2005

Identification of interface residues involved in protein-protein and protein-DNA interactions from sequence using machine learning approaches

Changhui Yan
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/rtd
Part of the Artificial Intelligence and Robotics Commons, Biology Commons, and the Molecular Biology Commons

Recommended Citation
Identification of interface residues involved in protein-protein and protein-DNA interactions from sequence using machine learning approaches

by

Changhui Yan

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
Drena Dobbs, Co-major Professor
Robert L. Jernigan
Kai-Ming Ho
David Fernández-Baca

Iowa State University
Ames, Iowa
2005
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
This is to certify that the doctoral dissertation of

Changhui Yan

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Committee Member

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Co-major Program

Signature was redacted for privacy.

For the Co-major Program
TABLE OF CONTENTS

CHAPTER 1. GENERAL INTRODUCTION 1
Introduction 1
A survey of related studies 2
Dissertation organization 15
References 17

CHAPTER 2. A TWO-STAGE CLASSIFIER FOR IDENTIFICATION OF PROTEIN-PROTEIN INTERFACE RESIDUES 29
Abstract 29
Introduction 29
Methods and materials 31
Experiments and results 35
Discussion 42
Acknowledgements 44
References 44

CHAPTER 3. IDENTIFYING AMINO ACID RESIDUES INVOLVED IN PROTEIN-DNA INTERACTIONS FROM SEQUENCE 49
Abstract 49
Introduction 50
Methods and materials 52
Experiments and results 56
Discussion 62
Acknowledgements 65
References 65

CHAPTER 4. IDENTIFICATION OF DNA BINDING RESIDUES IN THE S SUBUNIT OF THE *M. jannaschii* TYPE I RESTRICTION-MODIFICATION SYSTEM 69
Abstract 69
Introduction 69
Methods and materials 71
Results 73
Discussion 80
Acknowledgements 81
References 82

CHAPTER 5. A DATABASE OF PROTEIN-PROTEIN INTERFACES 86
Abstract 86
Introduction 86
Methods and materials 88
Database features and accessibility 91
Acknowledgements 93
### References

<table>
<thead>
<tr>
<th>Chapter Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER 6. AN ANALYSIS OF PROTEIN-PROTEIN INTERFACES</td>
<td>96</td>
</tr>
<tr>
<td>Abstract</td>
<td>96</td>
</tr>
<tr>
<td>Introduction</td>
<td>97</td>
</tr>
<tr>
<td>Methods and materials</td>
<td>98</td>
</tr>
<tr>
<td>Results and discussion</td>
<td>99</td>
</tr>
<tr>
<td>Conclusions</td>
<td>113</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>114</td>
</tr>
<tr>
<td>References</td>
<td>114</td>
</tr>
</tbody>
</table>

| CHAPTER 7. CONCLUSIONS                           | 116  |
| Summary and discussion                           | 116  |
| Contributions                                     | 117  |
| Future work                                      | 119  |
| References                                       | 120  |
CHAPTER 1. GENERAL INTRODUCTION

INTRODUCTION

Protein-protein interactions and protein-DNA interactions are among the ubiquitous types of macromolecule interactions in biological systems. Revealing the mechanisms of protein-protein and protein-DNA interactions is crucial for understanding the functions of biological systems. Identification of amino acid residues that contribute to the specificity and affinity of such interactions is very important for understanding macromolecular functions and has broad therapeutic applications. Genome sequencing and proteomics projects have provided lists of macromolecules potentially present in several organisms (Morgan, 2001; Venter, et al., 2001). Elucidating the structures and interactions among these macromolecules has become a critical challenge in functional genomics.

Various experimental methods have been used to identify interface residues involved in protein-protein and protein-DNA interactions. In the structure determination approach, structures of complexes are determined primarily using X-ray crystallography (Smyth and Martin, 2000; Pusey, et al., 2005) or NMR spectroscopy (Wishart, 2005). Specific residue-residue contacts that mediate interactions can then be identified from the structures. In addition to biophysical methods for structure determination, site-directed mutagenesis (Cunningham and Wells, 1989; Ashkenazi, et al., 1990), chemical cross-linking (Back, et al., 2003; Trester-Zedlitz, et al., 2003), radiolytic protein footprinting (Guan, et al., 2003; Rashidzadeh, et al., 2003), and hydrogen-deuterium exchange (Mandell, et al., 1998; Hamuro, et al., 2003; Lanman, et al., 2003) have also been used to identify interface residues. One common disadvantage of these experimental methods is that they are laborious and time-consuming. Therefore, identification of interaction sites solely depending on these methods cannot catch up with the pace at which protein sequences are determined. By July 21, 2005, the PIR non-redundant reference protein database (PIR-NREF) (Wu, et al., 2002), a comprehensive collection of protein sequences, contained 2,316,856 entries. At that time, the Protein Data Bank (PDB) (Berman, et al., 2000) contained only 31,823 structures and the Database of Interaction Proteins (DIP) (Xennarios, et al., 2002) contained
only 17,556 proteins. Hence, computational methods that can identify interface residues quickly and accurately are urgently needed.

Structure-based computational methods that can be used to identify interface residues from protein structures include protein docking (Schneidman-Duhovny, 2004), Evolutionary Trace (Lichtarge and Sowa, 2002), patch analysis (Jones and Thornton, 1997a; 1997b), homology modeling (Lu, et al. 2002;2003), and structural motifs (Shulman-Peleg, et al., 2004). Computational methods that can identify interface residues using only the amino acid sequence of proteins as input have drawn much attention recently (Ofran and Rost, 2003b; Yan, et al., 2004a; Yan, et al., 2004b; Ahmad and Sarai, 2005; Terribilini, et al., Submitted; Yan, et al., Submitted-a; Yan, et al., Submitted-b, Gloor et al. 2005). In the sequence-based methods, residues are classified as either interface residues or non-interface residues based on local sequence information or sequence profiles resulting from multiple sequence alignments. Because sequence-based methods do not require information derived from protein structures, they have broader applications than other computational methods. Furthermore, they can be quickly applied to large sets of protein sequences. Thus, sequence-based methods provide a promising approach for identifying interface residues at a rate compatible with the rapid increase of protein sequences.

In this study, we focus on the development of computational approaches, specifically, machine-learning methods, for identification of amino acid residues involved in protein-protein interactions and protein-DNA interactions from sequence.

A SURVEY OF CURRENT METHOD

Following is a survey of the methods for identification of interface residues, with emphasis on computational approaches.
Biophysical, biochemical, and molecular biology methods

**X-ray crystallography**

If high-resolution structures of complexes are available, the interface residues can be easily identified from the structures. Today, X-ray crystallography is the gold standard for determining the structures of macromolecules. This technique can be used to obtain structures with very high resolution. It has been successfully used to solve the structures of very large complexes (Oda, *et al.*, 2000; Wimberly, *et al.*, 2000; Harms, *et al.*, 2001; Ben-Shem, *et al.*, 2003). In principle, there is no limit on the size of the structures that can be studied using X-ray crystallography. However, the difficulties in producing samples of large complexes in sufficient quantity and the difficulties associated with determining appropriate crystallization conditions samples limited the applications of X-ray crystallography.

**Nuclear magnetic resonance (NMR) spectroscopy**

NMR measures the responses of nuclear spins to an applied external magnetic field. Distances between atoms can be extracted from the measured signals and used as constraints to build three-dimensional structural models. This method is especially desirable for determining the structures of the proteins that are difficult to crystallize (e.g. integral membrane proteins). However, the application of NMR has been limited to small molecules, usually less than 25 kDa, because of signal overlap and fast relaxation of NMR signals from larger molecules. The use of larger magnetic fields together with such techniques as transverse relaxation-optimized spectroscopy (TROSY) (Pervushin, *et al.*, 1997) has increase the size limit of NMR to ~900 kDa (Fiaux, *et al.*, 2002; Riek, *et al.*, 2002).

Furthermore, the addition of residual dipolar couplings has made possible the accurate determination of helix curvature, domain orientation and stoichiometry of homomultimeric nucleic acid complexes (reviewed in MacDonald and Lu, 2002). Several NMR experimental methods have been developed specifically for identify residues involved in interactions. For example, chemical shift mapping is used to identify contact sites by detecting the changes in chemical shifts upon binding of molecules (Pellecchia, *et al.*, 1999; Frickel, *et al.*, 2002). A method based on NMR was developed by Takahashi *et al.* (2000) to identify interface
residues in large protein-protein complexes based on cross-saturation phenomena and TROSY.

**Site-directed mutagenesis**

In site-directed mutagenesis, mutations are introduced to specific sites to allow identification of the residues that mediate interactions using functional assays. In a saturation mutagenesis, an ensemble of mutant proteins is generated in which each amino acid in the protein has been individually replaced by other amino acids (Myers, *et al.*, 1985; Chen, *et al.*, 1999). Alanine-scanning mutagenesis and cysteine-scanning mutagenesis, methods that systematically replace the wild type residue with alanine or cysteine, have been shown to be very powerful approaches for identifying residues critical for function (Cunningham and Wells, 1989; Ashkenazi, *et al.*, 1990; Frillingos, *et al.* 1998). However, the classic alanine-scanning and cysteine-scanning are laborious because a mutant at each position must be constructed and expressed. Combinatorial alanine-scanning, in which alanine substitutions are introduced in multiple positions, provides a quick way to identify residues that are important for protein functions (reviewed in Morrison and Weiss, 2001).

**Radiolytic protein footprinting and protease footprinting**

In radiolytic protein footprinting, amino acid side chains are oxidized using hydroxyl radicals (-OH) generated from millisecond exposure of aqueous solutions to synchrotron radiation (Guan, *et al.*, 2003; Rashidzadeh, *et al.*, 2003). The quantity of oxidation is measured using mass spectrometry and the results reveal the solvent accessibilities of individual residues. Interface residues are protected from the oxidation and have reduced oxidation rates. Thus, interface residues can be identified by comparing the oxidation rates of specific residues in unbound monomers versus in complexes.

Protease footprinting has been used to identify both protein-protein binding sites and protein-nucleic acids binding sites (Hori *et al.*, 1995; Bogenhagen 1993). In protease footprinting, complexes are subjected to limited proteolysis with a protease (e.g. trypsin or chymotrypsin). The binding sites can be identified by analyzing the resulting fragments using gel electrophoresis.
Hydrogen-deuterium exchange

In a deuterated environment, hydrogen atoms from protein backbone amides are exchanged for deuterium atoms. The exchange rates for individual residues can be measured using liquid chromatography mass spectrometry (LC-MS). Interface residues have lower exchange rates in complexes than in monomers because of their reduced solvent accessibility in complexes. Thus, interface residues can be identified by comparing the hydrogen-deuterium exchange rates in monomers and complexes (Mandell, et al., 1998; Hamuro, et al., 2003; Lanman, et al., 2003).

Chemical cross-linking

Chemical cross-linking is an approach to covalently link two molecules that in close proximity in space with cross-linking agents. The linked complexes are chemically or enzymatically digested and mass spectrometry (MS) is used to analyze the fragments resulting from the digestions and identify regions where proteins interact (Back, et al., 2003; Trester-Zedlitz, et al., 2003). This method irreversibly joins binding partners together, so it can be used to detect transient interactions.

Computational methods

Protein docking

Protein docking methods are used to predict the structures of complexes when the atomic structures of unbound monomers are available. Protein docking programs have been successfully used to identify the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase (Strynadka, et al., 1996) and the binding sites in a set of non-obligate hetero-complexes (Fernandez-Recio, et al., 2004; Fernandez-Recio, et al., 2005). Protein docking usually consists of two steps: (1) generating the possible conformations of the complexes and (2) identifying the complex structure that minimizes certain energy functions or has the best physical or chemical complementarity between the interacting proteins. For a successful docking method, the first step should be fast and effective in covering the conformational space, and the scoring function used in the second step should be fast enough to allow its
application to a large number of complex candidates and be effective in discriminating between native and non-native structures (reviewed in Halperin, et al., 2001). Protein docking methods have improved substantially recently, in part due to the Critical Assessment of Predicted Interactions (CAPRI) (Janin, et al., 2003) contest. The first docking method capable of performing large-scale docking calculations efficiently was developed by Katchalski-Katzir, et al. (1992). The method uses fast Fourier transformation to calculate a score indicating the extent of geometric match between the surfaces of the interacting molecules. Due to its computational efficiency, the method has become one of the most popular docking methods and has been extended to include electrostatic and solvation terms (Gabb, et al., 1997; Mandell, et al., 2001; Chen, et al., 2003; reviewed in Vajda and Camacho, 2004). Most docking methods start with a rigid-body docking in which the unbound monomers are considered as rigid bodies. However, conformational changes often occur during binding (Goh, et al., 2004). Flexible docking in which potential conformational changes in the molecules are taken into account has been introduced (Rosenfeld, et al., 1995; Gervasio, et al., 2005). In principle, docking methods can be applied to both protein-protein and protein-nucleic acid interactions. However, the application of docking methods to protein-nucleic acid interactions is more challenging, because DNA and RNA are highly flexible. While the binding-induced conformation changes in proteins are primarily located on side chains, global conformation changes, e.g. bending or unwinding, can occur in DNA and RNA backbones (reviewed in Sternberg, et al., 1998). Some docking methods have been developed to deal with protein-DNA docking by taking into consideration the flexibility of DNA (Knegtel, et al., 1994a; 1994b). Docking methods are traditionally used to predict pairwise interactions. A recent trend is to develop docking methods that can predict the complexes of multimolecular assemblies (Comeau and Camacho, 2005; Inbar, et al., 2005). Docking methods that can work with theoretical protein structure models (Tovchigrechko, et al., 2002), instead of experimentally determined atomic structures, have also been explored.

**Evolutionary trace (ET)**

The ET method detects functional residues by identifying residues whose variations correlate with the functional divergence in evolution. It uses a tree derived from a multiple
sequence alignment (MSA) to estimate function divergence in a set of proteins (Lichtarge and Sowa, 2002). At each branching point of the tree, trace residues are defined as the MSA positions that have identical residues within each branch but differing residues among the branches. The rank of a residue is given by the tree level at which it becomes a trace residue, with level 0 denoting the tree root. The residues with low ranks are assumed to correspond to points of fundamental divergence in function and thus are considered more important than those with high ranks. The rank of each residue is mapped onto the structure of a representative protein. Spatial clusters of residues with low ranks are considered as potential functional sites (Lichtarge, et al., 1996b). Variations of this method include adding weights to sequences (weighted ET) (Landgraf, et al., 1999), taking into account the physicochemical properties of residues to measure the variations (Armon, et al., 2001; Landau, et al., 2005), using better ways to construct the tree (del Sol Mesa, et al., 2003), using experimentally defined subtypes instead of a tree derived from MSA (Hannenhalli and Russell, 2000), and introducing tolerance to gaps in the MSA algorithm (Madabushi, et al., 2002). Because a tree is inadequate to take into account the evolutionary convergence and various evolution rates at different points, some derivatives of ET eliminate the tree from consideration. Landgraf et al. (2001) used a MSA to determine the evolutionary variation of each residue and find clusters of residues on the protein surface whose evolutionary variations deviate significantly from the average. Innis et al. (2004) developed a method to identify functional sites by searching for conserved functional groups on the protein surface. The ET method has been used successfully to identify binding sites in DNA binding domains, zinc binding domains, and other proteins (Lichtarge, et al., 1996b; Lichtarge, et al., 1996a; Lichtarge, et al., 1997). Experimental studies have shown the effectiveness of the ET method in guiding experimental analysis of functional sites (Sowa, et al., 2001). Despite its successes, one limitation of the ET method and its derivatives is that they depend on the availability of a sufficiently large and diverse set of functionally related protein sequences. The availability of a representative structure is also crucial for the success of the method, since the structure is needed to identify spatial clusters of residues with low ranks.
Patch-based analysis

Many studies have examined the differences between interfaces and the rest of the proteins. In general, interfaces are hydrophobic, planar, and have good complementarities in shape and electrostatics (Chothia and Janin, 1975; Jones and Thornton, 1997a; Lo Conte, et al., 1999; Jones, et al., 2000; Ofran and Rost, 2003a). Patch-based methods identify functional sites based on the observed structural, chemical, and physical properties of interfaces. Jones and Thornton (1997a; 1997b) investigated the properties of protein-protein interaction sites and developed an approach to identify interfaces by evaluating surface patches in terms of several parameters. Some methods derived from this approach encode surface properties as input to neural networks or other machine-learning algorithms to build classifiers for classifying surface patches into interface patches versus non-interface patches (Zhou and Shan, 2001; Fariselli, et al., 2002). The coupling of patch-based classifiers and docking methods has been used to generate predictions for CAPRI targets (reviewed in Wodak and Mendez, 2004).

Another approach related to patch-based methods focuses on hot spots, the residues that contribute the most to the free energy of binding (Cunningham and Wells, 1991; Clackson and Wells, 1995; Bogana and Thorna, 1998). This approach is based on the discovery by Cunningham and Wells (1991) that binding energy is not uniformly distributed across interfaces and a single residue can contribute a large fraction of it. Hot spots were originally identified using alanine-scanning mutagenesis (Cunningham and Wells, 1991). Computational methods have been developed to identify hot spots using Gaussian network model (Demirel, et al., 1998), structural alignments of interfaces (Keskin, et al., 2005) or using physical models (Kortemme and Baker, 2002).

Homology modeling

This approach uses the structure of a known complex to build the interacting model for two monomers. In the method developed by Aloy and Russell (2002; 2003), the interacting components are first assigned to Pfam domains. These domains are then matched against a database of complexes. Once a match is found, the interaction model can be built by homology modeling using the matched complex as a template. This method depends on the
availability of Pfam domains to which the interacting monomers can be reliably assigned. The MULTIPROSPECTOR algorithm developed by Lu et al. (2002; 2003) uses a threading method to identify putative folds for the interacting components. The folds are then matched against a database of known complexes. These methods have been evaluated for their capacity to identify interacting pairs. In principle, a detailed interaction model for the query monomers can be built using threading and modeling homology once a matched complex is identified (reviewed in Wodak and Mendez, 2004), and interface residues can be identified based on the model. However, homology modeling methods assume the reliability of extrapolating the interacting model from a complex to the homologs of its monomers, which is still debatable (reviewed in Wodak and Mendez, 2004). The study by Aloy et al. (2003) shows that proteins having similarity only in fold (i.e. without evidence of common ancestor) rarely share similar interactions. Thus there are some cases where the extrapolation is unreliable.

**Correlated mutations**

Contacting residues tend to mutate coordinately to maintain the functional and structural stabilities of proteins. Some methods have been developed to detect correlated mutations within a protein (Göbel et al., 1994; Pazos et al. 1997a; Pollock et al. 1999; Tillier and Lui 2003; Gloor et al. 2005). It has been shown that coevolutionary information is sufficient to specify sequences that fold into native structures (Socolich, et al. 2005), and conserved residues form physically connected networks that link distant functional sites (Suel, et al. 2003). Information about correlated mutations can be used to identify functionally important regions in proteins. Carettoni et al. (2003) combined the analysis of correlated mutations with phage-display to identify the binding sites involved in the homodimerization of *E. coli* FtsA. Yu et al. (In press) developed a surface patch ranking method for identification of the residues that determine the specificity of enzyme-substrate binding by exploring sequence conservation and correlated mutations in multiple sequence alignments.

When using correlated mutations to identify functionally important site, it is critical to distinguish correlated mutations due to functional constraints from those that occur for other reasons (Wang and Pllock 2005). Tillier and Lui (2003) developed a method to separate
functional correlations from phylogenetic correlations using multiple interdependency. Gloor et al. (2005) used information theory to identify non-conserved co-evolving positions and discovered two classes of co-evolving positions, with one of them corresponding to the critical region (e.g. active site, binding site) for protein function.

Correlated mutations among residues from two interacting proteins have also been explored in some studies. Based on the assumptions that interactions between proteins are of the same physical nature as interactions within a single polypeptide, and that residues involved in inter-protein contacts therefore may undergo similar correlated mutations, Pazos et al. (1997b) developed a method to identify interface residues by detecting correlated mutations in two interacting proteins. Based on a similar hypothesis, Jespers et al. (1999) identified the interface residues involved in the interaction between staphylokinase and plasmin by searching for correlated mutations in the two proteins and then used the results to guide the docking of the two proteins.

**Structural motifs**

If the structure of a protein is known, its functions can be inferred by comparing it to other proteins that adopt the similar fold (Orengo, et al., 1997). However, proteins can have similar folds while having different functions and functional sites (reviewed in Russell, 1998). Structural motifs, local 3-D patterns often correlated with specific functions, provide an alternative approach to predict functional sites. In this approach, structural motifs are manually defined or computationally generated, and then query proteins are scanned to search for the occurrence of the structural motifs. Once a match is found, functional sites can be assigned to the query proteins based on the matched structural motif. Structural motifs have been successfully used to recognize catalytic sites and protein-DNA binding sites (Shanahan, et al., 2004; Torrance, et al., 2005). Several different methods have been developed to derive structural motifs. Wallace et al. (1997) described a geometric hashing algorithm (TESS) for deriving 3-D templates for motifs. Russell (1998) reported a method to detect structural motifs automatically by pairwise comparisons of protein structures. The method can detect new structural motifs even in the absence of sequence or fold similarity. Binkowski et al. (2003) developed an approach for detecting structural motifs by evaluating
the similarities of surface pockets and voids in sequence and spatial arrangement. Wangikar et al. (2003) developed a method (DRESPAT) to detect the recurrence of side-chain patterns in protein families using all-against-all pairwise comparisons of protein structures. When tested using a set of SCOP superfamilies, the algorithm was able to discover known and novel patterns. Currently, most available structural motifs are from enzymes whose functions are experimentally well-defined, e.g. those in the PROCAT database (Wallace, et al., 1996; Wallace, et al., 1997). Automated methods that can effectively discover structural motifs correlated with specific functions are needed.

Methods for recognizing known structural motifs in protein structures have also been developed. Artymiuk et al. (1994) developed a method (ASSAM) for the detection of user-defined patterns of side-chains in a protein structure using a subgraph-isomorphism algorithm. Kleywegt (1999) developed a program (RIGOR) that scans a protein structure to search for structural motifs from a motif database and a program (SPASM) that detects user-defined structural motifs in a structure database. Barker and Thornton (2003) used a constraint-based algorithm (JESS) to detect the occurrence of structural patterns in protein structures. Shulman-Peleg et al. (2004) proposed a method to recognize the surface regions of one protein that are similar to the binding sites of another protein. The method achieves high efficiency and speed by using low-resolution surface representation, hashing triangles of physicochemical properties, and applying hierarchical scoring schemes.

**Sequence-based classifiers**

Since the atomic structures of most proteins are still unknown, computational methods that can identify interface residues from sequence are in urgent need. Sequence-based classifiers classify residues into interface versus non-interface residues based on sequence patterns or physicochemical properties of residues. Based on their observation that proline residues have a high frequency of appearing at the sequences flanking interface residues, Kini and Evans (1996) developed a method to identify interface residues by detecting the presence of "proline brackets." Eisenberg et al. (1982) developed a method to distinguish transmembrane helices, surface-seeking helices and helices from globular proteins by plotting their hydrophobic moments versus the mean hydrophobicity of their residues. Gallet
et al. (2000) extended Eisenberg's method and developed a fast method to predict protein interaction sites from sequence. Ofran and Rost (2003b) used a neural network classifier to identify interface residues using as input a window of 9 amino acid residues centered on the target residue. In our work, we developed a sequence-based two-stage classifier to identify interface residues (Yan, et al., 2004a). The classifier consists of a support vector machine (SVM) classifier and a Bayesian classifier. In the first stage, the SVM classifier classifies surface residues into interface residues and non-interface residues based on their sequence neighbors. In the second stage, a Bayesian classifier is used to refine the output of the SVM classifier based on the observation that interface residues form clusters on sequences. Recently we used similar approaches to identify amino acids in protein-DNA and protein-RNA binding sites based on local sequence information (Terribilini, et al., In press; Yan, et al., Submitted-a; Yan, et al., Submitted-b). Despite their broad applicability, the performance of current sequence-based classifiers is still relatively weak. In our studies, we have explored the use of information besides sequence information (e.g. solvent accessibility) and demonstrated that other information, when available, can improve the performance of sequence-based classifiers. A promising approach to exploiting multiple information is to build ensemble classifiers based on individual classifiers that predict interface residues using different properties (Sen, et al., 2004).

In summary, we have presented a survey of the methods for identification of interface residues involved in protein-protein and protein-DNA interactions. Table 1 shows a summary of the methods. Among the methods we listed, experimental biophysical, biochemical and molecular biology methods can identify interface residues with high accuracy, but they are time-consuming. Structure-based methods require the structures of query proteins as input, but the structures are still not available for most proteins. Sequence-based classifiers have broader applications than structure-based methods, but the performance of current sequence-based classifiers is relatively weak.

In this study, we have developed sequence-based classifiers to identify interface residues. The results have shown the feasibility of identifying interface residues from sequence. We have also used various approaches to improve the performance of sequence-based classifiers. In a two-stage method, we use a Bayesian method to model the distribution
of interface residues in protein sequences and used it to refine the output of the first-stage classifier which takes a window of sequence as input. The results show that adding the second stage can improve the performance significantly. We have also explored information besides sequence to improve the performance of sequence-based classifiers. The results show that the performance of sequence-based classifiers can be improved by using solvent accessibility and sequence entropy of the target residue as additional inputs.
Table 1. Methods for identifying interface residues.

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Advantage and disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein docking</td>
<td>Structure</td>
<td>An enormous amount of computation is required to search for the conformation that minimizes a scoring function. Current methods cannot effectively deal with conformation changes during binding.</td>
</tr>
<tr>
<td>Evolutionary trace</td>
<td>Structure</td>
<td>Its application depends on the availability of a sufficiently large and diverse set of functionally related protein sequences. The success of the method relies on the accuracy of multiple sequence alignment.</td>
</tr>
<tr>
<td>Patch-based analysis</td>
<td>Structure</td>
<td>Systematic analysis is still needed to search for the features that can effectively identify binding sites.</td>
</tr>
<tr>
<td>Structural motifs</td>
<td>Structure</td>
<td>Currently most available structure motifs are from enzymes. Automated methods that can effectively discover structural motifs correlated to functions are needed.</td>
</tr>
<tr>
<td>Homology modeling</td>
<td>Structure</td>
<td>It depends on the reliability of extrapolating interacting model from a complex to its component homologs.</td>
</tr>
<tr>
<td>Correlated mutations</td>
<td>Sequence</td>
<td>Effective methods are needed to distinguish the correlated mutations resulting from functional correlations from those occurring for other reasons.</td>
</tr>
<tr>
<td>Sequence-based classifiers</td>
<td>Sequence</td>
<td>Sequence-based classifiers have broader applications than structure-based methods, but the performance of current sequence-based classifiers is relatively weak.</td>
</tr>
</tbody>
</table>
Table 1. (Continued)

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>Advantage and disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental biophysical</td>
<td>Depend on the methods</td>
<td>These methods usually have greater reliability than computational methods. However, they are laborious and time-consuming. Annotations of interaction sites solely depending on these methods cannot catch up with the pace at which protein sequences are determined.</td>
</tr>
<tr>
<td>biochemical and molecular</td>
<td></td>
<td></td>
</tr>
<tr>
<td>biology methods</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DISSERTATION ORGANIZATION

In this study, we aim to develop machine-learning methods for identification of amino acid residues involved in protein-protein interactions and protein-DNA interactions. We focus on the methods using sequence information alone and build classifiers that can classify residues into interface and non-interface residues based on sequence information. To facilitate the studies, we have developed a database of protein-protein interfaces and systematically analyzed the characteristics of the interfaces. Following is the outline of this study.

Chapter 1: This chapter presents the problems we address, a survey of current studies, and the outline of the dissertation.

Chapter 2: We have developed a two-stage method consisting of a Support Vector Machine (SVM) and a Bayesian classifier for predicting which surface residues of a protein participate in protein-protein interactions. The SVM classifier identifies interface residues based on their sequence neighbors and the Bayesian classifier exploits the fact that interface residues tend to form clusters in the primary amino acid sequence. The results have been published in the journal *Bioinformatics* (Yan, et al., 2004a). Changhui Yan carried out the computational experiments, and drafted the manuscript; Drena Dobbs and Vasant Honavar contributed to experimental design, discussions, and manuscript preparation.
Chapter 3: We present a machine-learning approach for identification of amino acid residues involved in protein-DNA interactions. We start with a Naive Bayes classifier trained to predict whether a given amino acid residue is a DNA-binding residue based on its identity and the identities of its sequence neighbors. Our results indicate the feasibility of identifying interface residues based on local sequence information. The performance of the classifier can be improved by using solvent accessibility and sequence entropy of the target residue as additional inputs. Changhui Yan carried out the computations, and drafted the manuscript; Michael Terribilini requested and composed the results from Sarai's group and contributed to discussions and manuscript reviews; Feihong Wu contributed to discussions; Robert Jernigan contributed to discussions and manuscript reviews; Drena Dobbs and Vasant Honavar contributed to experimental design, discussions, and manuscript preparation.

Chapter 4: We present the results of identifying DNA binding sites on the S subunit from M. jannaschi type I restriction-modification system using the classifiers developed in chapter 3. The predictions form four major patches on the S subunit. Changhui Yan carried out the experiments, and drafted the manuscript; Jae-Hyung Lee performed threading experiments using FUGUE2; Robert Jernigan contributed to discussions and paper reviews; Drena Dobbs and Vasant Honavar contributed to experimental design, discussions, and manuscript preparation.

Chapter 5: We have developed a database of protein-protein interfaces. The database consists of all the interfaces derived from the PDB. It provides friendly tools for users to extract interface information. Changhui Yan conceived the project, created the dataset, designed the structure of the database and the user interfaces, and drafted the manuscript; Feihong Wu contributed to the implementation of the database and the user interfaces; Robert Jernigan contributed to discussions and paper reviews; Drena Dobbs and Vasant Honavar contributed to computational design, discussions, and manuscript preparation.

Chapter 6: We present an analysis of protein-protein interfaces using datasets obtained from the database developed in chapter 5. The datasets are much larger than datasets that have been used in previous studies. The results reveal the differences between interfaces and the rest of the proteins in residue composition, conservation, hydrophobicity, and secondary
structure. The study also reveals the differences between homo-interfaces and hetero-interfaces. Changhui Yan conceived the project, performed the experiments and the analyses of results, and drafted the manuscript; Feihong Wu contributed to discussions; Robert Jernigan contributed to discussions and manuscript reviews; Drena Dobbs and Vasant Honavar contributed to experimental design, discussions, and manuscript preparation.

Chapter 7: This chapter includes the summary of this study, the contributions, and future work.

REFERENCES


CHAPTER 2. A TWO-STAGE CLASSIFIER FOR IDENTIFICATION OF PROTEIN-PROTEIN INTERFACE RESIDUES

A paper published in Bioinformatics

Changhui Yan, Drena Dobbs, and Vasant Honavar

ABSTRACT

The ability to identify protein-protein interaction 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. We have developed a two-stage method consisting of a Support Vector Machine (SVM) and a Bayesian classifier for predicting which surface residues of a protein participate in protein-protein interactions. This approach exploits the fact that interface residues tend to form clusters in the primary amino acid sequence. Our results show that the proposed two-stage classifier outperforms previously published sequence-based methods for predicting interface residues. We also present results obtained using the two-stage classifier on an independent test set of 7 CAPRI (Critical Assessment of PRedicted Interactions) targets. The success of the predictions is validated by examining the predictions in the context of the 3-dimensional structures of protein complexes.

INTRODUCTION

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’ (Ho et al., 2002; Giot
et al., 2003; Li et al., 2004). 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 structures protein complex structures using X-ray and NMR methods lags far behind the number of known protein sequences. Hence, there is a need for development of reliable computational methods for identifying protein-protein interface residues (Teichmann et al., 2001; Valencia and Pazos, 2002; 2003). Identification of protein-protein interaction sites and detection of specific amino acid residues that contribute to the specificity and strength of protein interactions is an important problem with broad applications ranging from rational drug design to the analysis of metabolic and signal transduction networks.

Protein-protein interfaces have been a topic of study for several years (Chothia and Janin, 1975, Jones and Thornton, 1996; Lo Conte et al., 1999; Ofran and Rost, 2003a). Based on the different characteristics of known protein-protein interaction sites, several methods have been proposed for predicting these sites. These include methods based on the presence of "proline brackets" (Kini and Evans, 1996), patch analysis using a 6-parameter scoring function (Jones and Thornton, 1997), properties associated with interface topology (Valdar and Thornton, 2001), analysis of the hydrophobicity distribution around a target residue (Gallet et al., 2000), charge distribution on interfaces (Sheinerman et al., 2002), multiple sequence alignments (Pazos, et al., 1997; Valencia et al., 2003), structure-based multimeric threading (Lu et al., 2003), docking methods (Halperin et al., 2002), using potentials that describe protein-protein interactions (Keskin et al., 1998), analysis of characteristics of spatial neighbors of a target residue using neural networks (Zhou and Shan, 2001; Fariselli et al., 2002; Ofran and Rost, 2003b). Our recent work has focused on an analysis of sequence neighbors of a target residue using an SVM method (Yan et al., 2003).

In our previous report, we used an SVM to identify interface residues using sequence neighbors of a target residue (Yan et al., 2003). Here we report a two-stage classifier consisting of a support vector machine (SVM) and a Bayesian network classifier that identifies interface residues primarily on the basis of sequence information. The two-stage method achieved 72% accuracy with a correlation coefficient of 0.3 when tested on a set of
77 proteins using five-fold cross validations. Experiments on the same dataset demonstrated that the two-stage method outperforms the previously published sequenced-based method of Gallet et al. (2000).

CAPRI (http://capri.ebi.ac.uk/) is a community wide experiment to assess the capacity of protein-docking methods to predict protein–protein interactions. In each round of CAPRI, structures of protein-protein complexes are predicted based on structures of the unbound components. CAPRI targets present interesting test cases for evaluation of computational methods for prediction of interface residues. A two-stage classifier which was trained using the 77 proteins in our dataset was tested performance on CAPRI targets. The results were evaluated in the context of 3-dimensional structures of protein complexes.

METHODS AND MATERIALS

Datasets

We extracted individual proteins from a set of 70 protein-protein heterocomplexes used in the study of Chakrabarti & Janin (2002). After removal of redundant proteins and molecules with fewer than 10 residues, we obtained a data set of 77 individual proteins with sequence identity less than 30%. These proteins represent six different categories of protein-protein interfaces, classified according to the scheme of Chakrabarti and Janin (2002). The six categories and the number of representatives in each category are: Antibody-antigen (13), Protease-inhibitor (11), Enzyme complexes (13), Large protease complexes (7), G-proteins, cell cycle, signal transduction (16) and Miscellaneous (17). Because of the low level of sequence identity, the resulting data set is more challenging than the data sets used in previous studies by our group (Yan et al., 2003) as well as by other authors (Ofran and Rost, 2003b). The list of 77 proteins is available at http://www.public.iastate.edu/~chhvan/ISMB2004/list.html.

Definition of surface residue and interface residues

The definition of interface residues used in this study is based on the reduction of solvent accessible surface area (ASA) upon complex formation. ASA was computed for each
residue in the unbound molecule (MASA) and in the complex (CASA) using the DSSP program (Kabsch and Sander, 1983). A residue is defined to be a surface residue if its MASA is at least 25% of its nominal maximum area as defined by Rost and Sander (1994). A surface residue is defined to be an interface residue if its calculated ASA in the complex is less than that in the monomer by at least 1Å² (Jones and Thornton, 1996). Surface residues were extracted and divided into interface residues and non-interface residues, using structural information from PDB files. We obtained a total of 2340 positive examples corresponding to interface residues and 5091 negative examples corresponding to non-interface residues.

**Analysis of interface residue neighborhoods**

Let $P_{\text{actual}}$ be the observed probability that a given neighbor of an interface residue is also an interface residue. Let $P_{\text{background}}$ be the probability that this position has an interface residue by chance. The log likelihood of the residue for this position belonging to an interface is given by $\log_2(P_{\text{actual}}/P_{\text{background}})$. Positive values for likelihood indicate that the residue under consideration has probability greater than that expected by chance of being an interface residue. Negative likelihood indicates the opposite. A likelihood of 0 indicates that the probability that the residue is likely to be an interface residue is the same as what we would expect based simply on the fraction of residues in the data set that are interface residues.

**The two-stage classifier**

In designing the two-stage classifier we exploit the observation that interface residues tend to form clusters on amino acid sequence (Ofran and Rost, 2003b). In the first stage, a SVM classifier is trained to identify interface residues based on the identities of neighboring residues of the target residue. The input to the SVM is an encoding of the identities of 9 contiguous amino acid residues, corresponding to a window containing the target residue and 4 neighboring residues on either side of the target residue. Each of the 9 residues in the window is represented by a 20-bit vector (with one bit for each letter of the 20-letter amino acid alphabet). Thus, the SVM classifier accepts $9 \times 20 = 180$-bit vector as input and produces
a Boolean output (with 1 denoting an interface residue and 0 denoting a non interface residue). Our study used the SVM in the Weka package from the University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/~ml/weka/) (Witten and Frank, 1999). The package implements John C. Platt's (1998) sequential minimal optimization (SMO) algorithm for training a support vector classifier using scaled polynomial kernels.

In the second stage, a Bayesian network classifier is trained to identify interface residues based on the class labels (1 for interface or 0 for non-interface) of its neighbors. The inputs for Bayesian classifier are the class labels of the 8 residues surrounding the target residue (4 on each side). The Bayesian network classifier is trained to output the most likely class label for the target residue given the class labels of its neighboring residues. We used is the BayesNetB from the Weka package, which implements hill climbing algorithm to learn the Bayesian network structure (Buntine, 1991). (We found that on this data set, the Naive Bayes classifier performs as well as a more complex classifier that models the dependencies among the neighboring residues).

Let C be a Binary random variable that denotes the class label (1 for an interface residue, 0 for a non-interface residue) for the target residue. Let Z be a vector-valued random variable that denotes the input to the 2-stage classifier (i.e., a Binary encoding of the target residue and its sequence neighbors). The two stage classifier classifies the target residue as an interface residue if:

\[
\frac{P(C=1|Z)}{P(C=0|Z)} > \theta
\]

The schematic of the 2-stage classifier is shown in Figure 1. If \( \theta = 1 \), this procedure corresponds to assigning the most probable class label (maximum a posteriori classification) for the target residue. Varying \( \theta \) corresponds to trading off specificity against sensitivity of interface residue prediction (see Figure 4 under Experimental Results). We choose \( \theta \) so as to maximize the correlation coefficient (see below) which measures the agreement between the actual and predicted class labels on the training data. The resulting classifier is then used to predict whether or not a target residue is likely to be an interface residue based on its identity and the identities of its 8 sequence neighbors.
Five-fold cross-validations

The classifier was evaluated using five-fold cross-validations. The examples from the 77 proteins were randomly divided into five subsets. In each round of experiment, four subsets were used for training and the remaining subset was used as test set. Negative examples in the training set were randomly deleted so that the training set had equal numbers of positive and negative examples. There were five rounds of such experiments in a five-fold cross-validation.

Performance measures

Let TP = the number of true positives (residues predicted to be interface residues that actually are interface residues); FP = the number of false positives (residues predicted to be interface residues that are in fact not interface residues); TN = the number of true negatives; FN = the number of false negatives; N = TP+TN+FP+FN (the total number of examples). Then we have:

\[
\text{Sensitivity}^+ = \frac{TP}{TP + FN} ; \quad \text{Specificity}^+ = \frac{TP}{TP + FP} ;
\]

\[
\text{Accuracy} = \frac{TP + TN}{N} ; \quad \text{and} \quad \text{Correlation Coefficient} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}.
\]
Sensitivity\(^+\) (sensitivity for interface residue class) measures the fraction of interface residues that are identified as such. Specificity\(^+\) (specificity for the interface residue class) measures the fraction of the predicted interface residues that are actually interface residues. Accuracy of a classifier measures the estimated probability of correct predictions. Correlation coefficient (CC) is a measure of how well the predicted class labels correlate with the actual class labels. It ranges from -1 to 1 where a correlation coefficient of 1 corresponds to perfect predictions, and a correlation coefficient of 0 corresponds to random guessing. Note that the commonly used measure of accuracy is not a particularly useful measure for evaluating the effectiveness of a classifier when the distribution of samples over different classes is unbalanced (Baldi et al., 2000). Average values of specificity and sensitivity are given by:

\[
\text{Average Specificity} = \frac{1}{2} (\text{Specificity}^+ + \text{Specificity}^-) \\
\text{Average Sensitivity} = \frac{1}{2} (\text{Sensitivity}^+ + \text{Sensitivity}^-)
\]

**EXPERIMENTS AND RESULTS**

**Interface residues tend to form clusters on amino acid sequences**

Of\'fran et. al. (2003b) investigated the sequence neighborhood of protein-protein interface residues in a set of 333 proteins and reported that 98% of protein-protein interface residues have at least one additional interface residue within 4 positions N or C terminal and 74% have at least 4. Among the 77 proteins we used here, 44 are also in Of\'fran dataset. For the 77 proteins, we obtained similar results: 97% of interface residues have at least one additional interface residue, and 70% of the interface residues have at least 4 interface residues within 4 positions on either side. For each interface residue, we analyzed the likelihood that its sequence neighbors are also interface residues. The results are shown in Figure 2. Close neighbors of an interface residue have a high likelihood of being interface residues. The closer a sequence neighbor is to an interface residue, the greater is its likelihood of being an interface residue. When the distance increases to 16 residues, the likelihood drops to 0. The observation that the interface residues tend to form clusters on the primary sequence...
suggests the possibility of detecting protein-protein interface residues from local sequence information.

Based on the results shown in Figure 2, a window size of 9 contiguous residues centered on the target residue was empirically determined to be optimal (data not shown) for constructing the two-stage classifier.

![Figure 2. The likelihood that positions neighboring interface residues also contains interface residues. Position 0 is an interface residue. Negative positions are on N terminal side of this target residue, positive positions are C terminal. Positive likelihood means that the position has higher probability than random of also being an interface residue.](image)

**Classification of surface residues from 77 proteins into interface residues and non-interface residues**

The two-stage classifier was evaluated using the dataset of 77 proteins in a five-fold cross validation experiment. Table 1 shows the classification performance as measured by correlation coefficient, accuracy, specificity and sensitivity. The correlation coefficient was maximized by choosing $\theta = 1$. The resulting classifier achieved an overall accuracy of 72% with a correlation coefficient of 0.30. The standard deviation of accuracy is 2% and that of correlation coefficient is 0.04. 58% of the residues predicted to be interface are actually interface residues, and 39% of interface residues are identified as such. We also investigated the fraction of interface residues in each protein that are correctly identified by
the classifier. Our results show that in 65 out of 77 (84%) proteins the classifier can recognize at least 20% of interface residues.

To examine whether the two-stage method learns sequence characteristics that are predictive of target residue functions, we ran a control experiment in which the class labels were randomly shuffled to destroy the attributes-class relationship in the original dataset. The correlation coefficient obtained on the class label shuffled data set is -0.01 (as compared to 0.30 on the original data set) indicating that the two-stage classifier performs significantly better than a random predictor (correlation coefficient = 0) (Table 1).

Table 1. Classification performance on a dataset of 77 proteins based on five-fold cross validation.

<table>
<thead>
<tr>
<th></th>
<th>Two-stage method</th>
<th>Gallet’s method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original dataset</td>
<td>Randomized dataset</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.30</td>
<td>-0.01</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.72</td>
<td>0.53</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.58</td>
<td>0.31</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.39</td>
<td>0.37</td>
</tr>
</tbody>
</table>

1 Class labels were not shuffled (i.e., these are original class labels extracted from PDB structure files).

2 Class labels were randomly shuffled for all the examples before training and testing the classifiers.

Comparison with Gallet’s method

Previously Gallet et al. (2000) published a method to identify interface residues using an analysis of sequence hydrophobicity based on earlier work of Eisenberg et al. (1984). For direct comparison, we evaluated Gallet method using five-fold cross validation on the same dataset that was used to evaluate our two-stage classifier. We used an input window size of 5
for the Gallet method, which is the window size reported to perform best (Gallet et al., 2000). The results shown in Table 1 indicate that the two-stage method achieves a much higher accuracy, correlation coefficient, and specificity* than Gallet method, thereby outperforming Gallet method in overall classification, although the Gallet method achieves slightly higher value sensitivity*. Notably, the correlation coefficient for the Gallet’s method is -0.02 – very close that of a random predictor.

Two-stage Classifier yields substantially more accurate interface residue predictions than the one-stage SVM classifier

Previously we reported an SVM method to identify interface residues (Yan et al., 2003). The two-stage method reported here combines an SVM and a Bayesian classifier. Table 2 shows the performance enhancement achieved by the two-stage method. Comparison of the performance shows that the Bayesian method (the second stage) helps improve the classification: correlation coefficient increases from 0.19 to 0.30, accuracy increases from 0.66 to 0.72, and specificity* increases from 0.44 to 0.58; although sensitivity* decreases slightly from 0.43 to 0.39. Thus we conclude that exploiting the distribution of interface and non-interface residues in the neighborhood of an interface residue can significantly improve the performance of classifiers for identifying interface residues.

Table 2. The performance of two-stage and one-stage classifier.

<table>
<thead>
<tr>
<th></th>
<th>SVM method</th>
<th>Two-stage method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.19</td>
<td>0.30</td>
</tr>
<tr>
<td>coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.66</td>
<td>0.72</td>
</tr>
<tr>
<td>Specificity*</td>
<td>0.44</td>
<td>0.58</td>
</tr>
<tr>
<td>Sensitivity*</td>
<td>0.43</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Figure 3. Representative prediction results on the 77 proteins. The target protein (for which the predictions are made) in each complex is shown in green, with residues of interest shown in spacefill and color coded as follows: red, interface residues identified as such by the classifier (true positives); yellow, interface residues missed by the classifier (false negatives), and blue, residues incorrectly classified as interface residues (false positives). For clarity, interface residues for the partner protein in each complex (gray wireframe) are not shown. A1, B1 are the predictions of SVM method. A2, B2 are the corresponding predictions of two-stage method on the same proteins. A1, A2: Predictions on BARSTAR from PDB 1brs; B1, B2: Predictions on SEB from PDB 1seb; Structure diagrams were generated using RasMol (http://www.openrasmol.org/) .

Evaluation of the predictions in the context of three-dimensional structures

To further evaluate the performance of the classifier, we examined predictions in the context of the three-dimensional structures of heterocomplexes. Two representative prediction results are shown in Figure 3. For comparison, the prediction results for both the SVM method alone (the first stage) and two-stage method are shown. The 1st and 7th best (out of 77 proteins) predictions (in term of correlation coefficient) are shown in 3A, B respectively. Figure 3A1, B1 are the predictions of SVM method. Figure 3A2, B2 are the corresponding predictions of two-stage method on the same proteins. Figure 3A1, A2 show the predictions on BARSTAR from PDB 1brs, which is the complex of BARNASE and BARSTAR. On BARSTAR the SVM method identified 8 interface residues with 1 false positive (Figure 3A1), whereas two-stage method identified 16 interface residues with 0 false positives (Figure 3A2). Figure 3B1, B2 show the predictions on SEB from an MHC protein-
antigen complex (PDB 1seb), which is the structure of SEB bound by HLA-DR1. On SEB the SVM method identified 12 interface residues but with 20 false positives (Figure 3B1), whereas two-stage method identified 13 interface residues with only 7 false positives (Figure 3B2). The results show that the two-stage classifier can successfully identify interface residues with fewer false positives than the SVM classifier above. The correctly identified interface residues (residues in red) form contiguous patches on surface. With this level of success, such predictions could be valuable for guiding experimental investigations into the roles of specific residues of a protein in its interaction with other proteins or for limiting search space for docking studies.

**Specificity-sensitivity tradeoff**

In some situations (e.g., identification of critical interface residues for site-specific mutagenesis) it is desirable to predict interface residues with very high specificity. This requirement can be met by modifying the parameters used by the two-stage classifier. In the results presented so far the two-stage classifier labels a target residue as an interface residue if \( p(1|z)/p(0|z) > 1 \). As noted above, we can calibrate the cutoff to increase the specificity of interface residue predictions (specificity\(^*\)) at the expense of reduced coverage (sensitivity\(^*\)). Figure 4 shows the Specificity\(^*\) v/s Sensitivity\(^*\) plot of the predictions when different cutoffs are used. When we increased the cutoff to 8, the specificity of interface residue predictions (specificity\(^*\)) increases to 0.85 and sensitivity\(^*\) decreases to 0.05. That is, 85% of the residues predicted to be interface residues are actually interface residues and although only 5% of the interfaces residues are identified as such. Alternatively, if it is important to identify more potential interface residues (even at expense of confidence), 60% interface residues can be identified with 50% specificity\(^*\).

**Evaluation of the two-stage classifier on CAPRI targets**

To further evaluate the two stage classifier, we used the our dataset of interface and non-interface residues from the 77 proteins as a training set and used the resulting classifier to identify interface residues in CAPRI targets. At the time this study was performed, 7 CAPRI targets (target 01 through target 07) were available. A representative result is shown in
Figure 5: the prediction on Fab HC63 in target 03 which is the complex of Fab HC63 and hemagglutinin. On Fab HC63, the two-stage method identified 10 interface residues with 10 false positives.

Figure 4. Specificity* v/s Sensitivity* plot of the two-stage method.

Figure 5. Test results on Fab HC63 in CAPRI target 03. Fab HC63 is shown in green, with residues of interest shown in spacefill and color coded as follows: red, true positives; yellow, false negatives, and blue, false positives. For clarity, interface residues for hemagglutinin (gray wireframe) are not shown. Structure diagrams were generated using RasMol (http://www.openrasmol.org/).
DISCUSSION

Development of accurate and robust computational methods for identification of protein-protein interface residues from amino acid would contribute to elucidation of protein sequence-structure function relationships, with the attendant benefits in a number of applications including drug design. Several approaches for predicting of interface residues from amino acid sequence, protein structure, or both have been explored with varying degrees of success. Methods that predict interface residues from amino acid sequence alone, or using amino acid sequence along with the structure of the target protein (but not the structure of the complex it forms with another protein) are of interest because relatively few experimentally determined structures of protein-protein complexes are currently available. In this paper, we have described a machine learning approach to constructing a two-stage classifier for classifying protein surface residues into interface and non-interface residues. The first stage consists of an SVM classifier. A Bayesian classifier is used at the second stage. The Bayesian classifier exploits the observation that interface residues tend to form contiguous or nearly contiguous clusters along the protein sequence. When trained and tested using five-fold cross validation on a non-redundant set of 77 proteins (with sequence identity below 30%) selected from hetero complexes, the method achieved 72% accuracy with a correlation coefficient of 0.3, 66% average specificity and 65% average sensitivity. The specificity of interface residue predictions (Specificity') was 58% and sensitivity (Sensitivity') was 39%. Our results also show that the two-stage classifier that combines the SVM method with the Bayesian network classifier achieves better performance (correlation coefficient = 0.3, accuracy = 0.72) than a single stage SVM classifier (correlation coefficient = 0.19, accuracy = 0.66).

It is worth noting that the two-stage classifier trained using our method on a subset of 77 proteins also performed reasonably well in terms of identifying interface residues of CAPRI targets despite the fact that no information from the CAPRI targets was used in training the classifier. Taken together our experiments show that the two-stage approach which exploits the observation that interface residues tend to form contiguous or nearly contiguous clusters on protein sequences significantly outperforms the SVM classifier.
To the best of our knowledge, the methods proposed by Gallet (2000) and Ofran (2003b) represent the only fully sequence-based approaches to prediction of interface residues that have been evaluated on datasets consisting of more than a handful of proteins. These two methods predict interface residues by directly classifying all residues (including surface as well as core residues) into interface residues and non-interface residues whereas the methods reported in this paper classify surface residues into interface residues and non-interface residues. This is especially useful in cases where the structure of the target protein is known although the structure of the complex(es) formed by it with one or more other protein(s) is unknown. For direct comparison, we implemented the Gallet method and used it to classify the same dataset of surface residues used here into interface residues and non-interface residues. The results of our experiments show that the two-stage method presented here outperforms Gallet method on this dataset. Further comparisons of the method of Gallet and of Ofran, with and without a second stage Bayesian classifier, with the methods described in this paper, on a broader range of data sets is clearly of interest.

Two points should be emphasized in evaluating the significance of these and other interface prediction results. First, it is important to note that the numbers of true positive, false positive, true negative and false negative predictions taken together provide all the relevant information for evaluating a classifier. Specificity, Sensitivity, Accuracy, and Correlation Coefficient offer different ways to summarize these four numbers into a single measure of performance. As noted by Baldi et al. (2000), each of these measures, taken alone, yields only partial information about classifier performance. This problem is exacerbated when the data set has unequal numbers of positive examples and that of negative examples. For example, if 80% of the residues are non-interaction residues, then a predictor that always predicts a residue to be a non-interaction residue will have an accuracy of 0.80 (80%). However, such a predictor is useless for correct identification of interface residues. In such a scenario, correlation coefficient is a much better indicator of the performance of a method. In this context, it is worth noting that Gallet’s method shows a negative correlation coefficient that is close to zero (random prediction) on the data set used in this study.
Second, it should be pointed out that because any given protein can interact with multiple partners, some residues identified as false positives in performance assessment of our method, as well as the methods proposed by Gallet et al (2000) and Ofran and Rost (2003b), could in fact be residues that actually participate in contacts with protein(s) other than their known partners in the PDB file (or CAPRI targets).

Mucchielli-Giorgi et al. (1999) and Naderi-Manesh et al. (2001) have reported an accuracy of 85% in identifying surface residues based on amino acid sequence information using techniques for predicting solvent accessibility of residues. This raises the possibility of coupling our method with surface residue predictions to identify interface residues based on sequence information alone: first classify all residues into surface residues and core residues; then classify surface residues into interface residues and non-interface residues.

Evolutionary information in sequences has been used in sequence-based methods to identify interface residues (Pazos et al., 1997; Valencia and Pazos, 2003). It would be interesting to explore whether methods that exploit evolutionary information along with sequence identity (or biophysical properties of amino acid residues) would yield more accurate identification of interface residues from amino acid sequences. Alternative approaches to exploiting knowledge of the structure (or the predicted structural properties) of the target protein may also result in more accurate prediction of interface residues.

ACKNOWLEDGEMENTS

This research was supported in part by grants from the National Science Foundation (0219699), the National Institutes of Health (GM066387), and the Iowa State University Plant Sciences Institute.

REFERENCES


CHAPTER 3. IDENTIFYING AMINO ACID RESIDUES INVOLVED IN PROTEIN-DNA INTERACTIONS FROM SEQUENCE

A paper submitted to *BMC Bioinformatics*

Changhui Yan, Michael Terribilini, Feihong Wu, Robert L. Jernigan, Drena Dobbs, and Vasant Honavar

ABSTRACT

Understanding the molecular details of protein-DNA interactions is critical for deciphering the mechanisms of gene regulation. We present a machine learning approach for identification of amino acid residues involved in protein-DNA interactions. We start with a Naïve Bayes classifier trained to predict whether a given amino acid residue is a DNA-binding residue based on its identity and the identities of its sequence neighbors. Our results indicate the feasibility of identifying interface residues based on local sequence information. The performance of the classifier can be improved by using solvent accessibility and sequence entropy of the target residue as additional inputs. Using a non-redundant dataset of 56 proteins, the classifier achieves 77% overall accuracy with a correlation coefficient of 0.30, 39% specificity and 52% sensitivity in identifying interface residues as estimated by leave-one-out cross-validation. Examination of the predictions in the context of 3-dimensional structures of proteins demonstrates the effectiveness of this method in identifying DNA-binding sites from sequence information. In 59% of the proteins, the classifier recognizes the interaction sites by correctly identifying at least half of the interface residues, and in 95% of the proteins, at least 20% of the interface residues are correctly identified. Analysis of the trained classifiers suggests that the proposed methods may offer useful insights into sequence correlates of protein-DNA interactions.
INTRODUCTION

Protein-DNA interactions play a pivotal role in gene regulation. The ability to identify amino acids that contribute to the specificity and affinity of the interactions can significantly improve our understanding of macromolecular functions and has broad applications in drug discovery (Blancafort, et al., 2004; Ghosh and Papavassiliou, 2005). Discovery of the principles of protein-DNA interactions has been a topic of wide interest for many years (Pabo and Sauer, 1992). Understanding these principles requires experimental analysis of the structures of protein-DNA complexes in order to understand the molecular details of specific residue-residue contacts that mediate protein-DNA recognition (Laity, et al., 2001; Muller, 2001; Lawson, et al., 2004). In addition to biophysical methods for structure determination, biochemical and molecular genetic approaches have been widely used to identify DNA-binding sites on proteins and to investigate the interaction modes between proteins and DNA. For example, alanine-scanning mutagenesis has been used to identify the amino acids important for target recognition by the m5C methyltransferase (Radlinska, et al., 2005) and to distinguish specific amino acids important for DNA binding and transcription activation by SoxS (Griffith and Wolf, 2002). Methods for precisely identifying protein-DNA contacts by coupling photochemical crosslinking with mass spectrometry have also been developed recently (e.g., Geyer, et al., 2004).

With proteomics and genomics projects producing protein sequence data at increasing rates, computational tools that can identify DNA-binding sites on a large scale and with high accuracy are urgently needed. Hence, there has been significant recent interest in developing computational methods for identification of amino acid residues that participate in protein-DNA interactions based on combinations of sequence, structure, evolutionary information, and chemical or physical properties. For example, Jones et al. (2003) analyzed residue patches on the surface of DNA-binding proteins and used electrostatic potentials of residues to predict DNA-binding sites. They recently applied this method to the identification of three specific classes of DNA-binding proteins, based on the presence of solvent accessible DNA-binding structural motifs (Shanahan, et al., 2004). In related work, Tsuchiya et al. (2004) used a structure-based method to identify protein-DNA binding sites based on
electrostatic potentials and surface shape, and Keil et al. (2004) trained a Neural Network classifier to identify patches likely to be DNA-binding sites based on physical and chemical properties of the patches. Neural Network classifiers have also been used to identify protein-DNA interface residues based on a combination of sequence neighbor and structure information (Ahmad, et al., 2004). More recently, Ahmad and Sarai have proposed a sequence-based method for predicting DNA-binding residues that incorporates sequence alignment profiles into the input (Ahmad and Sarai, 2005).

In this paper, we present a machine learning approach to identify amino acid residues that are involved in protein-DNA interactions. First, a Naïve Bayes classifier is trained to predict whether or not a residue is an interface residue. The input to the classifier consists of the identities of the target residue and 4 sequence neighbors on each side of the target residue. The classifier is trained and tested using a non-redundant set of 56 proteins extracted from protein-DNA complexes that were used in the study of Jones et al. (2003). The classifier achieves 77% overall accuracy with a correlation coefficient of 0.25, and 37% specificity with 43% sensitivity in identifying interface residues as evaluated by leave-one-out cross-validation. The performance improves when either relative solvent accessibility or sequence entropy of the target residue is included as an additional input to the classifier. When both relative solvent accessibility and sequence entropy are included, the Naïve Bayes classifier achieves 77% overall accuracy with a correlation coefficient of 0.30, 39% specificity and 52% sensitivity. Examination of the predictions in the context of 3-dimensional structures of proteins demonstrates that the predictions correctly indicate the locations of DNA-binding sites. When the Naïve Bayes classifier is trained using only sequence identities as input, the prediction results overlap with 34 of the 37 PROSITE (Hulo, et al., 2004) DNA-binding motifs that appear in the dataset. This raises the possibility of automated identification of potential DNA-binding motifs using a Naïve Bayes classifier trained to predict DNA-binding residues.
METHODS AND MATERIALS

Dataset A

We used a dataset of 56 double-stranded DNA (dsDNA) binding proteins used in a previous study by Jones et al. (2003). This dataset (which we will refer to as dataset A) is derived from 427 protein-DNA complexes with resolution better than 3.0 Å that were extracted from the Nucleic Acid Database (NDB) (Berman, et al., 1992). Proteins in the complexes were clustered into homologous families and a non-redundant representative set of 56 proteins was obtained (Jones, et al., 2003).

Dataset B

The second dataset used in our study consists of 62 DNA binding proteins (PDNA-62). This dataset (which we will refer to as dataset B) was used in a previous study by Ahmad and Sarai (2005). There is an overlap of 15 proteins between dataset A and B. We evaluated the Naïve Bayes classifier on both datasets and the Naïve Bayes classifier achieved comparable performance in the two datasets.

Definition of interface residues

Interface residues are defined as described in Jones et al. (2003). Accessible surface area (ASA) was computed for each residue in the unbound protein (in absence of DNA) and in the protein-DNA complex using NACCESS (Hubbard, 1993). A residue is defined to be an interface residue if its ASA in the protein-DNA complex is less than its ASA in the unbound protein by at least 1Å². The 56 proteins from dataset A have 12,665 residues in total and 1,752 of them are interface residues. The 62 proteins in dataset B have 12,803 residues in total and 1,508 of them are interface residues. The interface residues in dataset B were provided by Ahmad and Sarai (2005). In their study, the residues within a distance of 3.5 Å from DNA are defined as interface residues (Ahmad, et al., 2004).
Naïve Bayes classifier

We used the Naïve Bayes implementation in the Weka package from the University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/~ml/weka/) (Witten and Frank, 1999). For each input target residue, the classifier produces a Boolean output (with 1 denoting an interface residue and 0 denoting a non-interface residue). The Naïve Bayes classifier assumes independence of the attributes given the class. The Naïve Bayes classifier performs as well as more sophisticated methods on many classification tasks (Buntine, 1991). For an input\(X = x_1, x_2, \ldots, x_n\), a Naïve Bayes classifier assigns it a class label by optimizing the posterior: \(c = \arg \max_c P(c \mid X = x_1, x_2, \ldots, x_n) = \arg \max_c P(c) \prod_{i=1}^{n} P(x_i \mid c)\). In the case of two class classification (\(c \in \{0, 1\}\)), this is equivalent to determining \(c\) by comparing the ratio likelihood with a parameter \(\theta\) as in equation (1).

\[
\frac{P(c = 1 \mid X = x_1, x_2, \ldots, x_n)}{P(c = 0 \mid X = x_1, x_2, \ldots, x_n)} = \frac{P(c = 1) \prod_{i=1}^{n} P(x_i \mid c = 1)}{P(c = 0) \prod_{i=1}^{n} P(x_i \mid c = 0)} > \theta
\]

\(c\) is predicted to be 1 if the ratio likelihood is greater than \(\theta\), and 0 otherwise. In a standard Naïve Bayes classifier, \(\theta\) takes the value of 1. The predictions of Naïve Bayes classifier are biased in favor of the majority class when the dataset consists of unequal numbers of examples for the two classes. Hence, we trained \(\theta\) to optimize classification performance on training data. We used leave-one-out cross-validation to train and test the classifier. In each round of experiment, all proteins except one were used as the training set and the remaining protein was used to test the classifier. In the training stage, the conditional probability table \(P(x_i \mid c)\) and prior probability \(P(c)\) were estimated using the training set. To determine \(\theta\), the classifier was applied to the training set and different values of \(\theta\) ranging from 0.01 to 1 were tested, in increments of 0.01. The value of \(\theta\) for which the classifier yields the highest correlation coefficient was used to make predictions on the test set.
Naïve Bayes classifier using only local sequence identity as input

The input to the Naïve Bayes classifier contains the identities of \(2n+1\) residues in the form of \(X = (x_{i-n}, x_{i-n+1}, ..., x_{i-1}, x_{i}, x_{i+1}, ..., x_{i+n}, x_{i+n+1})\), where \(x_{i}\) is the identity of target residue, \(x_{i-n}, x_{i-n+1}, ..., x_{i-1}\) and \(x_{i+1}, ..., x_{i+n}, x_{i+n+1}\) are the identities of \(n\) residues on each side of the target residue. Different values of \(n\) from 1 to 10 were tried and the best performance was obtained when \(n=4\) (corresponding to a window size of 9). A training example is an ordered pair \((X, c)\), where \(c \in \{0, 1\}\). 1 indicates that the target residue (the residue in the center of the input window) is an interface residue and 0 indicates that target residue is not an interface residue. For a test example \(X\), the classifier outputs 1 (i.e., \(X\) is predicted to be an interface residue) or 0 (i.e., \(X\) is predicted to be a non-interface residue) as the class label of \(X\).

Naïve Bayes classifier using local sequence identity plus additional features

To improve the classifier’s performance, features in addition to sequence identity were explored. These features include relative solvent accessibility (rASA), sequence entropy, secondary structure, electrostatic potential and hydrophobicity. When a feature of the target residue is added into the input of identities as additional information, a training example is encoded as \(X = (x_{t-n}, x_{t-n+1}, ..., x_{t-1}, x_{t}, x_{t+1}, ..., x_{t+n}, x_{t+n+1}, f_t, c)\), with \(f_t\) replaced by corresponding feature of the target residue, and \(x_t\) and \(c\) are defined as above. A test example is a vector \(X = (x_{t-n}, x_{t-n+1}, ..., x_{t-1}, x_{t}, x_{t+1}, ..., x_{t+n}, f_t)\). When a selected feature of the input window (e.g., the target residue and its sequence neighbors) is used as input, a training example is in the form of \(X = (f_{i-n}, f_{i-n+1}, ..., f_{i-1}, f_i, f_{i+1}, ..., f_{i+n}, c)\) and a test example is a vector of \(X = (f_{i-n}, f_{i-n+1}, ..., f_{i-1}, f_i, f_{i+1}, ..., f_{i+n})\), where \(f_i\) is the corresponding feature of residue \(i\). rASA of each residue (in the absence of DNA) was computed using NACCESS (Hubbard, 1993). Entropy of each sequence position was extracted from the HSSP database (http://www.cmbi.kun.nl/gv/hssp/). The entropy is normalized to the range of 0-100, with lower entropy values corresponding to more conserved sequence positions. Secondary structure for each residue was extracted from the PDB database.
Electrostatic potential for each atom was calculated using Delphi (Rocchia, et al., 2001; Rocchia, et al., 2002), using parameters based on the study of Jones et al. (2003). The electrostatic potential for each residue was calculated in a similar way as the study of Jones et al. (2003): the electrostatic potential of an atom is set to 0 if its solvent accessibility is less than 1 Å² and the electrostatic potential of a residue is the average over all its atoms. Hydrophobicity of each residue is obtained from the consensus normalized hydrophobicity scale derived by Eisenberg et al. (1984).

**Performance measures**

To evaluate the performance of the classifier, accuracy, correlation coefficient (CC), specificity and sensitivity were used as performance measures. These measures are defined as described in Baldi et al. (2000). Accuracy = \( \frac{TP + TN}{N} \); 
CC = \( \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FN)}} \); 
Sensitivity = \( \frac{TP}{TP + FN} \); 
Specificity = \( \frac{TN}{TN + FP} \), where:

- \( TP \) = the number of true positives (residues predicted to be interface residues that actually are interface residues);
- \( FP \) = the number of false positives (residues predicted to be interface residues that are in fact not interface residues);
- \( TN \) = the number of true negatives (residues predicted to be non-interface residues that actually are non-interface residues);
- \( FN \) = the number of false negatives (residues predicted to be non-interface residues that are in fact interface residues);
- \( N = TP + TN + FP + FN \) (the total number of examples).

Sensitivity is the fraction of positive examples (interface residues) that are predicted as such by the classifier. Specificity is the fraction of positive predictions (residues predicted to be interface residues) that are actually interface residues. Accuracy is the fraction of overall predictions that are correct. Correlation coefficient measures the correlation between predictions and actual class labels. In the evaluation of classifiers, we consider all the measures mentioned above, with an emphasis on correlation coefficient, which is believed to be a better measure than the others when the numbers of positive and negative examples are unequal (Baldi, et al., 2000).

The Receiver Operating Characteristic curve (ROC curve) is a plot of the "hit rate" \( \frac{TP}{TP + FN} \) versus the "false alarm rate" \( \frac{FP}{TN + FP} \) (Baldi, et al., 2000). It shows the
tradeoff between hit rate and false alarm rate when different threshold values are used for
the classifier.

**Identifying PROSITE motifs in protein sequences**

The PROSITE motif database was downloaded from PROSITE (http://us.expasy.org/prosite/). Protein sequences were scanned using the ps-scan program (http://us.expasy.org/prosite/) to identify motifs. Frequently matching (unspecific) patterns and profiles were omitted by setting the "-s" and "-r" options of ps-scan.

**EXPERIMENTS AND RESULTS**

**Identification of interface residues based on local sequence information**

A Naïve Bayes classifier was trained to identify amino acid interface residues in protein-DNA complexes based on local protein sequence information. It was evaluated using dataset A, which contains the 56 proteins used in the study of Jones et al. (2003). Leave-one-out cross-validation experiments were used to evaluate the performance of each classifier. Several input encodings based on local sequence information were tried, with input consisting of: (a) the identities of 9 amino acid residues, corresponding to a window containing the target residue and 4 neighboring residues on each side of the target residue; (b) the identities of 9 amino acid residues and the relative accessible surface area (rASA) of the target residue; (c) the identities of 9 amino acid residues and the sequence entropy of the target residue; and (d) the identities of 9 amino acid residues, the rASA and entropy of the target residue.

Table 1 shows that the classifier using amino acid identities as input achieved an overall accuracy of 77% with a correlation coefficient of 0.25, 37% of the residues predicted to be interface are actually interface residues, and 43% of interface residues are correctly identified. Adding either rASA or entropy of the target residue into the input improved the performance (Table 1). When both were added, the Naïve Bayes classifier achieved an overall accuracy of 77% with a correlation coefficient of 0.30, 39% specificity, and 52% sensitivity. In 59% (33 of 56) of the proteins, the classifier recognizes the interaction site by
identifying at least half of the interface residues, and in 95% (53 of 56) of the proteins, at least 20% of the interface residues are correctly identified.

Other features of the target residue, including secondary structure, electrostatic potential and hydrophobicity, were also explored. Adding these features of the target residue into the input of identities did not improve the performance (data not shown). When a feature (other than identity) of the target residue and its sequence neighbors was used instead as input, the classifiers achieved performance lower than that of the classifier using identities as input (data not shown).

Table 1. Prediction of interface versus non-interface residues using a Naive Bayes classifier.

<table>
<thead>
<tr>
<th></th>
<th>Sequence-based</th>
<th>Sequence/structure-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Identities (ID)</td>
<td>ID + entropy</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>Correlation coefficient</td>
<td>0.25</td>
<td>0.29</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>37</td>
<td>37</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>43</td>
<td>53</td>
</tr>
</tbody>
</table>

a Input contains only the identities of 9 amino acid residues (the target residue and its 4 sequence neighbors on each side). b Sequence entropy of the target residue position is added into the input of identities. c rASA of target residue is added into the input of identities. d Both rASA and entropy of target residue are added into the input of identities.

Evaluation of the predictions in the context of 3-dimensional structures of proteins

To further evaluate the performance of the classifier, we examined predictions in the context of the 3-dimensional structures of the protein-DNA complexes. A Naïve Bayes classifier was trained using as input the identities of 9 residues plus rASA and entropy of the
target residue (ID + rASA + entropy). The 2nd and 19th best predictions (in terms of correlation coefficient) out of 56 proteins are shown in Figures 1 and 2, respectively.

Figure 1 shows the predicted and actual interface residues on the Pit-1 transcription factor from PDB structure lau7, which is a complex of Pit-1 and DNA (Jacobson, et al., 1997). On Pit-1, the classifier identified 30 interface residues, correctly indicating the location of the binding site. Pit-1 is a member of POU domain family of transcription factors. It consists of a POU-specific domain and a homeodomain that contact DNA with helix-turn-helix (HTH) motifs (Herr, et al., 1988; Jacobson, et al., 1997). 10 residues from Pit-1 form direct contacts with base pairs. Residues Ser-43, Gln-44, Thr-45 and Arg49 from the POU-specific domain and Arg-5, Arg-46, Asn-51 and Gln-54 from the homeodomain contact with the base pairs by hydrogen bonds (Jacobson, et al., 1997). All these residues except Asn-51 and Gln-54 were correctly identified as interface residues by the Naïve Bayes classifier. Residues Val-47 and Cys-50 make van der Waals contacts with the base pairs. They were missed by the classifier. Figure 2 shows the predicted and actual interface residues on the λ-Cro repressor protein from PDB structure 6cro, which is a complex of λ-Cro and the operator sequence (Albright and Matthews, 1998). On Cro, the classifier correctly identified 10 interface residues. Note that the predictions form a contiguous patch on protein surface and again, correctly indicate the location of the binding site.

ROC curve

In some situations (e.g., identification of critical interface residues for site-specific mutagenesis), it is desirable to predict interface residues with high precision at the cost of reduced coverage. In other situations, discovering more potential interface residues might be more useful. These different requirements can be met by modifying the threshold θ used by the Naïve Bayes classifier in this study. The Naïve Bayes classifier predicts a residue to be an interface residue if

\[ \frac{P(c = 1 \mid X = x_1x_2...x_n)}{P(c = 0 \mid X = x_1x_2...x_n)} > \theta \]

Different values of θ ranging from 0.01 to 1 with increments of 0.01 were tried. Figure 3 shows the Receiver Operating Characteristic curve (ROC curve) for the identification of interface residues.
Figure 1. Comparison of predicted versus actual binding sites on Pit-1 from PDB complex 1au7, the 2nd best out of the 56 proteins in terms of correlation coefficient. A: Predicted interface residues are shown in red. B: Actual interface residues are shown in red. Figures are generated using PyMOL (DeLano, 2002).

Figure 2. Comparison of predicted versus actual binding sites on λ-Cro protein from PDB 6cro, the 19th best out of the 56 proteins in terms of correlation coefficient. A: Predicted interface residues are shown in red. B: Actual interface residues are shown in red. Figures are generated using PyMOL (DeLano, 2002).

Figure 3. Receiver Operating Characteristic curve (ROC curve) for interface residue identification.
Figure 4. Comparison of actual and predicted DNA-binding site residues for Pit-1 (1au7A) with PROSITE POU domain motifs. PROSITE motifs POU_1 and POU_2 (bottom row) cover many of the actual interface residues (first row below sequence). Note that the predictions of Naïve Bayes classifier overlap with the PROSITE motifs, but more closely correspond to the actual interface residues.

Naïve Bayes classifier using only local sequence identities as input can discover DNA binding motifs

The results summarized above show that a Naïve Bayes classifier trained on a set of DNA binding proteins can successfully identify protein-DNA interface residues from amino acid sequence. This raises the question as to how the sequence features picked out by Naïve Bayes classifier relate to known DNA binding motifs. To explore this question, we use the ps_scan program to search for PROSITE motifs in the 56 proteins of dataset A. PROSITE motifs were found in 28 of the 56 proteins (a total of 44 hits). Of these 44 hits, 37 overlap with actual protein-DNA binding sites in dataset A. When the Naïve Bayes classifier was trained and tested using the identities of the target residues and 4 of its sequence neighbors on each side as input, the predictions produced substantially overlap with 34 of the 37 PROSITE DNA-binding motifs (Figure 4). It is worth noting that 28 of the 56 proteins, contain no PROSITE motif whose annotation suggests a role in protein-DNA interactions. PROSITE motifs cover more than 50% of interface residues in only 16% (9 out of 56) of the proteins and cover at least 20% of interface residues in only 30% (17 out of 56) of the proteins. By comparison, a Naïve Bayes classifier using only local sequence information as input identifies at least 50% of the interface residues in 43% (24 out of 56) of the proteins and at least 20% of the interface residues in 93% (52 out of 56) of the proteins. These results
raise the possibility of automated identification of potential DNA-binding motifs using a Naïve Bayes classifier trained to predict protein-DNA binding residues.

**Comparison with previously published methods**

To facilitate comparison of our approach with previously published methods, our studies focused on two datasets of protein-DNA complexes used in related studies by other groups. Our dataset B corresponds to the set of 62 DNA-binding proteins (PDNA-62) provided by the Sarai group, which they have used to develop neural networks (Ahmad, et al., 2004) or classifiers based on position specific scoring matrices (PSSMs) (Ahmad and Sarai, 2005) for predicting DNA-binding sites. To our knowledge, their study in PSSM-based classifiers is the only previously published study in which the performance of a DNA-binding site prediction method has been reported on "per residue" basis and only sequence-based information is used as input. We trained and tested a Naïve Bayes classifier based on this dataset using as input the identities of 9 amino acid residues. No further refinement (e.g., based on structure-derived features) was applied to the Naïve Bayes classifier predictions. The performance of the Naïve Bayes classifier was compared with that PSSM-based classifier developed by Ahmad and Sarai (Table 2). In the Ahmad and Sarai study, datasets with different levels of redundancy were used to derive PSSMs and the method obtained best performance with the PSSMs derived from PIR (Ahmad and Sarai, 2005). Table 2 shows the comparison of Naïve Bayes classifier performance with Ahmad and Sarai method’s best performance. The results show that the Naïve Bayes classifier achieves slightly better performance in all measures except sensitivity. It is worth noting that the PSSM method requires multiple alignments of homologous sequences for each protein, whereas only local sequence information is required by the Naive Bayes classifier. Hence, the Naïve Bayes classifier can be efficiently updated as new data become available.

It would be interesting to compare the performance of the methods presented here with that of Jones et al. (2003), in which electrostatic potentials (derived from structural information) were used in addition to the protein sequence to identify protein-DNA binding sites. Our dataset A is the same protein-DNA complex dataset used in their study, but Jones et al. evaluated their method in terms of surface patches and no individual residue
performance data were reported. Thus, at present, with the results reported here cannot be directly compared those of Jones et al. (2003).

Table 2. Comparison of prediction performance using Naïve Bayes classifier versus PSSM-based classifier (Ahmad & Sarai).

<table>
<thead>
<tr>
<th>Method</th>
<th>Naïve Bayes classifier</th>
<th>Ahmad and Sarai method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.26</td>
<td>0.23</td>
</tr>
<tr>
<td>Coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>80</td>
<td>66</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>29</td>
<td>21</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>48</td>
<td>68</td>
</tr>
</tbody>
</table>

a Naïve Bayes classifier using identities of 9 amino acids (the target residues and 4 of its sequence neighbors on both sides) as input.

b Ahmad method using PSSMs derived from PIR (Ahmad and Sarai, 2005).

DISCUSSION

In this study we present a computational method for identifying interface residues involved in protein-DNA interactions. Naïve Bayes classifiers were trained to predict whether a residue is an interface residue, based on local amino acid sequence information alone or on a combination of sequence and structure-derived information. Features including sequence entropy, relative solvent accessibility (rASA), secondary structure, electrostatic potential and hydrophobicity were explored to improve the performance. Performance was improved when either rASA or entropy of target residue was added to the input. When both were included, the classifier achieved 77% overall accuracy with a correlation coefficient of 0.30, 39% specificity and 52% sensitivity. Thus, this study demonstrates that it is feasible to identify protein-DNA interface residues based solely on local sequence information, but that
incorporating additional features derived from structural information can enhance performance.

Ahmad and Sarai (2005) used a PSSM-based neural network classifier to identify interface residues in protein-DNA interactions. Their method requires as input both the sequence of the query protein information and the information derived from multiple sequence alignment. Our results show that when tested on the same dataset, the Naïve Bayes classifier using only sequence information achieves performance comparable to that reported by Ahmad and Sarai. The interface residue predictions of the Naïve Bayes classifier overlap with those of PSSM-based classifier (data not shown), but each classifier detects some interface residues that are not detected by the other. This indicates an ensemble classifier comprising Naïve Bayes and PSSM-based classifiers may detect more interface residues than either individual method.

As might be expected, interface residue prediction by Naïve Bayes classifiers was enhanced when the sequence entropy and relative solvent accessibility of the target residue were included as additional input, indicating that these features are correlated with residue functionality. Incorporating additional structure-derived information such as electrostatic potential, hydrophobicity or secondary structure of the target residue as additional input, however, did not improve the performance in this study. This should not be taken to mean that these features are not useful predictors of a residue’s functionality. On the contrary, electrostatic potential has been shown to be useful in identification of protein-DNA interface residues (Jones, et al., 2003; Shanahan, et al., 2004). The fact that this information does not improve performance in our study may relate to the input encoding or the classification methods we used. In this study, we simply added features to the amino acid identity input as additional attributes and assume that these attributes are independent, given the class. This assumption almost certainly does not hold. Systematic analysis is needed to identify features that are useful for identification of interface residues and develop methods of representing them in input. Jones and Thornton (1997) analyzed six features of surface patches in protein-protein interaction sites and developed an approach to identify protein-protein interfaces based on the scores combining the six features. Sen et al. (2005) developed an ensemble method to identify protease-inhibitor binding sites based on sequence, structure and
evolution information. Similar approaches should be tested in the prediction of protein-DNA interaction sites.

Protein sequence motifs (defined here as sequence segments associated with specific protein functions or structural families) are often used to identify potential DNA binding domains. Discovery of such motifs requires alignment of protein sequences that are known to have the same or similar functions. This requires expertise to identify a suitable set of sequences and to manually adjust the multiple sequence alignments. In this study, no DNA-binding PROSITE motifs were found in 28 of 56 proteins. In the remaining proteins, 37 PROSITE motifs were found to overlap with actual protein-DNA binding sites. The predictions of the Naive Bayes classifier trained to identify interface residues using residue identities as input significantly overlapped with 34 of these 37 motifs. This suggests the possibility of identifying novel sequence motifs that correspond to protein-DNA interfaces by using a classifier trained to identify protein-DNA interfaces.

In previous work, we have used similar approaches to identify interface residues involved in protein-protein interactions (Yan, et al., 2004a; Yan, et al., 2004b) and protein-RNA interactions (Terribilini, et al., Submitted). Here we show that it is also feasible to identify interface residues involved in protein-DNA interaction using only local sequence information, although classification performance is enhanced slightly when relative solvent accessibility and sequence entropy of target residue are included in input. With the level of success achieved in this study (77% accuracy), such predictions should be useful for guiding experimental investigations into the roles of specific residues of a protein in its interaction with DNA, e.g., by localizing candidate residues for alanine-scanning mutagenesis (Griffith and Wolf, 2002; Radlinska, et al., 2005). Moreover, analysis of the binding site "rules" generated by classifiers may provide valuable insight into the protein-DNA recognition code responsible for the specificity and affinity of protein-DNA interactions in living cells.
ACKNOWLEDGEMENTS

This Research was supported in part by a grant from the National Institutes of Health (GM 066387) to VH, DD, and RLJ. We wish to thank Ahmad and Sarai for kindly providing their dataset (dataset B in this study). We thank O. Yakhnenko and D. Caragea for providing comments on the manuscript.

REFERENCES


Hubbard, S.J. (1993) NACCESS. Department of Biochemistry and Molecular Biology, University College, London.


CHAPTER 4. IDENTIFICATION OF DNA BINDING RESIDUES IN THE S SUBUNIT OF THE *M. jannaschii* TYPE I RESTRICTION-MODIFICATION SYSTEM

A paper to be submitted to *Proteins*

Changhui Yan, Jae-Hyung Lee, Robert L. Jernigan, Vasant Honavar, and Drena Dobbs

**ABSTRACT**

Restriction-modification (R-M) systems play important roles in the recognition and elimination of foreign DNA. In type I R-M systems, the S subunit determines the specificity of DNA recognition, but the interaction mode between S subunit and DNA is still unknown. Here we use a Naïve Bayes classifier to identify the DNA binding residues in the S subunit of the *M. jannaschii* type I R-M system. The predictions form four major patches on the protein surface. Three of them overlap with the protein-DNA interaction sites predicted in a previous study by Kim *et al.* (2005). The results can be used to guide the design of mutagenesis studies to identify the critical residues that contribute to the specificity and affinity of DNA binding in type I R-M systems.

**INTRODUCTION**

Restriction-modification (R-M) systems are enzyme systems found in bacteria that recognize and degrade foreign DNA (Bertani and Weigle, 1953). The typical R-M system consists of a modification enzyme (M subunit) that methylates specific bases on the target DNA sequence, and a restriction enzyme (R subunit), which is an endonuclease that cleaves DNA at specific sites (Smith, *et al.*, 1972). The M subunit sets "imprint" on the DNA that belonging to bacteria itself by methylating specific bases, and the R subunit degrades foreign DNA that lacks the "imprint" (Murray, 2000). R-M systems are classified into three
types (type I, II and III) based on composition, cofactor requirements, target sequences and
the positions of DNA cleavage sites (Wilson and Murray, 1991; Dreier, et al., 1996; Murray,
2000; Szczelkun, 2000). However some systems that do not fit into the 3 conventional types
have been reported (Janulaitis, et al., 1992; Cesnaviciene, et al., 2001; Jurenaite-Urbanaviciene, et al., 2001; Lepikhov, et al., 2001). Type I R-M system is distinguished
from types II and III in that it cuts DNA at a random position far away from the recognition
site, while type II cuts at specific site and type III cuts at a position close to the recognition
site. Type I R-M system is a hetero-oligmeric complex consisting of a specificity subunit (S
subunit) that recognizes specific DNA sequence, a modification subunit (M subunit) that
methylates the DNA sequence recognized by S subunit and a restriction subunit (R subunit)
that cuts DNA. The stoichiometry of type I R-M system has been suggested to be R2M2S1
(Sain and Murray, 1980). R and M subunits are relatively conserved within the family of
type I R-M system, but the S subunit has two variable target recognition domains (TRDs)
that recognize two specific DNA sequences of 3–5 bp separated by 6–8 bp (Yuan, 1981;
Gough and Murray, 1983; Endlich and Linn, 1985; Kannan, et al., 1989; Szczelkun, et al.,
1996). The two TRDs are separated by a relatively conserved region (CR). Another
conserved region (DCR) is located at the C terminal of the S subunit. CR and DCR are
believed to be sites that involved in protein-protein interactions between S subunit and other
subunits (Kneale, 1994).

Argos (1985) used sequence alignment to study the S subunit sequences of type I R-M
systems and suggested that S subunit consists of repeating DNA binding domains. Sturrock
and Dryden (1997) used a combination of sequence alignment and secondary structure
prediction to analyze 51 TRDs from S subunits and showed that TRDs of type I R-M
systems have a common tertiary structure. O'Neil et al. (1998) applied random mutagenesis
method to localize the protein-DNA interface on the S subunit of EcoKI, the first R-M
system identified in E. coli K-12, and discovered 5 residues that are likely at protein-DNA
interface: Gly-91, Lys-92, His-95, Ser-103 and Phe-107. Recently a crystal structure of the S
subunit (NP_247095, gi|15669898) from M. jannaschii, the first crystal structure of S
subunit of type I R-M system, has been solved by Kim et al. (2005). The structure is
available at the Protein Data Bank (PDB) with the id of 1YF2. The structure shows that the
S subunit contains two globular TRDs separated by a long α-helical conserved region (CR). Using Dali program (Holm and Sander, 1993), Kim et al. found that although no structural homology is found for the overall structure of the S subunit, the two TRDs (TRD1, residue 1–168 and TRD2, residue 209–378) are similar to the DNA binding domain of TaqI-MTase. By aligning TRD1 and TRD2 with TaqI-MTase/DNA complex, Kim et al. suggested that TRD1 and TRD2 each have four loops involved in DNA binding. So far no crystal structure of the S subunit/DNA complex has been obtained.

In our previous study, we trained a Naive Bayes classifier to identify protein-DNA interface residues using as input the amino acid identity of the target residue, the identities of 4 sequence neighbors of target residues on each sides, the relative accessibility of the target residue, and the entropy of the target residue derived from HSSP profile (http://www.cmbi.kun.nl/gv/hssp/) (Yan, et al., Submitted). The Naive Bayes classifier is trained using a dataset of 56 double-stranded DNA (dsDNA) binding proteins that were previously used in the study of Jones et al. (2003). In this study, we apply the classifier to identify DNA-binding residues in the S subunit of the type I R-M system from M. jannaschii (which we will refer to as S subunit in this study).

MATERIALS AND METHODS

Training dataset

We used a dataset of 56 double-stranded DNA (dsDNA) binding proteins that were previously used in the study of Jones et al. (2003). In their study, 427 protein-DNA complexes that have resolution better than 3.0 Å were extracted from Nucleic Acid Database (NDB). Proteins in the complexes were clustered into homologous families and a non-redundant representative set of 56 proteins was obtained (Jones, et al., 2003). We used the align program from the fasta2 package (ftp://ftp.ebi.ac.uk/pub/software/unix/fasta/) (Pearson and Lipman, 1988; Pearson, 1990) to exam the similarity between these proteins and the S subunit. The results show that none of the 56 proteins share sequence identify higher than 20% with the S subunit.
Naïve Bayes classifier

We used the Naïve Bayes in the Weka package from the University of Waikato, New Zealand (http://www.cs.waikato.ac.nz/~ml/weka/) (Witten and Frank, 1999). For each input target residue, the classifier produces a Boolean output (with 1 denoting an interface residue and 0 denoting a non-interface residue). The Naïve Bayes classifier assumes independence of the attributes given the class. The Naïve Bayes classifier performs as well as more sophisticated methods on many classification tasks (Buntine, 1991). For an input \( X = x_1x_2,...,x_n \), a Naïve Bayes classifier assigns it a class label \( c \) by optimizing the posterior: 

\[
\hat{c} = \arg\max_c P (c \mid X = x_1x_2...x_n) = \arg\max_c P (c) \prod_{i=1}^{n} P (x_i \mid c).
\]

In the case of two class classification (\( c \in \{0, 1\} \)), this is equivalent to determining \( c \) by comparing the ratio likelihood with a parameter \( \theta \) as in equation (1).

\[
\frac{P (c = 1 \mid X = x_1x_2...x_n)}{P (c = 0 \mid X = x_1x_2...x_n)} = \frac{P (c = 1) \prod_{i=1}^{n} P (x_i \mid c = 1)}{P (c = 0) \prod_{i=1}^{n} P (x_i \mid c = 0)} > \theta \tag{1}
\]

c is predicted to be 1 if the ratio likelihood is greater than \( \theta \), and 0 otherwise. In a standard Naïve Bayes classifier, \( \theta \) takes the value of 1. The predictions of Naïve Bayes classifier are biased in favor of the majority class when the dataset consists of unequal numbers of examples for the two classes. Hence, we trained \( \theta \) to optimize classification performance on training data. In the training stage, the conditional probability table \( P (x_i \mid c) \) and prior probability \( P (c) \) were estimated using the training set of 56 proteins. To determine \( \theta \), the classifier was applied to the training set and different values of \( \theta \) ranging from 0.01 to 1 were tested, in increments of 0.01. The value of \( \theta \) for which the classifier yields the highest correlation coefficient was used to make predictions on the S subunit.

Input to Naïve Bayes classifier

The input to the Naïve Bayes classifier contains the identities of \( 2n+1 \) residues, the relative solvent accessible area (rASA) of the target residue, and the sequence entropy of the target residues in the form of \( X = (x_{r-n}, x_{r-n+1},...,x_{r-1}, x_r, x_{r+1},...,x_{r+n}, e_r) \), where \( x_i \) is the
identity of the target residue, \( x_{r-n}, x_{r-n+1}, \ldots, x_{r-1} \) and \( x_{r+1}, \ldots, x_{r+n-1}, x_{r+n} \) are the identities of \( n \) residues on each side of the target residue, \( r_i \) is the rASA of the target residue computed using NACCESS (Hubbard, 1993) and \( e_i \) is the sequence entropy at the position of the target residue extracted from HSSP database (http://www.cmbi.kun.nl/gv/hssp/). The entropy is normalized to the range of 0-100, with lower entropy values corresponding to more conserved sequence positions. Different values of \( n \) from 1 to 10 were tried and the best performance was obtained when \( n=4 \) (corresponding to a window size of 9). A training example is an order pair \((X, c)\), where \( c \in \{0, 1\} \). 1 indicates that the target residue (the residue in the center of the input window) is an interface residue and 0 indicates that the target residue is not an interface residue. For a test example \( X \), the classifier outputs 1 (i.e., \( X \) is predicted to be an interface residue) or 0 (i.e., \( X \) is predicted to be a non-interface residue) as the class label of \( X \).

RESULTS

The predictions form patches on protein surface and overlap with the predictions of previous study

The S subunit consists of 425 amino acid residues. Naïve Bayes classifier predicts 67 residues to be interface residues. Figure 1A shows the predictions on the protein surface. In the figure we can see that the predicted interface residues form some contiguous patches on the surface. The 4 largest patches are: patch 1 consisting of 14 residues, patch 2 consisting of 8 residues, patch 3 and patch 4 consisting of 7 residues each. Each of the rest patches consists of no more than 4 residues. In their study, Kim et al. (2005) found that although no structural homology is found for the overall structure of the S subunit, the two target recognition domains (TRD1, residue 1–168 and TRD2, residue 209–378) are similar to the DNA binding domain of TaqI-MTase. By aligning the structures of TRD1 and TRD2 with the structure of TaqI-MTase/DNA complex, Kim et al. proposed an interaction model in which 4 loops from TRD1 and 4 loops TRD2 contact with DNA. Among the 4 largest patches resulting from the predictions of the Naïve Bayes classifier, patch 1, 3, and 4...
overlap with the DNA-binding sites in the model proposed by Kim et al. (2005). Figure 1B shows the predictions of the Naïve Baye classifier on the protein surface with TRD1 and TRD2 superimposed with TaqI-MTase/DNA complex. The figure shows that patch 1, 3, and 4 contact with the DNA from the superimposed TaqI-MTase/DNA complex.

**Patch 1**

There are two target recognition domains (TRDs) in the S subunit. The largest patch (patch 1) formed by the predicted interface residues locates on the surface of the second target recognition domain (TRD2) and includes Thr-243, Thr-244, Ser-246, Thr-247, Lys-248, Lys-249, Arg-305, Ala-306, Asn-321, Gln-322, Gly-357, Ser-358, Thr-359, and Lys-361. In a previous study, Kim et al. (2005) aligned the structures of the TRDs with the structure of the DNA binding domain from 7ti-MTase and proposed an interaction model for the S subunit and DNA. We used DaliLite program (Holm and Park, 2000) to align TaqI-MTase/DNA complex (PDB 1g38) with TRD2. The resulting alignment has an rmsd of 3.3 Å. Figure 2A shows that patch 1 forms a contiguous surface patch contacting with the superimposed DNA from the superimposed TaqI-MTase complex. Examination of the secondary structure shows that patch 1 covers four loops that are close to the superimposed DNA (Figure 2B). In Kim’s model (2005), four loops from TRD2 are predicted to contact with DNA: loop 1 (Thr-262–Asp-265), loop 2 (Arg-305–Pro-307), loop 3 (Asn-321–Gly-323), and loop 4 (Phe-360–Glu-362) (Figure 2C). Comparison of Figure 2B and Figure 2C shows that patch 1 overlaps with loop 2, 3 and 4. Patch 1 covers an extra loop (loop 5: Thr-243–Lys-249, shown in magenta in Figure 2C) that is not predicted to contact with DNA by Kim et al. (2005).

In this study, loop 5 (Thr-243–Lys-249) is predicted to contact with the DNA. This result is inconsistent with that of the study by Kim et al. (2005). We analyzed the structural alignment between TRD2 and TaqI-MTase/DNA complex and found that loop 5 is located near the DNA, with the closest distance being 3.6 Å (Figure 3A). Further investigation shows that loops 1–4 from TRD2 do not align perfectly with their counterparts from TaqI-MTase (Figure 3B). In Figure 3B, loop 1–4 from TRD2 are shown in red and their counterparts (named loop I, loop II, loop III and loop IV respectively) from TaqI-MTase in
the structural alignment are shown in yellow. The figure shows that loop 2 and 4 protrude out and are closer to the bound DNA than loop II and IV, and loop 1 and 3 are at lower positions than loop I and III. This suggests that in the structural alignment the DNA binding cleft of TRD2 tilts toward the left compared to the DNA-binding cleft of TaqI-MTase. If we adjust the structural alignment to make better alignment between loop 1-4 and loop I-IV by tilting TaqI-MTase/DNA toward the left, the DNA will be moved closer to loop 5 (in the direction shown by arrow in Figure 3B). Note that loop 5 consists of Thr-243, Thr-244, Pro-245, Ser-246, Thr-247, Lys-248, Lys-249. All these residues except Pro-245 have polar side chains. The hydroxyl group (-OH) from Thr and Ser, and the amino group (-NH²) from Lys are able to form hydrogen bonds with the DNA. Therefore, we hypothesize that loop 5 is directly involved in the interaction between TRD2 and DNA.

**Figure 1. The predictions of interface residues.** A: The predictions are shown on the protein surface with the four largest patches marked with circles. The S subunit is shown in green with the predicted interface residues shown in red. B: The predictions are shown on the protein surface with TRD1 and TRD2 superimposed with TaqI-MTase/DNA complex. DNA from the TaqI-MTase/DNA complex is shown in blue. TaqI-MTase is not shown in the figure. Figures are generated using PyMOL (DeLano, 2002). The sequence of the protein is shown below the figures with predicted interface residues shown in red and residues corresponding to patches 1-4 highlighted with different colors.
Figure 2. **Patch 1 from the predictions.** A: Patch 1 forms a contiguous patch (red) contacting with the DNA (blue) from the superimposed TaqI-MTase/DNA complex. (TaqI-MTase is not shown). B: Patch 1 covers four loops that are close to the superimposed DNA (blue). The residues in patch 1 are colored in red and labeled with their sequence numbers. C: The four loops predicted to contact with DNA by Kim et al. (2005) are shown in yellow. Magenta is the loop (loop 5) that is covered by patch 1 but is not predicted to contact with DNA by Kim et al. (2005). Figures are generated using PyMOL (DeLano, 2002).

Figure 3. **Loop 5 from the predictions.** A: In the structural alignment of TRD2 and TaqI-MTase/DNA complex, loop 5 (red) is close to DNA (blue). Green: TRD2; red: loop 5; blue: DNA from the TaqI-MTase/DNA complex. TaqI-MTase is not shown. B: DNA binding cleft on TRD2 is not perfectly aligned with that of TaqI-MTase in the structural alignment. Loop 1–5 from TRD2 are shown in red. Loop I–IV from TaqI-MTase are shown in yellow. Blue is the DNA from TaqI-MTase/DNA complex. Figures are generated using PyMOL (DeLano, 2002).
Patch 3 and 4

Patch 3 consists of Gly-97, Gln-147, Thr-148, Thr-149, Gin-150, Lys-151 and Asn-152. Patch 4 consists of Thr-37, Lys-39, Asn-79, Asn-80, Ser-81, Asn-82 and Gln-113. They are on the surface of the first target recognition domain (TRD1). We aligned the structure of TRD1 with the TaqI-MTase/DNA complex using DaliLite (Holm and Park, 2000), resulting in a superimposition with rmsd of 3.4 Å. Figure 4A shows that patch 3 and 4 form two contiguous surface patches contacting with the DNA from the superimposed TaqI-MTase/DNA complex. Figure 4B shows that patch 3 and 4 cover five loops that are close to the superimposed DNA. In the study of Kim et al. (2005), four loops from TRD1 are predicted to contact with DNA: loop 6 (Lys-55–Asp-58), loop 7 (Tyr-96–Ser-98), loop 8 (Asn-112–Ala-114) and loop 9 (Gln-150–Asn-152) (Figure 4C). Compared with the predictions by Kim et al. (2005), patch 3 overlaps with loop 7, 9, and patch 4 overlaps with loop 8. In addition, patch 4 includes two loops that are not in the predictions by Kim et al. (2005): loop 10 (Thr-37–Glu-43) and loop 11 (Asn-79–Asn82).

TRD1 and TRD2 share a sequence identity of 37% in the sequence and they have similar structure. We used DaliLite to make structure alignment between TRD1 and TRD2. The resulting alignment has an rmsd of 2.0 Å. In the structural alignment, the predictions on TRD1 correlate well with that on TRD2, except that loop 11 from TRD1 is predicted to interact with DNA and its counterpart (Glu-288–Asn-291) from TRD2 is not (Figure 5A). We aligned TRD1 with TaqI-MTase/DNA complex using DaliLite (Holm and Park, 2000), resulting in a superimposition with rmsd of 3.4 Å. We then aligned TRD2 to TRD1, resulting in an alignment of TRD1, TRD2, and TaqI-MTase/DNA complex (Figure 5B). Figure 5B shows that loop 11 protrudes toward the DNA and is closer to the DNA than its counterpart from TRD2. The closest distance between loop 11 and DNA is 1.6 Å, and that between its counterpart and DNA is 4.0 Å. So it is possible that loop 11 from TRD1 contacts with DNA and its counterpart from TRD2 does not.

O’Neil et al. (1998) used random mutagenesis approach to identify the protein-DNA binding sites on the S subunit of E. coli type I R-M system (EcoKI). In their study, 5 residues, Gly-92, Lys-92, His-95, Ser-103 and Phe-107, were predicted to be at the protein-DNA interface. We used the threading program FUGUE2 (Shi, et al., 2001) to align the
sequence of EcoKI S subunit to the structure of the S subunit from *M. jannaschii*. In the alignment Gly-91, Lys-92, His-95, Ser-103 and Phe-107 from EcoKI S subunit are aligned to Gly-100, Glu-101, Ile-104, Thr-111 and Ile-115 from TRD1 respectively. On TRD1, Gly-100 and Glu-101 are close to loop 7, and Thr-111 and Ile-115 directly connect to loop 8. Ile-104 is on a helix between loop 7 and loop 8.

**Patch 2**

Patch 2 includes Lys-209, Gly-212, His-213, Ser-214, Arg-215, Phe-216, Lys-217, and Lys-218. Zinkevich *et al.* (1992) showed that the mutation of Ala-204 to Thr in EcoYA S subunit affected the subunit assembly and resulted in a restriction-deficient and modification temperature-sensitive phenotype. They suggested that this mutation altered a domain in the S subunit that is essential for the binding with the R subunit. When the sequence of EcoKI S subunit is aligned to the structure of the S subunit from *M. jannaschii* using FUGUE2 (Shi, *et al.*, 2001), Ala-204 from EcoKI S unit is aligned with Gly-210 from the *M. jannaschii* S subunit. Gly-210 is directly contacted to patch 2 on the surface of the S subunit (Figure 6A). This indicates that patch 2 may overlap with the interaction sites between the S subunit and the R subunit. Kim *et al.* (2005) proposed a subunit assembly model for the *M. jannaschii* type I R-M system. In that model, S subunit and R subunit interact in the sites corresponding to patch 2 identified by the Naïve Bayes classifier (Figure 6B). Here a Naïve Bayes classifier trained to identify protein-DNA interface residues discover potential sites corresponding to protein-protein interactions. This suggests that the protein-DNA interactions and protein-protein interactions share some common mechanisms.
Figure 4. Patch 3 and 4 from the predictions. A: In the structural alignment of TRD1 and TaqI-MTase/DNA complex, patch 3 and 4 (red) form two contiguous patches contacting to the superimposed DNA (blue). B: Patch 3 and 4 cover 5 loops that are close to the DNA (blue) from the superimposed TaqI-MTase/DNA complex. Predicted interface residues are shown in red and labeled with their sequence numbers. C: The four loops (loop 6-9) that are predicted to interact with DNA by Kim et al. (2005) are shown in yellow. Magenta are the loop (loop 10-11) that are covered by patch 3 and 4 but are not predicted to contact with DNA by Kim et al. (2005). Figures are generated using PyMOL program (DeLano, 2002).

Figure 5. Loop 11 from the predictions. A: The predictions on TRD1 correlate well with those on TRD2, except that the loop 11 (Asn-79–Asn82) from TRD1 is predicted to interact with DNA and its corresponding part from TRD2 is not. TRD1 and TRD2 are aligned using DaliLite (Holm and Park, 2000). TRD1 is colored in gray with the predicted interface residues on it colored with yellow. TRD2 is colored in green with predicted interface residues in red. B: Loop 11 (red) from TRD1 protrudes toward the DNA and is closer to the DNA than its counter part (yellow) from TRD2. Green, TRD1; gray, TRD2;
blue, DNA from the TaqI-MTase/DNA complex. Figures are generated using PyMOL (DeLano, 2002).

**Figure 6. Patch 2 from the predictions.** A: Patch 2 (red) is shown on the surface of the S subunit from *M. jannaschii*. Gly-210 that is aligned to Ala-204 from EcoKI in the threading is shown in blue. B: Patch 2 overlaps with the sites that S units interacts with R units in the subunit assembly model by Kim *et al.* (2005). R subunits in the model are represented by circles. Figures are generated using PyMOL (DeLano, 2002).

**DISCUSSION**

In this study we applied a Naïve Bayes classifier to identify DNA binding sites in the S subunit of *M. jannaschii* type I R-M system. The classifier takes as input the identities of nine residues, corresponding to the target residue and 4 neighbor residues on each side of the target residue, and relative solvent accessibility and sequence entropy of the target residue. The predictions form contiguous patches on the surface of the S subunit. The 1\textsuperscript{st} (consisting of 14 residues), 3\textsuperscript{rd} (consisting of 7 residues) and 4\textsuperscript{th} (consisting of 7 residues) largest patches overlap with the DNA binding sites from the interaction model proposed by Kim *et al.* (2005). The predictions also reveal new DNA binding residues that are not included in the model by Kim *et al.* (2005). The results can be used to guide the design of mutagenesis studies to identify the critical residues that contribute to the specificity and affinity of DNA binding in type I R-M systems.
Protein-DNA interactions play pivotal role in gene regulations. Discovery of the principles of protein-DNA interactions has been a topic of wide interest for many years (Pabo and Sauer, 1992). Identifying the DNA binding sites using experimental studies lags far behind the pace at which protein and DNA sequences are produced by proteomic and genomic projects. Computational methods that can identify DNA-binding sites in a large scale and with high accuracy are urgently needed. In our previous study (Yan, et al., Submitted), we explored different information and developed a Naïve Bayes classifier to identify amino acid residues involved in protein-DNA interactions. The classifier achieved a performance of 77% overall accuracy with a correlation coefficient of 0.30 based on leave-one-out cross-validation using a set of 56 proteins. In this study, we apply the Naïve Bayes classifier to identify DNA binding sites in the S subunit from *M. jannaschii* using sequence information, accessibility and sequence entropy as input. A few other computational methods have been developed to identify DNA binding amino acid residues based on various information, including sequence, structure, evolutionary information, or physiochemical properties (Jones, et al., 2003; Keil, et al., 2004; Shanahan, et al., 2004; Tsuchiya, et al., 2004; Ahmad and Sarai, 2005). It will be interesting to compare the predictions of different methods and to develop ensemble methods that can take advantage of the complementarities among different methods.

In this study, the Naïve Bayes classifier is trained using a non-redundant set of 56 proteins bound to double-stranded DNA that are in previous study. These proteins share sequence identity less than 20% with the S subunit and none of them belongs to type I R-M systems. A Naïve Bayes classifier trained using a set of proteins that are not homologous to the S subunit can predicted potential DNA-binding sites. This indicates that different categories of protein-DNA interactions share some common mechanisms.

**ACKNOWLEDGEMENTS**

This Research was supported in part by a grant from the National Institutes of Health (GM 066387) to VH, DD, and RLJ.
REFERENCES


Hubbard, S.J. (1993) NACCESS. Department of Biochemistry and Molecular Biology, University College, London.


CHAPTER 5. A DATABASE OF PROTEIN-PROTEIN INTERFACES

A paper to be submitted to Nucleic Acids Research

Changhui Yan, Feihong Wu, Robert L. Jernigan, Drena Dobbs, and Vasant Honavar

ABSTRACT

Protein-protein interactions play vital roles in biological functions. Elucidating the details of protein-protein interactions is essential for understanding the metabolism network and signal transduction pathways. In various studies, identifying interface residues and retrieving the structural, biophysical and chemical properties of the interfaces on a large scale are often needed. Tools that can accomplish these tasks are needed.

Here, we present a database of protein-protein interfaces. It consists of all the protein-protein interfaces derived from the Protein Data Bank (PDB). It provides convenient tools for users to identify interface residues and retrieve information about the interfaces. Using this database, users can quickly retrieve interface information with a user-specified definition of interface residues and user-customized constraints on structure resolution, species, and protein function. Batch retrieval is available for users to download interface information for a list of interfaces. Currently, we are developing a Gene Ontology (GO)•• based query system, which will allow users to define a function group using GO terms and retrieve information about the interfaces belonging to the group.

INTRODUCTION

The mechanism of protein-protein interactions has been a topic of study for a long time. Various aspects of protein-protein interfaces – including residue composition, residue contact preferences, and interface size – have been analyzed in previous studies using
different datasets of protein-protein complexes (Chothia and Janin, 1975; Jones and Thornton, 1997a; LoConte et al., 1999; Nooren and Thornton, 2003; Ofran and Rost, 2003a; Caffrey et al., 2004; Halperin et al., 2004; Zhang and Palzkill, 2004; Keskin et al., 2005; Reichmann et al., 2005). Due to the difficulties in obtaining large representative sets of protein-protein interfaces and the difficulties in extracting interface information on a large scale, most previous studies were based on small datasets. Large representative datasets are required for the fair evaluation of the results. Furthermore, different definitions of interface residues were used in different studies. This makes it difficult to have a direct comparison of the results. In their study, Jones and Thornton (1997a) considered a residue an interface residue if its solvent accessibility is reduced by more than $1\AA^2$ during the formation of the complexes. In the study by Ofran and Rost (2003a), a residue is considered an interface residue if it is within a certain distance from the interacting protein. Fariselli et al. (2002) defined interface residues based on the distance between $\alpha$-carbon atoms. A database that allows users to obtain large representative sets of interfaces and to extract interface features with flexibilities in defining interface residues is highly demanded. Such a database will also benefit the development of computational methods for identification of interface residues. Because identifying interface residues using experimental methods can not catch up with the pace at which protein sequences are being determined, computational methods have been used to identify interface residues (Jones and Thornton, 1997b; Del et al., 2003; Ofran and Rost, 2003b; Bradford and Westhead, 2004; Carugo and Franzot, 2004; Liang et al., 2004; Neuvirth et al., 2004; Nissink and Taylor, 2004; Panchenko et al., 2004; Sen et al., 2004; Yan et al., 2004a; Yan et al., 2004b; Fernandez-Recio et al., 2005; Joughin et al., 2005). Usually, a computational method uses a set of protein-protein interfaces to search for the features that can be used to distinguish interface residues from non-interface residues. A representative dataset is needed for the success of the method. This need calls for a database that can provide representative sets of interfaces.

Although several protein-protein interface databases have been developed, none of them have satisfied the needs outlined above. Here, we present a database that consists of all the protein-protein interfaces derived from the PDB database. The database allows users to obtain large representative sets of interfaces and to retrieve interface features with a user-
specified definition of interface residues and user-specified constraints on structure resolution, interaction type, and protein function.

METHODS AND MATERIALS

Collection of protein-protein interfaces from the Protein Data Bank (PDB)

One crucial problem in collecting protein-protein interfaces is to distinguish the real interfaces that correspond to protein-protein contacts from those resulting from crystal compact. Usually, an interface is considered a crystal contact if the buried area is less than a certain cutoff (Henrick and Thornton, 1998). Here, we adopt the results from the Protein Quaternary Structure (PQS) (Henrick and Thornton, 1998) – a protein quaternary structure database – to reduce the chance of recruiting crystal contacts. During the development of the PQS, if a PDB structure contained multiple copies of quaternary structures, it was split into multiple PQS files with each file containing one independent quaternary structure. In this study, if a PDB structure had corresponding PQS entries, it was split into independent quaternary structures as described in PQS. Otherwise, we assumed that the PDB structure contained only one quaternary structure. Then, within each quaternary structure, a pair of protein chains is considered interacting if the buried area on one chain is at least 200Å². The solvent accessibility of residues were computed using the NACCESS program (Hubbard, 1993; Gutteridge et al., 2003). This process is illustrated in Figure 1. Using this method, we obtained a set of 25,953 interfaces including 27,392 protein chains.

Non-redundant sets of sequences

The PDB protein chains were clustered based on sequence similarity and the clustering information is available at (ftp://ftp.rcsb.org/pub/pdb/derived_data/NR/). The cd-hit program (Li et al., 2001) was used to generate the clusters at identity levels ≥40%, and the blastclust program (Altschula et al., 1990) was used to cluster sequences at the identity level of 30%. Here, we use the clustering results from the PDB to remove redundant sequences. Table 1 shows the numbers of protein chains in the non-redundant sets at different identity levels.
Non-redundant sets of interfaces

Because two protein chains can interact at different locations by adopting different orientations, there may be multiple interfaces between two given chains. In this study, the similarity between two interfaces is defined based on both sequence similarity and residue composition of the interfaces. The similarity between interface AB (the interface between chain A and chain B) and interface CD (the interface between chain C and chain D) is higher than \( \alpha \) if any of the following two conditions is satisfied:

1. (Sequence identity (A, C) >\( \alpha \)) \&\& (Sequence identity (B, D) >\( \alpha \)) \&\& (Composition similarity (AB, CD) > 60%)

2. (Sequence identity (A, D) >\( \alpha \)) \&\& (Sequence identity (B, C) >\( \alpha \)) \&\& (Composition similarity (AB, CD) > 60%)

The composition similarity between two interfaces AB and CD is given by

\[
\text{Similarity} = \left( 1 - \frac{\sum_{i} \text{abs}(n_{AB}^i - n_{CD}^i)}{\sum_{i} n_{AB}^i} \right) \left( 1 - \frac{\sum_{i} \text{abs}(n_{AB}^i - n_{CD}^i)}{\sum_{i} n_{CD}^i} \right)
\]

with \( n_{AB}^i, n_{CD}^i \) denoting the numbers of residues \( i \) in AB and CD respectively. Table 2 shows the numbers of non-redundant interfaces with different cutoffs.
PDB structures that contain at least two protein chains with at least 20 residues in each chain

Are there corresponding PQS entries?

Yes → Split the PDB structure into independent quaternary structures as described in PQS

No → The PDB structure is assumed to contain only one quaternary structure

Set of quaternary structures

For every quaternary structure {
  For every pair of chains {
    Calculate the buried area on each chain using NACCESS
    If the buried area on one chains >= 200 Å² {
      The interface between the pair is considered a real protein-protein contact
    } else {
      Ignore the interaction between the two chains
    }
  }
}

25,953 interfaces including 27,392 chains

Figure 1. The process of collecting protein-protein interfaces.
Table 1. Protein chain numbers in the non-redundant sets of proteins.

<table>
<thead>
<tr>
<th>Sequence identity cutoff (%)</th>
<th>Chain number</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>2,968</td>
</tr>
<tr>
<td>50</td>
<td>3,198</td>
</tr>
<tr>
<td>70</td>
<td>3,620</td>
</tr>
<tr>
<td>90</td>
<td>4,196</td>
</tr>
<tr>
<td>100</td>
<td>7,005</td>
</tr>
</tbody>
</table>

Table 2. Interface numbers in the non-redundant sets of interfaces.

<table>
<thead>
<tr>
<th>Sequence identity cutoff (%)</th>
<th>Number of interfaces</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>5,985</td>
</tr>
<tr>
<td>50</td>
<td>6,008</td>
</tr>
<tr>
<td>70</td>
<td>6,223</td>
</tr>
<tr>
<td>90</td>
<td>6,630</td>
</tr>
<tr>
<td>100</td>
<td>9,211</td>
</tr>
</tbody>
</table>

DATABASE FEATURES AND ACCESSIBILITY

User-specified definition of interface residues

The database allows users to define interface residues using any of the three types of definition: (a) Definition based on the distance between α-carbon atoms, that is, a residue is an interface residue if its α-carbon is within a certain distance from any α-carbon of the interacting chain; (b) Definition based on the closest distance between heavy atoms, that is, a residue is an interface residue if any of its heavy atoms is within a certain distance from any heavy atom of the interacting chain; and (c) Definition based on the reduction of solvent accessibility, that is, a residue is an interface residue if its solvent accessible surface is reduced more than a certain value during the formation of the complex.
Retrieval of interface features

Users can search and retrieve the features of a specific interface by providing the IDs of the two interacting protein chains, or the features of all the interfaces in a PDB structure by providing the structure ID, or the features of all the interfaces that a certain protein involved by providing the protein name.

Batch retrieval of information for a set of interfaces

Users can submit a list of interfaces or a list of protein chains and retrieve interface information in a batch. The returned results will be sent to the email provided by users.

Gene Ontology (GO)* based query

One goal of the Gene Ontology Annotation (GOA) (http://www.ebi.ac.uk/GOA/) project is to map gene products to the Gene Ontology (GO) resources. As a part of the GOA, the PDB-GOA project has assigned GO terms to the protein chains in the PDB. Using the results from the PDB-GOA, we are developing a GO-based query system that allows users to use GO terms to define the set of proteins that they are interested in and to retrieve interface information for the set in a batch.

Non-redundant sets of interfaces and non-redundant sets of protein chains

The database allows users to obtain non-redundant sets of interfaces (or proteins) either from a redundant set provided by the users or from the whole database. Constraints on sequence similarity, structure resolution, experimental methods, species, and function can be set by users.

Accessibility

The database is available for the public at http://einstein.cs.iastate.edu.
ACKNOWLEDGEMENTS

This Research was supported in part by a grant from the National Institutes of Health (GM 066387) to Vasant Honavar, Drena Dobbs, and Robert Jernigan.

REFERENCES


CHAPTER 6. AN ANALYSIS OF PROTEIN-PROTEIN INTERFACES

A paper to be submitted to Journal of Molecular Biology

Changhui Yan, Feihong Wu, Robert L. Jemigan, Drena Dobbs, and Vasant Honavar

ABSTRACT

Analyzing the characteristics of protein-protein interfaces to identify the features that can distinguish interface residues from non-interface residues is very important for understanding the mechanisms of protein-protein interactions. Here, we analyze the characteristics of protein-protein interfaces using the largest dataset available in the Protein DataBank (PDB). We start by a comparison of interfaces with protein core and non-interface surface. The results show that interfaces differ from protein core and non-interface surface in residue composition, entropy, and secondary structure. Since interfaces, protein core, and non-interface surface have different solvent accessibilities, it is important to investigate whether the observed differences are due to the difference in solvent accessibility. We excluded the effect of solvent accessibility by comparing interfaces with the samples of residues that were randomly chosen from the overall residues and had the same solvent accessibility as the interfaces. This strategy reveals some interface characteristics that are not observable by comparing interfaces with protein core and non-interface surface. When we repeated the analysis using three datasets with different constraints on redundancy and structure quality, similar results were obtained. This indicates the significance of the results. We also investigated the differences between hetero-interfaces and homo-interfaces.
INTRODUCTION

Various aspects of protein-protein interfaces, including size, planarity, hydrophobicity, electrostatics, conservation, residue composition, and contact preferences, have been studied based on different sets of protein-protein complexes (Janin and Chothia, 1988; Janin and Chothia, 1990; Young, et al., 1994; Jones and Thornton, 1996; Lo Conte, et al., 1999; Bahadur, et al., 2003; Nooren and Thornton, 2003; Ofran and Rost, 2003; Prasad Bahadur, et al., 2004; Ponstingl, et al., 2005). Some contradictory results have been reported in different studies, in part, due to the difference in the datasets used.

We extracted all the protein-protein interfaces from the Protein Data Bank (PDB) (Berman, et al., 2000) and removed redundant interfaces so that there were no identical interfaces in the dataset. The resulting dataset, which consists of 7,151 pairs of interacting chains with at least 20 amino acids in each chain, is much larger than any other dataset that was used in previous studies.

Each protein in the dataset was divided into three disjoint sections: interface, protein core, and non-interface surface. Comparisons show that the three sections are significantly different in residue composition, entropy, and secondary structure. Since interfaces, protein core, and non-interface surface have different solvent accessibilities, it is unknown whether these differences are due to the difference in functionality or the difference in solvent accessibility. To exclude the effect of solvent accessibility, we compared interfaces with the samples that were randomly chosen from the overall residues and had the same solvent accessibility as the interfaces. The results reveal some interface characteristics that cannot be detected by comparing interfaces with protein core or non-interface surface. Other properties of interfaces including size and contact preferences were also investigated. We repeated the analysis using three datasets with different constraints on sequence similarity and structure resolution, and similar results were obtained. We divided interfaces into hetero-interfaces and homo-interfaces based on the sequence similarity of the interacting chains. Comparisons show significant differences between the two types of interface in residue composition, entropy, secondary structure, size, and contact preferences.
METHODS AND MATERIALS

Dataset100, Dataset30, and Dataset30_3

All the protein complexes in the PDB that consist of at least two protein chains with at least 20 amino acids in each chain were obtained. If a PDB complex contains multiple copies of quaternary structures according to the Protein Quaternary Structure (PQS) (Henrick and Thornton, 1998), it was split into multiple files with each file containing one independent quaternary structure. The buried area on each chain was computed using the NACCESS program (Hubbard, 1993; Gutteridge, et al., 2003). A pair of protein chains is considered interacting if the buried area on one chain is at least 200 Å². A dataset of interacting pairs was thus obtained from the set of quaternary structures. Then, sequence similarity information was obtained from the sequence clusters from PDB (ftp://ftp.rcsb.org/pub/pdb/derived_data/NR/). First, redundant data were removed so that there were no identical pairs in the dataset. The resulting dataset, which we will refer to as Dataset100, consists of 7,151 pairs of interacting chains. Dataset100 was further processed, so that the identity between any two pairs was less than 30%. The resulting dataset, which we will refer to as Dataset30, consists of 2,981 pairs of interacting chains. Then, all the structures having resolution >3 Å were removed from Dataset30. The resulting dataset, which we will refer to as Dataset30_3, consists of 2,654 pairs of interacting chains.

Protein core, interface, and non-interface surface

We defined residue contacts as described in Ofran and Rost (2003). Two residues were considered contacting if the closest distance between their heavy atoms is less than 6 Å. The residues that had at least one contact with residues from the interacting chain were considered interface residues. Non-interface residues were divided into two groups: protein core that consists of residues with relative solvent accessibility (rASA) <25% and non-interface surface that consists of residues with rASA ≥ 25%. The rASA of residues was calculated using the NACCESS program (Hubbard, 1993; Gutteridge, et al., 2003).
Hetero-interfaces and homo-interfaces

An interface is a homo-interface if the two interacting chains have a sequence identity greater than 95%. Otherwise, it is a hetero-interface. We used Dataset100 to compare the properties of hetero-interfaces and homo-interfaces. Dataset100 contains 4,126 homo-interfaces and 3,025 hetero-interfaces.

Interface propensity (raw interface propensity, RIP) and contact preferences

Let $F_i$ be the number of residue $i$ in the dataset, $f_i$ be the number of residue $i$ in the interfaces, $w_i = f_i / \sum_i f_i$, and $W_i = F_i / \sum_i F_i$. The interface propensity of residue $i$ is given by $\log_2 \left( \frac{w_i}{W_i} \right)$. Propensities for protein core and non-interface surface are computed similarly with $w_i$ replaced by the fraction of residue $i$ in protein core and non-interface surface respectively. Let $C_{ij}$ be the number of interface-crossing contacts formed by residue $i$ and $j$. The contact preference between residue $i$ and $j$ is given by $\log_2 \left( \frac{C_{ij}}{\sum_i C_{ij}} \right) \left( \frac{w_i \times w_j}{w_i \times w_j} \right)$.

Normalized interface propensity (NIP)

Residues were randomly extracted from the overall residues so that the resulting samples had the same solvent accessibility as interfaces. Let $s_i$ be the number of residue $i$ in the samples, and $S_i = s_i / \sum_i s_i$. The normalized interface propensity of residue $i$ is given by $\log_2 \left( \frac{w_i}{S_i} \right)$, where $w_i$ is defined as above.

RESULTS AND DISCUSSION

Characteristics of interfaces

Proteins were divided into three disjoint sections: protein core, interfaces, and non-interface surface. Interface properties including residue composition, secondary structure, entropy, contact preferences, and size were analyzed using Dataset100.
Residue composition

Figure 1A compares the residue compositions of protein core, interfaces, and non-interface surface. Residues are placed in the order of increasing hydrophobicity based on the Kyte and Doolittle hydropathy index (1982). The comparisons show that protein core has the most hydrophobic residues (e.g. Met, Cys, Phe, Ile, Leu, and Val) and non-interface surface has the least. This indicates that hydrophobic residues are preferred in protein core and unfavorable on non-interface surface. The opposite trend is observed for hydrophilic residues (e.g. Arg, Lys, Glu, and Asp). For most residues (except Gly, His, and Tyr), the distribution among the three sections is either protein core > interfaces > non-interface surface or protein core < interfaces < non-interface surface. For each residue, we calculated its propensities for the three sections separately. Figure 1B shows that all the residues have opposite propensities for protein core and non-interface surface, and, in most cases, the propensity for interfaces are is between those for protein core and non-interface surface.

Entropy

Entropy values were extracted from the HSSP database (http://www.cmbi.kun.nl/gv/hssp/). The entropy shows the conservation at each residue position. It was normalized to the range of 0-100 with low entropy values corresponding to conserved positions. Figure 2 compares the entropy distributions of protein core, interfaces, and non-interface surface. The comparisons show that protein core has the most residues in the low entropy region (entropy <40), and non-interface surface has the least. In the high entropy region, the opposite trend is observed. The results indicate that the trend of conservation is protein core > interfaces > non-interface surface. In a study based on a small set of transient protein-protein complexes, Nooren et al. (2003) showed that interface residues are more conserved than surface residues. Consistent results are obtained here using a large dataset.
**Figure 1. Residue composition.** A. Residue compositions of protein core, interfaces, and non-interface surface. B. Propensities for the protein core, interfaces, and non-interface surface. Residues are placed in the order of increasing hydrophobicity based on the Kyte and Doolittle hydropathy index (1982).

**Figure 2. The entropies of protein core, interfaces, and non-interface surface.**

**Secondary structure**

We considered eight classes of secondary structure as defined by the DSSP program (Kabsch and Sander, 1983). Figure 3 compares the secondary structure composition of protein core, interfaces, and non-interface surface. The comparisons show that non-interface surface has the most residues in S (Bend) and T (Turn), and the protein core has the least. The opposite trend is observed for E (Extended strand). No obvious location preferences were observed for the other types of secondary structure.
Figure 3. The secondary structure compositions of protein core, interfaces and non-interface.

Contact preferences

Figure 4 shows the preferences for residue contacts. In the figure, positive preferences are shown in red, negative ones are in blue, and neutral ones are in green. Residues are placed in the order of increasing hydrophobicity. Figure 4 shows that the contacts between hydrophobic residues have high preferences. These highly preferred contacts make a red region in the lower-right corner of figure 4. The fact that Cys-Cys contact has one of the highest preferences indicates that disulfide bonds have an important role in protein-protein interactions. The contacts between residues with opposite charges (Arg-Asp, Arg-Glu, Lys-Asp, and Lys-Glu) also have high preferences. These contacts form several red entries near the upper-left corner of figure 4. These results are consistent with the previous claim that disulfide bonds, salt-bridges, and hydrophobic interactions are the main forces in protein-protein interactions (McCoy, et al., 1997; Sheinerman, et al., 2000; Glaser, et al., 2001; Ofiran and Rost, 2003). The face-to-face arrangement of two aromatic rings was found to be favorable for interactions in a previous study (Glaser, et al., 2001). Here, high preferences are observed for the contacts between different aromatic residues. The interaction between a proline ring and an aromatic ring can resemble the interaction between two aromatic rings (Glaser, et al., 2001). In this study, positive preferences are also observed for the contacts between Pro and aromatic residues (Tyr, Trp, and Phe).
Figure 4. Residue contact preferences. Residues are placed in the order of increasing hydrophobicity based on the Kyte and Doolittle hydropathy index (1982).

Figure 5. Interface size.

Interface size

Interface size was calculated separately for each side of interfaces. Figure 5 shows that interface sizes span a broad range and the distribution has a peak value at 600–800 Å\(^2\). The average size of interfaces is 1227 Å\(^2\). 14% of the interfaces in the dataset have a size in the range of 600–800 Å\(^2\). In a study based on a set of 75 hetero-complexes, Lo Conte et al. (1999) found that most interfaces have a total buried area (that is, the sum of the buried area from both sides of the interfaces) in the range of 1600 (±400) Å\(^2\), which is roughly
equivalent to 800 (±200) Å² from each side of the interfaces. Here, about 25% of the (one-side) interfaces have a size in the range of 800 (±200) Å²

**Are the differences due to the difference in solvent accessibility or the difference in functionality?**

By the definitions, protein core residues have a relative solvent accessibility (rASA) less than 25%, non-interface surface residues have a rASA equal to or greater than 25%, and interface residues have a rASA ranging from 0% to 100%. The results from above have shown the differences among protein core, interfaces and non-interface surface. However, since these three sections have different accessibilities, it is unknown whether these differences are due to the difference in accessibility or the difference in functionality. To exclude the effect of accessibility, we randomly extracted samples of residues from the overall residues so that the resulting samples had the same rASA distribution as interfaces and then compared interfaces with the samples. Five samples were extracted from the overall residues. The sample size is about 60% of the size of the overall residues. Figure 6 shows the rASA of the samples and interfaces.

![Figure 6. Relative solvent accessibilities of the samples and interfaces. Mean values for the samples are displayed with standard deviations shown as bars.](image)

**Residue composition and interface propensity**

Figure 7 compares the residue compositions of the samples and interfaces. The comparisons show that interfaces have more aromatic residues (Tyr, Try, and Phe) and
hydrophobic residues (Cys, Met, Ile, Leu, and Val) than the samples. Residues with medium hydrophobicity (Ser, Thr, Gly, and Ala) are underrepresented in interfaces. All the charged residues, except Arg, are underrepresented in interfaces. Interfaces have more His than the samples.

We calculated the interface propensities of residues based on the residue composition of the samples, that is, $\text{propensity } (i) = \log_2 \left( \frac{w_i}{S_i} \right)$, where $S_i$ is the fraction of residue $i$ in the samples and $w_i$ is the fraction of residue $i$ in interfaces. We named this propensity normalized interface propensity (NIP), since the samples can be considered as a version of the overall residues that are normalized according to the accessibility distribution of interfaces. The results are shown in figure 8 with residues placed in the order of increasing hydrophobicity. Figure 8 shows that NIP reveals the trend that hydrophobic residues have high preferences for interfaces and hydrophilic residues are not preferred in interfaces. On the right-hand side (the hydrophobic end) of figure 8, residues have highly positive propensities for interfaces and Cys have the highest preference overall. On the left-hand side (the hydrophilic end), residues (except Arg and His) have negative propensities. This indicates that the interfaces are more hydrophobic than the rest of the proteins. This is consistent with the results presented by Young et al. (1994). Figure 8 also shows that aromatic residues have high propensities for interfaces.

![Figure 7](image.png)

**Figure 7. Residue compositions of the samples and interfaces.** The mean values for the samples are displayed with standard deviations shown as bars. The residues are placed in the order of increasing hydrophobicity.
We compared NIP with the interface propensities (raw interface propensities, RIP) that were calculated based on the overall composition, that is, raw interface propensity (i) = \log_2(W_i/W_{\text{overall}}), where W_i is the fraction of residue i overall, and w_i is the fraction of residue i in interfaces. While NIP reveals the trend that hydrophobic residues have high preferences for interfaces and hydrophilic residues are unfavorable in interfaces, this trend cannot be revealed by RIP (Figure 9). Figure 9 shows that many residues have opposite signs in RIP and NIP. Ile, Val, Leu, and Met have high positive NIP but negative RIP. Asn, Asp, Gln, and Glu have negative or neutral NIP, while the corresponding values in RIP are positive or neutral. Cys and aromatic residues (Try, Trp, and Phe) have highly positive NIP but only weakly positive RIP. The difference between RIP and NIP is that in NIP interfaces are compared with samples that have the same accessibility distribution as the interfaces, while in RIP interfaces are compared with the overall residues whose accessibility is different from that of interfaces. The differences in the values of RIP and NIP indicate that accessibility affects the distribution of residues. Therefore, it is crucial to exclude the effect of accessibility when searching for the features that can distinguish interfaces from the rest of the proteins.
Figure 9. Comparison of the normalized interface propensities (NIP) and the raw interface propensities (RIP).

Contradictory results in interface propensities have been reported in different studies. For example, some studies showed that Ile, Val, and Leu have highly positive propensities for interfaces (Jones and Thornton, 1996; Lo Conte, et al., 1999; Bahadur, et al., 2003), while the study of Ofran and Rost (2003) showed that these residues have negative or weakly positive propensities for the inter-protein interfaces. Our results show that the three residues have highly positive propensities when evaluated using NIP and negative propensities when evaluated using SIP. In Ofran and Rost's study, interface propensities were calculated using SWISS-PROT as background, so the results are similar to that based on RIP in this study, which is calculated using overall residues as background. In the studies by Jones and Thornton (1997), Lo Conte et al. (1999), and Bahadur et al. (2004), interface propensities were calculated based on the accessible surface area of residues, and the results are similar to that based on NIP in this study.

Entropy

The entropies of the samples and interfaces are compared in Figure 10. The results show that interfaces have more residues with low entropies (conserved) than the samples. This indicates that interfaces are more conserved than the rest of the samples. The result from a previous section (shown in figure 2) has shown that protein core is more conserved than interfaces, which in turn are more conserved than non-interface surface. Here, figure 10
shows that after excluding the effect of accessibility, interfaces are more conserved than the rest of the proteins.

Figure 10. Entropies of the samples and interfaces. The mean values for the samples are displayed with standard deviations shown as bars.

Secondary structure

The secondary structure contents of the samples and interfaces are shown in figure 11. Compared with the samples, interfaces have slightly more residues in E (Extended strand) and H (π helix) and fewer residues in S (Bend) and T (Turn). Despite this, there are no significant differences between interfaces and the samples in terms of secondary structure. Although the results from a previous section (shown in figure 3) show some differences in secondary structure content among protein core, interfaces, and non-interface surface, here, figure 11 shows that interfaces do not differ from the rest of the proteins in secondary structure content after excluding the effect of accessibility. This suggests that the differences in secondary structure content among protein core, interfaces, and non-interface surface are due to the different accessibilities of the three parts instead of the different functionalities.
In summary, to exclude the effect of solvent accessibility, we compared interfaces with the randomly extracted examples that have the same accessibility distribution as the interfaces. The results show that hydrophobic residues and aromatic residues have high propensities for interfaces; hydrophilic residues (except Arg and His) have negative propensities for interfaces; and interfaces are more conserved than the rest of the proteins.

Are the results consistent across different datasets?

So far, the results reported are based on Dataset100, in which there are no identical interfaces. However, some interacting pairs in Dataset100 are still highly similar in sequence. To reduce the effect of biased data, Dataset100 was further processed by removing redundant data so that the identity between any two pairs was less than 30%. The resulting dataset (Dataset30) consists of 2,981 pairs of interacting chains. Then, all the structures having resolution >3 Å were removed from Dataset30. The resulting dataset (Dataset30_3) consists of 2,654 pairs of interacting chains. We analyzed interface properties using these three datasets. Figure 12 shows that the results obtained using the three datasets are consistent. Here, similar results are obtained using three datasets with different constraints on sequence redundancy and structural quality. This indicates the statistical significance of the results.
Homo-interfaces vs. hetero-interfaces

Some studies have shown that different types of interface have different characters (Jones and Thornton, 1996; Bahadur, et al., 2003). We divided Dataset100 into hetero-interfaces and homo-interfaces based on the sequence identity between the interacting pairs and compared the characteristics of the two types of interface (Figure 13). Figure 13A shows the interface propensities of residues. The results show that in both homo-interfaces and hetero-interfaces hydrophobic residues (Ile, Val, Leu, Phe, Cys, and Met) have high interface propensities and hydrophilic residues (Lys, Asn, Asp, Gln, and Glu) have negative propensities. This suggests that both types of interface are more hydrophobic than the rest of the proteins. Figure 13A also shows that Cys and aromatic residues (Phe, Trp, and Tyr) have higher propensities for hetero-interfaces than for homo-interfaces. The comparisons also show that hydrophobic residues (Ile, Val, Leu, and Met) have higher propensities for homo-interfaces than for hetero-interfaces, and charged residues (except Arg) show the opposite trend. This indicates that homo-interfaces are more hydrophobic than hetero-interfaces. This result is consistent with the results of previous studies (Jones and Thornton, 1996; Bahadur, et al., 2003). Figure 13B shows that hetero-interfaces have more residues with low entropy (conserved) than homo-interfaces, suggesting that hetero-interfaces are more conserved than homo-interfaces. Comparisons in secondary structure content (Figure 13C) show that hetero-interfaces have more coils (_) and extended strands (E) and fewer α-helixes (H) than homo-interfaces. Figure 13D shows the size distributions of hetero-interfaces and homo-interfaces. Both types of interface have a peak value at 600-800 Å². However, there are more large-size homo-interfaces than large-size hetero-interfaces. 63% of the homo-interfaces have a size larger than 800 Å², while only 53% of the hetero-interfaces have a size larger than 800 Å². The average size of the homo-interfaces is 1,311 Å², and the average size of the hetero-interfaces is 1,112 Å². This result is consistent with the claim of a previous study that homo-interfaces are larger than hetero-interfaces (Bahadur, et al., 2003). Figure 13E–F show that the contacts between residues with opposite charges (Arg–Asp, Arg–Glu, Lys–Asp, and Lys–Glu) and the contacts between hydrophobic residues (the red regions at the lower-right corners of Figure 13E–F) have high preferences in both types of interface. Compared with the homo-interfaces, hetero-interfaces show higher preferences for the
contacts involving Cys or aromatic residues (Phe, Tyr, and Trp). The columns and rows having these residues in Figure 13E are redder than the corresponding entries in Figure 13F.

CONCLUSIONS

Different from any previous study, this study uses the largest set of interfaces that are available. Dataset100 consists of 7,151 interfaces; Dataset30 consists of 2,981 interfaces; and Dataset30_3 consists of 2,654 interfaces. Each of these datasets is much larger than any other dataset that was used in previous studies. The three datasets have different constraints in sequence redundancy and structure resolution. The fact that similar results were obtained based on the three datasets indicates the significance of the results.

We divided proteins into three disjoint sections and analyzed their properties. The results show that: (1) protein core, interfaces, and non-interface surface are significantly different in residue composition, entropy, and secondary structure; (2) interface sizes span a broad range and about 25% of interfaces have a buried area from each side of the interfaces in the range of 600-1000 Å²; and (3) the Cys-Cys contact, the contacts between residues with opposite charges, the contacts between different aromatic residues, and the contacts between hydrophobic residues are preferred in the interfaces.

In particular, we compared interfaces with the samples that had the same solvent accessibility as the interfaces. This strategy excludes the effect of solvent accessibility on the distributions of residues, secondary structure, and entropy. The results show that: (1) hydrophobic residues and aromatic residues are preferred in interfaces and hydrophilic residues (except Arg) are unfavorable; (2) interfaces are more conserved than the rest of the proteins; and (3) interfaces do not differ from the rest of the proteins in secondary structure content.

We also analyzed the differences between homo-interfaces and hetero-interfaces. The results show that: (1) homo-interfaces are more hydrophobic and larger than hetero-interfaces; (2) hetero-interfaces are more conserved than homo-interfaces; and (3) hetero-interfaces have higher preferences for the contacts involving Cys, Phe, Trp, and Tyr.
ACKNOWLEDGEMENTS

This Research was supported in part by a grant from the National Institutes of Health (GM 066387) to VH, DD, and RLJ.

REFERENCES


Hubbard, S.J. (1993) NACCESS. Department of Biochemistry and Molecular Biology, University College, London.


CHAPTER 7. CONCLUSIONS

SUMMARY AND DISCUSSION

Identification of interface residues involved in protein-protein and protein-DNA interactions is critical for understanding the functions of biological systems, and has broad applications. Computational methods that can accurately identify interface residues from sequence are urgently needed. In this study, we have developed a two-stage classifier to identify interface residues involved in protein-protein interactions. The classifier achieves 72% accuracy with a correlation coefficient of 0.3 when tested on a set of 77 proteins using five-fold cross-validations, which is the best performance ever reported for sequence-based methods for the identification of protein-protein interface residues. We have developed a Naïve Bayes classifier to identify residues involved in protein-DNA interactions. The results based on leave-one-out experiments show that a Naïve Bayes classifier using only residue identities as input can correctly indicate the locations of interaction sites and the performance of the classifier can be improved by using additional information. We have applied the Naïve Bayes classifier to identify the DNA binding sites on the S subunit of M. jannaschii type I R-M system. The prediction consists of 4 major patches on the protein surface. Among them, three overlap with the prediction made by Kim et al. (2005) based on structure alignments. Due to the difficulties in collecting large representative datasets and the difficulties in extracting interface information on a large scale, all the previous studies in developing computational methods to identify interface residues used only small datasets. However, large and representative datasets are critical for the training and fair evaluations of machine-learning classifiers. We have developed a database of protein-protein interfaces. The database consists of all the protein-protein interfaces derived from the PDB database. It allows users to quickly extract interface information with a user-defined function group and a user-provided definition of interface residues. We have analyzed the characteristics of interfaces using datasets obtained from the database. The datasets are much larger than any other dataset that has been used in previous studies. The results reveal the differences between interfaces and the rest of the proteins in residue composition, conservation,
hydrophobicity, and secondary structure. The results also reveal the differences between homo-interfaces and hetero-interfaces.

CONTRIBUTIONS

We have developed a two-stage classifier to identify interface residues involved in protein-protein interactions

The first stage of the classifier takes the identities of a window of 9 amino acids that center at the target residue as input, and the second-stage classifier exploits the observation that interface residues form clusters on sequence to refine the predictions of the first-stage classifier. The results show that the second-stage classifier improves the performance significantly. This study provides valuable examples of customizing sophisticated machine learning algorithms for the task of identifying interface residues. The data representations in this study and the second-stage classifier that exploits the distribution of interface residues in sequence provide original examples for encoding domain knowledge to solve biology problems using machine-learning approaches. In this study, we, for the first time, point out the pitfalls of using accuracy as the sole performance measure in interface residue predictions and evaluate the performance of classifiers using a collection of performance measures including accuracy, correlation coefficient, sensitivity and specificity.

We have developed Naïve Bayes classifiers for identification of DNA-binding residues

The classifiers use local amino acid sequence information alone or a combination of sequence and structure-derived information as input. The results show that the classifier using only sequence information as input can correctly identify the locations of DNA binding sites and achieves 77% overall accuracy with 0.25 correlation coefficient, 37% specificity, and 43% sensitivity. The classifier using both sequence and structure-derived information as input achieves 77% overall accuracy with 0.30 correlation coefficient, 39% specificity, and 52% sensitivity. This study demonstrates that it is feasible to identify protein-DNA interface residues based solely on local sequence information, and using
additional information derived from structure can enhance the performance. We searched for DNA binding sites in the S subunit of *M. jannaschii* type I R-M system using the classifier that takes sequence, relative solvent accessibility (rASA), and entropy as input. The prediction significantly overlaps with the DNA binding sites predicted in a previous study by Kim *et al.* (2005) and, in addition, reveals three potential DNA binding loops that have not been identified before.

**We have developed a database that contains all the protein-protein interfaces in the PDB database**

The database contains all the pairwise interfaces from the PDB database that have at least 20 amino acid residues in each chain and at least 200 Å² buried area at one side. For each interface, the database provides the information about interface residues, sequence entropy, solvent accessibility, structure resolution, species, and Gene Ontology annotations. Particularly, we provide the tools that allow users to extract non-redundant sets of interfaces belonging to specific functions, which can be specified by users using Gene Ontology terms. Using the database, users will be able to extract interface information quickly, on a large scale, using a user-specified definition of interface residues. This database can be very helpful for various studies, e.g. the development of computational methods for identification of interface residues and the analysis of the characteristics of interfaces.

**We have systematically analyzed the characteristics of protein-protein interfaces using large datasets**

This study provides a comprehensive analysis of protein-protein interface characteristics in residue composition, secondary structure, hydrophobicity, contact preferences, conservation and size. Particularly, for the first time, we excluded the effect of solvent accessibility and searched for the characteristics that can distinguish the interfaces from the rest of the proteins. The results show that hydrophobic residues and aromatic residues have high propensities for the interfaces; hydrophilic residues, except Arg and His, have negative propensities for the interfaces; interfaces are more conservative than the rest of the proteins; and there are no obvious differences in the secondary structure content between the
interfaces and the rest of the proteins. This study also reveals the differences between homo-interfaces and hetero-interfaces. The results show that homo-interfaces are more hydrophobic, less conservative, and larger than hetero-interfaces.

**FUTURE WORK**

The problem of identifying interface residues from sequence is extremely challenging. The performance of the sequence-based classifiers from this study and other studies is still weak. Following are some potential directions for future studies.

**To systematically analyze the properties of interfaces to find the features that can effectively distinguish the interface residues from the non-interface residues**

We have used the information about a window of nine amino acid residues as the input to the classifiers. One limitation of this approach is that it cannot provide sufficient global information. Using global information should be able to improve the performance of the classifiers. For example, in the two-stage classifier, a Bayesian classifier is used to explore the distribution of interface residues and refine the output of the first-stage classifier. However, a position specific distribution model based on the entire sequence will be more effective in modeling the distribution of interface residues in sequence. The performance of the classifier can be further improved by using such a model that can capture the distribution of interface residues in the entire sequence.

**To improve the classification performance by customizing the machine learning algorithms and the data representations**

This goal focuses on the development of data mining algorithms that are better suited for the specific problem of identifying interface residues. Data representations that can effectively encode domain knowledge for the algorithms will also be explored. In the study of identifying DNA-binding residues, attributes (e.g. residue identity, sequence entropy and rASA) were simply included in the input under the assumption that these attributes were independent given the class. This assumption may not hold. Systematic analysis is needed to
search for data representations that can effectively encode domain knowledge for specific learning algorithms.

To develop ensemble methods based on individual classifiers that identify interface residues using different information

Various aspects of interfaces including size, planarity, hydrophobicity, electrostatics, conservation, residue composition, and contact preferences have been studied based on different sets of protein-protein complexes (Janin and Chothia, 1988; Janin and Chothia, 1990; Young, et al., 1994; Jones and Thornton, 1996; Lo Conte, et al., 1999; Bahadur, et al., 2003; Nooren and Thornton, 2003; Ofran and Rost, 2003a; Prasad Bahadur, et al., 2004; Ponstingl, et al., 2005). Although differences in some of these features have been found between interfaces and the rest of the proteins, none of them alone can sufficiently distinguish interface residues from non-interface residues. One promising approach to identify interface residues using information from multiple features is to build individual classifiers that identify interface residues using a single feature and then build an ensemble classifier based on the individual classifiers. Sen et al. (2005) developed an ensemble method to identify protease-inhibitor binding sites by taking simple votes among the classifiers that exploited sequence, structural, or evolutionary information. Similar approaches that use more features and more sophisticated ensemble methods should be tested.

To assess the effectiveness of the classifiers in the context of specific biological problems

The successes (and failures) in making predictions that can be validated using experimental techniques will be used to guide the refinement of the algorithms.

REFERENCES


