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):181–99.

[128] Nguyen, C., Gardiner, K. J., and Cios, K. J. (2007). A hidden markov model for predicting protein interfaces. *J Bioinform Comput Biol.*, 5(3):739–53.

[129] Nooren, I. and Thornton, J. (2003a). Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol*, 325(5):991–1018.

[130] Nooren, I. M. and Thornton, J. M. (2003b). Diversity of protein-protein interactions. *EMBO J.*, 22(14):3486–92.

[131] Norel, R., Petrey, D., Wolfson, H. J., and Nussinov, R. (1999). Examination of shape complementarity in docking of unbound proteins. *Proteins*, 36(3):307–17.

[132] Oda, Y., Saeki, K., Takahashi, Y., Maeda, T., Naitow, H., Tsukihara, T., and Fukuyama, K. (2000). Crystal structure of tobacco necrosis virus at 2.25 a resolution. *J Mol Biol.*, 300(1):153–69.

[133] Ofran, Y. and Rost, B. (2003a). Analysing six types of protein-protein interfaces. *J Mol Biol*, 325(2):377–87.

[134] Ofran, Y. and Rost, B. (2003b). Predicted protein-protein interaction sites from local sequence information. *J Mol Biol*, 544(1-3):236–9.

[135] Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P. E., Raoult, D., and Claverie, J. M. (2000). Selfish dna in protein-coding genes of rickettsia. *Science*, 290:347–350.

[136] P, C. and J., J. (2002). Dissecting protein-protein recognition sites. *Proteins*, 47:334–43.

[137] Pal, A., Chakrabarti, P., Bahadur, R., Rodier, F., and Janin, J. (2007). Peptide segments in protein-protein interfaces. *J Biosci*, 32(1):101–11.

[138] Pervushin, K., Riek, R., Wider, G., and Wthrich, K. (1997). Attenuated t2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to nmr structures of very large biological macromolecules in solution. *Proc Natl Acad Sci*, 94:12366–12371.

[139] Phizicky, E., Bastiaens, P. I. H., Zhu, H., Snyder, M., and Fields, S. (2003). Protein analysis on a proteomic scale. *nature*, 422:208–215.

[140] Pintar, A., Carugo, O., and Pongor, S. (2002). Cx, an algorithm that identifies protruding atoms in proteins. *Bioinformatics*, 18(7):980–4.

[141] Porollo, A. and Meller, J. (2007). Prediction-based fingerprints of proteinprotein interactions. *Proteins*, 66:630–645.

[142] Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18(Suppl 1):71–7.

[143] Qin, S. and Zhou, H.-X. (2007). meta-ppisp: a meta web server for protein-protein interaction site prediction. *Bioinformatics*, 23(24):3386–3387.

[144] Rackovsky, S. and Scheraga, H. A. (1977). Hydrophobicity, hydrophilicity, and the radial and orientational distributions of residues in native proteins. *Proc Natl Acad Sci U S A*, 74(12):52485251.

[145] Rappsilber, J., Siniossoglou, S., Hurt, E. C., and Mann, M. (2000). A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.*, 72:267–75.

[146] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555.

[147] Rawlings, N., Tolle, D., and Barrett, A. (2004). Merops: the peptidase database. *Nucleic Acids Res*, 32:D160–4.

[148] Res, I., Mihalek, I., and Lichtarge, O. (2005). An evolution based classifier for prediction of protein interfaces without using protein structures. *Bioinformatics*, 21(10):2496–501.

[149] Robert D. Finn, M. M. and Bateman, A. (2005). ipfam: visualization of protein-protein interactions in pdb at domain and amino acid resolutions. *Bioinformatics*, 21(3):410–412.

[150] Rosenfeld, R., Vajda, S., and DeLisi, C. (1995). Flexible docking and design. *Annu. Rev. Biophys. Biomol. Struct.*, 24:677–700.

[151] Rost, B. and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Proteins*, 216-26:20(3).

[152] S, A., MM, G., and A, S. (2004). Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4):477–86.

[153] S, F. and O, S. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–6.

[154] Saigo, H., Vert, J., Ueda, N., and Akutsu, T. (2004). Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–9.

[155] Sander, C. and Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68.

[156] Sayle, R., Mueller, A., Grossman, G., Molinaro, M., Bernstein, H. J., Chigbo, C., Chachra, R., and Yamanishi, M. (2005). Openrasmol: Molecular graphics visualisation tool. *http://www.openrasmol.org*.

[157] Schölkopf, B., Tsuda, K., and Vert, J.-P. (2003). *Kernel methods in computational biology. MIT Press*.

[158] Sen, T. Z., Kloczkowski, A., Jernigan, R. L., Yan, C., Honavar, V., Ho, K.-M., Wang, C.-Z., Ihm, Y., Cao, H., Gu, X., and Dobbs, D. (2004). Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. *BMC Bioinformatics*, 5(1):205.

[159] Sharan, R. and Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24:427–433.

[160] Sheinerman, F. B., Norel, R., and Honig, B. (2000a). Electrostatic aspects of protein-protein interactions. *Curr Opin Struct Biol*, 10(2):153–9.

[161] Sheinerman, F. B., Norel, R., and Honig, B. (2000b). Electrostatic aspects of protein-protein interactions. *Curr Opin Struct Biol*, 10(2):153–9.

[162] Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol.*, 310(1):243–57.

[163] Shimada, I. (2005). Nmr techniques for identifying the interface of a larger proteincprotein complex: Cross-saturation and transferred cross-saturation experiments. *Methods in Enzymology*, 394:483–506.

[164] Shoemaker, B. A. and Panchenko, A. R. (2007a). Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol.*, 3(3):e42.

[165] Shoemaker, B. A. and Panchenko, A. R. (2007b). Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol.*, 3(4):e43.

[166] Smith, G. R. and Sternberg, M. J. (2002). Prediction of protein-protein interactions by docking methods. *Current Opinion in Structural Biology*, 12(1):28–35.

[167] Sternberg, M. J., Gabb, H. A., and Jackson, R. M. (1998). Predictive docking of proteinłprotein and proteinłdna complexes. *Current Opinion in Structural Biology*, 8(2):250–256.

[168] Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255.

[169] Szilagyi, A., Grimm, V., Arakaki, A. K., and Skolnick, J. (2005). Prediction of physical protein-protein interactions. *Phys Biol.*, 2(2):S1–16.

[170] Takahashi, H., Nakanishi, T., Kami, K., Arata, Y., and Shimada, I. (2000). A novel nmr method for determining the interfaces of large protein?protein complexes. *Nature Structural Biology*, 7:220–223.

[171] Takashi, I., Tomoko, C., Ritsuko, O., Mikio, Y., Masahira, H., and Yoshiyuki, S. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98:4569–4574.

[172] Terribilini, M., Lee, J.-H., Yan, C., Jernigan, R. L., Carpenter, S., Honavar, V., and Dobbs, D. (2006). Identifying interaction sites in "recalcitrant" proteins: predicted protein and RNA binding sites in Rev proteins of HIV-1 and EIAV agree with experimental data. *Pacific Symposium on Biocomputing*, 11:415–426.

[173] 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.

[174] 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. (1999). A comprehensive analysis of proteincprotein interactions in saccharomyces cerevisiae. *Nature*, 403:623–627.

[175] Uetz, P., Titz, B., and Cagney, G. (2008). *Experimental methods for protein interaction identification and characterization*. Springer-Verlag.

[176] Valdar, W. and Thornton, J. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, 42(1):108–24.

[177] Vanschoenwinkel, B. and Manderick, B. (2004). Substitution matrix based kernel functions for protein secondary structure. *The Procedding of International Conference on Machine Learning and Applications*.

[178] Vazquez A, Flammini A, M. A.-V. A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.*, 21:697C700.

[179] Vert, J.-P. (2005). Kernel methods in genomics and computational biology. *Technical Report HAL:ccsd-00012124, October 2005*.

[180] Wang, B., Chen, P., Huang, D.-S., jing Li, J., Lok, T.-M., and Lyu, M. R. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett.*, 580:380–384.

[181] Wells, J. A. (1991). Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol*, 202:390–411.

[182] Wimberly, B. T., Brodersen, D. E., William M. Clemons, J., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T., and Ramakrishnan, V. (2000). Structure of the 30s ribosomal subunit. *Nature*, 426:327–339.

[183] Winter, C., Henschel, A., Kim, W. K., and Schroeder, M. (2006). Scoppi: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, 34:D310–4.

[184] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco*.

[185] Wu, F., Towfic, F., Dobbs, D., and Honavar, V. (2007). Analysis of protein-protein dimeric interfaces. *International Conference on Bioinformatics and Biomedicine*.

[186] 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*, 28:289–91.

[187] Yan, A. and Jernigan, R. L. (2005). How do side chains orient globally in protein structures? *Proteins*, 61(3):513–22.

[188] Yan, C., Dobbs, D., and Honavar, V. (2004a). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, Suppl 1:I371–I378.

[189] Yan, C., Honavar, V., and Dobbs, D. (2004b). Identification of interface residues in protease-inhibitor and antigen-antibody complexes: a support vector machine. *Neural Comput. Applic.*, 13:123–129.

[190] Yan, C., Terribilini, M., Wu, F., Jernigan, R. L., Dobbs, D., and Honavar, V. (2006). Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, in press.

[191] Yan, C., Wu, F., Dobbs, D., Jernigan, R., and Honavar, V. (2007). Characterization of protein-protein interfaces. *Protein*, in press.

[192] Yook, S. H., Oltvai, Z. N., and Barabasi, A. L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928–942.

[193] Young, L., Jernigan, R., and Covell, D. (1994). A role for surface hydrophobicity in protein-protein recognition. *Proteins*, 3(5):717–29.

[194] Zhou, H. and Shan, Y. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44(3):336–43.

[195] Zhou, H.-X. and Qin, S. (2007). Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*.

[196] Zhu, H. Y. X., Greenbaum, D., Karro, J., and Gerstein, M. (2004). Topnet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res*, 32:328–337.