Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By David La

Entitled Computational Models of Mutations for Predicting and Classifying Protein-Protein Interaction Sites

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Daisuke Kihara
Chair

Michael Gribskov

Michael Zanis

Carol Post

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Daisuke Kihara

Approved by: Peter J Hollenbeck 12/07/2011
Head of the Graduate Program Date

Graduate School Form 20
(Revised 9/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:

Computational Models of Mutations for Predicting and Classifying Protein-Protein Interaction Sites

For the degree of Doctor of Philosophy

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Executive Memorandum No. C-22,* September 6, 1991, *Policy on Integrity in Research.**

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

David La

Printed Name and Signature of Candidate

12/05/11

Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/c_22.html

COMPUTATIONAL MODELS OF MUTATIONS FOR PREDICTING AND CLASSIFYING PROTEIN-PROTEIN INTERACTION SITES

A Dissertation

Submitted to the Faculty

of

Purdue University

by

David La

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2011

Purdue University

West Lafayette, Indiana

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

To My Awesome Family

ACKNOWLEDGMENTS

I would like to express immeasurable gratitude to my major advisor, Daisuke Kihara, for his advice and encouragement throughout my graduate study. Also, I would like to thank my committee members, Michael Gribskov, Michael Zanis, and Carol Post for their helpful expert advice and support. In particular, I would like to give special thanks to Michael Zanis for his insight in evolution and phylogenetics, which has helped inspire initial work in this thesis.

Also, I would like to give special thanks to my labmates, Sael Lee, Bin Li, Meghana Chitale, Juan Esquivel, Chao Yuan, Muyi Liu, Hao Chen, Yifeng Yang, and Troy Hawkins for providing a wonderful and lively research environment. I appreciate all of the fascinating chats with them on various occasions. Finally, I would like to thank our post-doctoral researchers, Vishwesh Venkatraman and Mateusz Kurcinski, for their insightful discussions and brilliance.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

PPI	Protein-Protein Interaction
PBI	Protein Binding Interface
PDB	Protein Data Bank
3D	Three Dimension
CE	Combinatorial Extension
ASA	Accessible Surface Area
ASP	Atomic-based Desolvation Parameters
RMSD	Root Mean Square Deviation
MSA	Multiple Sequence Alignment
CMA	Correlated mutation analysis
LRM	Logistic Regression Model
NN	Neural Networks
SVM	Support Vector Machine
PPV	Positive Predictive Value
MCC	Matthews Correlation Coefficient
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
Y2H	Yeast-2-Hybrid
TAP	Tandem Affinity Purification
CoIP	Co-Immunoprecipitation
SPR	Surface Plasmon Resonance
NMR	Nuclear Magnetic Resonance

ABSTRACT

La, David. Ph.D., Purdue University, December 2011. Computational models of mutations for predicting and classifying protein-protein interaction sites. Major Professor: Daisuke Kihara.

Protein-protein interaction residues are largely responsible for mediating many critical functions in the cell, such as inhibitory effects through enzyme-inhibitor interaction, initiating immune response by an antibody-antigen interaction, and regulation of cell-signaling proteins. Currently, various methods are available for predicting protein-protein interaction sites, these methods allows a residue-level understanding of the protein-binding phenomena presented by the global construction protein-protein interaction networks. In this thesis, protein-protein interaction sites are predicted using phylogenetic substitution models of amino acid mutations at protein interfaces:

1) Predicting Protein-Protein Interaction Sites using Phylogenetic Substitution Models: Protein-protein are critical for maintaining many different biological functions in the cell. In particular, these processes involve functionally important amino acid residues that are traditionally accepted as conserved in sequence throughout evolutionary time. However, protein-protein interaction sites exhibit higher sequence variation than other functional regions, such as those that correspond to catalytic sites and ligand-binding sites. Consequently, the semi-conservation of protein-protein interaction sites pose significant challenges in the current protein-protein interface prediction methods. To approach this problem, we developed a phylogenetic framework to capture the mutational behavior of essential protein-protein binding residues. Through the comprehensive analysis of functionally diverse protein families, we discover key amino acid substitution patterns that are characteristic of protein-protein interfaces. We demonstrate the contrast between interface and non-interface substi-

tution models shows mutational biases imposed on protein-protein binding residues. Based on this analysis, we have developed a novel method, BindML, which utilizes these evolutionary models to predict protein-protein binding sites on protein structures even without knowledge of their interacting partners. When assessed on a large benchmark of protein complexes, our method performs better compared to alternative methods for protein binding interface prediction. The conceptual novelty of this method is that it detects semi-conserved mutations rather than conventional conservation in protein family sequences, thus aimed to open a new direction in protein sequence analysis.

2) Prediction and Classification of Permanent and Transient Protein-Protein Interfaces: Proteins interact with each other in different ways for specific functional consequences. Our current research direction involves the development of a new method to classify mutation patterns of protein-protein interaction sites into permanent and transient types. The permanent type of interactions requires tight binding between proteins to assemble strong complexes. For example, enzyme-inhibitor, antigen-antibody, and large homo-oligomeric enzyme structures all compose of proteins that are required to be permanently bound in order to correctly carry out their functions. In contrast, transient type protein-protein interactions can readily dissociate after binding. Examples of transient interactions include proteins involved in signaling pathways, in which binding of transient proteins (such as protein kinases and G-proteins) induces conformational changes that allow protein function (and hence pathways) to switch on and off allowing strict and precise control of cellular activity. Although there are many studies that have already explored the differences in these two types of interactions at the level of the protein structure, in this study we develop amino acid substitution models to differentiate the differences between permanent and transient type interfaces primarily using sequence information. We built highly discriminative substitution models that can be used to classify protein interface predictions into permanent and transient interaction types. A detailed understanding of the mutational constraint differences between permanent and transient

protein complexes should help elucidate critical amino acid substitution preferences that are useful for annotating protein binding interface predictions of structures and sequences of unknown function.

3) 3D-SURFER Software for high-throughput protein surface comparison and analysis: A web-based tool, 3D-Surfer, has been developed to facilitate high-throughput comparison and characterization of proteins based on their surface shape. As each protein is effectively represented by a vector of 3D Zernike descriptors, comparison times for a query protein against the entire PDB take, on average, only a couple of seconds. The web interface has been designed to be as interactive as possible with displays showing animated protein rotations, CATH codes and structural alignments using the CE program. In addition, geometrically interesting local features of the protein surface, such as pockets that often correspond to ligand binding sites as well as protrusions and flat regions can also be identified and visualized.

1. INTRODUCTION

In the recent years, the biology community has been fascinated by snapshots of the complex inner workings of cellular activities from various different angles enabled by the advancement of high throughput experiments. For the first time, the construction and vivid visualization of protein-protein interaction networks provided maps of physical protein-protein binding which are important for maintaining many critical cellular events involving cell signaling, gene regulation, and metabolism [1, 2, 3]. Alternatively, genome sequencing projects and structural genomics projects continue to accumulate sequence and structure information of individual proteins. Therefore, the urgent need in the post-genomic era is to capitalize these two types of data to provide useful and practical information to the biology community. Protein-protein interaction sites play a crucial role in proteomics. Therefore, a greater understanding of protein-protein interaction sites ultimately bridges the gap between protein-protein interaction data and structural genomic information, substantiating protein-protein interaction data as a critical platform for advanced molecular recognition research and experimental design. In this chapter, we review current computational methods for predicting protein-protein interaction sites from sequence and structure with an extendedf discussion to the future direction of this field. Protein-protein interaction data from high throughput experiments, such as Yeast-2-Hybrid (Y2H), Tandem Affinity Purification (TAP) and Co-Immunoprecipitation (CoIP), yields networks typically representing complex biological processes such as cell signaling, gene regulation, and metabolism. At a molecular level, these physical protein-protein binding events are commonly classified into three major groups in the context of their function: (1) enzyme-inhibitor complexes, (2) antibody-antigen interactions, and (3) other types of interactions [4, 5, 6]. This classification has been proved quite useful to atomic-level analysis such as protein-protein binding site prediction and protein-protein docking.

An example structural complex for each interaction class is shown in Figure 1.1. The class of enzyme-inhibitor interactions involves the inhibition or regulation of an enzyme function by physical binding. It has been shown that there is a general tendency for enzyme-inhibitor interfaces to be more evolutionarily conserved than other surface residues [7, 8, 9]. Further, antibody-antigen interactions are generally established through hydrophobic contacts involving regions of the hyper-variable loops involved in the effective recognition of complementary antigenic proteins. Interactions classified as other, include permanent homo- or hetero-oligomers (permanently bound) and other transient (temporarily bound) complexes involved in cell signaling, e.g., protein kinase interactions, G-coupled receptors and many others. Classification of permanent and transient interaction is also important as the binding sites of the two classes exhibit notable differences.

In this chapter, we review PPI site prediction methods from several angles including the features they use, the accuracy of current methods, and intrinsic limitation of PPI prediction and potential directions for improvement. As protein-protein interactions have drawn much attention in understanding complex nature of cellular activities, the importance of PPI site prediction has been increased because it can bridge protein-protein interaction networks determined by high throughput experiments and protein structure information solved by structural genomic projects. With the increasing availability of new protein structures and sequences and understanding of biophysics of protein-protein interaction, we expect to see advanced PPI site prediction methods to be applicable on a more diverse dataset of PPI sites.

1.1 Features of Protein-Protein Interaction Sites

There are a broad range of sequence-based and structure-based features which can be used to predict protein-protein interaction (PPI) sites of a protein. A PPI site prediction method aims to predict amino acid residues which bind to another protein (i.e. closer than a threshold value to the binding partner) from the primary sequence

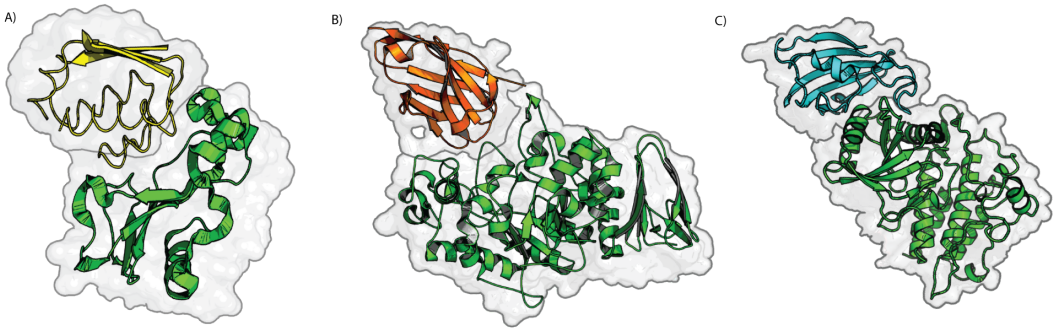

Fig. 1.1. Examples of the different groups of protein-protein interactions: **(A)** PDBID: 1BRS, a Barnase-barstar complex **(B)** PDBID: 1KXQ, an alpha-amylase to antibody interaction, and **(C)** PDBID: 1B6C, a kinase to isomerase interaction, representing the all other types of interactions.

or a set of the sequence and the tertiary structure of the protein of interest. Residues at the PPI sites that make significant energetic contribution to binding constitute hot spots (Bogan, Thorn 1998). In principle, hot spots on a protein surface are to be predicted by biophysical considerations [10,11,12,13,14] but they can be also generally predicted by the current methods of PPI site prediction, which use sequence-based and structure-based features.

Features useful for predicting PPI sites can be generally classified into three different categories: (1) Sequence-based features: information that is directly taken from sequences without structural information, (2) Low-resolution structural features: sequence data that make use low resolution spatial data to find neighboring residue so that surface clusters or patches can be readily identified, and (3) High resolution structural features: structural properties that require high resolution atomic detail. In the following sections, we will review various critical protein features that fall into these general categories. Some of the features are used alone but it is more common to combine several features to construct a composite score.

1.1.1 Sequence-based features

As whole genomes continue to become available, protein sequence information is almost always available and can be readily applied for a target protein even when its structure has not experimentally solved or accurately predicted. Furthermore, as genomic sequence projects continue to deposit sequences at a rapid rate, methods that make use of sequences features will become more important.

Residue Propensity

It is clear that certain residues are more preferred at PPI sites than other residues [15, 16] and thus information related their amino acid composition biases should be incorporated as an important feature in PPI site prediction methods [13, 17, 18, 19]. In particular, methods make use of the PPI site residue propensities, which measures the percent occurrence of amino acid residues at PPI sites over the percent occurrence amino acid residues from the background. Studies have shown that permanent complexes have high propensity for aromatic and hydrophobic residues at PPI sites compared to sites of all other surfaces [20]. On the other hand, transient complexes generally have more polar residues at regions that interact with other proteins [21, 22, 23]. Local surface patches of aromatic residues is indicative of π-π stacking interactions between two complexes or cation-π interactions [24]. π-π stacking involves an aromatic ring-to-ring interaction and cation-π interactions involve an aliphatic pair interaction (tyrosine to lysine residue contacts). Although rarely found at protein surfaces, exposed cysteine residues are likely to be involved in forming a disulfide bridge with an adjacent protein, which helps stabilize the resulting complex.

Sequence Conservation

When a set of homologous sequences are available, a multiple sequence alignment of the sequences will provide additional value in a sequence analysis. Residues

are conserved if they have structural and/or functional importance, and residues in PPI site show some degree of conservation [25]. Sequence conservation is commonly represented in the form of sequence profile. Sequence profiles represents the amino acid occurrences at individual sequence positions in a given protein family alignment. Specifically, common sequence profiles take form of a position specific scoring matrix (PSSM), which describes the probability of each of the standard 20 amino acids at each position of a sequence in question. Given a query sequence, PSSMs can be constructed by using PSI-BLAST [26], where the significant search results are used to generate the profile. Machine learning techniques, e.g., Neural Network and Support Vector Machines, are commonly applied for predicting PPI sites using PSSMs as input [27, 28, 29]. The Shannon Entropy [30] is alternative approach for measuring sequence conservation among protein families [31]. The form of an entropy score involves the following: a given K number of amino acid types (typically 20 from the number standard amino acids), N total number of residues in the column, n_a number of type a residues, and the f_a frequency of type a residue within all N residues in a given column ($f_a = n_a/N$), we use equation 1.1 to define Shannon entropy.

$$S = -\sum_{a}^{N} f_a log_2 f_a \tag{1.1}$$

For each column, the value of S ranges from zero to a maximum theoretical value of $\log_2 K$ when all amino acids are equally frequent. Larger values of S indicate higher entropy or residue-type diversity and lower values indicate less residue-type diversity and more evolutionary sequence conservation. Further, residues may be classified into chemically similar groups. A common classification would be to specify K=6, where residues are grouped into aliphatic, aromatic, polar, positively charged, negatively charged, and unusual conformations [32]. Such a classification would consider chemically similar amino acid residues into the calculation of sequence entropy.

However, various studies have shown that sequence conservation alone is not very discriminative for PPI sites, because residues of several other classes, such as those locating at protein cores and small chemical ligand binding sites of enzymes, are equally

well or more conserved than PPI sites [33, 34, 35, 36, 37, 38, 39, 40, 41, 21]. In addition, the quality of the MSA may be problematic, especially for protein families containing sequences of large evolutionary divergence, where multiple sequence alignment becomes challenging. In such cases, multiple structural alignments would provide better quality MSAs, if requisite structures are available. Nevertheless, with the assumption that the MSA is of sufficient quality, sequence conservation information should be used in combination with other features.

Correlated Mutation Analysis

Correlated mutation analysis (CMA) is a technique for detecting compensatory sequence variations among protein family sequence alignments. The assumption is that interacting residues share co-evolutionary mutation patterns. When a mutation occurs in a sequence position, its spatially neighboring residues will also change to compensate for this mutation in a complementary manner. CMA was first used to measure the relationship between these mutations and intra-residue contact pairs in proteins structures [42], where residue positions that have high correlated mutations are likely to be also spatially close to each other. Similarly, correlated mutations are also used for identifying inter-residue contacts where they correspond to protein-protein interaction sites [43, 44]. Pazos and Valencia further applied CMA to detect interacting proteins from a large set of protein pairs [45]. The extent of correlated mutation for two MSA positions can be calculated using equation 1.2. Given an all-to-all pair-wise comparison of sequence k and l in a multiple sequence alignment, the mutational relationship between two alignment positions i and j are evaluated by a correlation coefficient shown in equation 1.2:

$$r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{(s_{ikl} - \bar{s_i})(s_{jkl} - \bar{s_j})}{\sigma_i \sigma_j} \tag{1.2}$$

Where N is all the possible number of pair-wise comparisons between a sequence k and j, $\bar{s_i}$ and $\bar{s_j}$ are mean values of N^2 mutation scores in the position i and j

respectively, while σ_i and σ_j are standard deviations of the N^2 similarity scores in the position i and j respectively. s_{ikl} is the similarity score of position i in sequence k and l, and s_{jkl} is the similarity score of position j in sequence k and l. A similarity matrix provides the individual similarity scores for the pair-wise mutations represent the level in which the physiochemical properties between mutations are conserved.

1.1.2 Low-resolution Structural Features

Low-resolution structural features involve the minimal use information that do not consider the fine details of the small variations in side chain and backbone conformations. Low-resolution structures, which do not have atomic details, are still very useful to identify surface residues which are spatially clustered. This is especially applicable to low-resolution structures that be obtained by protein structure prediction [46, 47, 48]. When a low-resolution structure for a target protein is available, PPI site prediction can be made with patch-based methods. A patch is defined either as (1) a spatial region of a given surface residue and its neighboring residues with a defined cut-off distance value, essentially capturing residues into a sphere, or (2) a fixed number of residues that form a cluster of residues around a single site of interest. The first patch definition can be simply implemented and extracts a uniform circular shape on the protein surface. If PPI site residues cluster into shapes that are non-circular or irregular in shape, the second patch definition is more appropriate.

Hydrophobic Patch

A typical example of using a patch, is to consider the average hydrophobicity of patches, because PPI sites, particularly permanent PPI sites, generally consists of hydrophobic residues [49,50]. For a given patch, all residues in the patch are assigned a hydrophobicity value [51]. Taking the average hydrophobicity scale value of all the residues in the patch defines hydrophobicity. All possible patches are then evaluated

for a given structure and the patch is the most hydrophobic would be designated as the PPI site prediction.

Secondary Structure

Some preferences in secondary structures at PPI sites are observed for certain classes of proteins. Protein binding site regions of transient homo-complexes have been found to favor β-sheets [17]. β-sheets of one protein may bridge β-sheets of another, extending the β-sheet structure from one protein interface to another adjacently bound partner. To a lesser extent, there are also PPI sites that mainly consist of α-helical regions. For example, coiled-coil regions largely consist of the interaction between the hydrophobic regions in α-helices and are common in transcription factors that form complexes to regulate gene expression. Lastly, intrinsically disordered regions are commonly found to be involved in the transient interactions of hub proteins [52, 53, 54, 33, 55].

1.1.3 High-resolution structural features

Atomic structural detailed enables consideration of biophysical properties (e.g., electrostatic and desolvation energies) and the precise geometric shapes describing surface regions (e.g cavities and protrusions). In principle, computing these features requires a protein structure solved by X-ray crystallography and nuclear magnetic resonance (NMR).

Atom Group Propensity

Similar to the amino acid residue propensity discussed in the sections sequence-based features, propensity of atom groups to be on PPI site can be considered. Kufareva et al. classified heavy atoms in amino acids into 32 types according to their chemical features including charge and sp-, sp2-, or sp3-hybridization and used partial

least squares regression to computed the probability to find each atom group on the interface. Using an optimal probability threshold value, surface patches are predicted to be a PPI site or not [56].

Desolvation Energy

PPI sites consist of residues that are composed of atoms that contribute to the general hydrophobic character of the protein surface that is involved protein-protein association. Rather than evaluating hydrophobicity on a per residue basis, atom based desolvation could more precisely discriminate between PPI sites and non-PPI sites because desolvation features are not limited to only hydrophobic residues. Therefore, atomic based desolvation parameters (ASP), which evaluate the free energy of transferring from water to a protein-protein interface, have been optimized for protein-protein docking application [10]. These parameters were originally determined from water/octanol transfer experiments [57]. Desolvation energies are particularly discriminative for predicting interacting regions of antibodies [58]. However, prediction performance on antigens compared to antibodies is considerably less discriminative. Nevertheless, use of desolvation energies still results in prediction accuracy that is noticeably better than using sequence conservation alone as a feature [58].

Relative Accessibility Surface Area

The average relative accessible surface area (ASA) of residues in a local patch is found to have discriminating power for patches at PPI site from other surface patches [20], i.e. PPI site patches tend to have larger relative ASA. ASA is determined using a probe sphere with an approximate size to that of a water molecule to trace the protein surface in three-dimensions. Note that the relative ASA of a residue in a protein structure is defined by the ratio of the observed ASA of the residue relative to the reference ASA. And the reference ASA of an amino acid residue X is obtained by calculating the ASA of Gly-X-Gly peptide in extended conformations [59]. This

interesting observation indicates that a patch with residues that are more exposed than generally expected are preferred to be covered by binding to another protein.

Molecular Surface Shape

Protein molecular surfaces have been defined in terms of cavities, protrusions, and flatness. These abstract molecular surface shapes are important for identifying unique and functionally important regions [60]. Given that the protein surface is discretized into a uniform grid of voxels. One way to define cavities and protrusions is to determine the visibility of each voxel on the surface. Visibility is defined as the number of directions in which a surface voxel that can cast rays without touching surrounding surface voxels. The lower the visibility, the deeper the position of a voxel in a the cavity. Conversely, the greater the visibility, the higher a voxel will be positioned near the peak of the protrusion. As mentioned in the previous section of relative ASA [20, 61], it was found that PPI sites are geometrically flat, in general. The flatness of a surface patch is defined as the least square fit of a plane to atoms of residues in the patch. Jones and Thornton examined another geometrical measure, named protrusion. Further, the cavity and protrusion is defined as the average cavity or protrusion of each residue in the patch from the surface of the protein. Thus intuitively, the flatness indicates an overall geometrical shape of the surface of a patch, and the protrusion indicates local ruggedness of the patch.

Side Chain Energy

Liang et al. used a composite score for evaluating the energy of side chain conformation (e.g., rotamer) of surface residues, called the side chain energy, and showed that residues at PPI sites have higher side-chain energy than other surface residues [13]. The side chain energy is calculated as a linear combination of several energetic terms including side-chain atom surface contact area ($S_{contact}$), overlap volume ($V_{overlap}$) hydrogen bonding energy (E_{hbond}), electrostatic interaction energy

(E_{elec}), hydrophobic solvent accessible surface (S_{pho}), hydrophilic solvent accessible surface (S_{phi}), difference between the fraction of buried hydrophilic atom surfaces from rotamers in the presence and isolation of surrounding residues ($\Delta {F_{phi}}^{30}$), solvent exclusion volume around charged atoms ($V_{exclusion}$), frequency of observed rotamers (f_1), frequency of conformations of a given backbone residue (f_2), the presence of a disulfide bridge (N_{ss} bond), and the difference between the rotamer free energy in solvent and the denatured state (ΔG_{ref}). The side chain energy for a given residue (R_i) is given in the following equation: (3) The weights are optimized so that each term with reference values best fit the observed energetically favorable residue conformations. The higher side-chain energy of PPI sites for unbound monomer implies that the free energy of two interacting proteins declines significantly upon association.

$$\begin{aligned} E_{sidechain}(R) = & -0.143 S_{contact} + 0.724 V_{overlap} + 1.72 E_{hbond} + 28.6 E_{elec} \\ & - 0.0467 \Delta S_{pho} + 0.0042 \Delta S_{phi} + 1.14 \Delta (F_{phi})^{30} \\ & + 7.95 V_{exclusion} - 0.919 \ln(f_1 f_2) - 4.3 N_{ssbond} - \Delta G_{ref} \end{aligned} \tag{1.3}$$

B-factor

B-factors or temperature-factors of an X-ray crystal structure represent the atomic positional variation that quantify the average displacement of X-ray scattering caused by thermal motion. Higher *B*-factors are generally found at regions of proteins that are flexible (e.g., loops or disordered regions), whereas lower *B*-factors usually correspond to regions that are less flexible (e.g., the compact core of the protein). Hydrophobic protein cores are tightly packed and exhibit less flexibility and hence lower *B*-Factors. Similarly, PPI sites of bound complexes also have low *B*-factors because they are not exposed to bulk solvent and exhibit less thermal motion. Interestingly, it was found that the *B*-factors of unbound PPI sites also exhibit lower *B*-factor values compared to non-protein binding regions [17]. Neuvirth and Schreiber suggested that the preference for a low *B*-factor might be due to the potential of binding sites to be involved in crystal contacts. Nevertheless, as the formation of crystal contacts at a certain position would suggest physicochemical differences of the involved sur-

face patches, the B-factor is still useful for predicting PPI sites of an X-ray crystal structure.

Water Molecule

It is common that a X-ray crystallographic structure of a protein-protein bound complex contains water molecules at its PPI site, forming water-mediated polar interactions [15]. Surface shape complementarity is often better established when water facilitates close atomic packing to fill voids and spaces at PPI sites. Neuvirth and Schreiber further found that PPI sites of unbound monomers have a higher water content compared to other surface regions [17]. This observation may be rationalized by the polar residues found at PPI sites, which provide the coordination points for water.

Electrostatic Potential

Burgoyne and Jackson tested peak and average values of electrostatic potential for local surface regions as a predictor of PPI site [58]. The electrostatic potential is computed by the finite difference Poisson-Boltzmann calculation. It was found that electrostatic potentials showed some predictive capacity for enzymes but did not show general trends in the prediction of protein-protein interfaces. The results imply that electrostatic complementarity play an important role in orienting enzymes and inhibitor complexes.

1.2 Existing Methods for Predicting PPI Sites

In the previous section we have seen various features that can be used for predicting PPI sites. In most of the prediction methods, several features are combined to capture different characteristics of PPI sites to improve the prediction accuracy. Due to the fact that various methods robustly use a combination of features to improve

PPI site discrimination, the number protein features used for predicting PPI sites outnumber the methods that use them. As examples of how features are combined in actual prediction methods, in this section we overview currently available PPI site prediction methods. The features these methods use are summarized in Figure 1.2 and their corresponding websites are listed in Table 1.1.

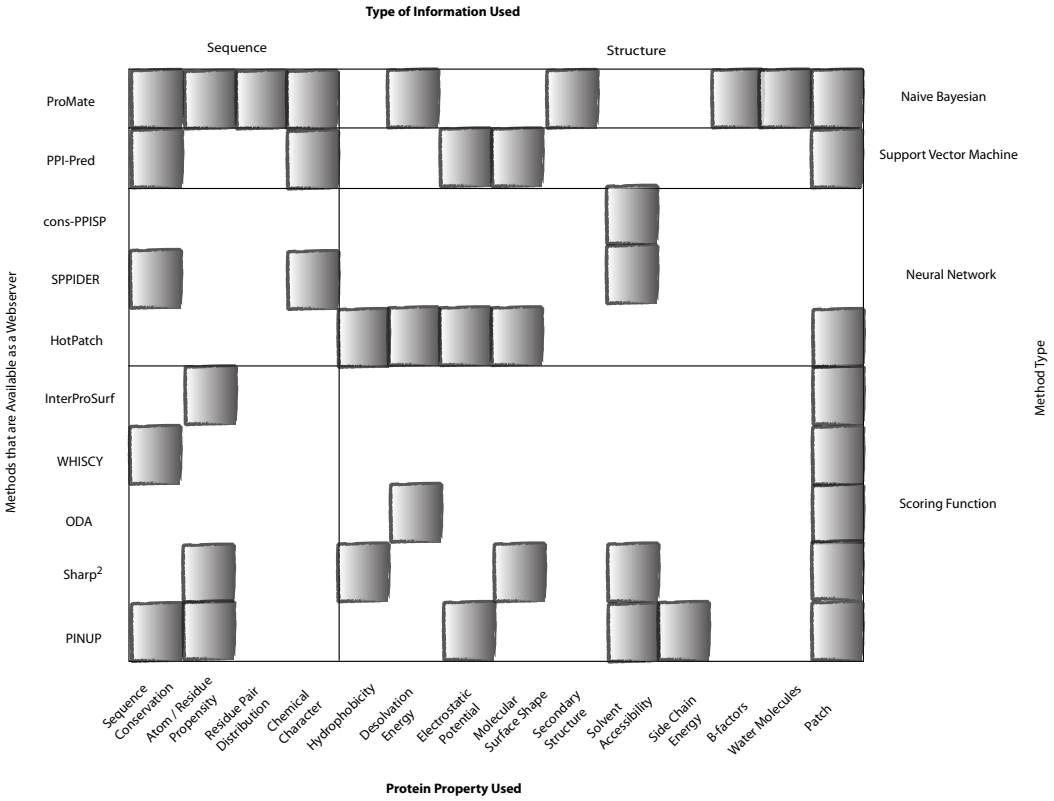

Fig. 1.2. Comparison between methods and the sequence- or/and structure-based features they use.

ProMate utilizes a naïve Bayesian to integrate the following features [17]: sequence level features include evolutionary sequence conservation, single residue frequency, paired-residue frequency, and residue chemical character (aromatic, positively charged, negatively charged, hydrophobic). Structure-based information includes de-

Table 1.1
Available resources for protein binding interface prediction.

Available Methods	**URL**
Class 1: Use sequence and structural information	
Promate	`http://bioportal.weizmann.ac.il/promate/`
PPI-Pred	`http://www.bioinformatics.leeds.ac.uk/ppi_pred/`
Cons-PPISP	`http://pipe.scs.fsu.edu/ppisp.html`
SPPIDER	`http://sppider.cchmc.org/`
Sharp^2	`http://www.bioinformatics.sussex.ac.uk/SHARP2/`
PINUP	`http://sparks.informatics.iupui.edu/PINUP/`
InterProSurf	`http://curie.utmb.edu/prosurf.html`
WHISCY	`http://nmr.chem.uu.nl/Software/whiscy/index.html`
Class 2: Use only structural information	
Optimal Docking Area	`http://www.molsoft.com/oda.cgi`
HotPatch	`http://hotpatch.mbi.ucla.edu/`
Class 3: Integrate predictions from various servers	
Meta-PPISP	`http://pipe.scs.fsu.edu/meta-ppisp.html`

solvation energies, secondary structure elements, B-factors and water molecules. All of such sequence- and structure-based information is derived from a defined patch encompassed by a radius of 10 Å drawn from a center of a given protein surface residue. Compared to other methods for PPI site prediction, this approach uses the largest number of features. An extended server, called ProMateus, allows the user to upload custom features to the naïve Bayesian approach to add to the set of fea-

tures of used in ProMate [62]. PPI-Pred uses a support vector machine (SVM) to integrate sequence-based features, including sequence conservation, sequence profile, the chemical character of a given residue, and structure-based features, which include electrostatic potential and molecular surface shape data ([63]. This approach uses features similar to that found in ProMate except it does not consider B-factors and position of water molecules. PPISP (Protein-Protein Interaction Site Predictor) method uses a position specific sequence profile derived from PSI-BLAST and solvent accessibilities of known PPI sites as input for neural network (NN) training [27]. To further improve the prediction performance, multiple PPISP NNs are trained and combined into a consensus-scoring framework called cons-PPISP.

Porollo and Meller found that surface residues that are incorrectly predicted to be part of the core of the protein from their sequence characteristics frequently correspond to PPI sites. They have implemented this interesting idea into a PPI site prediction program, SPPIDER (Solvent accessibility-based Protein-Protein Interaction site Identification and Recognition) [29]. SPPIDER basically makes use of the difference between the predicted residue solvent accessibility from their sequence-based features and actual residue solvent accessibilities computed from the structure of the input protein into a NN architecture. The sequence-based features used include the measure of expected number of neighboring residue contacts, sequence conservation (entropy), residue frequencies and sequence profiles from PSI-BLAST, sequence conservation of chemical property, size, and hydrophobicity. InterProSurf solely relies on residue PPI site propensity. It identifies clusters of residues on a protein surface that are observed at high frequency in a representative dataset of PPI sites [18, 19]. The Sharp^2 (Solvation potential, Hydrophobicity, Accessible surface area, Residue interface propensity, and Planarity and Protrusion) method measures three geometric features, e.g., surface planarity, average protrusion of residues, average relative accessible surface area of residues, along with three other residue-based features, namely, solvation potential, hydrophobicity, interface propensity [64]. These six features in relation to PPI sites are examined in their previous work [61]. The three geometrical

features are similar to those explained in the previous section. Each parameter is normalized so that the resulting values range from 0 to 100 and an average value for all parameter values is the representative score for a given patch in question. PINUP (Protein Interface residUe Prediction) uses the side chain energy explained in the previous section, linearly combined with a residue conservation score and residue PPI site propensity [13]. Further, Liang et al. reported that the residue PPI site propensity score performed the best among the three individual terms and the combination of the three showed improvement on top of it. The Optimal Docking Area method computes the docking desolvation energy of a given surface patch [10]. The desolvation energy of a patch is a sum of parameterized per-atomic contributions (σ_i) in proportion to the solvent accessible surface area (ASA): (4) where i is an atom included in the patch. The atomic solvation parameter, σ_i, was determined by using octanol/water transfer energies measured from experiment [65] and further optimized for protein-protein docking [66]. For each surface point distributed on a protein surface, the desolvation energy of a series of patches defined by a surface sphere with different radii is computed. Then the lowest energy among the patches is assigned to the surface point. Finally the patch of a surface point that give the lowest desolvation energy is considered to be a putative PPI site. Hotpatch designed to identify various functional sites on a protein surface, including catalytic sites of particular enzyme classes, metal ion binding sites, DNA/RNA interaction sites, and PPI sites [50]. It integrates several features in a NN architecture depending on which sites are going to be predicted. When predicting PPI sites, it integrates residue-based hydrophobicity with other structure-based features, including electrostatic potential, charge, surface shape, residue-based hydrophobicity and atomic-based desolvation energies. WHISCY (WHat Information does Sequence Conservation Yield) uses sequence conservation alone to predict PPI sites [7]. The conservation scoring function used by WHISCY is a pair-wise comparison of sequences in a multiple sequence alignment using a modified Dayhoff amino acid substitution matrix [67]. Further, the WHISCY-MATE option provided by the server combines the conservation scores with ProMate

predictions. The resulting predictions are then used to guide protein-protein docking strategies (Dominguez et al. 2003). Meta-PPISP takes a consensus approach [68,69]; it combines raw scores from ProMate, PINUP and Cons-PPSIP through linear regression to produce a final score. In general meta-server approaches can combine strengths of component methods thus successful in many prediction methods in bioinformatics. Like wise Meta-PPISP showed a better performance over the three component methods in their paper.

1.3 Prediction Performance of Current Methods

We have overviewed useful features for predicting PPI sites and how existing methods combine those features. To understand the current status of PPI methods, we will discuss the prediction accuracy. We will review two recent benchmark studies by Zhou and Qin [69] and Huang and Schroeder [70], because they assessed the performance of most of the methods discussed in the previous section. The prediction performance is commonly measured in terms of the sensitivity (the proportion of correctly predicted residues at PPI sites relative to the actual residues at PPI sites) and the positive predictive value (PPV) (the proportion of the correctly predicted PPI site residues relative to the total number of predicted PPI site residues). The PPV is plotted relative to the sensitivity due to the tradeoff in sensitivity and the PPV. Zhou and Qin prepared two benchmark datasets; one dataset consists of 35 proteins in the enzyme and inhibitor category from the ZDOCK protein-protein docking benchmark dataset [5] and another one which consists of 25 docking prediction targets from CAPRI (Critical Assessment of Prediction of Interactions) prediction contest (`http://www.ebi.ac.uk/msd-srv/capri/`). The CAPRI dataset contains eight proteins that do not belong to the enzyme/inhibitor category, e.g., antibody/antigen or other immune system complexes. Six webservers in Table 1.1, namely, cons-PPISP, ProMate, PINUP, PPI-Pred, SPPIDER, and Meta-PPISP were benchmarked. In the 35 enzyme dataset, PPI-Pred, SPIDDER, cons-PPISP, ProMate, PINUP, and meta-

PPISP showed the PPV of 27, 33, 36, 36, 48, and 50%, respectively, at a sensitivity of 50%. These methods performed uniformly worse on the other dataset of 25 CAPRI targets, probably because proteins in the CAPRI dataset are more diverse. At a sensitivity of 30%, the methods showed specificities of 23, 25, 26, 26, 28, and 31, respectively (the methods are placed in the same order). Note that the purpose of showing the actual numbers is to have a general idea of the performance of currently available methods but not to rank webservers. Actually a fair comparison of existing methods is difficult, due to the fact that some of the tested proteins may be used in training the methods. We can conclude from this benchmark study is that the current prediction methods show a PPV of around 27 to 50% at the sensitivity of 50% and that a better performance can be achieved by taking a meta-server approach through Meta-PPISP. A comparison of the benchmark performances (measured by sensitivity verses PPV curves) for each PPI site prediction method can be found in the review by Zhou and Qin [69]. In attempt to formulate an alternative meta-server solution, Huang and Schroeder developed MetaPPI which combines five individual PPI site prediction methods including PPI-Pred, PINUP, PPISP, ProMate, and SPPIDER. Furthermore, they reported success rates for these 5 methods in comparison to their MetaPPI method on the similarly derived dataset of 62 protein-protein complexes taken from the combined the ZDOCK benchmark dataset and the CAPRI targets. The performance of PPI site prediction methods evaluated here is quite different from the above previous review in that the datasets were classified by the protein interaction categories rather than from the source in which the protein complexes were acquired. Therefore, they used two datasets consisting of 20 complexes classified as enzyme-inhibitors and the remaining 42 complexes that were classified as other types complexes. Huang and Schroeder reported success rate of five of the PPI site prediction methods mentioned in this chapter and one meta-server method they call MetaPPI. Success rate is defined as the percent complexes in the dataset with PPI site predictions of 50% PPV. For the enzyme-inhibitor category, the methods including PPI-Pred, PINUP, PPISP, ProMate, SPPIDER, and MetaPPI performed with 45,

52, 55, 36, 23, 70% success rate respectively. While for the category of other types of complexes, PPI-Pred, PINUP, PPISP, ProMate, SPPIDER, and MetaPPI performed with 28, 15, 25, 13, 10, and 44% success rate respectively. Enzyme-inhibitor complexes can be generally predicted with greater success than other types of complexes. More information on the benchmark results can be found in the review by Huang and Schroeder [70]. Huang and Schroeder, similar to the review by Zhou and Qin, also reported the sensitivity and PPV of each of the 5 methods and compared the performance to their MetaPPI method. For the test cases which were included both categories in the dataset, PPI-Pred, PINUP, PPISP, ProMate, SPPIDER, and MetaPPI provided the average sensitivity of 33, 29, 29, 12, 42, and 25% respectively. Using the same benchmark dataset, PPI-Pred, PINUP, PPISP, ProMate, SPPIDER, and metaPPI performed with the average PPV of 31, 31, 33, 24, 29, 45% respectively. The performance described here might not seem very good. However, encouragingly, these studies showed that a PPI prediction with 42-45% of the sensitivity and the PPV in PPI site prediction can potentially be used to guide protein-protein docking.

1.4 Discussion

Current methods use a combination of sequence- and structure-based features but there still much room for improvement. A part of the difficulty of PPI site prediction comes from the fact that different classes of protein-protein interaction have different characteristics of interaction. For example, PPI sites of enzymes are known to have relatively well conserved residues, while sequence conservation information is less useful in predicting PPI sites of antigens in antibody/antigen interactions, where epitope regions might not always be conserved in sequence. Further, sequence variation at antigenic sites may not exhibit any biological pressure in terms of their sequence evolution (because the antibody is evolving to binding to the antigen rather than the reverse), it is important to understand the biophysical properties that allow antibodies to be specific in recognizing antigenic proteins. Similarly, PPI sites of

permanent complexes and those of proteins that undergo transient interaction with several different proteins may use different biophysical mechanisms for establishing binding, such as those that lead to tightly controlled association and dissociation in signaling proteins. Therefore, developing class specific prediction methods separately for different classes of PPI may improve prediction performance. Another potential reason for the limited accuracy of current PPI methods is that usually prediction is made without using information from the binding partner. Of course, it is not trivial to specify the PPI sites of a docking partner (except for a case of a PPI site of an antibody), but incorporating some information about the partner may make significant improvement, and if implemented an intelligent way, will contribute an improvement in the accuracy of PPI prediction. Finally, the effect of water at the PPI site, which may play a critical role in protein-protein binding, is not well understood biophysically and still limited in use for PPI site prediction. As mentioned above, water molecules are frequently found in the PPI sites of crystal structures of binding proteins. In addition, recent molecular dynamic simulations indicates that residues at PPI sites exhibit less water molecule traffic (less water molecules that are constantly associating and dissociating with residues) than other sites, showing that the hydrogen bonds at interfaces to water are more stable than those that are located away from the binding interface [71]. A general understanding of interplay between water molecules and a PPI site would provide very important information for predicting PPI sites. Although current PPI site prediction methods have limitations, with more structure and sequence data of protein complexes becoming available, the performance of prediction methods should only improve in the future.

2. EXPERIMENTAL METHODS FOR DETERMINING BINDING CONSTANTS IN PROTEIN-PROTEIN INTERACTIONS

Many critical functions in the cell are mediated by the precise coordination of diverse types of protein-protein interactions. An important aspect of this diversity lies in the spectrum of binding affinities of many different interacting proteins. For example, enzyme to protein-based inhibitor interactions or multimeric enzyme complexes require strong binding and will be permanently bound, whereas protein-protein interactions involved in signal transduction or membrane trafficking are weaker, thus are critical for coordinating association and dissociation to maintain and control of biological processes. The classification of protein-protein complexes into permanent and transient types using their experimentally determined binding affinities will be later discussed in chapter 4. Equilibrium binding constants are generally used to measure the extent in which proteins interact with each other. Various experimental methods for determining *equilibrium* dissociation constant K_d and *equilibrium* association constant K_a will be described in this chapter.

2.1 Protein-Protein Binding Constants

The K_d constant is a general measure of the propensity in which protein pairs (receptor and ligand proteins) to interact [72]. The K_d value is calculated given the concentration of receptor protein [R] and ligand protein [L] using equation 2.1. Conversely, is the equilibrium association constant $[K_a]$ is calculated using equation 2.2. Here, lower K_d values mean stronger interactions, whereas lower K_a values represent weaker interactions.

In calculating K_d or K_a, [RL] represents the concentration of bound receptor and ligand complex, whereas [R] and [L] refer to the concentration of the unbound receptor and unbound ligand respectively. Further, the association rate constant of receptor and ligand is k_a [R][L], while the dissociation rate constant is k_d [RL]. When the binding rate of RL equals unbinding rate of RL at equilibrium, the K_d is calculated as k_d/k_a. An introduction to the various experimental methods for determining equilibrium K_d values for protein-protein interactions will be presented in the following sections.

$$K_d = \frac{[R][L]}{[RL]} \tag{2.1}$$

$$K_a = \frac{[RL]}{[R][L]} \tag{2.2}$$

2.2 Experimental Methods for Determining Binding Constants

Various methods can be used for determining K_d values for protein-protein interactions. Common methods described in the literature include those that use chromatography, various methods in fluorescence, calorimetry, and stopped-flow fluorimetry. In addition, an increasingly popular method using surface plasmon resonance will also be reviewed.

2.2.1 Protein Affinity Chromatography

Estimates of K_d can be determined by protein affinity chromatography [73]. Columns with different concentrations of covalently attached protein receptors are initially prepared. The solution of interacting ligand proteins are then loaded on to the columns, washed with buffer (ten times the volume of the column) and eluted with sodium dodecyl sulfate (SDS), an anionic detergent that denatures protein structures. For ideally strong interaction between protein pairs, a larger fraction of ligand proteins will be bound to the immobilized receptor proteins in columns of higher concen-

tration. The population of bound ligand proteins would immediately decrease for columns with lower concentrations of covalently tethered receptor proteins. Therefore, the detection of the protein-protein interactions by chromatography is dependent on the concentrations of bound proteins. This method assumes that the ligand is in equilibrium during the elution of the receptor. Also, binding of solid-phase proteins and liquid-phase proteins are assumed to be the same as interactions occurring in solution. Nevertheless, affinity chromatography agrees well with other methods (mentioned later in this chapter) for determining the K_d.

2.2.2 Fluorescence Spectrum

Fluorescence is a highly sensitive method for the detecting of changes in protein binding through the use of tryptophan residues [74]. On the formation of the bound complex, the fluorescence emission spectrum can change through the wavelength shift of the peak fluorescence emission or fluorescence intensity, from which the dissociation constant can be determined. However, this technique relies on proteins with tryptophan, which rarely occurs in proteins. Alternatively, the attachment of fluorescence tags can be used to measure protein interactions with greater sensitivity.

2.2.3 Fluorescence Polarization or Anisotropy

Fluorescence can also be used to monitor the rotational motion of molecules by using plane-polarized light for excitation, and then measuring the emission at parallel and perpendicular planes [75]. Rotational correlation times are dependent on the molecular size ($\tilde{1}$ ns per 2,400 Da). When two proteins form a complex, the molecular size increase, which in turn also increase the rotational correlation time. Therefore, the binding affinity can be determined. Further, fluorescent anisotropy is commonly performed on proteins with covalently attached tags, which increases the fluorescence signal intensity and the fluorescence duration of the excited state. However, the

addition of tags may induce artificial changes to the protein structure, which would affect binding.

2.2.4 Fluorescence Resonance Energy Transfer

Fluorescence resonance energy transfer (FRET) is a useful method that can be used for detecting protein-protein association and dissociation [76]. FRET is an transition in energy between donor and acceptor fluorophores when they are in close in proximity (around 100 Å). The protein pairs of interest would have a donor and acceptor fluorophore attached to the surface respectively. For example, green fluorescent protein (GFP) and yellow fluorescent protein (YFP), are frequently used as acceptor and donor fluorophores. Furthermore, different concentrations of ligand proteins that are bound to receptor proteins would result in different levels of energy transfers between donor and acceptor fluorophores, and thus can be used to estimate the K_d value. The attachment of fluorophores has the same disadvantages as those described for fluorescence polarization or anisotropy.

2.2.5 Stopped-Flow Fluorimetry

The stopped-flow device is a rapid mixing instrument can be used to study protein-protein interactions [77]. Given a receptor protein solution and ligand protein solution tagged with fluorophores, they are mixed and florescence is measured once binding occurs. The time between the end of mixing and the starting of binding kinetics is referred to as the deadtime, which generally lasts around 1-2 milliseconds. The stopped-flow fluorimeter is capable of measuring the association constant and dissociation constant in the time scale of milliseconds. The stopped flow technique is advantageous because it's simple and it gives low error. The disadvantage is that it requires consistent time control for precise florescence measurements.

2.2.6 Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) is a highly quantitative method for measuring the thermodynamic parameters in protein-protein interactions and is becoming increasingly important for structural studies [78, 79]. This method measures the precise heat changes that occur during interactions between proteins in solution. Unlike fluorescence methods, ITC does not require an attached tag or immobilized protein. The absorption or production of heat is a natural property for any biochemical reaction. Conveniently, in addition to the K_d, information regarding the enthalpy, entropy, and the stoichiometry for binding in interacting protein pairs are also provided. Unlike florescence methods, another advantage of the ITC method is free of artifacts, where it does not require a florescence tag to be attached to proteins. However, ITC requires high concentration of interacting protein pairs.

2.2.7 Surface Plasmon Resonance

Surface plasmon resonance (SPR) can be used to monitor protein-protein interactions in real time by measuring changes in the resonance angle of light affected by the gold surface [80]. The resonance angle is the angle of light that excites surface plasmons on the gold surface and when the minimal light intensity of reflection is reached. The receptor protein of interest is attached to a dextran polymer and a solution of the ligand protein flows through a cell. When the ligand protein interacts with the tethered receptor protein, the resonance angle changes, in turn altering the refractive index. The extent of change in the refractive index reflects the concentration of bound complexes. Therefore, the amount of protein-protein interactions can be measured using the monitored change in the refractive index. A great advantage of the SPR method is that the direct measurement of the on rate (k_a) and off rate (k_d) of protein-protein interactions can be measured in real time.

In particular, the equilibrium binding constant can be determined using two values: (1) increase in resonance units and (2) decrease in resonance units. The increase

in resonance units is measured as a function of time by transferring the solution of ligand proteins across the covalently tethered receptor proteins, until the measurement of resonance units stop changing. On the other hand, the decrease in resonance units is monitored with respect to time by washing a buffer without the ligand protein through the flow cell. The two stages involved in measuring these two resonance units allows for continuous recording of resonance units with time and can be easily repeated for different concentrations of ligand proteins (after the dextran surface has been reconstructed). Subsequently, the data that is gathered can then be graphed (ligand concentration verses stable resonance units), resulting in two curves that plots the increasing and decreasing resonance units, where the k_a and k_d can be calculated respectively. Therefore, the K_d can then be calculated using k_d/k_a.

3. A NOVEL METHOD FOR PREDICTING PROTEIN-PROTEIN INTERACTION SITES USING PHYLOGENETIC SUBSTITUTION MODELS

3.1 Introduction

Protein-protein interactions (PPIs) mediate many critical biological processes in the cell. The complexity of the interactions can be seen through the recent construction of large-scale PPI maps, which reveal the intricate functional interplay between many proteins in pathways and the formation of oligomeric complexes [2, 3, 1]. At the same time, genome sequencing projects and the structural genomic initiatives are rapidly accumulating individual protein sequences and structures [81, 82]. With the increasing availability of individual protein data and their interactions, it is becoming more essential to locate protein binding interfaces (PBIs) of many interacting proteins to bridge the gap between the global view of PPI networks and high-resolution scrutiny of amino acids that structurally form protein complexes. PBI prediction is indispensable toward this end, where it can help substantiate PPI data as a critical platform for enhanced molecular recognition research and experimental design.

Various ideas have been explored in predicting PBI in proteins, which utilize structural- and/or sequence-based features. Several structure-based methods use relative area of solvent accessibilities based on the observation that interacting residues are generally more exposed on the protein surface [17, 20, 28, 29, 63]. The other structural features used include surface shape [20, 63, 61], the crystallographic temperature factors [17], and the propensity of the secondary structure at PBI [17]. Several studies utilize specific physiochemical properties and energetic features. These features include desolvation energies [10] and complementary residue-residue charges that establish salt bridges across interfaces [83]. These features have been used by machine

learning techniques such as neural networks [84,85] and naive Bayesian classifiers [17] for PBI site prediction. The increasing value of sequence and structural features employed by PBI site prediction methods are discussed in recent reviews [86,87]. Although amino acid composition at the PBI is biased compared to non-PBI (NPBI) sites [20,83,88,89], it was reported that PBI sites are less conserved than protein cores and functional sites in proteins, such as ligand binding sites and catalytic sites [33,34]. Hence, conventional sequence conservation can be used as one of the features of PBI sites [33, 34, 90], but it is not sufficient to be used alone for prediction [34, 91]. Other than sequence conservation, phylogenetic information [36] and correlated mutation [43] has been used for predict PBI but only for specific biological instances.

Here, we have developed a novel PBI site prediction method, named Binding site prediction by Maximum Likelihood (BindML), which utilizes mutational constraints that are found in known PBI and NPBI sites. The mutation patterns of known PBI sites and NPBI sites are captured in the form of amino acid substitution models (i.e. amino acid similarity matrices). BindML uses these substitution models with a likelihood-based phylogenetic tree inference method to compute and compare the likelihood that the mutation pattern of a local protein surface patch follows that of PBI and NPBI by constructing trees. There exist sequence-based functional site prediction methods which use phylogenetic trees [36,35,92,93,94]. However, our approach is entirely different from those methods because existing methods use phylogenetic trees to examine subgroups where conserved residues are found [36,95,96,97] or to find mutation patterns which are in agreement to the tree of the protein family [92,94,93]. In contrast, our method, BindML, specifically determines whether a multiple sequence alignment (MSA) from local surface patches on the structure of the query protein exhibit mutation patterns that follow PBI or NPBI, which is achieved by evaluating the likelihood that phylogenetic trees constructed for the MSA of the local patch follow the sequence evolution of PBI and NPBI substitution models. Remarkably, BindML performs well in comparison to existing methods, which combine various sequence and structural information into machine learning frameworks. The impact

of our method is broad, since it can be easily extended for predicting other specific types of functional sites, such as DNA, RNA, or membrane binding sites in proteins. The conceptual novelty of this method is that it detects constrained sequence variations rather than conventional conservation in protein family sequences, thus aimed to open a new direction in protein sequence analysis.

3.2 Methods

3.2.1 Dataset of Protein Complexes

Multiple sequence alignments (MSAs) of protein complexes were taken from the iPFAM database of protein-protein interactions, which provided an initial dataset of 2733 pair-wise protein-protein interactions in the PFAM database (rel. 12.0) [98]. The interactions in iPFAM are based on structural evidence found from complexes in the Protein Data Bank (PDB). We applied the following eight criteria to filter the iPFAM entries: (1) the unaligned seed sequences for each corresponding PFAM family were used. (2) We selected PFAM families with 20 to 100 seed sequences yielding 748 PFAM families. (3) PFAM families consisting domain sequences were replaced with their corresponding full-length sequences from UniProt [99]. A representative PDB structure was then selected from each PFAM family given by the association in iPFAM. When a PFAM family has more than one structure representative, we selected one of them arbitrarily. (4) Protein structures that do not have any observable interacting partners in their PDB files were removed. (5) Complexes were eliminated if they were classified as monomers bound by crystal contacts in the PQS definition [100]. (6) Proteins with their PDB entries that have non-standard amino acids, too small proteins where the entire structure are part of PPI sites, short protein sequences (shorter than 40 amino acids), and obsolete PDB files were filtered out. (7) PDB structures with antibody-antigen and protein-DNA/RNA interactions were removed. (8) In the final dataset, PFAM families with redundant representative structures with $\geq$35% sequence identity were filtered out. Given that MSAs in

PFAM may not have the PDB structure as a part of the alignment, we employed MUSCLE (ver. 3.6) [101] with default parameters to compute MSAs from PFAM unaligned sequences and one sequence from the selected PDB structure. This procedure forms the final MSA dataset of 505 families. The dataset can be accessed at `http://kiharalab.org/bindml/families.tar.gz`.

3.2.2 Substitution Models

Amino acid substitution models reflect the ratio of pair-wise amino acid substitutions observed in MSAs and the same amino acid pairs appearing by chance. Substitution models for PBI and NPBI were constructed from the MSAs in the filtered iPFAM dataset. Protein surface residues were defined as those which have larger than 10% of the relative solvent accessible area in comparison with the value in the tripeptide with glycines on both sides [59]. Among the protein surface, residues at PBI were defined as those that are closer than 5 Å to any residues in the protein docking partner, otherwise residues were defined as NPBI. The observed substitutions for PBI and NPBI were counted at gapless positions in the set of pairwise set of alignments following the JTT procedure [102]. The values in the substitution matrices were calculated using the BLOSUM method [103]. The PBI and the NPBI substitution models are given Table 3.2A and 3.3B, respectively.

3.2.3 The BindML Algorithm

BindML computes phylogenetic trees for a MSA at a local surface patch using the substitution models of PBI and NPBI with associated likelihood scores. A patch is predicted to be PBI if the likelihood for PBI is sufficiently significant. A flow chart of the BindML algorithm is illustrated in Figure 3.1. The method takes a PDB structure of a target protein and a MSA of its family including the target sequence. For each surface residue, a patch is defined as neighboring residues that are within the sphere of a certain radius. The β-carbon of a given amino acid (α-carbon is used for glycine)

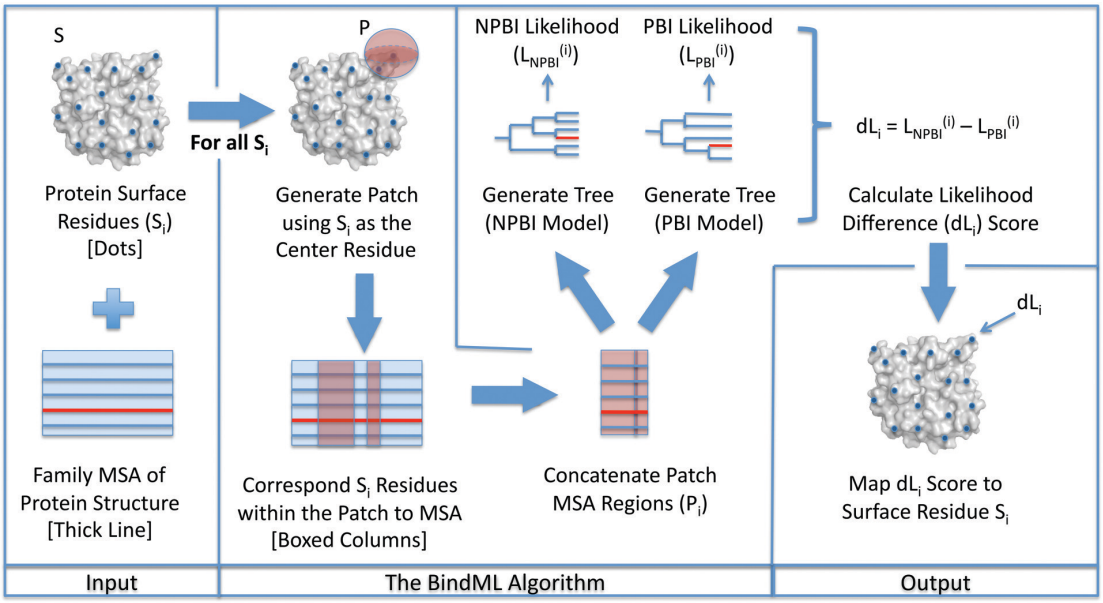

Fig. 3.1. Flowchart of the BindML algorithm.

is selected as the representative point when computing the distance between amino acids. For a patch, all corresponding residues in the MSA are concatenated together. As will be shown later in Results, a radius of 15 Å was found to perform the best.

We employed our modified version of the PHYML (ver. 2.4.5) program [104] to compute the likelihood that a patch MSA comes from PBI and NPBI by constructing phylogenetic trees with either the PBI or NPBI substitution models. Our modified version of PHYML hardcodes the PBI and NPBI substitution models. PHYML is a maximum likelihood method that builds a phylogenetic tree for a given MSA and adjusts the topology and branch length simultaneously. The algorithm starts from an initial tree constructed by the BIONJ method [105], a fast distance-based algorithm, and improves it by branch-length optimization and tree swapping procedures. BIONJ has been shown to outperform other distance-based tree inference methods [106]. To compute the likelihood of a surface patch as PBI/NPBI, we the MSA of the patch as an input into PHYML and select the PBI/NPBI substitution model and the amino acid frequency distribution of PBI/NPBI, which are used as the equilibrium frequencies.

Values in the PBI/NPBI substitution model are shifted so that all the values in the matrices are positive, which is required by PHYML.

Due to the large number of residue points on the protein surface that are needed to compute the likelihood values, we used the initial tree topology computed by BIONJ rather than the optimized tree with the maximized likelihood. This shortcut speeds up the computation dramatically, by about 15 times, and nevertheless, the prediction accuracy did not deteriorate significantly. PHYML computes the likelihood of having the input patch MSA following the PBI/NPBI substitution model given the initial tree topology. Finally, the difference of the likelihood under PBI and NPBI substitution models provides a score used to predict PPI sites. For a patch MSA, P_i, which has residue i at the center,

$$L_{NPBI} = \log\{Prob(P_i, T_i^{NPBI}|M_{NPBI})\} \tag{3.1}$$

$$L_{PBI} = \log\{Prob(P_i, T_i^{PBI}|M_{PBI})\} \tag{3.2}$$

$$dL = L_{NPBI} - L_{PBI} \tag{3.3}$$

where M_{NPBI} and M_{PBI} is the substitution model of NPBI and PBI, respectively, and T_i^{NPBI} and T_i^{PBI} are tree generated with M_{NPBI} and M_{PBI}, respectively, for the input patch MSA. Note that T_i^{NPBI} and T_i^{PBI} are not necessarily identical. The distance likelihood (dL) score is the difference between the log likelihood of the patch MSA being NPBI and PBI (Eqn. 3). Once the dL scores for all surface residues are computed, these scores are recast into Z-scores and a threshold is placed. Lower (negative) Z-scores indicate more likelihood of PBI mutation patterns, while higher Z-scores correspond to less likelihood of following the PBI substitution model. Any residues with a dL score that is equal to or smaller than a given Z-score threshold value are predicted to be included in a PPI site. A web server implementation and stand-alone program for BindML is available at `http://kiharalab.org/bindml/`.

3.2.4 Evaluation of the Prediction Performance

The prediction performance of PBI residues was evaluated mainly using the area under the curve (AUC) of the receiving operator characteristic (ROC) [107]. A ROC curve plots the false positive rate relative to the true positive rate over various scoring threshold values of a method. The overall AUC of a method was computed as the average of ROCs for every protein family in the dataset. In addition to the AUC, we also provide the sensitivity, the specificity, and the positive predictive value (PPV). True positives (TP) are the true binding interfaces residues predicted correctly, true negatives (TN) are non-protein-protein interactions sites correctly classified, false positives (FP) are false predictions of protein-protein interaction sites, and false negatives (FN) are protein-protein binding sites that are not predicted. The sensitivity is the fraction of correctly predicted PBI residues over all the true PBI residues. The specificity is the fraction of true negatives among all residues predicted to be NPBI. Positive Predictive Value (PPV) is defined as the true positives among those residues predicted to be PBI. The Matthews Correlation Coefficient (MCC) measures the correlation between observed and predicted PBI and NPBI, where a value of -1 represent inverse prediction, 0 mean random prediction, and 1 is a perfect prediction.

$$Sensitivity\ (True\ Positive\ Rate) = \frac{TP}{TP + FN} \tag{3.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3.5}$$

$$False\ Positive\ Rate = 1 - Specificity \tag{3.6}$$

$$PPV = \frac{TP}{TP + FP} \tag{3.7}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{3.8}$$

3.2.5 Other Methods used for Comparison

We compared BindML with two existing methods, ProMate [17] and cons-PPISP [85]. ProMate is based on a nave Bayesian method that uses a combination of various sequence-based (single amino acid distribution, amino acid pairs distribution, and evolutionary conserved positions) and structural-based properties (non-regular secondary structure length, secondary structure, hydrophobic patch rank, and water molecules). Cons-PPISP uses sequence conservation and relative solvent accessibilities of residues as features in a consensus neural network, which is a combination of multiple differently parameterized neural networks. Web servers of the two methods were used (ProMate: `http://bioinfo.weizmann.ac.il/promate/`; cons-PPISP: `http://pipe.scs.fsu.edu/ppisp.html`). We disregarded entries in the iPFAM dataset when either ProMate or cons-PPISP servers did not return prediction results by batch processing. Thus, 449 entries in the iPFAM dataset were used for the performance comparison with BindML.

3.3 Results

3.3.1 Comparison of PBI and NPBI Substitution Models

BindML exploits the differences in the amino acid substitutions at PBI and those at NPBI for protein binding site prediction. First, we compared the amino acid frequency at PBI and NPBI sites. Differences in the amino acids frequencies at PBI and NPBI were counted in multiple sequence alignments (MSAs) in the iPFAM dataset (Fig. 3.2A & 3.2B). Several amino acids showed noticeable differences in frequency at PBI and NPBI. Aromatic residues and hydrophobic residues were more abundant at PBI than NPBI. On the other hand, NPBI sites were richer in charged residues as compared with PBI sites. These results are consistent with previous studies [17, 20, 88, 108]. Next, we compared amino acid substitutions at PBI and NPBI sites (Fig. 3.3). We also included the BLOSUM35 matrix [103] in comparison

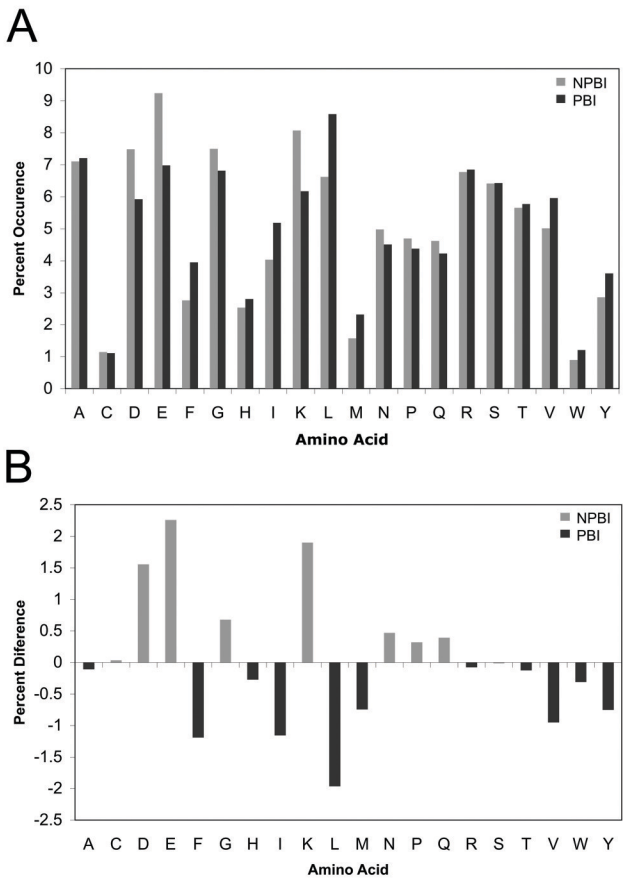

Fig. 3.2. **(A)** Distribution of the frequencies of the standard amino acids in the dataset used in this study. Black bars correspond to known protein binding interfaces (PBI) and gray bars represent non-protein binding interfaces (NPBI). **(B)** The difference of the amino acid frequency at the NPBI and protein binding interfaces PBI. The percentage occurrence of each amino acid at PBI is subtracted from the corresponding value at NPBI.

to clarify the distinction between the newly constructed NPBI and PBI substitution models. Among the series of BLOSUM matrices, we chose BLOSUM35 because it is generated with a 35% sequence identity threshold, which is the same similarity cutoff that we used to compute the PBI and the NPBI models.

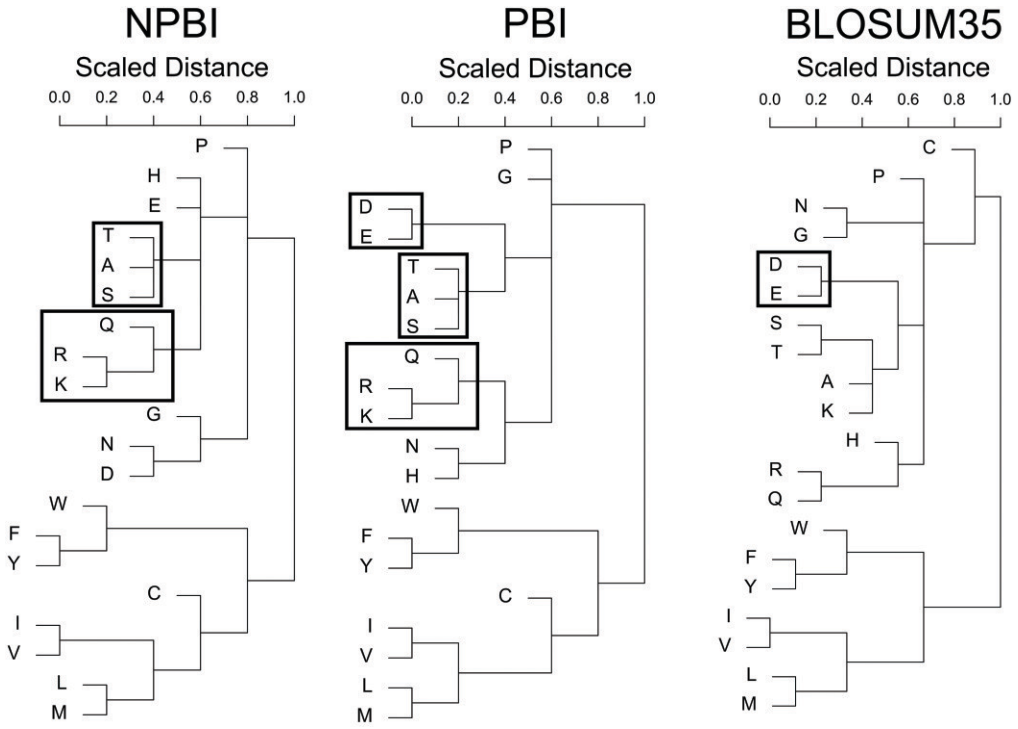

Fig. 3.3. Cluster dendrograms of amino acid substitutions from the NPBI and PBI models, as well as BLOSUM35. Common subclasses of the hydrophilic substitutions are shown in boxes.

Although the PBI and the NPBI models were more correlated with each other (a correlation coefficient of 0.906) relative to the BLOSUM35 (the correlation coefficient values with the PBI and the NPBI are 0.721 and 0.726, respectively), they were different with statistical significance when subject to a Kolmogorov-Smirnov distribution test (D=0.171, p=0.004) [109]. In Figure 3.3, the three substitution models were compared by complete linkage clustering of the log-odd values of the substitution models. To compute the distance of an amino acid pair, the log-odds score of the pair was negated and shifted by a constant value to have all the positive distances. Then, the distance values were normalized to a range between 0 and 1.

All three models showed clear division of aromatic and hydrophobic amino acids from the rest of the residues. Hydrophobic amino acids on protein surfaces appeared

to have conservative mutations between each other (i.e. a smaller distance) even at NPBI sites, implying that they could possibly correspond to other types of functionally important regions that may not directly involve PPI [110]. BLOSUM35 classified cysteine with the polar groups, while NPBI and PBI placed cysteine in the hydrophobic cluster of residues. The similarities between the NPBI and PBI models were observed for the sub-classification of polar amino acid substitutions. The clustering of polar residues in the NPBI and PBI models had common arginine ↔lysine↔gluatamine and threonine↔alanine↔serine substitution marked in boxes in Figure 3.3. When comparing PBI with BLOSUM35, there was only one instance of a common subcluster, aspartic acid↔glutamic acid, which did not appear in the NPBI.

There were several differences between overall NPBI and PBI. The two common subtrees, arginine↔lysine↔gluatamine and threonine↔alanine↔serine, were grouped with different amino acids in the PBI and NPBI. There was also clear distinction in the substitution preference of asparagines and glycine. Taken together, we observed distinct amino acid substitutions found at protein binding interfaces, which would be useful in identifying protein binding regions of the protein surface.

3.3.2 Example of Protein Binding Site Prediction by BindML

To begin with discussing PBI prediction results, we present an example of protein binding site prediction by BindML for a homo-dimer structure, triosephosphate isomerase (PDB ID: 7TIMA). A MSA for this protein was taken from the PFAM database [98] (PFAM ID: PF00121). In Figure 3.4A, the Z-score of each surface amino acid is indicated as a vertical bar, where smaller Z-scores are more likely to be included in PBI. To illustrate how the predicted residues change with different threshold values, predictions made using four Z-score threshold values, -2.0, -1.5, -1.0, and -0.5 are shown. As shown in the mapped structures (Fig. 3.4B), using the Z-score -2.0 and -1.5 did not yield any over predictions. The specificity was 1.0 for the two cases,

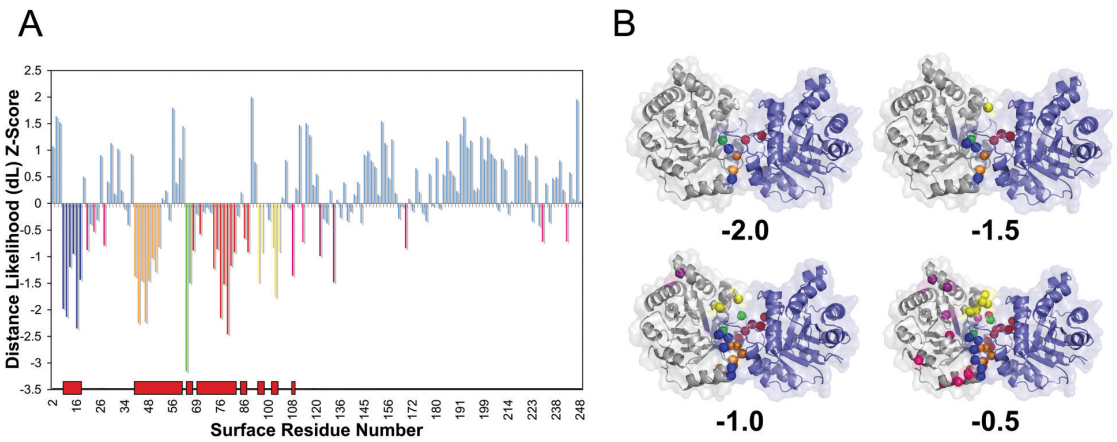

Fig. 3.4. An example of protein binding interface prediction by BindML (PDB ID: 7TIMA). **(A)** The colored vertical bars in the graph show residues predicted with a Z-score value at or below threshold values, -2.0, -1.5, -1.0, or -0.5. Each range of nearby signals in sequence are colored in green, red, blue, orange, yellow and purple bars corresponds to the first, second, third, forth, fifth and sixth highest scoring regions, respectively. The remaining signals are colored in pink. The red blocks along the x-axis indicate the correct interface regions. **(B)** The predicted residues using the four different threshold values are shown in the same colors on the structure.

and the sensitivity was 0.212 and 0.303 for the Z-score value of -2.0 and -1.5, respectively. Increasing the Z-score threshold to -1.00 resulted in predicting more residues at the true interface (sensitivity: 0.636, specificity: 0.984) with a single false positive residue. Further increasing the Z-score threshold to -0.50 provided a good trade-off in the sensitivity (0.849) and the specificity (0.893). Obviously, more residues were predicted as PBI when the Z-score threshold was raised, however, predictions were mostly concentrated on the true PBI sites. The strongest signal (colored in green with the lowest Z-score of -3.151 corresponding to glutamine 64) was located at the most central part of the true PBI. The second strongest signals (including a significant Z-score of -2.455 for asparagine 78) spanned over the residue range from 68 to 87 (with some gaps) colored in red, which corresponds to a protruding binding loop.

The third significant region, residue 12-17 (blue) and the fourth, residue 43-46, 48, 49, 51, 52 (orange) with the strongest Z-score of -2.230, corresponded to loop regions, which bind the protruding loop of the symmetrically bound partner (this loop is an equivalent region that was previously shown in red). Finally, the fifth signal included phenylalanine 102 (colored in yellow with the Z-score of -1.751), which is a part of a loop that follows a short six-residue helix. Overall, the prediction performance on 7TIM showed an outstanding AUC value of 0.936.

3.3.3 Effect of Different Alignment and Prediction Sphere Sizes

BindML extracts partial MSAs by employing a sphere that is centered at a particular surface residue (the alignment sphere) and another sphere to map the final PBI prediction to residues within it (the prediction sphere). Figure 5 shows the effect of the changing sphere sizes.

The radius of the alignment sphere was changed from 5 to 30 Å (Fig. 3.5A). Approximately two residues were included when the radius of five is used, while it increased to 25 and 101 residues for the radius of 15 and 30 Å, respectively. The best prediction performance was observed when using a sphere size of 15 Å in terms of all the performance metric we used except for the sensitivity. A sphere of 15 Å radius covered roughly 700 Å^2 of the protein surface area, which corresponded to about one third of the average PBI surface area in the iPFAM dataset (2200 Å^2).

Figure 3.5B shows the performance of BindML with different prediction sphere sizes. With the prediction sphere size of 0 Å, PBI prediction was assigned just to the center residue of the sphere. With larger sphere sizes, all the residues included in the sphere were predicted as a PBI site. The results showed that simply assigning prediction to the center residue (i.e. sphere size of 0 Å) performed the best in terms of the AUC value (0.623). Based on these results, we used the alignment sphere of 15 Å radius and the prediction sphere of 0 Å in the subsequent benchmark studies.

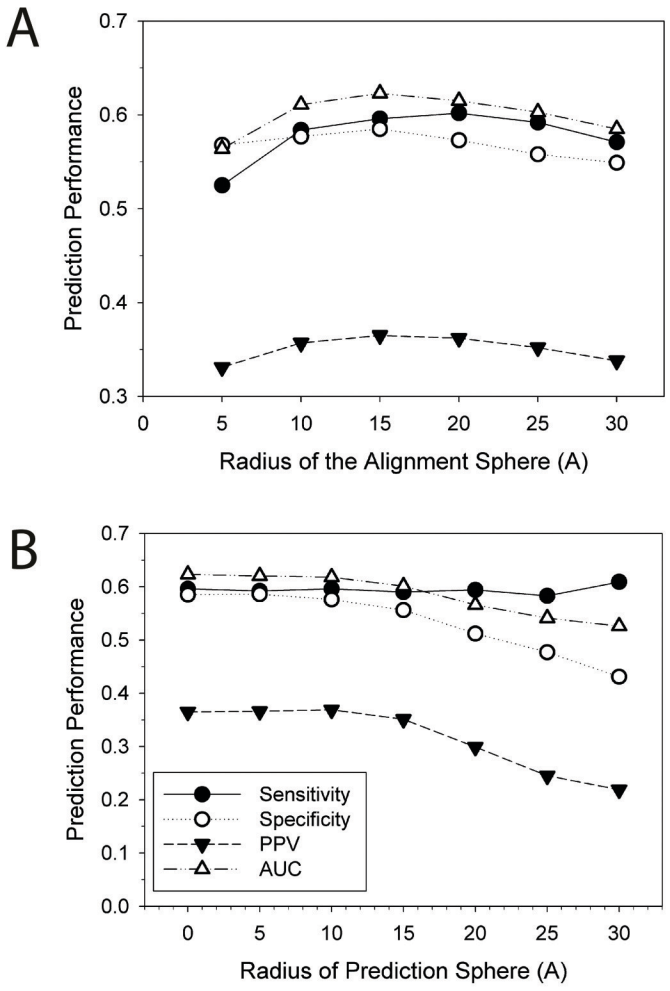

Fig. 3.5. AUC relative to the size of the alignment sphere **(A)** and the prediction sphere **(B)**. One entry, 1AVZC, is discarded from the iPFAM dataset because the use of a sphere of 30 Å radius centered at every surface residue captures the entire protein. The 35% sequence identity was used as the threshold values for computing the PBI/NPBI substitution models. These results show the optimal Z-score threshold value is chosen for each prediction sphere size that gives the closest point to the true positive and the false positive rate of 1 and 0 on the ROC curve, respectively. Therefore, the sensitivity does not simply increase when larger prediction sphere sizes are used.

3.3.4 Prediction performance of BindML

We benchmarked BindML by a five-fold cross validation using the iPFAM dataset. The whole dataset of 505 proteins are randomly split into five subsets. The list of proteins in the five subsets is provided in Table 3.1. The training set, which consists of four fifth of the entire dataset, was used to compute the PBI and the NPBI substitution models and also to determine the optimal sequence identity threshold value of sequence pairs for computing the substitution models. Sequences which are closer than the threshold value to any other sequence in the dataset were pruned out. Then, the computed PBI/NPBI substitution models were used in making predictions on the remaining one fifth of the dataset (testing set). This process was repeated five times by changing the testing and training sets. Both training and testing sets yielded an average AUC value of 0.624, indicating that the PBI and the NPBI models were very well generalized for protein-protein binding site prediction. The AUC values for each of the five training and the corresponding testing sets were given in Table 3.4. Systematically, multiple sequence identity threshold values were evaluated to determine the optimal substitution models for each training set. The threshold values for the five training datasets were consistent, where 35% worked best for the three training sets and 30% was optimal for the remaining two sets. Further, to eliminate the influence of similar proteins more thoroughly, we excluded proteins in the control dataset with equal to or higher than 25% in the sequence identity to those found in the training dataset, which still resulted in very consistent AUCs between the control (0.623) and training (0.624) models (the two rightmost columns in Table 3.4). We have also compared the global (built using the entire iPFAM dataset, Table 3.2 and 3.3) and each substitution model built using each training dataset. We found that they are very consistent, showing an average correlation coefficient of 1.000 for PBI and 0.984 for NPBI models. In Table 3.5, we examined the effect of sequence identity threshold values used to compute the substitution models on the prediction performance of BindML. As the purpose was to investigate the effect of the param-

eter, the entire iPFAM dataset was used. Raising the threshold value increased the number of sequences included in each family, while lowering the threshold eliminated more sequences. The sensitivity and the specificity was computed using a threshold Z-score which gave the closest point in the ROC curve to the true positive and the false positive rate of 1 and 0 (i.e. the corner of the ROC curve). A 35% sequence identity yielded the highest predictive performance with the highest AUC value (0.623), the positive predict value (0.365), and the sensitivity (0.596). Because of its best performance, we have chosen 35% sequence identity in the subsequent analyses.

List of the five test datasets (named by the representative structure PDB ID dash Chain ID of the iPFAM MSA dataset) used in the cross validation benchmark.

Test Set 1	**Test Set 2**	**Test Set 3**	**Test Set 4**	**Test Set 5**
1A81-C	1A02-F	1AVZ-B	1A2K-B	1A02-N
1ADU-B	1A0H-B	1AVZ-C	1A4I-B	1A0H-A
1AJK-B	1A7H-B	1AYO-B	1A4Y-B	1A2O-B
1AV1-D	1AB8-B	1B01-B	1AOH-B	1AIP-D
1B6C-B	1AG9-B	1B4A-D	1ATL-A	1AIS-A
1BNC-B	1B4U-D	1C0M-C	1B33-K	1AVF-J
1BVY-A	1B6S-A	1CFZ-A	1BMT-A	1B27-B
1CF7-A	1DVR-B	1CXZ-B	1C4Z-B	1B4F-C
1CVS-D	1ECE-B	1D0Q-B	1CB7-C	1BVS-C
1CY9-B	1FS1-B	1E2T-C	1DHK-B	1C4Z-D
1DKG-B	1FXK-C	1EXB-A	1E5L-B	1CQZ-A
1DT5-G	1G5Q-D	1FL7-B	1EAI-D	1DYO-B
1EBD-C	1G8J-A	1FP3-A	1F3M-B	1EI6-D
1EO2-A	1GL2-C	1GPZ-B	1FA0-A	1EPF-C
1G71-B	1H65-A	1I2M-A	1G60-B	1F0K-A
1GK4-B	1HE8-A	1IJD-C	1GXJ-A	1F7T-B
1GZM-B	1HZZ-A	1IS2-A	1HKQ-A	1FLE-I
1HUX-B	1I3Q-A	1IVY-B	1HRU-A	1FOH-D
1ICT-D	1I4J-B	1JBO-A	1HX8-B	1ICI-B
1IES-B	1IDS-C	1JBQ-B	1IB6-D	1ID3-D
1II6-A	1IHX-C	1JF5-A	1IBJ-C	1IIG-B
1II8-A	1II4-B	1JK9-C	1IHJ-A	1IPH-C
1IIM-B	1IJK-C	1JO0-B	1IL0-A	1IQ6-A
1IJX-C	1IK7-B	1JTK-A	1IQP-F	1IW7-M
1IQA-A	1IKN-C	1JXA-C	1IT0-A	1JAT-B
1IRJ-C	1IMA-B	1JZ0-H	1IUN-A	1JC4-A

1ITQ-A	1IO4-B	1JZT-A	1IW7-B	1JEQ-A
1ITZ-A	1IR2-3	1K1X-A	1J1D-E	1JEQ-B
1IU1-A	1IW7-O	1K6D-B	1J3W-D	1JJC-A
1IUG-B	1IXE-A	1KEK-A	1JCQ-B	1JJK-P
1IVH-B	1J9C-L	1KFV-A	1JD1-B	1JKT-B
1IWE-B	1J9Q-B	1KIB-C	1JEH-B	1JL9-B
1JFI-B	1JR1-B	1KIU-L	1JG8-B	1JMK-C
1K41-A	1JSC-A	1KT8-A	1JJ4-A	1JQK-A
1K4E-B	1KBU-A	1KU2-A	1JKF-A	1K1D-C
1K82-B	1KEE-C	1KU6-B	1JMA-B	1K3R-A
1KIJ-A	1KJY-A	1KYO-W	1JS1-Z	1K83-B
1KPA-A	1KV3-D	1KYW-F	1JSU-A	1K83-C
1KTJ-B	1KXP-D	1L0L-A	1JYQ-B	1K9O-I
1L5A-C	1LDJ-B	1L1O-A	1K3E-B	1KAM-B
1L8X-A	1LM8-C	1L7V-B	1K5Q-B	1KB2-B
1LFD-B	1LNW-C	1LFD-A	1K66-B	1KHD-D
1LTX-B	1LPA-B	1LL0-H	1K6Y-D	1KSX-B
1M2O-A	1M56-G	1LXY-B	1KP0-A	1KYO-M
1M7E-A	1MDU-A	1M0S-A	1KPK-B	1L6R-A
1MUU-B	1MG2-G	1MB2-E	1L1O-F	1L7J-B
1N7K-B	1MPX-D	1MB4-A	1L3L-B	1L9X-B
1NBF-D	1N0E-C	1MHD-B	1L4I-B	1LDK-B
1NK1-A	1N1C-B	1MHY-D	1L5H-B	1LGB-A
1NKQ-B	1N69-A	1N12-C	1LX5-B	1M2T-B
1NP3-C	1NCH-B	1N1B-A	1M10-B	1M3D-B
1NVM-E	1NG9-B	1N1E-A	1M2A-B	1M4Z-B
1NX4-A	1NLY-B	1NFH-B	1M2O-D	1M5B-B
1NY5-B	1NNQ-B	1NHE-C	1M56-B	1M5X-B
1O4U-A	1NUB-B	1NXM-A	1M7G-B	1M63-G

1O6C-B	1O7A-D	1NYT-D	1MCZ-B	1M8P-A
1O6V-A	1O7K-B	1O91-B	1NOY-B	1N5W-E
1O7D-C	1OCC-C	1O94-D	1NYS-D	1NAW-B
1O9S-B	1ORT-F	1O9Y-D	1O61-B	1NBC-B
1OED-B	1PB0-B	1OEY-D	1O7D-D	1NBF-B
1OGP-C	1PYI-B	1OKR-B	1OC0-A	1NCF-A
1OGY-O	1Q5H-C	1OXM-B	1OFG-C	1NI4-B
1OJV-B	1QFH-B	1OY5-C	1OV9-A	1NW1-B
1OLZ-A	1QMI-A	1P8J-C	1PJQ-A	1NZI-B
1ORD-B	1QO7-A	1PBI-B	1QAE-B	1O7D-A
1P51-D	1QW8-B	1QB2-B	1QDV-D	1OCC-B
1P9L-A	1R27-A	1QDN-C	1QVE-B	1OTV-B
1PBW-B	1R28-B	1QHB-C	1QZX-A	1OW3-B
1PGJ-B	1R61-B	1QIU-C	1R30-A	1OYP-D
1QLV-A	1RP3-D	1QMG-C	1RV1-A	1P27-B
1R0R-I	1S4M-A	1R1U-D	1RZ0-D	1PEG-A
1R45-D	1S80-E	1R4M-D	1SN8-B	1PG5-A
1R4C-C	1SC5-A	1R4W-B	1TCM-B	1QSC-A
1R4P-A	1SPP-B	1R71-D	1TE5-B	1R4A-C
1R56-D	1SR9-A	1RD5-B	1U07-B	1RWT-H
1R5J-B	1T8Q-B	1RTY-B	1UFL-C	1S3S-F
1RCU-A	1T9G-S	1S3O-B	1ULI-B	1S5L-C
1SBW-A	1TEZ-A	1SCJ-A	1UQT-B	1S7M-C
1SHS-C	1TGZ-B	1SFC-A	1UT4-B	1S9A-A
1SI9-B	1TO0-C	1SHY-B	1V9D-D	1SFX-B
1T11-B	1TVE-A	1SQU-B	1VC1-B	1SGJ-B
1T6S-B	1TX0-B	1T3E-B	1VH1-D	1T8T-A
1TEJ-B	1U1Z-B	1T5E-D	1VIO-A	1TF0-A
1TH8-B	1U69-C	1T72-G	1VKP-B	1TLU-B

1TQQ-C	1UF4-B	1T92-B	1VL6-C	1TNR-A
1U0S-A	1UTC-B	1TD5-D	1VLD-U	1TUI-B
1UAD-D	1UW4-D	1TV8-A	1VME-A	1TYG-C
1V5X-B	1VKI-B	1U0L-A	1VQR-B	1UD0-D
1VL0-B	1VQS-E	1UC2-B	1XCB-E	1V5V-A
1VM6-B	1WE3-A	1UKV-Y	1XXH-A	1VL4-B
1VPV-A	1XBR-B	1UUJ-B	1ZAK-B	1VSC-A
1W1W-C	1Y8T-B	1VPZ-A	2MTA-C	1W36-B
1YJ8-A	1YEM-B	1YIF-B	2PKA-B	1WDL-C
2OCC-D	1YLA-B	2AYQ-A	2PSP-B	1Y97-A
2PGH-B	2AZU-B	2BGN-D	2QIL-C	2AY1-A
2PRO-B	2PHI-A	2BKJ-B	2QR2-B	2BEK-B
2RLN-E	2PJR-F	2PMT-C	2UUG-B	2PGT-A
2SQC-A	2RAM-A	2SEM-A	3YGS-C	2R1R-A
2TMK-B	2SFP-A	2USH-A	4OTA-O	2SHK-A
2VAB-A	2TPS-B	4SGB-E	6ALD-C	2SPC-A
2VAO-B	6ADH-B	830C-A	6EBX-A	6COX-B

Table 3.2
Log odds amino acid substitution matrix for PBI.

PBI	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	3	-1	-1	-1	-1	-1	-1	-1	-2	-1	-1	-1	-1	-2	-1	0	0	-2	-2	0
R	-1	3	-1	-2	-1	0	-1	-2	-1	-2	-2	1	-2	-3	-2	-1	-1	-2	-2	-2
N	-1	-1	3	0	-2	-1	-1	-1	0	-2	-2	0	-1	-3	-1	0	-1	-3	-2	-2
D	-1	-2	0	3	-2	-1	1	-1	-1	-3	-3	-1	-3	-3	-1	-1	-2	-3	-3	-2
C	-1	-1	-2	-2	6	-2	-3	-2	-1	-1	-1	-2	-1	-1	-2	0	-1	-2	-1	-1
Q	-1	0	-1	-1	-2	3	0	-1	0	-2	-1	0	0	-2	-1	-1	-1	-2	-2	-2
E	-1	-1	-1	1	-3	0	3	-1	-1	-2	-2	-1	-2	-3	-1	-1	-1	-3	-2	-2
G	-1	-2	-1	-1	-2	-1	-1	3	-2	-3	-3	-2	-3	-3	-2	-1	-2	-2	-3	-2
H	-2	-1	0	-1	-1	0	-1	-2	4	-3	-2	-1	-1	-1	-2	-1	-2	-1	0	-2
I	-1	-2	-2	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	-1	-2	-1	1
L	-1	-2	-2	-3	-1	-1	-2	-3	-2	1	3	-2	1	0	-2	-2	-1	-1	-1	0
K	-1	1	0	-1	-2	0	-1	-2	-1	-2	-2	3	-1	-2	-1	-1	-1	-2	-2	-2
M	-1	-2	-1	-3	-1	0	-2	-3	-1	1	1	-1	4	-1	-2	-2	-1	-1	-1	0
F	-2	-3	-3	-3	-1	-2	-3	-3	-1	-1	0	-2	-1	4	-2	-2	-2	0	1	-1
P	-1	-2	-1	-1	-2	-1	-1	-2	-2	-2	-2	-1	-2	-2	4	-1	-1	-2	-3	-2
S	0	-1	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-2	-2	-1	3	0	-3	-2	-1
T	0	-1	-1	-2	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	0	3	-2	-2	0
W	-2	-2	-3	-3	-2	-2	-3	-2	-1	-2	-1	-2	-1	0	-2	-3	-2	6	1	-2
Y	-2	-2	-2	-3	-1	-2	-2	-3	0	-1	-1	-2	-1	1	-3	-2	-2	1	4	-1
V	0	-2	-2	-2	-1	-2	-2	-2	-2	1	0	-2	0	-1	-2	-1	0	-2	-1	3

Table 3.3
Log odds amino acid substitution matrix for NPBI.

NPBI	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	3	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1	-1	-2	0	0	0	-2	-2	0
R	-1	3	-1	-2	-2	0	-1	-2	0	-2	-1	1	-1	-2	-1	-1	-1	-1	-1	-1
N	-1	-1	3	0	-2	-1	-1	-1	0	-2	-2	0	-2	-2	-1	0	0	-2	-1	-2
D	-1	-2	0	3	-2	-1	0	-1	-1	-3	-2	-1	-2	-3	-1	-1	-1	-3	-2	-2
C	-1	-2	-2	-2	6	-2	-2	-2	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1
Q	0	0	-1	-1	-2	3	0	-1	0	-1	-1	0	-1	-2	-1	-1	-1	-2	-2	-1
E	0	-1	-1	0	-2	0	2	-1	-1	-2	-2	0	-1	-2	-1	-1	-1	-2	-2	-1
G	-1	-2	-1	-1	-2	-1	-1	3	-2	-3	-3	-2	-2	-3	-2	-1	-2	-2	-2	-2
H	-1	0	0	-1	-1	0	-1	-2	4	-1	-1	-1	-1	0	-1	-1	-1	-1	1	-1
I	-1	-2	-2	-3	-1	-1	-2	-3	-1	3	1	-1	1	0	-2	-2	-1	-1	-1	2
L	-1	-1	-2	-2	-1	-1	-2	-3	-1	1	3	-1	1	0	-2	-2	-1	0	-1	0
K	-1	1	0	-1	-2	0	0	-2	-1	-1	-1	2	-1	-2	-1	-1	-1	-2	-2	-1
M	-1	-1	-2	-2	-1	-1	-1	-2	-1	1	1	-1	4	0	-2	-1	0	-1	-1	0
F	-2	-2	-2	-3	-1	-2	-2	-3	0	0	0	-2	0	4	-2	-2	-1	1	2	-1
P	0	-1	-1	-1	-2	-1	-1	-2	-1	-2	-2	-1	-2	-2	4	-1	-1	-2	-2	-1
S	0	-1	0	-1	-1	-1	-1	-1	-1	-2	-2	-1	-1	-2	-1	3	1	-2	-1	-1
T	0	-1	0	-1	-1	-1	-1	-2	-1	-1	-1	-1	0	-1	-1	1	3	-2	-2	0
W	-2	-1	-2	-3	-1	-2	-2	-2	-1	-1	0	-2	-1	1	-2	-2	-2	6	1	-2
Y	-2	-1	-1	-2	-1	-2	-2	-2	1	-1	-1	-2	-1	2	-2	-1	-2	1	4	-1
V	0	-1	-2	-2	-1	-1	-1	-2	-1	2	0	-1	0	-1	-1	-1	0	-2	-1	3

a) Proteins in the control dataset which have equal to or more than 25% sequence identity to any of proteins in the training set were removed.

Table 3.4
Performances of each of the five cross validation datasets.

Dataset	**Sequence Identity Threshold (%)**	**Control AUC**	**Training AUC**	**Control AUC (25% cutoff) [a)]**	**Training AUC (25% cutoff)**
1	35	0.643	0.617	0.637	0.617
2	35	0.599	0.630	0.597	0.630
3	35	0.625	0.627	0.622	0.627
4	30	0.658	0.616	0.662	0.616
5	30	0.596	0.631	0.596	0.631
Average		*0.624*	*0.624*	*0.623*	*0.624*

Patches on the surface are generated using a scanning sphere size of 15Å and a defining instance of a given predicted site as the single central residue in the sphere. The line shown in bold text highlights parameters used with the best threshold value.

Table 3.5
Effect of different sequence identity percentage cutoff values for PBI and NPBI Substitution Models. The cutoff that gives the highest overall benchmark performance is shown in bold text.

Percent Identity	**Sensitivity**	**Specificity**	**PPV**	**AUC**
60	0.496	0.486	0.276	0.481
55	0.464	0.490	0.265	0.463
50	0.546	0.445	0.277	0.482
45	0.571	0.435	0.283	0.489
40	0.516	0.487	0.282	0.489
35	**0.596**	**0.585**	**0.365**	**0.623**
30	0.596	0.581	0.363	0.622
25	0.584	0.584	0.361	0.617
20	0.585	0.584	0.362	0.617
15	0.584	0.586	0.362	0.617
10	0.585	0.585	0.362	0.618

3.3.5 Distribution of Individual Performances

On closer inspection of individual predictions by BindML, we analyzed the distribution of AUCs for all proteins in the iPFAM dataset (Fig. 3.6A). The AUC values were taken from the testing dataset in the cross validation (Table 3.4). Although we reported an average AUC of 0.624 in the previous section, the performance of individual prediction distributed widely from almost perfect prediction, 0.956, to predictions below 0.5 (Fig. 3.6A). The peak of the AUC distribution was between 0.70 and 0.75 (consisting 68 proteins), and the second most frequent AUC values were in the range

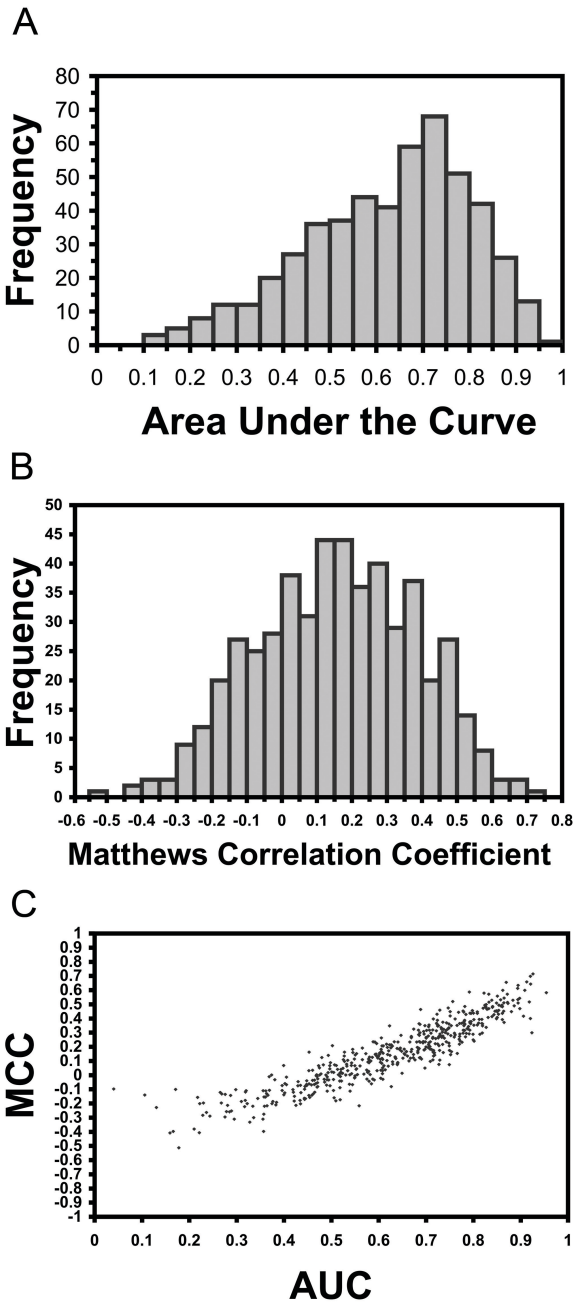

Fig. 3.6. Distribution of performances for proteins in the iPFAM dataset for **(A)** the AUC values, **(B)** the MCC values, **(C)**, and the correlation between AUC and MCC values.

of 0.75 to 0.80 (51 proteins). Predictions above AUC of 0.5 shared 86.9%. In addition, we computed the distribution of MCC values using a Z-score threshold of -0.5, which provided good balance of sensitivity and specificity for many proteins in the iPFAM benchmark dataset (Figure 3.6B). The average MCC value was 0.156. As shown in Figure 3.6C, AUC and MCC correlate well with a correlation coefficient of 0.918. Thus, MCC of the predictions essentially provides the same picture as the AUC distribution. The distribution revealed that there were cases where BindML predicted poorly. The cause of poor predictions depends on cases: In the case of cellular receptor HVEA/HVEM interacting with an envelope glycoprotein of herpes simplex virus (1JMAB), the sequence profile at the binding interface was too conserved for BindML to detect PBI specific interaction pattern. Also, it is not easy for BindML to predict PBI sites that are too small or narrow, given that it uses an alignment sphere of 15 Å radius. An example for this case (1S70A) will be discussed in the next section. We further examined prediction performance by BindML on homo- and hetero-protein complexes, because several studies have shown differences in the sequence and structural composition of homo- and hetero-complexes [20, 61, 111, 112, 113]. 65% of our dataset are composed of homo-complexes, whereas the remaining 35% are hetero-complexes. The PBI prediction performance for homo-complexes was higher (AUC: 0.638) than hetero-complexes (AUC: 0.599). Thus, BindML can pick up PBI sites for both using a general PBI/NPBI substitution model, but provides better performance for homo- than hetero-complexes.

3.3.6 Prediction Examples

Figure 3.7 provides five examples of BindMLs predictions. We used a Z-score threshold of -0.5. These five structural complexes from heterogeneous protein families consist of one complex taken from the iPFAM dataset (Fig. 3.7A, 1KT8A), two targets used in the Critical Assessment of Prediction of Interactions (CAPRI) 49 (Figs. 3.7C & E, 2B3TB and 1HWZA), and two other additional structural examples

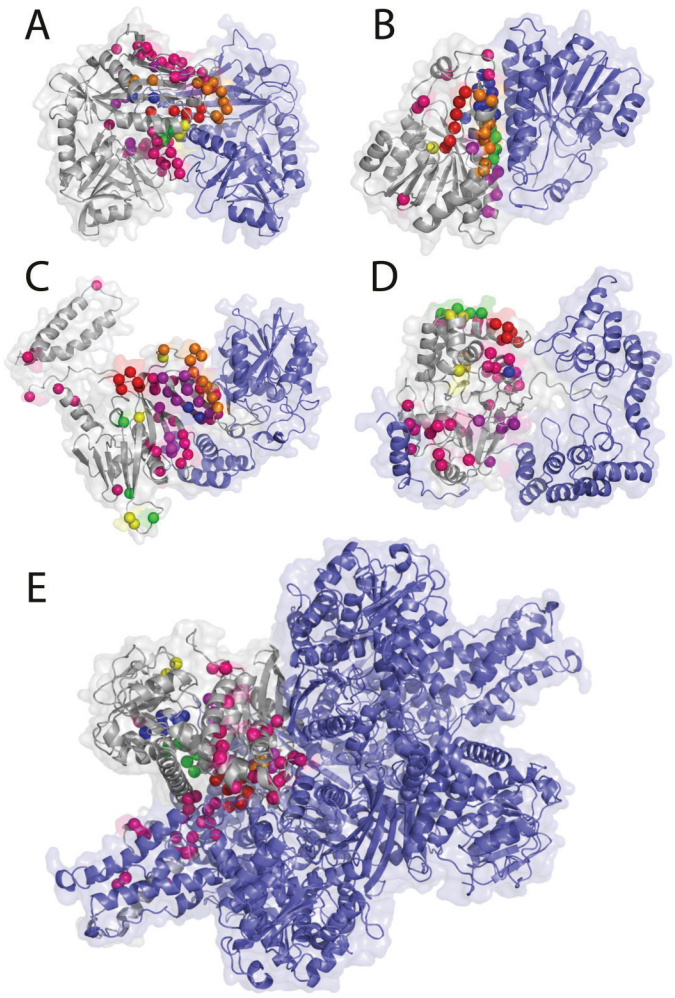

Fig. 3.7. Additional prediction examples: **(A)** amino acid transferase (PDB: 1KT8A), **(B)** alcohol dehydrogenase (1A4UA), **(C)** peptide chain release factor 1 complexed with methyltransferase hemK (2B3TB), **(D)** protein serine/threonine phosphatase complexed with smooth muscle myosin phosphatase (1S70A), **(E)** hexameric glutamate dehydrogenase (1HWZA). The target chain subject to the prediction is shown in gray. The same color scheme is used to rank the strength of the cluster of signals as in Figure 3.4.

(Figs. 3.7B & 3.7D, 1A4UA, 1S70A). The first two shown are successful examples of enzymes (Figs. 3.7A & 3.7B). Figure 3.7A is the result for amino acid transferase (1KT8A), which forms a homo-dimer complex. The AUC value of this prediction was

high, 0.847, and the sensitivity and the specificity are 0.636 and 0.859, respectively. In Figure 3.7B, binding sites for the homo-dimer alcohol dehydrogenase (1A4UA) is shown. The predicted residues corresponded very well at the α-helical regions of the dimer interface (AUC: 0.800, sensitivity: 0.569, specificity: 0.889). The next example, Figure 3.7C, is peptide chain release factor 1 complexed with methyltransferase hemK (2B3T). The prediction is made for peptide chain release factor 1 (2B3TB). Several false positives on the opposite side from the binding interface were observed but the overall performance was decent with an AUC value, sensitivity, specificity value of 0.656, 0.444, and 0.777, respectively. We also show the prediction for serine/threonine phosphatase (1S70A), a CAPRI target (T14) structure, where BindML did not perform well (Fig. 3.7D). BindML prediction was reasonably specific, with the specificity of 0.667; however, a poor sensitivity (0.225) dropped the overall AUC value to 0.476. The main reason for the poor result comes from prediction at the binding region to the N-terminal tail of binding partner, myosin phosphatase, which wraps around the phosphatase structure. The interacting region to the tail forms an elongated shape, which is not advantageous for BindML to scan by using a sphere. The last example is the prediction on a large homo-hexameric complex of Glutamate dehydrogenase (1HWZA) (Fig. 3.7E). The A-chain interacts with the B, D, E, F chains forming the binding interface of 64 residues. BindML predicts this large binding interface well with the AUC value of 0.711 and the specificity of 0.789.

3.3.7 Comparing highly conserved regions to BindML predictions

To illustrate that BindML is not simply identifying conserved regions, we compare BindML prediction with those by considering naive sequence conservation (the percentage of conserved residues at each position in the MSA) and also with prediction by the ConSurf method [38, 40, 41]. ConSurf utilizes a phylogenetic tree to calculate conservation rate along the evolution for a MSA of a protein family. We employed the

web server implementation of ConSurf (`http://consurf.tau.ac.il/`). For ConSurf predictions, we only considered solvent exposed residues.

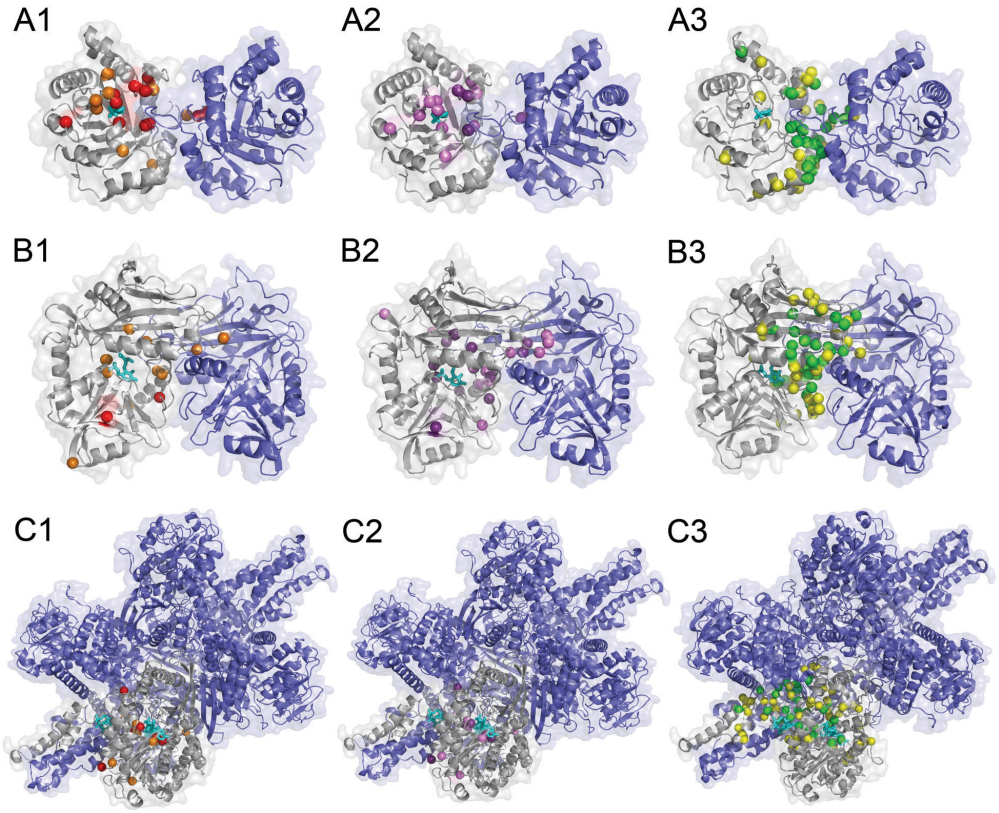

Fig. 3.8. Comparison of residues selected by naive sequence conservation, ConSurf, and BindML. **(A)** Triosephosphate isomerase homo-dimer (7TIM), **(B)** Amino acid transferase homo-dimer (1KT8), **(C)** Glutamate dehydrogenase (1HWZ). Ligand molecules binding to these proteins are shown in cyan. For the three proteins, residues which are assigned with a significantly high score by the three methods are shown. A1, B1, C1, residues with high conservation; A2, B2, C2, residues with a high score by ConSurf; A3, B3, C3, residues identified by BindML. For sequence conservation, residues which are conserved in more than 90% (70%) of sequences are shown in red (orange) (A1, B1, C1). As for ConSurf, residues detected with a Z-score of -1.3 (-1) or lower is shown in purple (violet) (A2, B2, C2). Residues identified with a Z-score of -1 (-0.5) or lower by BindML are shown in green (yellow) (A3, B3, C3).

Figure 3.8 shows predictions by the three methods. The first protein, triosephosphate isomerase homo-dimer (7TIM), binds phosphoglycolohydroxamate (PGH) at its well conserved ligand binding pocket. Thus, strong sequence conservation selected mostly residues in contact with the ligand or residues surrounding the binding site (Fig. 3.8A1). This is also true for the ConSurf prediction (Fig. 3.8A2). Both methods do not identify almost any PBI site residues. In contrast, predictions provided by BindML concentrated on residues at PBI site (Fig. 3.8A3). With the Z-score of -1.0, the prediction was very specific to the PBI site, with a specificity of 0.992 and a sensitivity of 0.516. Raising the Z-score threshold to -0.5 yielded an increased in sensitivity of 0.849 (specificity: 0.885). The next examples are predictions for amino acid transferase homo-dimer (1KT8) which bind N-[O-phosphono-pyridoxyl]-isoleucine in the ligand binding cavity. Again, naive conservation predictions (Fig. 3.8B1) as well as ConSurf (Fig. 3.8B2) identified ligand binding residues and not PBI site residues. In comparison, the predicted residues by BinML were located in different places in the structure, where they were mostly covering PBI sites (Fig. 3.8B3). Using the Z-score threshold value of -1.0, the sensitivity and the specificity were 0.40 and 0.941, respectively. Further increasing the Z-score threshold to -0.5 resulted in increased sensitivity (0.636) and specificity (0.859). The difference in predictions by BindML relative to the other two methods of sequence conservation is further apparent when the AUC value of the ROC curve for ligand binding residues and PBI site residues are compared (Fig. 3.9). For both of the cases of triosephosphate isomerase (Fig. 3.9A) and amino acid transferase (Fig. 3.9B), BindML showed lower AUC value for ligand binding residues compared to naive sequence conservation and ConSurf (Figs. 3.9A1 & 3.9B1). BindMLs AUC value for ligand binding sites for the second case remained high (0.728) (Fig. 3.9B1), but this is due to six residues that were located at the PBI site that also interact with the ligand molecule. When the AUC value for PBI sites were computed (Figs. 3.9A2 & 3.9B2), BindML showed significantly higher predictive performance than the other two methods.

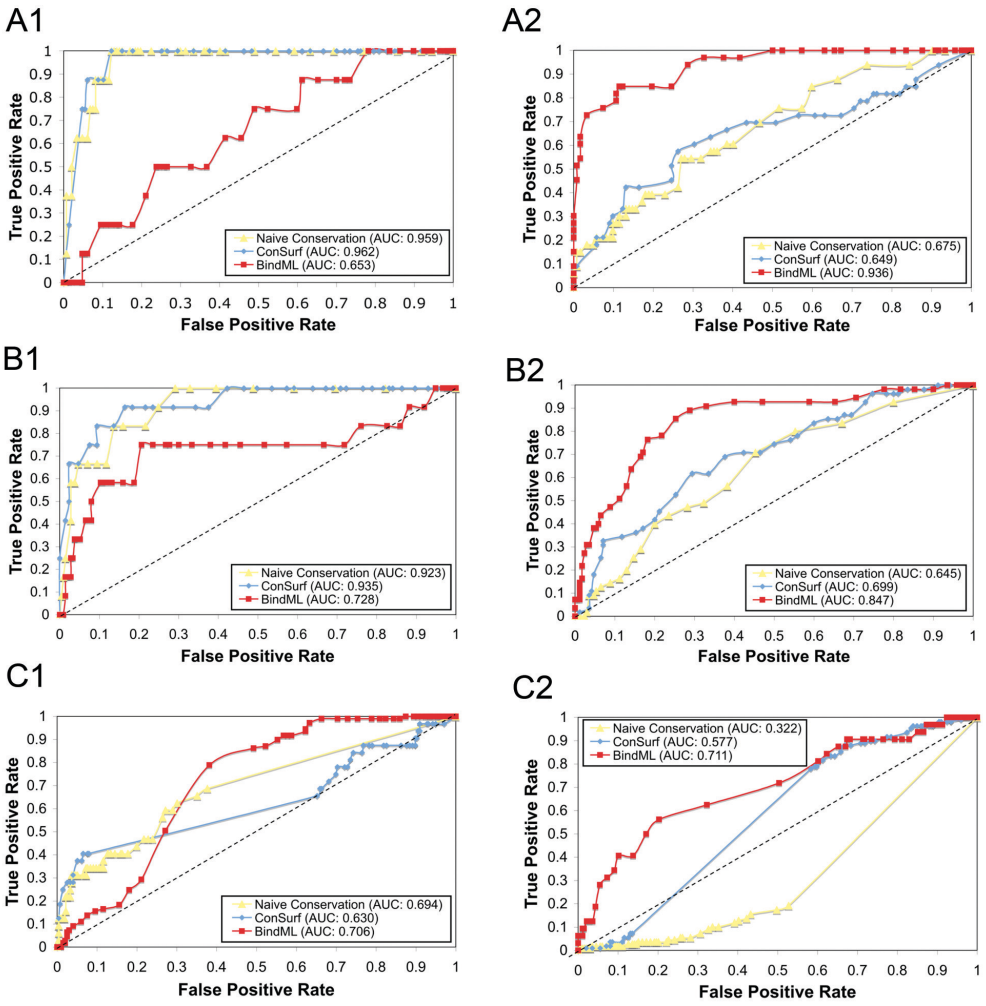

Fig. 3.9. ROC curves of ligand binding residue prediction and PBI residue prediction by sequence conservation (open circles), ConSurf (filled diamonds) and BindML (filled circles). **(A)** Triosephosphate isomerase homo-dimer (7TIM), **(B)** Amino acid transferase homo-dimer (1KT8), **(C)** Glutamate dehydrogenase (1HWZ). A1, B1, C1, ligand binding site; A2, B2, C2, PBI site prediction. Ligand binding residues are defined as those which are within 5.0Å to the ligand molecule.

The last examples in Figure 3.9 are predictions for hexameric glutamate dehydrogenase (1HWZ). This six-chain complex has three ligands bound to the active site, NADPH, glutamate, and GTP. Ligand binding residues for all three ligands were evaluated for the chain A. Although the sensitivity was low, the naive residue

conservation and ConSurf identified ligand binding residues very specifically. The specificity by the naive sequence conservation and ConSurf was 0.978 and 0.996 with the threshold value of 70% and -1.0, respectively. On the other hand, prediction by BindML provided more PBI site residues. Again, the predictive difference by BindML is obvious by considering the ROC curves (Fig. 3.9C). The ROC curve of naive sequence conservation (Fig. 3.9C, yellow line) indicates that the PBI site of this protein chain is not conserved at all in comparison to the other surface residues. In summary, Figures 3.8 and 3.9 clearly illustrate that BindML identifies the mutation patterns of PBI sites, which cannot be simply captured by the methods of sequence conservation. Naive sequence conservation (and ConSurf) can evidently identify ligand binding sites well, but not PBI sites.

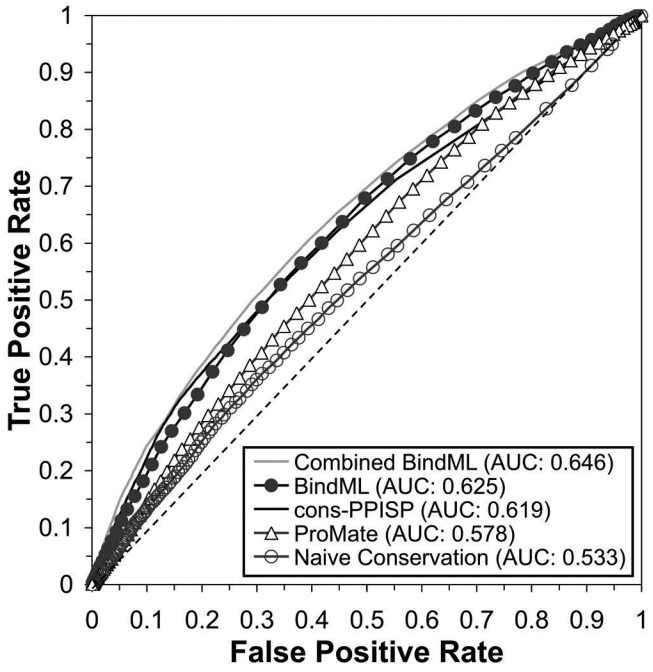

Fig. 3.10. ROC performances for BindML, cons-PPISP, ProMate, and the nave conservation on the iPFAM dataset. Combined BindML is an ensemble approach that combines BindML and cons-PPISP. The dashed diagonal line is the expected performance of random predictions (AUC value of 0.5).

3.3.8 Comparison with Other Existing PBI Prediction Methods and Sequence Conservation

In this section we compare BindML with two existing PBI prediction methods (cons-PPISP [85] and ProMate [17]) and naive sequence conservation on the iPFAM dataset (Fig. 3.10). The results for BindML were taken from the cross-validation test of the entire iPFAM dataset. For the false positive rates between 0.00 and 0.30, cons-PPISP performed slightly higher than BindML with approximately 1 percent difference in the true positive rate. Therefore, the AUC up to 20 percent false positives (AUC_{20}) shows the following order of cons-PPISP, BindML, ProMate, and naive conservation with 0.043, 0.035, 0.027, and 0.026 respectively. However, overall, BindML showed the highest ROC value, 0.625. Cons-PPISP, ProMate, and naive conservation followed in this order with an AUC value of 0.619, 0.578, and 0.533, respectively. The AUC value of BindML is significantly higher than that of cons-PPISP when subject a Hanley and McNeil test (p-value: 4×10^{-5}) [114], which compares two ROC curves by calculating the standard error and the difference in AUCs. This is remarkable because BindML primarily uses sequence mutation patterns, while the cons-PPISP and ProMate methods both combine sequence and structural information using a more elaborate machine learning framework. Naive sequence conservation performed worst in predicting PBI sites confirming that BindML does not simply identify conserved regions in MSAs.

As illustrated in the previous section with Figures 3.8 & 3.9, naive conservation captured more ligand binding residues than BindML. Using 174 proteins that have both ligands and PBI sites in the dataset, AUC computed for residues which bind ligands but not at PBI sites was 0.582 for naive conservation while 0.549 for BindML.

To seek for further improvement of the prediction accuracy, we took an ensemble approach [115, 116, 117] (or meta-server approach), which integrates independent prediction by BindML and cons-PPISP. The score of the ensemble approach, named Combined BindML in Figure 3.10, is simply the average of the rescaled BindML

score and the cons-PPISP score. The new Combined BindML method performed with consistently higher true positive rate at all false positive rate values across the entire ROC Curve in comparison to the original BindML and the other methods in the figure. The AUC of the Combined BindML is 0.646, which is higher than that of original BindML (0.625) and cons-PPISP (0.619) with statistical significance (p-value < 0.000001). Moreover, the AUC20 of Combined BindML is 0.047, highest among the methods compared.

3.3.9 Prediction Results for Unbound Structures

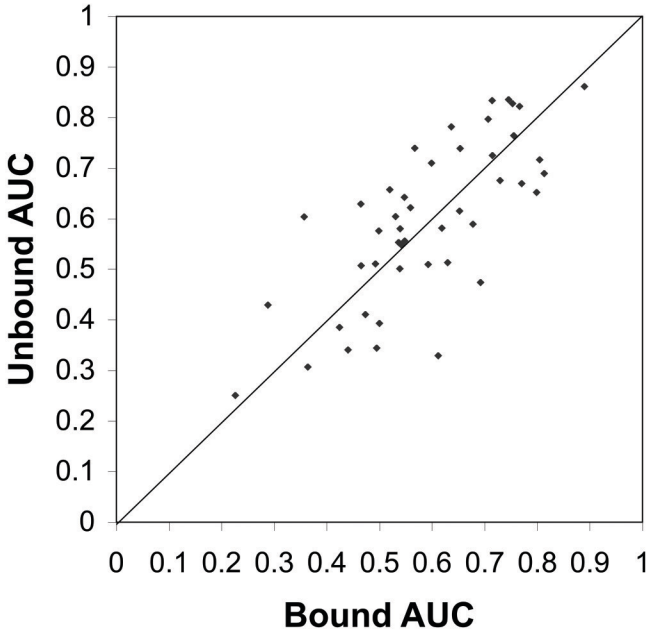

Fig. 3.11. Comparison of AUC of PBI prediction for bound and unbound form of proteins. 46 structures from protein-protein docking benchmark dataset 4.0 were used.

Finally, we apply BindML to predict PBI sites of proteins in unbound conformation. Out of 112 structures of bound and unbound form of protein structures in a protein-protein docking benchmark dataset ver. 4.0 compiled by Z. Weng group [115],

46 pairs were selected, whose sequence are not included in the MSAs in the iPFAM dataset used to determine parameters and to compute PBI and NPBI substitution models. PBI residues were defined on the bound form of proteins. The average root mean square deviation (RMSD) of PBI sites between bound and unbound form of the proteins was 1.43Å.

Figure 3.11 compares the AUC values of bound and unbound conformation of each protein in the dataset. Overall, the prediction for unbound structures does not deteriorate, rather, showing even slightly better results for the unbound form than bound form in this dataset. The correlation in AUC values of bound and unbound predictions was 0.734. The average AUC for bound form structures was 0.587 while that of unbound form was 0.596. Out of the total of 46 cases, 26 unbound predictions (56.5%) are better than bound case predictions. Thus, BindML, is tolerant to usual level of conformation change from unbound to bound forms, partly because it is using a large alignment sphere of the 15 radius to extract a MSA of a surface region. As an example, we compared corresponding surface patches between bound and unbound conformations of cysteine protease ATG4B complexed with microtubule-associated proteins (1A/1B) (bound structure: 2Z0E; unbound: 1V49A & 2D1IA). The RMSD of the bound and unbound structures at the interaction interface is 2.15Å, and the AUC of predicted PBI site was very similar, 0.781 and 0.782 (the average of ligand and receptor proteins) for bound and unbound structures, respectively. For this complex, the bound and unbound structures have 239 corresponding surface patches, each of which contains on average 22.6 residues.

As expected, the conformation change does not have significant impact on the residues included in corresponding patches between bound and unbound conformation: Among the 239 corresponding patches, 86.2% have less than 5 residue difference, and 69.0% has less than 3 residue difference. Results of more detailed analysis for this protein complex and another one are shown in Tables 3.6, 3.7 and 3.8. Of course it is still possible that proteins that undergo significant conformational changes from unbound to bound forms may pose challenges to BindML. However, note that such

drastic conformational change is rather rare according to our current knowledge based on the structure database. The average RMSD of PBI sites between bound and unbound form of structures in another independent dataset of 124 bound and unbound complexes [4] is 1.36Å, which is consistent with the dataset we used.

Table 3.6
The RMSD values of bound and unbound structures and the corresponding AUC values of PBI prediction. We examined consistency of residues included in surface patches of bound and unbound conformations of two protein complexes, 1H9D (corresponding unbound structures: 1EANA, 1ILFA) and 2Z0E (unbound structures: 2D1A, 1V49A). Note that AUC values are averaged over the two chains (ligand and receptor proteins).

	RMSD at PBI Region	**AUC of Bound Structures**	**AUC of Unbound Structures**
1H9D	1.32 Å	0.730	0.817
2Z0E	2.15 Å	0.781	0.782

3.4 Discussion

We have reported a novel computational method, BindML, for predicting PBI sites by capturing their characteristic mutation patterns. BindML extracts the MSA of a local surface patch on the query protein structure and computes phylogenetic trees using PBI and NPBI substitution models, whose likelihood scores are then compared. A great advantage of BindML is that the procedure is very general; therefore it can be readily applied for identifying other types of sites in proteins that have distinctive mutation patterns. For example, interaction sites to other molecules, such as DNA, RNA, or membrane can be predicted by computing site specific substitution models from known structural complexes. However, BindML is disadvantageous for structures that have no sequence homologs, where no MSA can be constructed for PBI

Table 3.7
The percentage of common residues in each corresponding surface patches between the bound and unbound structures of 1H9D (corresponding unbound structures: 1EANA, 1ILFA)

	Ligand Protein		**Receptor Protein**	
	Bound	**Unbound**	**Bound**	**Unbound**
	1H9DA	**1EANA**	**1H9DB**	**1ILFA**
Number of corresponding surface patches between bound and unbound structures	79		142	
Average number of residues in a patch	25.038	25.562	23.827	23.407
Percent of patches with 0 residue difference	0.051		0.136	
Percent of patches with 1 residue difference	0.190		0.358	
Percent of patches with 2 residue difference	0.304		0.519	
Percent of patches with 3 residue difference	0.418		0.630	
Percent of patches with 4 residue difference	0.557		0.753	
Percent of patches with 5 residue difference or more	0.684		0.815	

site predictions. Further, the quality of the MSA can potentially affect the accuracy of PBI predictions. In addition, protein complexes that undergo structure conformational changes from unbound to bound forms may pose challenges to BindML, if the conformation change is large enough to significantly shift residues in a given patch from which a MSA is extracted. Further improvement of prediction accuracy is expected by several future developments. More sensitive PBI/NPBI substitution models

Table 3.8
The percentage of common residues in each corresponding surface patches between the bound and unbound structures of 2Z0E (unbound structures: 2D1IA, 1V49A)

	Ligand Protein		**Receptor Protein**	
	Bound	**Unbound**	**Bound**	**Unbound**
	2Z0EB	**1V49A**	**2Z0EA**	**2D1IA**
Number of corresponding surface patches between bound and unbound structures	80		159	
Average number of residues in a patch	22.886	22.988	22.736	22.063
Percent of patches with 0 residue difference	0.125		0.145	
Percent of patches with 1 residue difference	0.400		0.333	
Percent of patches with 2 residue difference	0.600		0.535	
Percent of patches with 3 residue difference	0.750		0.660	
Percent of patches with 4 residue difference	0.850		0.786	
Percent of patches with 5 residue difference or more	0.913		0.836	

could be obtained by employing sophisticated sequence weighting scheme [4, 115] and pseudo-counts [116], rather than the current scheme, which simply uses the BLOSUM method [103] on a sequence set pruned out by the global sequence identity. Instead of the BLOSUM procedure, utilizing thorough techniques for maximizing the likelihood of phylogenetic trees can further optimize the estimation of the amino acid replacement rates [117, 118, 119]. Lastly, including additional structural and/or en-

ergetic features of PBIs, such as residue accessible surface area, surface shape, the secondary structure, by using machine learning framework is certainly expected to improve the accuracy of detection of PBI sites. The present method was developed and evaluated based on our current knowledge of protein-protein interaction sites obtained from the databases. However, there may be many interactions of proteins that are not yet known. Therefore, it is possible that the false positives evaluated in our predictions may actually capture unknown interacting surface regions. The limitation of our current knowledge is an inevitable problem for developing and evaluating prediction methods. It will be necessary to reevaluate and retune our method as well other existing methods a few years later when we have more information about protein-protein interactions. As protein sequence and structure information continue to accumulate at an increasing rate, there is an urgent need for developing methods for annotating functions to new proteins and specifically to their local sites, where these functions are carried out. Most of the popular existing methods are still based on the traditional principle of conservation in sequences and structures. In this work, we showed that mutation patterns forms a rich source of information for identifying functional sites of proteins. In contrast to finding conserved regions, through the development of BindML, we propose a new direction of sequence analysis, which aim to capture mutational constraints or structure of variation in protein sequences. Thus, the current work is conceptually very different from the conventional methods and has broader applicability, where specific local mutational signatures can be classified for newly sequenced proteins. New directions for sequence analyses will become more crucial as the amount of sequence information awaiting our interpretation continue to rapidly grow, particularly by recent new generation sequencing techniques. We believe analyzing hidden structures of sequence variation will open up new directions for biological sequence analysis.

4. COMPUTATIONAL CLASSIFICATION OF PERMANENT AND TRANSIENT PROTEIN-PROTEIN INTERFACE PREDICTIONS

4.1 Introduction

Protein-protein interactions are critical in biology because they mediate many critical biological functions in the cell. Proteins interact with each other in different ways for specific functional reasons. Permanent interactions require tight binding between proteins that form strongly bound complexes of high binding affinity. For example, enzyme-inhibitor, antigen-antibody, and large homo-oligomeric enzyme structures are all comprised of proteins that are required to be permanently bound in order to correctly carry out their functions. In contrast, transient-type protein-protein interactions have a physical mechanism for dissociation after binding, and thus help regulate protein activity at specific times. Examples of transient interactions include proteins involved in signaling pathways, in which binding of transient proteins (such as protein kinases and G proteins) induces conformational changes that allow protein function (and hence pathways) to switch on and off allowing strict and precise control of cellular activity. In a recent review, it was estimated that transient interactions make up a significant portion of protein-protein interaction networks [120]. Thus, distinguishing the two interaction types is essential for understanding and predicting the functions of proteins and has important implications for the origin of protein-protein interaction networks.

The importance of permanent and transient interactions in the functional cellular dynamics of diverse protein-protein interactions, despite works on analyzing the differences between permanent and transient protein interaction sites. In particular, protein structural comparisons have uncover differences between permanent and

transient-type interfaces [121,122,123,88,9,124,125,108,21]. Permanent type protein-protein interfaces interactions have a preference for hydrophobic-type residues, while PPI sites of transient type complexes are more polar [126]. Further, permanent interfaces tend to have fewer gaps in the alignment than transient interfaces [34]. In terms of the size of protein interfaces, transient complexes have smaller interfaces than permanent interfaces [125]. Permanent interfaces are more conserved in sequence than transient interfaces [21]. In a recent review, it has been suggested that dissociation constants (K_d) of permanent complexes are determined to be in the nM range or lower, while transient complexes have K_d in the μM range or higher [122].

In this work, we introduce a new method to classify mutation patterns of protein-protein interaction sites into permanent and transient types. Amino acid substitution models are constructed to differentiate differences between permanent and transient type interfaces. We build specific substitution models that can be used to classify protein interface predictions into permanent and transient interaction types. A detailed understanding of mutational constraint differences between permanent and transient protein complexes would help elucidate critical amino acid substitution preferences that are useful for annotating protein binding interface predictions of structures and sequences of unknown function. We further extend our previous work with BindML [127], a protein binding interface prediction method, to be able to distinguish permanent and transient protein binding sites. Structural classification of protein-protein interaction sites would enable further predictions of binding surfaces to be annotated in ways that would provide clues to their function.

4.2 Methods

4.2.1 Dataset of Permanent and Transient Complexes

Permanent and transient complexes are used to construct amino acid substitution models and benchmark the performance of protein-protein interface predictions and classifications. We initially consider 90 permanent protein structures (from 39 per-

Table 4.1
Dataset of 110 permanent PDB structures.

Permanent Protein Dataset					
1ATN-D	1EZU-A	1LPA-A	1T6B-X	2GOX-A	2TSC-A
1AVX-B	1F34-B	1LPA-B	1T6B-Y	2GOX-B	2VDB-A
1AY7-A	1FLE-I	1LYA-A	1TGS-I	2I25-N	2VDB-B
1AY7-B	1FSK-A	1LYA-B	1UTG-A	2I9B-A	3AAT-A
1BJ1-V	1GL1-I	1M10-A	1UUG-A	2I9B-E	3BP8-C
1BRS-A	1GPW-A	1M10-B	1VFB-A	2J0T-A	3ENL-A
1BVN-T	1GPW-B	1MSB-A	1VFB-B	2J0T-D	3HHR-A
1CDT-A	1GXD-A	1NB5-A	1VSG-A	2JEL-L	3HHR-B
1CHO-F	1GXD-C	1NSN-S	1WDW-B	2JEL-P	3HVT-A
1CHO-G	1HCF-B	1OC0-A	1WEJ-F	2NYZ-A	3SDP-A
1CSE-E	1I2M-A	1OC0-B	1WEJ-H	2O3B-A	3SGB-E
1DFJ-E	1I2M-B	1OPH-A	1XU1-A	2O3B-B	4CPA-I
1DFJ-I	1IBR-B	1PHH-A	1YPI-A	2OUL-B	4MDH-A
1DQJ-C	1IQD-C	1PPE-E	2ABZ-B	2OZA-A	5ADH-A
1EAW-B	1JIW-I	1PPE-I	2B42-B	2OZA-B	5HVP-A
1EER-A	1JPS-T	1PXV-A	2BTF-P	2PCH-A	
1EER-B	1JTG-A	1PYP-A	2CCY-A	2RUS-A	
1EMV-A	1JTG-B	1R0R-I	2CTS-A	2RVE-A	
1EMV-B	1KXQ-A	1STF-I	2GNS-A	2SNI-I	

manent complexes) [16] and 145 transient protein structures (from 45 complexes), in which we designate as the Jones, Nooren, and Thornton (JNT) dataset [9]. Further, we add 161 permanent structures (71 permanent complexes) defined as those with dissociation constant (K_d) values of 1×10^{-9} or lower and 78 transient structures (33 transient complexes) as those with weak K_d values of 1×10^{-6} and higher from the Affinities dataset [128]. The Affinities dataset is a database of protein complexes with assigned K_d values that have been experimentally determined. A summary of the K_d

Table 4.2
Dataset of 72 transient PDB structures.

Transient Protein Dataset					
1A15-B	1C1Y-A	1EWY-A	1HE8-A	1S1Q-B	2BTF-A
1A2K-A	1C1Y-B	1EWY-C	1I2M-B	1SCF-A	2BTF-P
1A78-A	1CEE-B	1F6M-A	1I4D-A	1TRZ-B	2FJU-B
1AK4-A	1CMI-A	1F6M-C	1IKN-A	1TX4-A	2OOB-A
1AK4-D	1CXZ-B	1FFW-A	1J2J-A	1US7-B	2PCB-A
1AKJ-A	1DOM-A	1FFW-B	1LFD-A	1WQ1-G	2PCB-B
1AKJ-B	1E4K-A	1FIN-A	1MQ8-A	1XQS-C	2TGP-I
1AKJ-D	1E4K-C	1FIN-B	1MQ8-B	1YCS-A	2TGP-Z
1BBH-A	1E96-A	1GCQ-B	1QA9-A	1Z0K-B	2TRC-P
1BEB-A	1EDH-A	1GCQ-C	1RRP-A	1ZBD-B	2VIS-A
1BI2-A	1EFU-A	1GOT-B	1RRP-B	1ZM4-B	2VIS-B
1BKD-S	1EFU-B	1HE1-A	1S1Q-A	2AQ3-A	3IL8-A

value ranges and the associated number of complexes is shown in Table 4.3. Next, we combine permanent and transient proteins from the JNT and Affinities dataset together and remove redundant structures with 30% or greater sequence identity, proteins that are annotated as monomers by PISA [129] database, which supersedes the PQS database [100]. PFAM sequences for the representative structures are necessary for building substitution models discussed later in this chapter. Therefore, proteins that did not have PFAM [130] assignments in the dataset are further filtered out. Our final non-redundant permanent and transient dataset consist of 110 permanent and 72 transient structures respectively and is shown in tables 4.1 and 4.2.

4.2.2 Generating Interface Specific Substitution Models

Previously, we generated substitution matrices for the protein binding interface (PBI) and non-protein binding interface (NPBI) regions of proteins to generally pre-

Table 4.3
Ranges of K_d values and their associated number of protein complexes in the Affinities Dataset.

Interaction Type	**Dissociation Constant (K_d)**	**Number of Complexes**
Permanent	$\leq 1\times10^{-12}$	13
Permanent	1×10^{-11}	7
Permanent	1×10^{-10}	21
Permanent	1×10^{-9}	30
Transient	1×10^{-6}	24
Transient	1×10^{-5}	8
Transient	1×10^{-4}	2

dict the location of protein surface binding interfaces without discriminating their specific types [127]. Using the same method, we generate permanent and transient PPI specific substitution models. We generate models for the permanent and transient for PBI (PERM and TRAN respectively) and NPBI (NPERM and NTRAN respectively). Given each protein structure in our permanent and transient protein dataset, we use the all seed sequences for each family provided by the PFAM database. Sequences from PFAM [130] and sequence associated the PDB structure [131] are then used to construct a multiple sequence alignment (MSA) using MUSCLE [101]. Counts of amino acid substitution at the PBI and NPBI are made from the MSAs using the JTT method [102], where mutations are counted for the most evolutionarily close pairs of sequences in each family. Log-odds of the substitution matrices are then calculated using the BLOSUM method [103]. Each log-odd value describes the ratio of the likelihood of two amino acids substitutions observed in evolution and substitutions expected by random chance. The resulting log odds matrices generated for PERM, NPERM, TRAN, and NTRAN are given in tables 4.4, 4.5, 4.6, and 4.7 respectively.

Table 4.4
Log-odds amino acid substitution matrix for permanent protein binding interface (PERM).

PERM	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	0	0	0	-2	-1	0
R	-1	3	-1	-1	-1	0	-1	-1	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1
N	-1	-1	3	0	-2	0	-1	-1	0	-1	-1	0	-1	-1	-1	0	0	-2	-1	-1
D	-1	-1	0	3	-2	-1	0	-1	-1	-2	-2	-1	-2	-2	-1	0	-1	-2	-2	-2
C	-1	-1	-2	-2	5	-2	-2	-2	-1	-1	-1	-2	-1	-1	-2	-1	-1	-2	-1	-1
Q	-1	0	0	-1	-2	3	0	-1	0	-1	-1	0	0	-1	-1	-1	-1	-2	-1	-1
E	0	-1	-1	0	-2	0	2	-1	-1	-1	-1	0	-1	-2	-1	-1	-1	-2	-2	-1
G	-1	-1	-1	-1	-2	-1	-1	3	-1	-2	-2	-1	-2	-2	-2	-1	-1	-2	-2	-2
H	-1	0	0	-1	-1	0	-1	-1	4	-1	-1	-1	-1	0	-1	-1	-1	-1	1	-1
I	-1	-1	-1	-2	-1	-1	-1	-2	-1	3	1	-1	1	0	-1	-1	0	-1	-1	1
L	-1	-1	-1	-2	-1	-1	-1	-2	-1	1	3	-1	1	0	-1	-1	-1	-1	0	0
K	-1	0	0	-1	-2	0	0	-1	-1	-1	-1	2	-1	-2	-1	-1	-1	-2	-1	-1
M	-1	-1	-1	-2	-1	0	-1	-2	-1	1	1	-1	4	0	-1	-1	0	-1	-1	0
F	-1	-1	-1	-2	-1	-1	-2	-2	0	0	0	-2	0	4	-2	-1	-1	0	2	0
P	0	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	3	0	-1	-2	-2	-1
S	0	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	2	0	-2	-1	-1
T	0	-1	0	-1	-1	-1	-1	-1	-1	0	-1	-1	0	-1	-1	0	3	-2	-1	0
W	-2	-1	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-1	0	-2	-2	-2	6	0	-1
Y	-1	-1	-1	-2	-1	-1	-2	-2	1	-1	0	-1	-1	2	-2	-1	-1	0	4	-1
V	0	-1	-1	-2	-1	-1	-1	-2	-1	1	0	-1	0	0	-1	-1	0	-1	-1	3

4.2.3 PPI Site Prediction Method

Previously, we introduced BindML [127], a PPI site prediction method that utilizes PBI and NPBI models. Here, we use the BindML framework with the permanent and transient substitution models. For a given protein structure and its protein family MSA, we predict PPI sites using the substitution models generated from the permanent dataset or transient dataset in the same manner as described previously using PBI/NPBI. PPI site predictions scores are derived from the difference in likelihoods

Table 4.5
Log-odds amino acid substitution matrix for permanent non-protein binding interface (NPERM).

NPERM	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	0	0	0	-2	-1	0
R	-1	3	-1	-1	-1	0	-1	-1	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1
N	-1	-1	3	0	-2	0	-1	-1	0	-1	-1	0	-1	-1	-1	0	0	-2	-1	-1
D	-1	-1	0	3	-2	-1	0	-1	-1	-2	-2	-1	-2	-2	-1	0	-1	-2	-2	-2
C	-1	-1	-2	-2	5	-2	-2	-2	-1	-1	-1	-2	-1	-1	-2	-1	-1	-2	-1	-1
Q	-1	0	0	-1	-2	3	0	-1	0	-1	-1	0	0	-1	-1	-1	-1	-2	-1	-1
E	0	-1	-1	0	-2	0	2	-1	-1	-1	-1	0	-1	-2	-1	-1	-1	-2	-2	-1
G	-1	-1	-1	-1	-2	-1	-1	3	-1	-2	-2	-1	-2	-2	-2	-1	-1	-2	-2	-2
H	-1	0	0	-1	-1	0	-1	-1	4	-1	-1	-1	-1	0	-1	-1	-1	-1	1	-1
I	-1	-1	-1	-2	-1	-1	-1	-2	-1	3	1	-1	1	0	-1	-1	0	-1	-1	1
L	-1	-1	-1	-2	-1	-1	-1	-2	-1	1	3	-1	1	0	-1	-1	-1	-1	0	0
K	-1	0	0	-1	-2	0	0	-1	-1	-1	-1	2	-1	-2	-1	-1	-1	-2	-1	-1
M	-1	-1	-1	-2	-1	0	-1	-2	-1	1	1	-1	4	0	-1	-1	0	-1	-1	0
F	-1	-1	-1	-2	-1	-1	-2	-2	0	0	0	-2	0	4	-2	-1	-1	0	2	0
P	0	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	3	0	-1	-2	-2	-1
S	0	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	2	0	-2	-1	-1
T	0	-1	0	-1	-1	-1	-1	-1	-1	0	-1	-1	0	-1	-1	0	3	-2	-1	0
W	-2	-1	-2	-2	-2	-2	-2	-2	-1	-1	-1	-2	-1	0	-2	-2	-2	6	0	-1
Y	-1	-1	-1	-2	-1	-1	-2	-2	1	-1	0	-1	-1	2	-2	-1	-1	0	4	-1
V	0	-1	-1	-2	-1	-1	-1	-2	-1	1	0	-1	0	0	-1	-1	0	-1	-1	3

calculated from phylogenetic trees generated using NPBI and PBI substitution models. The simple difference in NPBI and PBI tree likelihoods is defined as the distance likelihood (dL) score. Therefore, specific dL scores for permanent (dL_p) and transient (dL_t) predictions are calculated using equation 4.1 and 4.2 respectively.

$$dL_p = L_{NPERM} - L_{PERM} \tag{4.1}$$

$$dL_t = L_{NTRAN} - L_{TRAN} \tag{4.2}$$

Table 4.6
Log-odds amino acid substitution matrix for transient protein binding interface (TRAN).

TRAN	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	3	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	0	0	0	-1	-2	0
R	-1	3	-1	-1	-2	0	-1	-1	-1	-2	-1	0	-1	-2	-1	-1	-1	-2	-1	-1
N	-1	-1	3	0	-2	0	-1	0	0	-1	-1	0	-1	-1	-1	0	0	-1	-1	-1
D	-1	-1	0	3	-3	-1	0	-1	-1	-2	-2	-1	-2	-2	-1	-1	-1	-2	-2	-2
C	-1	-2	-2	-3	5	-2	-2	-1	-2	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1
Q	-1	0	0	-1	-2	3	0	-1	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1
E	-1	-1	-1	0	-2	0	3	-1	-1	-2	-1	-1	-2	-2	-1	-1	-1	-2	-1	-1
G	0	-1	0	-1	-1	-1	-1	3	-1	-2	-2	-1	-2	-2	-1	0	-1	-2	-2	-2
H	-1	-1	0	-1	-2	0	-1	-1	4	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1
I	-1	-2	-1	-2	-1	-1	-2	-2	-1	3	0	-1	0	-1	-1	-1	0	-1	-1	1
L	-1	-1	-1	-2	-1	-1	-1	-2	-1	0	3	-1	1	0	-1	-1	-1	-1	-1	0
K	-1	0	0	-1	-2	0	-1	-1	-1	-1	-1	3	-1	-2	-1	-1	-1	-2	-2	-1
M	-1	-1	-1	-2	-1	-1	-2	-2	-1	0	1	-1	4	0	-1	-1	-1	-1	-1	0
F	-1	-2	-1	-2	-1	-1	-2	-2	-1	-1	0	-2	0	4	-2	-1	-1	0	1	-1
P	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-2	4	0	-1	-1	-2	-1
S	0	-1	0	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	-1	0	2	0	-2	-1	-1
T	0	-1	0	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	-1	-1	0	3	-1	-1	0
W	-1	-2	-1	-2	-1	-1	-2	-2	-1	-1	-1	-2	-1	0	-1	-2	-1	6	0	-2
Y	-2	-1	-1	-2	-1	-1	-1	-2	0	-1	-1	-2	-1	1	-2	-1	-1	0	4	-1
V	0	-1	-1	-2	-1	-1	-1	-2	-1	1	0	-1	0	-1	-1	-1	0	-2	-1	3

L_{NPERM} and L_{NTRAN} represent the phylogenetic tree likelihood following the substitution models of permanent and transient non-protein interface regions respectively. Further, the phylogenetic tree likelihood following substitution models of permanent and transient protein interfaces are denoted as L_{PERM} and L_{TRAN} respectively.

Table 4.7
Log-odds amino acid substitution matrix for transient non-protein binding interface (NTRAN).

NTRAN	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	-1	-1	-1	0	0	0	0	-1	-1	-1	-1	-1	-1	0	0	0	-1	-1	0
R	-1	3	-1	-1	-1	0	-1	-1	0	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1
N	-1	-1	3	0	-1	0	0	0	0	-1	-1	0	-1	-1	-1	0	0	-2	-1	-1
D	-1	-1	0	3	-2	-1	0	-1	-1	-2	-2	-1	-1	-2	-1	0	-1	-2	-1	-2
C	0	-1	-1	-2	5	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	0	0
Q	0	0	0	-1	-1	3	0	-1	0	-1	-1	0	-1	-1	-1	0	-1	-1	-1	-1
E	0	-1	0	0	-2	0	2	-1	-1	-1	-1	0	-1	-2	-1	-1	-1	-2	-1	-1
G	0	-1	0	-1	-1	-1	-1	3	-1	-2	-2	-1	-2	-2	-1	-1	-1	-2	-2	-2
H	-1	0	0	-1	-1	0	-1	-1	4	-1	-1	-1	-1	0	-1	-1	-1	-1	1	-1
I	-1	-1	-1	-2	-1	-1	-1	-2	-1	3	1	-1	1	0	-1	-1	0	-1	-1	1
L	-1	-1	-1	-2	-1	-1	-1	-2	-1	1	3	-1	1	0	-1	-1	-1	0	0	0
K	-1	1	0	-1	-1	0	0	-1	-1	-1	-1	2	-1	-2	-1	-1	-1	-2	-1	-1
M	-1	-1	-1	-1	-1	-1	-1	-2	-1	1	1	-1	4	0	-1	-1	0	-1	-1	0
F	-1	-1	-1	-2	-1	-1	-2	-2	0	0	0	-2	0	4	-1	-1	-1	1	2	0
P	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	3	0	-1	-2	-2	-1
S	0	-1	0	0	0	0	-1	-1	-1	-1	-1	-1	-1	-1	0	2	0	-1	-1	-1
T	0	-1	0	-1	-1	-1	-1	-1	-1	0	-1	-1	0	-1	-1	0	3	-1	-1	0
W	-1	-1	-2	-2	-1	-1	-2	-2	-1	-1	0	-2	-1	1	-2	-1	-1	6	0	-1
Y	-1	-1	-1	-1	0	-1	-1	-2	1	-1	0	-1	-1	2	-2	-1	-1	0	4	-1
V	0	-1	-1	-2	0	-1	-1	-2	-1	1	0	-1	0	0	-1	-1	0	-1	-1	3

4.2.4 PPI Site Classification Method

Given a protein structure and its predicted interface, we attempt to classify this interface into either the permanent or the transient type using a Logistic Regression Model (LRM). LRM performs binary classification by fitting a given number of features to a logit function [132, 133]. Here, we will describe the use of LRM to classify permanent and transient interface predictions. We start by deriving the LRM train-

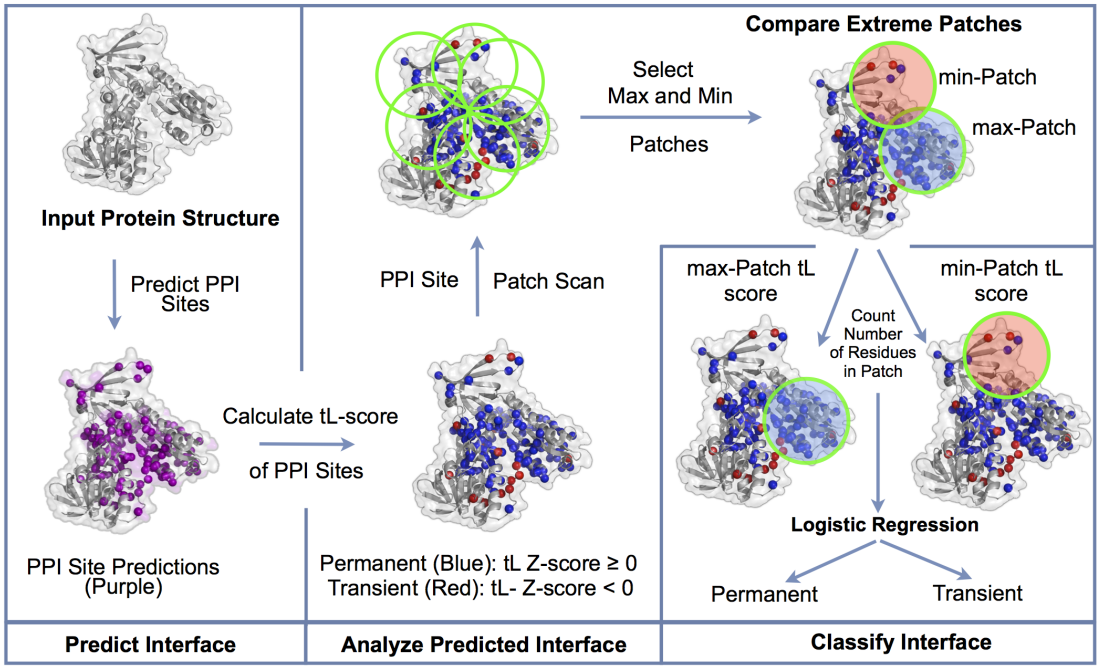

Fig. 4.1. Flowchart of the method for classifying protein interface predictions.

ing parameters by calculating the interface *type likelihood* (tL) score for each PPI site prediction using equation 4.3.

$$tL = L_{PERM} - L_{TRAN} \tag{4.3}$$

Further, tL-scores are recast into Z-scores. Intuitively, a tL Z-score above zero indicates a predicted PPI site is more like permanent, while a lower value below zero suggest that it is more transient. Using the tL-score, we derive various patch-based features used for training the LRM. We define a patch as a given PPI site prediction and other predictions within 15 Å distance, essentially casting surface patches that consist of residues that are a subset of the predicted protein interface. Further, we calculate the average tL-scores in each patch to find two distinct patches: (1) a patch with the lowest average tL-score (min-patch) and (2) a patch with the highest average tL-score (max-patch). From these two patches, we derive the following ten features:

1. Average tL Z-score of residues in the min-patch
2. Average tL Z-score of residues scoring above or equal to zero in the min-patch
3. Average tL Z-score of residues scoring below zero in the min-patch
4. Number of residues with tL Z-scores above or equal to zero in the min-patch
5. Number of residues with tL Z-scores below zero in the min-patch
6. Average tL Z-score of residues in the max-patch
7. Average tL Z-score of residues scoring above or equal to zero in the max-patch
8. Average tL Z-score of residues scoring below zero in the max-patch
9. Number of residues with tL Z-scores above or equal to zero in the max-patch
10. Number of residues with tL Z-scores below zero in the max-patch.

We train a LRM by leave-on-out cross-validation and a combination of the ten input features to classify an entire predicted protein interface as either permanent or transient, using the response variable for the permanent type residues set to one and transient set to zero. Given a predicted protein interface, we use the trained LRM to determine the logistic probability value. A probability value greater than or equal to a set constant threshold will classify a protein interface as permanent. Otherwise, a probability value lower than the set constant threshold will classify a protein interface as transient. A flowchart of the method for classifying protein interface predictions is shown in Fig. 4.1.

4.2.5 Evaluating PPI Site Prediction Performance

The prediction performance of PBI residues was evaluated mainly using the area under the curve (AUC) of sensitivity and specificity across multiple thresholds. True

positives (TP) are the true binding interfaces residues predicted correctly, true negatives (TN) are non-protein-protein interactions sites correctly classified, false positives (FP) are false predictions of protein-protein interaction sites, and false negatives (FN) are protein-protein binding sites that are not predicted. The sensitivity is the fraction of correctly predicted PBI residues over all the true PBI residues. The specificity is the fraction of true negatives among all residues predicted to be NPBI.

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.5}$$

4.2.6 Evaluating PPI Site Classification

PPI site predictions are evaluated in the using the area under the curve (AUC) of recording operating characteristic (ROC) [107] curve, which plots the false positive rate verses the true positive rate for various prediction thresholds. In evaluating protein interface classification, we consider the area under a curve that plots the true permanent classification rate (PCR) over the true transient classification rate (TCR). This is similar to an AUC for a curve that plots sensitivity over specificity in PPI site prediction, where sensitivity represents the true positive (PPI site) rate and specificity represents the true negative (non-PPI site) rate. However, we do not classify permanent and transient on a residue basis. Instead, we classify entire protein interface predictions are either permanent or transient. We use the following equations to calculate PCR and TCR:

$$PCR = \frac{T_{Perm}}{T_{Perm} + F_{Perm}} \tag{4.6}$$

$$TCR = \frac{T_{Tran}}{T_{Tran} + F_{Tran}} \tag{4.7}$$

where T_{Perm} and F_{Perm} represent the number of true and false permanent interfaces classified respectively. Furthermore, T_{Trans} and F_{Trans} represent the number of true and false transient interfaces classified respectively.

4.3 Results

4.3.1 Amino Acid Composition of Permanent and Transient Complexes

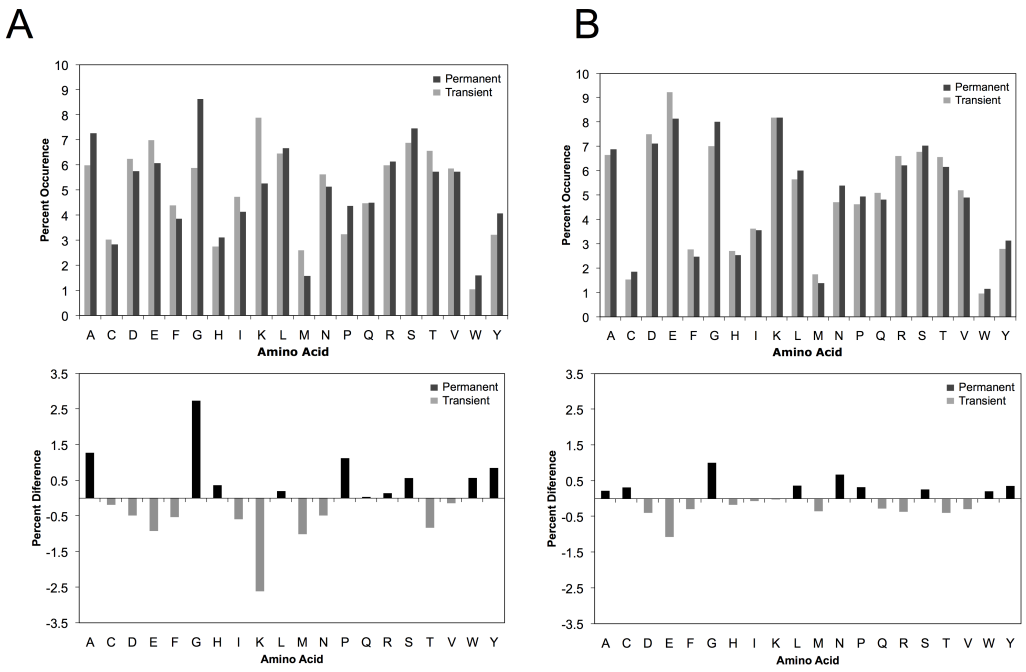

Fig. 4.2. Amino acid frequencies of **(A)** interface and **(B)** non-interfacing regions in permanent and transient complexes.

Permanent and transient protein interfaces have differences in their amino acid compositions, as reported in a previous study [125]. Here, we measure the composition of amino acid frequencies of MSAs (from PFAM sequences) at the interfacing regions and other surface regions. Amino acid frequencies and frequency differences is shown in Fig. 4.2. We find that permanent complexes prefer more hydrophobic

type residues than transient complexes. We observe preference for aromatic residues, namely phenylalanine, tryptophan, and tyrosine in permanent complexes. On the other hand, transient complexes prefer charged residues such as aspartic acid and glutamic acid, and a very distinct bias in composition for lysine residues. Interestingly, there is a very clear bias in glycine residue at permanent interfaces compared to transient ones.

We also include amino acid composition analysis on non-protein interfacing regions of permanent and transient complexes. Although we see much less differences compared to that of the interface regions, there is an appreciable distinction between permanent and transient compositions. In particular, we see further bias in glycine in permanent complexes, where as clear preference for aspartic acid and glutamic acid, which is similar to what was observed in interfacing regions. However, we do not see any clear difference in lysine residues bias for either complex types. This is interesting because it is possible that transient type interactions are complex, where regions outside the protein interface may help mediate transient type association and dissociation. A clear example of this include G-proteins where they require GTP to convert to GDP, which acts as a switch to promote binding and unbinding of signaling proteins [134].

4.3.2 Analysis of Substitution Models

We constructed amino acid substitution models for interfaces and non-interfaces of permanent (NPERM and PERM) and transient (TRAN and NTRAN) complexes. When comparing PERM with NPERM log-odds matrices, they are correlated (r=0.815) by Spearman rank correlation, but are significantly different when subjected to the Kolmogorov-Smirnov distribution test (D=0.110, p=0.016). Further, TRAN and NTRAN were comparatively less correlated (r=0.776), but differences were not significant (D=0.08, p=0.155). As shown later, this result also reflects the lower performance in PPI site prediction compared to that of models built from the permanent

dataset. However, we show that PERM and TRAN comparison have further less correlation (r=0.774) with each other, and is significantly different in their distribution (D=0.105, p=0.024). This is encouraging because significant differences (determined by the Kolmogorov-Smirnov distribution test) in interface amino acid substitutions in permanent and transient complexes shows its usefulness in PPI site classification. Intuitively, we argue that NPERM and NTRAN matrices are not significantly different (D=0.075, p=0.211), which shows that mutation signals of the PPI interfaces are more useful than non-interfaces for interface classification.

4.3.3 Prediction of Permanent and Transient Interfaces

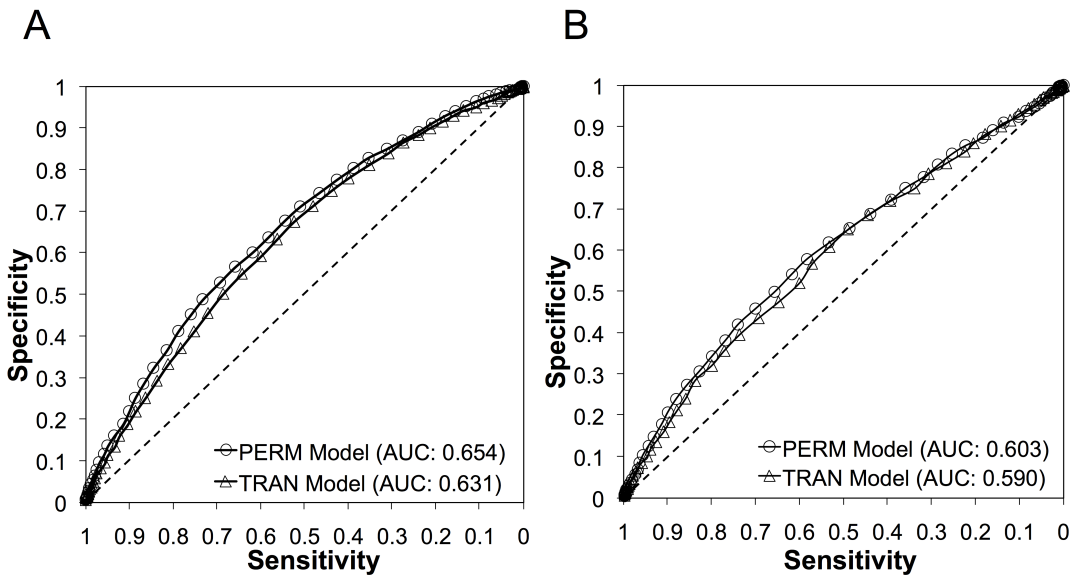

Fig. 4.3. The ROC curve for the overall benchmark results on the JNT dataset of permanent and transient protein complexes. **(A)** Permanent PPI site prediction performance is shown using the PERM/NPERM model in open circles, while **(B)** transient PPI site prediction performance using the TRAN/NTRAN model is shown in open triangles. The dashed line indicates expected performance of random predictions.

We used the BindML method with the NPERM/PERM and NTRAN/TRAN to predict PPI sites on the permanent and transient complexes in the JNT dataset. We find that the permanent protein dataset performs well in comparison to the transient protein dataset shown in Fig. 4.3. This result indicates that the BindML benchmark performance of permanent PPI sites are better predicted (AUC: 0.654) than transient PPI sites, which are comparatively more challenging (AUC: 0.603), but still with good performance. Nevertheless, predictions on both datasets do better than expected from random predictions. For both datasets, we find that the substitution models constructed using the permanent complex dataset performs better than the transient substitution models. The permanent models performs better by two percent AUC of the transient models for predicting permanent type protein interfaces. Further, the permanent model performs slightly better than transient models for the transient dataset as well, by nearly one percent AUC. This shows that permanent and transient substitution models perform nearly the same for transient interface prediction. In the following sections, we will generally use the permanent substitution models to predict protein-protein interfaces because we find that it performs the best for predicting for both permanent and transient proteins.

4.3.4 Examples of tL Z-scores on Protein Structures

As an example of the tL Z-scores, we provide structural mapping of 4MDH-A, a crystal structure of cytoplasmic malate dehydrogenase, which is a permanent complex. As clearly seen in Fig. 4.4A and 4.4B, the protein interface was well predicted (AUC: 0.828). For each of these PPI site predictions, we calculate the tL Z-score, where values above or equal to zero are permanent site predictions, whereas those that are lower are predicted as transient site predictions shown in Fig. 4.4C and D. We find that there are 84 permanent site predictions, where as only 18 transient type predictions. This shows that there is an overwhelming agreement for this complex is likely to be of a permanent type. However, for transient proteins, the distribution of

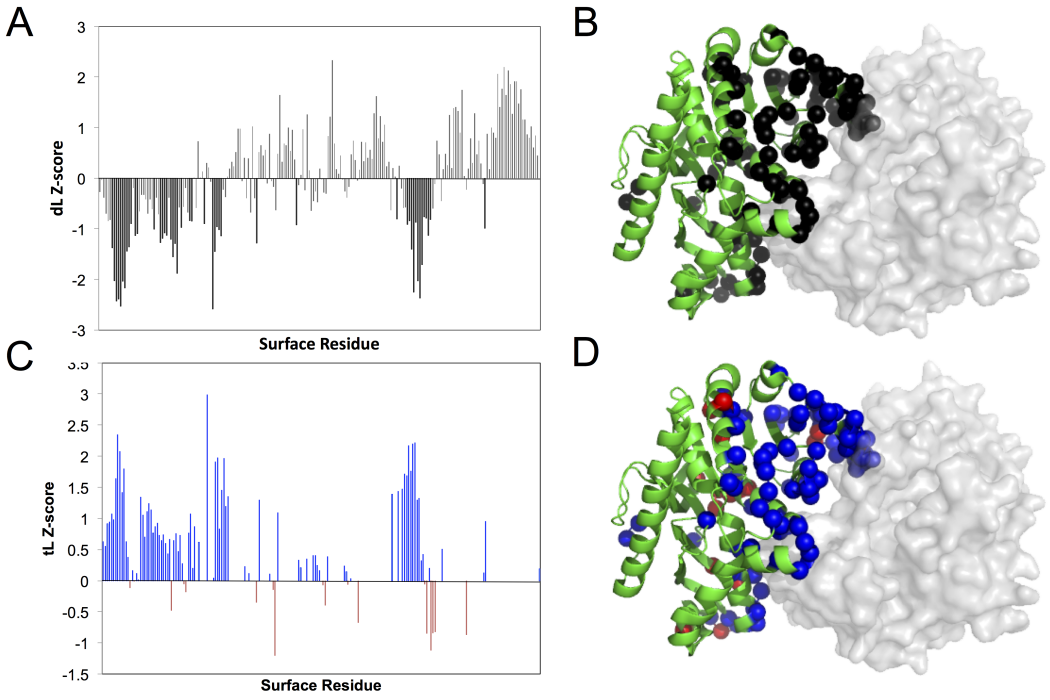

Fig. 4.4. Examples of tL Z-scores mapped to a structure of a permanent complex. PPI site predictions and classification is performed on the structure in green, while the interacting partner in translucent grey surface. **(A)** Distribution of dL-scores used to predict PPI sites, where the interface predictions are colored in black bars **(B)** Cytoplasmic malate dehydrogenase (PDBID: 4MDH-A), with interface predictions in the corresponding black spheres. **(C)** Distribution of tL Z-scores of the PPI site predictions, where blue are permanent site predictions while red are transient site predictions, as they are mapped to the **(D)** PDB structure by their corresponding colors.

tL Z-scores cannot be easily used alone for interface classification, as we will show in the following examples.

We show four additional examples in Fig. 4.5. We illustrate Ran-GPPNHP-RanBP1-RanGAP complex, which is considered to be permanent type interaction that has a well predicted protein interface (AUC: 0.767). This structure as 45 permanent site predictions and 10 transient site predictions, which strongly suggest that it is

in good agreement with the true permanent nature of the interface. Further, we show that for the permanent staphostatin-staphopain complex, we can predict much of the interface as permanent (63 permanent sites, whereas 10 transient sites). The interface for this complex can also be well predicted (AUC: 0.836). Additionally, we show a transient complex, solution structure of cdc42 in complex with the GTPase binding domain of wasp, which interface is predicted with specificity of 0.625 (AUC: 0.607). In this transient example, we see much more transient type site predictions contacting at the interface than those that are predicted to be permanent (9 transient and 18 permanent site predictions). However, there are 8 out of 9 transient type residues in contact with the interacting partner. In comparison, 11 out of 18 permanent type residue predictions are in contact with the interacting partner. This suggests that even though there are more permanent type predictions, there is a larger number of predicted transient residues are part of the true protein interface. Furthermore, another transient complex, bovine beta-lactoglobulin, shows much mixture of transient and permanent type predictions, but 30 transient site predictions are in close to the true interface, whereas 27 permanent site predictions are mixed with transient site predictions. The overall PPI site prediction of this example has an AUC of 0.734. This shows that transient type interfaces appear to be clustered around the true interface, despite an appreciable mixture of permanent site predictions at the periphery of the correct interface residues.

4.3.5 Classification of PPI Site Predictions

The overall LRM classification performance by leave-one-out cross-validation of permanent and transient protein interfaces is shown in Fig. 4.6. Using all ten features, the highest performance along the curve can be found using a LRM probability threshold of 0.629, where our method was able to achieve a permanent classification rate of 0.746 and transient classification rate of 0.762. We find a strong overall AUC value of 0.793 for the permanent and transient classification curve. We also tried

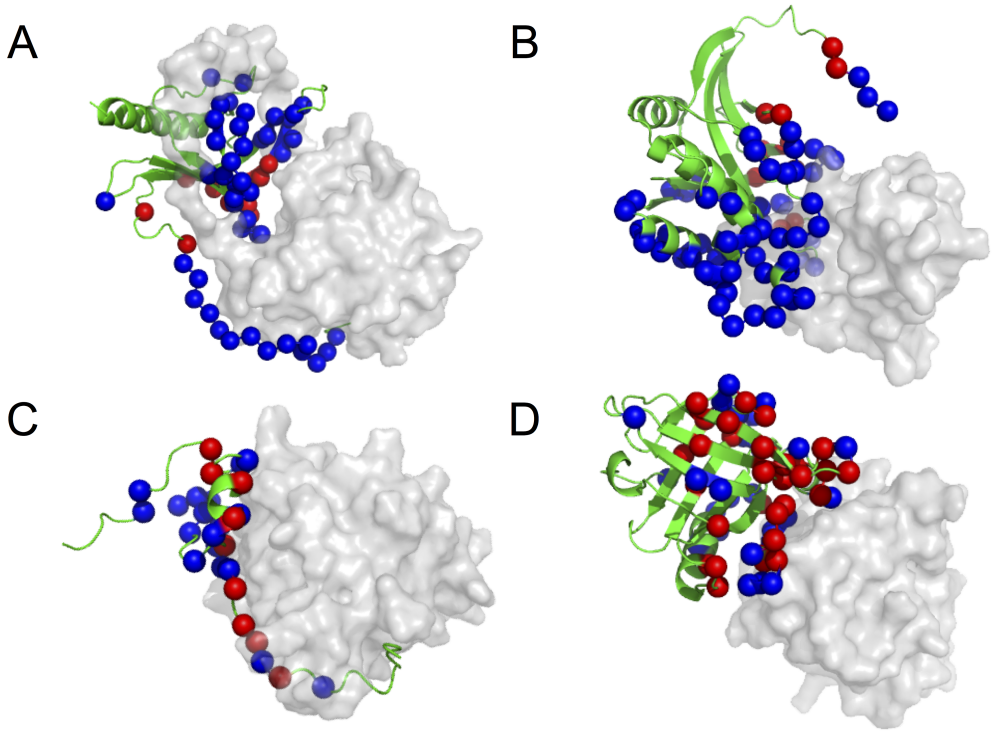

Fig. 4.5. Examples of tL Z-scores mapped to structures. PPI site predictions and classification is performed on the structure in green, while the interacting partner in translucent grey surface. **(A)** Permanent interaction (1K5D-B): structure of Ran-GPPNHP-RanBP1-RanGAP complex, **(B)** Permanent interaction (1PXV-A): staphostatin-staphopain complex, a forward binding inhibitor in complex with its target cysteine protease, **(C)** Transient interaction (1CEE-B): solution structure of cdc42 in complex with the GTPase binding domain of wasp, **(D)** Transient interaction (1BEB-A): bovine beta-lactoglobulin. Blue and red spheres represent the α-carbon of residues predicted as permanent and transient respectively.

different combinations of features. We compared the classification performance of features that only directly use tL Z-scores with features that are based on only counts of predicted permanent and transient residues. We find that residue counts serve best for classification of permanent and transient interfaces (AUC: 0.907), whereas the use of tL Z-scores was slightly worse than the combination of all features (AUC: 0.725). We also provide an ideal comparison in a situation in which the true protein

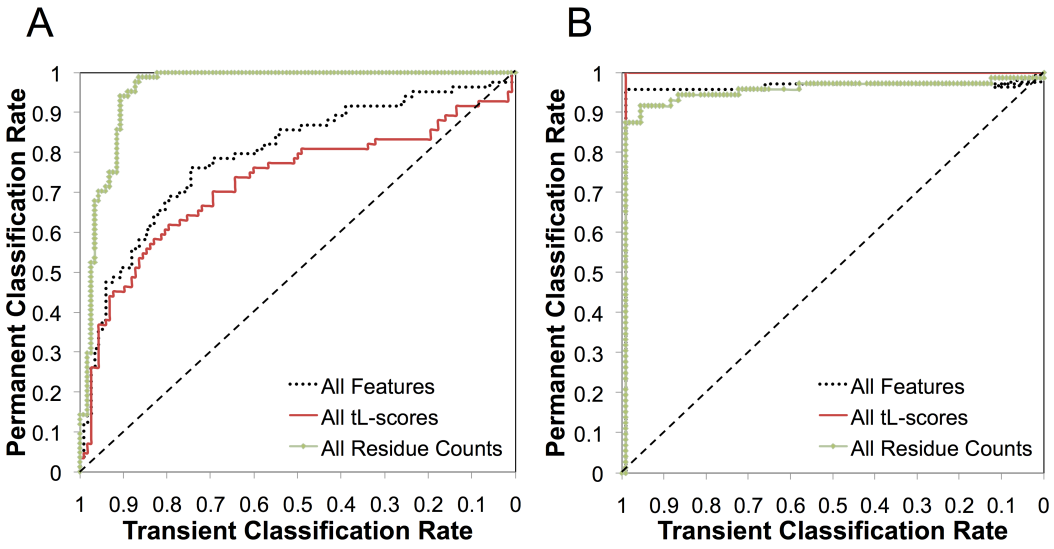

Fig. 4.6. Benchmark classification performances of the permanent and transient complex dataset for **(A)** predicted and **(B)** true interfaces by logistic regression. Lines with black dots show performances for all features used. Red lines represent performances when only using tL Z-scores directly. Green lines demonstrate performances when only using features based on counting the number of permanent and transient residue predictions.

Table 4.8
Classification performance of protein-protein interface that are predicted and known (true answer). Comparison of the best PCR and TCR classification performances for all tL Z-scores, all residue counts and all features.

	Predicted Interface			**Known Interface**		
Combination	**PCR**	**TCR**	**AUC**	**PCR**	**TCR**	**AUC**
All tL Z-scores	0.702	0.695	0.725	1.000	0.991	0.991
All Residue Counts	0.941	0.907	0.957	0.917	0.955	0.951
All Features	0.762	0.746	0.793	0.958	0.991	0.960

interface is known (to assume that the interface is perfectly predicted). We find that performances for all three combination of features (using all features, tL Z-scores,

and residue counts) performed with nearly perfect classification of permanent and transient interfaces with AUC at least greater than 0.950. In contrast to classification using the predicted interfaces, the tL Z-score based features performed with the near perfect classification (AUC: 0.991), while use of residue counts performed slightly worse (AUC: 0.951). Nevertheless, when the interface is known *apriori*, classification of permanent and transient interfaces by our method shows superior performance through all three combination of features used. The best performance in permanent and transient classification rates are shown in Table 4.8.

Furthermore, we show two examples of min- and max-patches and their relative location on the protein surface. Fig. 4.7 shows Tryptophan synthase complex with 4 site predictions in the min-patch and 14 predictions in the max-patch. Eight residues in the max-patch are in direct contact with the interacting partner, where as the min-patch has no contacting residues near the true interface. Since this complex is permanent, the max-patch was used by LRM to correctly classify this interface as permanent. Alternatively, we evaluate a transient interacting structures, the AlphaL I domain in complex with ICAM-1. Interestingly, we observe that the min-patch is closer to the true interface than the max-patch. Further, there are 3 of residues in the min-patch that are in direct contact with the interacting partner, while the max-patch does not have any residues that are part of the true protein-protein interface. This suggest that the min-patch can used by LRM as a strong indicator for classifying this protein as transient, even with over predictions from the max-patch.

4.4 Discussion

We have developed a new computational framework to classify PPI site predictions into permanent and transient types. Through a cross-validation benchmark of a diverse set of proteins from a large permanent and transient dataset, we find that our method preforms well classifying between the two types. When examined in detail, transient interfaces form a more complex mixture or of permanent and transient

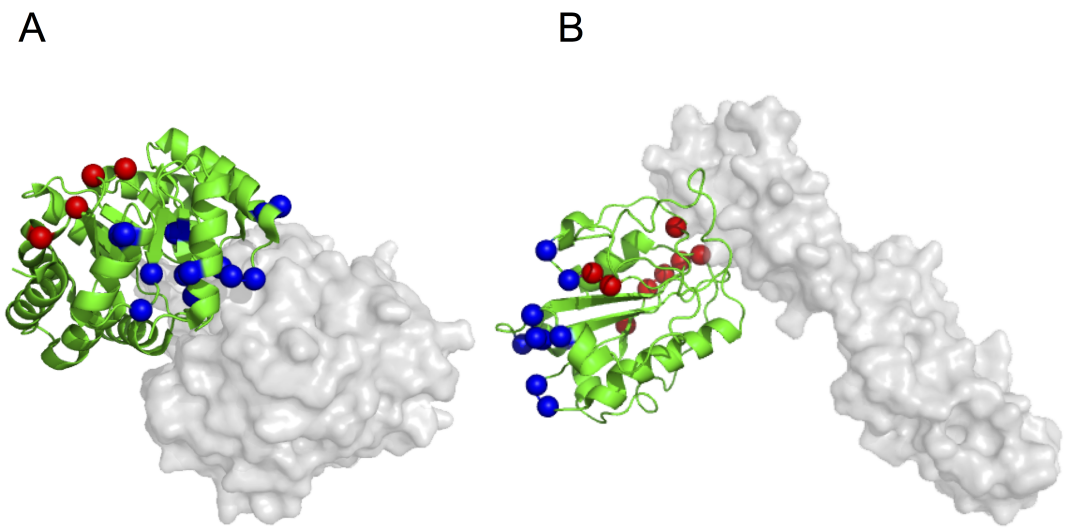

Fig. 4.7. Example of surface residues in min- and max-patches for **(A)** 1WDW-C, Tryptophan synthase complex from a hyperthermophile and **(B)** 1MQ8-B, AlphaL I domain in complex with ICAM-1. Residues in the min-patch are colored in red, while residues in the max-patch are colored in blue.

types residues. This is intuitive because transient interaction sites have residues that promote binding, and also for coordinating subsequent dissociation. Therefore, a mixture of residues types that help bind and stabilize the complex form would be just as important as those that aid precise and tightly controlled in dissociation.

Further, our methodology has broad applications in terms of classifying other type of interaction sites, such as those that are involved in protein-RNA, protein-DNA or protein-membrane binding. Further, specific substitution models for various types of ligand binding sites, such as those that bind metals or specific types of chemical ligands or cofactors, such as ATP/ADP, NAD/NADH, or GTP/GDP can be used by our proposed framework for their prediction and classification. This work marks the beginning of new future methods for not just identifying functional regions on proteins, but also provide key insight to what these regions can do.

5. 3D-SURFER: SOFTWARE FOR HIGH-THROUGHPUT PROTEIN SURFACE COMPARISON AND ANALYSIS

5.1 Introduction

Structural genomics initiatives for solving protein structures at high-throughput are continuing to rapidly progress, which is increasingly providing greater insight into the inner workings of the cellular machinery. Given the large amount protein structures that have been experimentally determined, there are still many structures with uncharacterized functions in the PDB [131]. Although numerous representations of proteins have been used, surface-based approaches have been found to be quite useful both by way of analysis and visualization [135, 136, 137]. Traditional protein structure comparison techniques make use of the pairwise alignment of protein C-α backbone or all atom structure representations. However, computing alignments has a high time complexity and is unsuitable for applications such as real-time structure database searches. To obviate this difficulty, methods such as 3D-BLAST encode the structure as a 1D sequence of alphabets [138]. Light Field Descriptors, on the other hand, create 2D projections (combination of 2D Zernike and Fourier coefficients) rendered from uniformly distributed points around a sphere that surrounds the protein [139]. More recently, the development of 3D moment-based shape representations have shown promising performance for large-scale comparisons [140]. Among these, the 3D Zernike descriptors (3DZD) have been found to be suitable for the efficient comparison of protein surfaces [141]. Unlike the previous two methods, which compare 1D or 2D representations, 3DZD are based on a 3D function expansion. Here, we present 3D-SURFER, a web-based environment for high-throughput protein surface comparison and analysis. The server compares a single protein surface against all protein structures in PDB in just a couple of seconds (over 130,000 single chain

structures, more than 55,000 total PDB entries, which are updated monthly). A performance comparison against other similar tools can be found in the previous work [141]. In addition, local geometrical characteristics of a protein, which represent potential ligand binding sites, can be identified by the VisGrid algorithm [60]. Results shown include visual aids in the form of animated rotations of proteins along with the associated CATH codes [142], and structure alignment calculations using the Combinatorial Extension (CE) algorithm [143].

5.2 Methods

5.2.1 3D Zernike Descriptors

3DZD are utilized for the efficient comparison of protein surfaces across the entire PDB. The calculation of the invariants starts by voxelizing the protein molecular surface that is triangulated by MSROLL version 3.9.3 [144]. The mesh is then discretized to generate a cubic grid. 3DZD, a vector of 121 numbers, is then computed for the protein surface represented by the grid voxels. A single protein represented as a vector can be compared with other structures simply using the Euclidean distance. An example of the retrieval by 3DZD is shown in Table 5.1. We have shown in our previous work that structure retrieval by 3DZD agrees well with main-chain comparison by CE [141]. Also it was found that surface comparison by 3DZD can identify functionally related proteins that cannot be discovered otherwise, due to distant evolutionary relationship [140].

a) Structural alignments calculated using CE.

Table 5.1
Top 10 results using the query *2MTA-A*.

Rank	PDB	CATH Code	Euclidean Distance	RMSD[a)] (Å)
1	1T5K-A	2.60.40.420	1.575	0.48
2	2IdQ-A	2.60.40.420	1.643	0.44
3	1T5K-B	2.60.40.420	1.777	0.77
4	1T5K-C	2.60.40.420	1.778	0.38
5	1AAN-A	2.60.40.420	1.840	0.55
6	2IDU-A	2.60.40.420	2.160	0.89
7	1ID2-C	2.60.40.420	2.201	0.60
8	1AAJ-A	2.60.40.420	2.215	0.53
9	1ID2-A	2.60.40.420	2.320	0.56
10	2IDT-A	2.60.40.420	2.322	1.00

5.2.2 Analyzing Local Surface Geometry

A protein can be interactively analyzed by VisGrid, which identifies geometric features of protein surfaces, e.g., pockets, protrusions, hollow spaces and flat regions [60], which are often associated with binding sites. VisGrid uses a novel visibility criterion, which essentially indicates the fraction of open directions for a given point on the protein surface. The three largest pockets and protrusions are reported. The Qhull program [145] is used to calculate volumes and surface area of the pockets identified.

5.2.3 Input

3D-SURFER takes a PDB ID as an input structure to compare against the entire PDB. PDB IDs may be followed by a character representing the chain. For example, if the PDB structure 2MTA and chain A is of interest, the text entry should be 2MTA-A. Alternatively, a custom PDB structure may be uploaded and utilized as the query. In either case, a search against the entire structure database is executed on-the-fly. Additionally, the user can specify two types of filtering: CATH filtering that avoids displaying structures with similar CATH levels, and length filtering, in charge of displaying proteins whose lengths are similar to the query structure.

5.2.4 Output

The right section of the results panel lists the structures identified as similar by 3DZD (Fig. 5.1). The distance of each retrieved protein to a query is shown after the label, EucD. CATH codes for each of the results are also displayed. Each reported result displays the corresponding PDB ID and is directly linked to the PDB web site. Root mean squared deviations (RMSD) values, calculated using CE, can be viewed by selecting the RMSD checkbox and visualized by clicking on the RMSD button. The protein surface analysis results can be viewed on the left panel (Jmol applet), which can be used to color the surface by clicking on the buttons called Cavity, Protrusion or Flat. The interface will render the surface in three different colors based on their rank in terms of geometric visibility: Red (1st), Green (2nd) and Blue (3rd). The volumes and surface areas of each region are also shown.

5.3 Summary

3D-SURFER provides a platform to perform both global and local structure analysis in real time. Similarity in global structure infers evolutionary relationship in many cases, which can give a clue for the function of the protein. We plan to in-

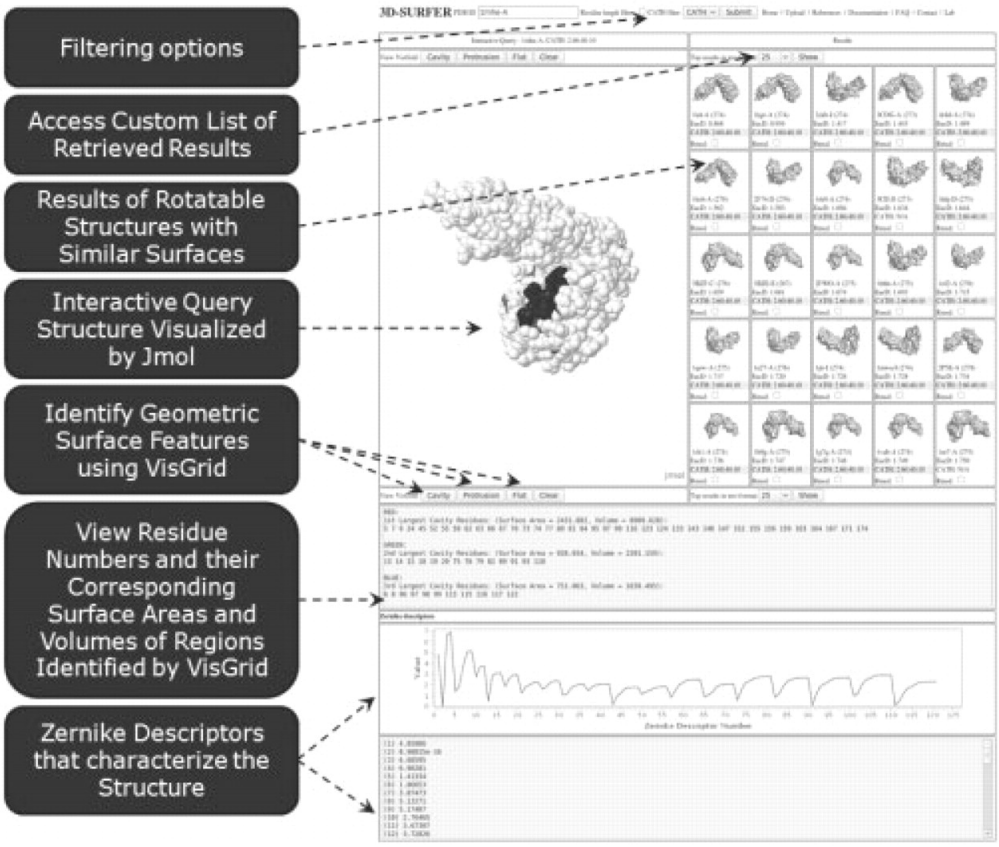

Fig. 5.1. The 3D-SURFER user interface.

corporate protein pocket database search into our platform in the future. In addition, protein surface properties such as electrostatic potentials, hydrophobicity and conservation will be integrated into 3D-SURFER for detailed analysis designed to assist investigating function of proteins. The webserver can be freely accessed at `http://kiharalab.org/3d-surfer/`.

5.4 Supported Platforms

All latest web browsers are supported. The Java plug-in, and appropriate configuration, is required for visualization using Jmol.

6. SUMMARY

Protein-protein interactions are important because they involve many critical (and extremely interesting) molecular processes in biology. A firm understanding of protein-protein interactions starts with the identification of residues most critical for mediating recognition and binding of proteins. The prediction of protein-protein interfaces in structures remains an important challenge, even with the various computational methods that have been developed for identifying PPI sites using combinations of sequence and structural information [87]. Various methods, and the sequence and structural features they use, for PPI site prediction are reviewed in the introductory chapter. Sequence conservation information is, by far, the most important and commonly used feature in the current protein-protein interaction site prediction methods. However, PPI sites are not well conserved compared to other functional regions such as small chemical ligand binding sites and active sites in enzymes [33, 34, 90]. Therefore, it is possible that other functional sites in proteins can be overpredicted as PPI sites using sequence conservation as a major feature alone. The problem stems from the semi-conservation of PPI sites, which poses a great challenge for the current prediction methods.

A direct method for measuring semi-conservation or mutational patterns is critical for improving PPI site predictions. In chapter 3 of this thesis, we introduce a new method for predicting protein-protein interaction sites by the use of amino acid substitution models (substitution matrices) that are constructed from known protein-protein interaction sites. We develop a unique phylogenetic-based framework, called BindML, to capture mutation patterns on patches of a given protein structure [127]. If a given mutation pattern of the patch matches that of known PPI-sites, our method can predict this region as participating in interactions with other proteins. We further tested our method on a large dataset of 505 non-redundant proteins through a five-

fold cross-validation procedure and show that our method performs well compared to alternative machine learning methods, which combine both sequence and structural information. This is very encouraging because our method provides a unique feature that is viewed from a new angle of information that is primarily taken from sequences. When comparing BindML with traditional sequence conservation, we show improvement in the performance for binding site prediction. Further, compared to using BindML alone, the combination of BindML and another strong alternative machine learning method result in further improvement in interface prediction.

Following the work previously described, the BindML framework was extended to predict and classify PPI sites into permanent or transient types. Permanent complexes are tightly bound together and become inseparable throughout their lifetime. For example, enzyme-inhibitor, antigen-antibody, and large oligomeric enzyme complexes are all required to be permanently bound. On the other hand, transient complexes can dissociate after binding as required by their functional requirement. Such examples of transient complexes include protein kinases and other cell signaling proteins. We have developed a computational method to classify results from PPI site prediction into permanent or transient types using BindML and an logistic regression model. By leave-one-out cross-validation, we tested our method tested our classification method on a large non-redundant dataset of 110 permanent and 72 transient proteins with their PPI site predictions and found very good performance for interface classification. Further, when our method is evaluated on the assumption that the true protein interface is known, the performance of our benchmark results show near perfect classification. This is encouraging because as the methods for interface prediction becomes increasingly better in the future, we would expect the results from our interface classification method to also improve. The ability to classify interface types provides greater functional implications to PPI sites that are predicted. Therefore, as more protein structures are become available in the future, our combined protein-protein interface prediction and classification strategy would be able to automatically identify PPI sites on protein surfaces and annotate them with their functions.

Lastly, protein surfaces provide an new perspective on how structures can be compared. By working with established methods of 3D-Zernike descriptors (3DZD) for protein surface analysis [141, 140], we have developed an integrated platform, called 3D-SURFER [146], to facilitate high-throughput screening of similar protein surfaces in the PDB [131]. The screening process starts with taking a query PDB structure as input and discretizing its molecular surface into a uniform grid of fixed dimensions. This grid is then used to compute 3DZD, which is a vector of coefficients (a string of 121 values) that compactly represent the general shape of the protein surface. Important features of 3DZD includes rotational and translational invariance, unlike traditional 3D comparison methods, which require superimposition of structures. Further, the simple comparison between the query protein surface and all surfaces in the database can be made by calculating a one-to-all pairwise euclidean distance between the input and precomputed 3DZD of all known protein structures in the PDB. The searching process is very computationally efficient, where it only takes minutes to do a comparison of a query against the entire PDB. In addition, 3D-SURFER also integrates other surface analysis methods into the results. In particular, the protruding, cavity, and flat surface regions can be identified using the Visibility Criterion [60] and visualized. The pocket volume of identified cavities can also be computed by using a convex hull algorithm. Also, the protein backbone alignment between query and the corresponding hits from retrieval can be computed using the Combinatorial Extension method [143].

LIST OF REFERENCES

LIST OF REFERENCES

[1] L. Giot, "A Protein Interaction Map of *Drosophila melanogaster*," *Science*, vol. 302, pp. 1727–1736, Dec. 2003.

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

[3] T. Ito, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 4569–4574, Mar. 2001.

[4] H. Hwang, B. Pierce, J. Mintseris, J. Janin, and Z. Weng, "Protein-protein docking benchmark version 3.0," *Proteins*, vol. 73, pp. 705–709, May 2008.

[5] J. Mintseris, K. Wiehe, B. Pierce, R. Anderson, R. Chen, J. Janin, and Z. Weng, "Protein-protein docking benchmark 2.0: An update," *Proteins*, vol. 60, pp. 214–216, June 2005.

[6] R. Chen, J. Mintseris, J. l. Janin, and Z. Weng, "A protein-protein docking benchmark," *Proteins*, vol. 52, pp. 88–91, May 2003.

[7] S. J. de Vries, A. D. J. van Dijk, and A. M. J. J. Bonvin, "WHISCY: What information does surface conservation yield? Application to data-driven docking," *Proteins*, vol. 63, pp. 479–489, Jan. 2006.

[8] I. M. A. Nooren, "NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions," *The EMBO Journal*, vol. 22, pp. 3486–3492, July 2003.

[9] I. M. A. Nooren and J. M. Thornton, "Structural characterisation and functional significance of transient protein-protein interactions," *Journal of Molecular Biology*, vol. 325, pp. 991–1018, Jan. 2003.

[10] J. Fernández-Recio, M. Totrov, C. Skorodumov, and R. Abagyan, "Optimal docking area: A new method for predicting protein-protein interaction sites," *Proteins*, vol. 58, pp. 134–143, Oct. 2004.

[11] A. Shulman-Peleg, M. Shatsky, R. Nussinov, and H. J. Wolfson, "Spatial chemical conservation of hot spot interactions in protein-protein complexes," *BMC biology*, vol. 5, no. 1, p. 43, 2007.

[12] X. Li, O. Keskin, B. Ma, R. Nussinov, and J. Liang, "Protein–Protein Interactions: Hot Spots and Structurally Conserved Residues often Locate in Complemented Pockets that Pre-organized in the Unbound States: Implications for Docking," *Journal of Molecular Biology*, vol. 344, pp. 781–795, Nov. 2004.

[13] S. Liang, "Protein binding site prediction using an empirical scoring function," *Nucleic Acids Research*, vol. 34, pp. 3698–3707, July 2006.

[14] L. F. Murga, M. J. Ondrechen, and D. Ringe, "Prediction of interaction sites from apo 3D structures when the holo conformation is different," *Proteins*, vol. 72, pp. 980–992, Feb. 2008.

[15] L. lo Conte, C. Chothia, and J. Janin, "The atomic structure of protein-protein recognition sites," *Journal of Molecular Biology*, vol. 285, pp. 2177–2198, Feb. 1999.

[16] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, pp. 13–20, Jan. 1996.

[17] H. Neuvirth, R. Raz, and G. Schreiber, "ProMate: A Structure Based Prediction Program to Identify the Location of Protein–Protein Binding Sites," *Journal of Molecular Biology*, vol. 338, pp. 181–199, Apr. 2004.

[18] S. S. Negi and W. Braun, "Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces," *Journal of Molecular Modeling*, vol. 13, pp. 1157–1167, Sept. 2007.

[19] S. S. Negi, S. S. Negi, C. H. Schein, C. H. Schein, N. Oezguen, N. Oezguen, T. D. Power, T. D. Power, W. Braun, and W. Braun, "InterProSurf: a web server for predicting interacting sites on protein surfaces," *Bioinformatics*, vol. 23, pp. 3397–3399, Oct. 2007.

[20] S. Jones, "Analysis of protein-protein interaction sites using surface patches," *Journal of Molecular Biology*, vol. 272, pp. 121–132, Sept. 1997.

[21] J. Mintseris, "Structure, function, and evolution of transient and obligate protein-protein interactions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 10930–10935, Aug. 2005.

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

[23] J. Janin, F. Rodier, P. Chakrabarti, and R. P. Bahadur, "Macromolecular recognition in the Protein Data Bank," *Acta crystallographica Section D, Biological Crystallography*, vol. 63, pp. 1–8, Jan. 2007.

[24] E. V. Pletneva, A. T. Laederach, D. B. Fulton, and N. M. Kostic, "The role of cation-pi interactions in biomolecular association. Design of peptides favoring interactions between cationic and aromatic amino acid side chains," *Journal of the American Chemical Society*, vol. 123, pp. 6232–6245, July 2001.

[25] I. Mihalek, I. Res, H. Yao, and O. Lichtarge, "Combining inference from evolution and geometric probability in protein structure evaluation," *Journal of Molecular Biology*, vol. 331, pp. 263–279, Aug. 2003.

[26] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389–3402, Sept. 1997.

[27] H. X. Zhou and Y. Shan, "Prediction of protein interaction sites from sequence profile and residue neighbor list," *Proteins*, vol. 44, pp. 336–343, Aug. 2001.

[28] H. Tjong, S. Qin, and H. X. Zhou, "PI2PE: protein interface/interior prediction engine," *Nucleic Acids Research*, vol. 35, pp. W357–W362, May 2007.

[29] A. Porollo and J. Meller, "Prediction-based fingerprints of protein-protein interactions," *Proteins*, vol. 66, pp. 630–645, Dec. 2006.

[30] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, Jan. 1948.

[31] W. S. J. Valdar, "Scoring residue conservation," *Proteins*, vol. 48, pp. 227–241, June 2002.

[32] L. A. Mirny and E. I. Shakhnovich, "Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function1," *Journal of Molecular Biology*, vol. 291, pp. 177–196, Sept. 1999.

[33] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, pp. 1875–1882, May 2007.

[34] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang, "Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?," *Protein Science*, vol. 13, pp. 190–202, Jan. 2004.

[35] O. Lichtarge, "An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families," *Journal of Molecular Biology*, vol. 257, pp. 342–358, Mar. 1996.

[36] O. Lichtarge, H. R. Bourne, and F. E. Cohen, "Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, pp. 7507–7511, July 1996.

[37] F. Glaser, Y. Rosenberg, A. Kessel, T. Pupko, and N. Ben-Tal, "The ConSurf-HSSP database: The mapping of evolutionary conservation among homologs onto PDB structures," *Proteins*, vol. 58, pp. 610–617, Dec. 2004.

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

[39] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal, "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues," *Bioinformatics*, vol. 18 Suppl 1, pp. S71–7, Jan. 2002.

[40] M. Landau, I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures," *Nucleic Acids Research*, vol. 33, pp. W299–W302, July 2005.

[41] A. Armon, D. Graur, and N. Ben-Tal, "ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information1," *Journal of Molecular Biology*, vol. 307, pp. 447–463, Mar. 2001.

[42] U. Göbel, C. Sander, R. Schneider, and A. Valencia, "Correlated mutations and residue contacts in proteins," *Proteins*, vol. 18, pp. 309–317, Apr. 1994.

[43] I. Halperin, H. Wolfson, and R. Nussinov, "Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families," *Proteins*, vol. 63, pp. 832–845, Feb. 2006.

[44] F. Pazos, "Correlated mutations contain information about protein-protein interaction," *Journal of Molecular Biology*, vol. 271, pp. 511–523, Aug. 1997.

[45] F. Pazos and A. Valencia, "In silico two-hybrid system for the selection of physically interacting protein pairs," *Proteins*, vol. 47, pp. 219–227, May 2002.

[46] D. Kihara, "*Ab initio* protein structure prediction on a genomic scale: Application to the Mycoplasma genitalium genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 5993–5998, Apr. 2002.

[47] D. Kihara and J. Skolnick, "Microbial genomes have over 72assignment by the threading algorithm PROSPECTOR_Q," *Proteins*, vol. 55, pp. 464–473, Mar. 2004.

[48] D. Kihara, "TOUCHSTONE: An *ab initio* protein structure prediction method that uses threading-based tertiary restraints," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 10125–10130, Aug. 2001.

[49] P. Lijnzaad and P. Argos, "Hydrophobic patches on protein subunit interfaces: characteristics and prediction.," *Proteins*, vol. 28, pp. 333–343, July 1997.

[50] F. K. Pettit, E. Bare, A. Tsai, and J. U. Bowie, "HotPatch: A Statistical A pproach to Finding Biologically Relevant Features on Protein Surfaces," *Journal of Molecular Biology*, vol. 369, pp. 863–879, June 2007.

[51] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, pp. 105–132, May 1982.

[52] Z. Dosztányi, J. Chen, A. K. Dunker, I. Simon, and P. Tompa, "Disorder and Sequence Repeats in Hub Proteins and Their Implications for Network Evolution," *Journal of proteome research*, vol. 5, pp. 2985–2995, Nov. 2006.

[53] A. K. Dunker, M. S. Cortese, M. S. Cortese, P. Romero, P. Romero, L. M. Iakoucheva, L. M. Iakoucheva, V. N. Uversky, and V. N. Uversky, "Flexible nets. The roles of intrinsic disorder in protein interaction networks," *FEBS Journal*, vol. 272, pp. 5129–5148, Oct. 2005.

[54] C. J. Oldfield, J. Meng, J. Y. Yang, M. Q. Yang, V. N. Uversky, and A. K. Dunker, "Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners," *BMC Genomics*, vol. 9, no. Suppl 1, p. S1, 2008.

[55] M. Higurashi, T. Ishida, and K. Kinoshita, "Identification of transient hub proteins and the possible structural basis for their multiple interactions," *Protein Science*, vol. 17, pp. 72–78, Jan. 2008.

[56] I. Kufareva, L. Budagyan, E. Raush, M. Totrov, and R. Abagyan, "PIER: Protein interface recognition for structural proteomics," *Proteins*, vol. 67, pp. 400–417, Feb. 2007.

[57] D. Eisenberg and A. D. McLachlan, "Solvation energy in protein folding and binding," *Nature*, vol. 319, pp. 199–203, Jan. 1986.

[58] N. J. Burgoyne, "Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces," *Bioinformatics*, vol. 22, pp. 1335–1342, Mar. 2006.

[59] S. Miller, J. Janin, A. M. Lesk, and C. Chothia, "Interior and surface of monomeric proteins," *Journal of Molecular Biology*, vol. 196, pp. 641–656, Aug. 1987.

[60] B. Li, S. Turuvekere, M. Agrawal, D. La, K. Ramani, and D. Kihara, "Characterization of local geometry of protein surfaces with the visibility criterion," *Proteins*, vol. 71, no. 2, pp. 670–683, 2008.

[61] S. Jones, "Prediction of protein-protein interaction sites using patch analysis," *Journal of Molecular Biology*, vol. 272, pp. 133–143, Sept. 1997.

[62] H. Neuvirth, U. Heinemann, D. Birnbaum, N. Tishby, and G. Schreiber, "ProMateus–an open research approach to protein-binding sites analysis," *Nucleic Acids Research*, vol. 35, pp. W543–W548, May 2007.

[63] J. R. Bradford and D. R. Westhead, "Improved prediction of protein-protein binding sites using a support vector machines approach," *Bioinformatics*, vol. 21, pp. 1487–1494, Apr. 2005.

[64] Y. Murakami, Y. Murakami, S. Jones, and S. Jones, "SHARP2: protein-protein interaction predictions using patch analysis," *Bioinformatics*, vol. 22, pp. 1794–1795, July 2006.

[65] J. Fauchere, "Hydrophobic parameters-pi of amino-acid side-chains from the partitioning of N-acetyl-amino-acid amides," *European Journal of Medicinal Chemistry*, vol. 18, pp. 369–375, Jan. 1983.

[66] J. Fernández-Recio, M. Totrov, and R. Abagyan, "Identification of protein-protein interaction sites from docking energy landscapes," *Journal of Molecular Biology*, vol. 335, pp. 843–865, Jan. 2004.

[67] M. O. Dayhoff, W. C. Barker, and L. T. Hunt, "Establishing homologies in protein sequences.," *Methods in Enzymology*, vol. 91, pp. 524–545, 1983.

[68] S. Qin and H. X. Zhou, "meta-PPISP: a meta web server for protein-protein interaction site prediction," *Bioinformatics*, vol. 23, pp. 3386–3387, Oct. 2007.

[69] H. X. Zhou and S. Qin, "Interaction-site prediction for protein complexes: a critical assessment," *Bioinformatics*, vol. 23, pp. 2203–2209, Sept. 2007.

[70] B. Huang, B. Huang, M. Schroeder, and M. Schroeder, "Using protein binding site prediction to improve protein docking," *Gene*, vol. 422, pp. 14–21, Oct. 2008.

[71] I. Mihalek, I. Res, and O. Lichtarge, "On Itinerant Water Molecules and Detectability of Protein–Protein Interfaces through Comparative Analysis of Homologues," *Journal of Molecular Biology*, vol. 369, pp. 584–595, June 2007.

[72] E. M. Phizicky and S. Fields, "Protein-protein interactions: methods for detection and analysis," *Microbiological Reviews*, vol. 59, pp. 94–123, Mar. 1995.

[73] T. Formosa, J. Barry, B. M. Alberts, and J. Greenblatt, "Using protein affinity chromatography to probe structure of protein machines.," *Methods in Enzymology*, vol. 208, pp. 24–45, 1991.

[74] A. Otto-Bruc, B. Antonny, T. M. Vuong, P. Chardin, and M. Chabre, "Interaction between the retinal cyclic GMP phosphodiesterase inhibitor and transducin. Kinetics and affinity studies.," *Biochemistry*, vol. 32, pp. 8636–8645, Aug. 1993.

[75] J. Weiel and J. W. Hershey, "Fluorescence polarization studies of the interaction of Escherichia coli protein synthesis initiation factor 3 with 30S ribosomal subunits.," *Biochemistry*, vol. 20, pp. 5859–5865, Sept. 1981.

[76] K. Kuroda, M. Kato, J. Mima, and M. Ueda, "Systems for the detection and analysis of protein–protein interactions," *Applied Microbiology and Biotechnology*, vol. 71, pp. 127–136, Mar. 2006.

[77] J. M. Holt and G. K. Ackers, "Kinetic trapping of a key hemoglobin intermediate.," *Methods in molecular biology (Clifton, N.J.)*, vol. 796, pp. 19–29, 2012.

[78] M. Pierce, "Isothermal Titration Calorimetry of Protein–Protein Interactions," *Methods (San Diego, Calif)*, vol. 19, pp. 213–221, Oct. 1999.

[79] R. J. Falconer and B. M. Collins, "Survey of the year 2009: applications of isothermal titration calorimetry.," *Journal of molecular recognition : JMR*, vol. 24, no. 1, pp. 1–16, 2011.

[80] M. Malmqvist, "Biospecific interaction analysis using biosensor technology.," *Nature*, vol. 361, pp. 186–187, Jan. 1993.

[81] A. E. Todd, R. L. Marsden, J. M. Thornton, and C. A. Orengo, "Progress of Structural Genomics Initiatives: An Analysis of Solved Target Structures," *Journal of Molecular Biology*, vol. 348, pp. 1235–1260, May 2005.

[82] J. M. Chandonia, "The Impact of Structural Genomics: Expectations and Outcomes," *Science*, vol. 311, pp. 347–351, Jan. 2006.

[83] D. Xu, C. J. Tsai, and R. Nussinov, "Hydrogen bonds and salt bridges across protein-protein interfaces," *Protein Engineering*, vol. 10, pp. 999–1012, Sept. 1997.

[84] P. Fariselli, F. Pazos, A. Valencia, and R. Casadio, "Prediction of protein–protein interaction sites in heterocomplexes with neural networks," *European Journal of Biochemistry*, vol. 269, pp. 1356–1361, Mar. 2002.

[85] H. Chen and H.-X. Zhou, "Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against NMR data," *Proteins*, vol. 61, pp. 21–35, Aug. 2005.

[86] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M. L. Tress, "Progress and challenges in predicting protein-protein interaction sites," *Briefings in Bioinformatics*, vol. 10, pp. 233–246, Dec. 2008.

[87] D. La and D. Kihara, "Discovering Protein-Protein Interaction Sites from Sequence and Structure," in *Biological Data Mining in Protein Interaction Networks* (X.-L. Li and S.-K. Ng, eds.), pp. 64–79, 2008.

[88] K.-i. Cho, K. Lee, K. H. Lee, D. Kim, and D. Lee, "Specificity of molecular interactions in transient protein-protein interaction interfaces," *Proteins*, vol. 65, pp. 593–606, Nov. 2006.

[89] Z. Hu, B. Ma, H. Wolfson, and R. Nussinov, "Conservation of polar residues as hot spots at protein interfaces," *Proteins*, vol. 39, pp. 331–342, June 2000.

[90] S. Jones and J. M. Thornton, "Searching for functional sites in protein structures," *Current Opinion in Chemical Biology*, vol. 8, pp. 3–7, Feb. 2004.

[91] A. J. Bordner, "Predicting small ligand binding sites in proteins using backbone structure," *Bioinformatics*, vol. 24, pp. 2865–2871, Oct. 2008.

[92] D. La and D. R. Livesay, "Predicting functional sites with an automated algorithm suitable for heterogeneous datasets.," *BMC Bioinformatics*, vol. 6, p. 116, 2005.

[93] D. La, B. Sutch, and D. R. Livesay, "Predicting protein functional sites with phylogenetic motifs," *Proteins*, vol. 58, pp. 309–320, Nov. 2004.

[94] D. Livesay and D. La, "The evolutionary origins and catalytic importance of conserved electrostatic networks within TIM-barrel proteins," *Protein Science*, Jan. 2005.

[95] O. Lichtarge and M. E. Sowa, "Evolutionary predictions of binding surfaces and interactions.," *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 21–27, 2002.

[96] S. Sankararaman and K. Sjölander, "INTREPID–INformation-theoretic TREe traversal for Protein functional site IDentification," *Bioinformatics*, vol. 24, pp. 2445–2452, Aug. 2008.

[97] A. Rausell, D. Juan, F. Pazos, and A. Valencia, "Protein interactions and ligand binding: From protein subfamilies to functional specificity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 1995–2000, Feb. 2010.

[98] R. D. Finn, M. Marshall, and A. Bateman, "iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions," *Bioinformatics*, vol. 21, pp. 410–412, Jan. 2005.

[99] The UniProt Consortium, "The Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 35, pp. D193–D197, Jan. 2007.

[100] K. Henrick and J. M. Thornton, "PQS: a protein quaternary structure file server," *Trends in Biochemical Sciences*, vol. 23, pp. 358–361, Sept. 1998.

[101] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, pp. 1792–1797, Mar. 2004.

[102] D. Jones, W. Taylor, and J. Thornton, "The rapid generation of mutation data matrices from protein sequences," *Bioinformatics*, Jan. 1992.

[103] S. Henikoff and J. Henikoff, "Amino Acid Substitution Matrices from Protein Blocks," *Proceedings of the National Academy of Sciences of the United States of America*, Jan. 1992.

[104] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, pp. 696–704, Oct. 2003.

[105] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data," *Molecular Biology and Evolution*, vol. 14, pp. 685–695, July 1997.

[106] V. Hollich, "Assessment of Protein Distance Measures and Tree-Building Methods for Phylogenetic Tree Reconstruction," *Molecular Biology and Evolution*, vol. 22, pp. 2257–2264, July 2005.

[107] M. Gribskov and N. L. Robinson, "Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching," *Computers & Chemistry*, vol. 20, no. 1, pp. 25–33, 1996.

[108] A. J. Bordner and R. Abagyan, "Statistical analysis and prediction of protein-protein interfaces," *Proteins*, vol. 60, pp. 353–366, May 2005.

[109] Y. Y. Tseng, "Estimation of Amino Acid Residue Substitution Rates at Local Spatial Regions and Application in Protein Function Inference: A Bayesian Monte Carlo Approach," *Molecular Biology and Evolution*, vol. 23, pp. 421–436, Sept. 2005.

[110] M. D. Kelly and R. L. Mancera, "A New Method for Estimating the Importance of Hydrophobic Groups in the Binding Site of a Protein," *Journal of Medicinal Chemistry*, vol. 48, pp. 1069–1078, Feb. 2005.

[111] A. J. Venkatakrishnan, E. D. Levy, and S. A. Teichmann, "Homomeric protein complexes: evolution and assembly," *Biochemical Society Transactions*, vol. 38, p. 879, Aug. 2010.

[112] R. Landgraf, I. Xenarios, and D. Eisenberg, "Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins1," *Journal of Molecular Biology*, vol. 307, pp. 1487–1502, Apr. 2001.

[113] E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann, "3D complex: a structural classification of protein complexes.," *PLoS Computational Biology*, vol. 2, p. e155, Nov. 2006.

[114] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29–36, Apr. 1982.

[115] H. Hwang, H. Hwang, T. Vreven, T. Vreven, J. Janin, J. Janin, Z. Weng, and Z. Weng, "Protein-protein docking benchmark version 4.0," *Proteins*, vol. 78, pp. 3111–3114, Aug. 2010.

[116] S. R. Sunyaev, F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan, and E. N. Kuznetsov, "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations.," *Protein Engineering*, vol. 12, no. 5, pp. 387–394, 1999.

[117] J. Pei and N. V. Grishin, "AL2CO: calculation of positional conservation in a protein sequence alignment," *Bioinformatics*, vol. 17, pp. 700–712, Aug. 2001.

[118] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler, "Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology.," *Computer Applications in the Biosciences*, vol. 12, no. 4, pp. 327–345, 1996.

[119] I. Holmes, "Using evolutionary Expectation Maximization to estimate indel rates," *Bioinformatics*, vol. 21, pp. 2294–2300, Mar. 2005.

[120] R. Krause, C. von Mering, P. Bork, and T. Dandekar, "Shared components of protein complexes–versatile building blocks or biochemical artefacts?," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 26, pp. 1333–1343, Dec. 2004.

[121] M. Gao and J. Skolnick, "Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 22517–22522, Dec. 2010.

[122] J. R. Perkins, I. Diboun, B. H. Dessailly, J. G. Lees, and C. Orengo, "Transient Protein-Protein Interactions: Structural, Functional, and Network Properties," *Structure*, vol. 18, pp. 1233–1243, Oct. 2010.

[123] J. Mintseris and Z. Weng, "Atomic contact vectors in protein-protein recognition," *Proteins*, vol. 53, pp. 629–639, Oct. 2003.

[124] Y. Ofran and B. Rost, "Analysing six types of protein-protein interfaces," *Journal of Molecular Biology*, vol. 325, pp. 377–387, Jan. 2003.

[125] S. Ansari and V. Helms, "Statistical analysis of predominantly transient protein-protein interfaces," *Proteins*, vol. 61, pp. 344–355, Aug. 2005.

[126] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, "Residue frequencies and pairing preferences at protein-protein interfaces," *Proteins*, vol. 43, pp. 89–102, May 2001.

[127] D. La and D. Kihara, "A novel method for protein-protein interaction site prediction using phylogenetic substitution models.," *Proteins*, Sept. 2011.

[128] P. L. Kastritis, I. H. Moal, H. Hwang, Z. Weng, P. A. Bates, A. M. J. J. Bonvin, and J. Janin, "A structure-based benchmark for protein-protein binding affinity," *Protein Science*, vol. 20, pp. 482–491, Feb. 2011.

[129] Q. Xu, A. A. Canutescu, G. Wang, G. Wang, M. Shapovalov, M. Shapovalov, Z. Obradovic, Z. Obradovic, R. L. Dunbrack Jr., and R. L. Dunbrack Jr., "Statistical Analysis of Interface Similarity in Crystals of Homologous Proteins," *Journal of Molecular Biology*, vol. 381, pp. 487–507, Aug. 2008.

[130] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Research*, vol. 38, pp. D211–D222, Dec. 2009.

[131] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235–242, Jan. 2000.

[132] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*. Wiley-Interscience, New York, 2000.

[133] J. Hilbe, *Logistic regression models*. Chapman & Hall/CRC, 2009.

[134] D. G. Lambright, J. Sondek, A. Bohm, N. P. Skiba, H. E. Hamm, and P. B. Sigler, "The 2.0 Å crystal structure of a heterotrimeric G protein," *Nature*, vol. 379, pp. 311–319, Jan. 1996.

[135] V. Venkatraman, L. Sael, and D. Kihara, "Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors.," *Cell Biochemistry and Biophysics*, vol. 54, no. 1-3, pp. 23–32, 2009.

[136] D. Fischer, R. Norel, H. Wolfson, and R. Nussinov, "Surface motifs by a computer vision technique: searches, detection, and implications for protein-ligand recognition.," *Proteins*, vol. 16, pp. 278–292, July 1993.

[137] M. Rosen, S. L. Lin, H. Wolfson, and R. Nussinov, "Molecular shape comparisons in searches for active sites and functional similarity.," *Protein Engineering*, vol. 11, pp. 263–277, Apr. 1998.

[138] J.-M. Yang and C.-H. Tung, "Protein structure database search and evolutionary classification.," *Nucleic Acids Research*, vol. 34, no. 13, pp. 3646–3659, 2006.

[139] J.-S. Yeh, D.-Y. Chen, B.-Y. Chen, and M. Ouhyoung, "A web-based three-dimensional protein retrieval system by matching visual similarity.," *Bioinformatics*, vol. 21, pp. 3056–3057, July 2005.

[140] L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, and D. Kihara, "Fast protein tertiary structure retrieval based on global surface shape similarity," *Proteins*, vol. 72, pp. 1259–1273, Mar. 2008.

[141] L. Sael, L. Sael, D. La, D. La, B. Li, B. Li, R. Rustamov, R. Rustamov, D. Kihara, and D. Kihara, "Rapid comparison of properties on protein surface," *Proteins*, vol. 73, pp. 1–10, July 2008.

[142] C. A. Orengo, J. E. Bray, D. W. A. Buchan, A. Harrison, D. Lee, F. M. G. Pearl, I. Sillitoe, A. E. Todd, and J. M. Thornton, "The CATH protein family database: a resource for structural and functional annotation of genomes.," *Proteomics*, vol. 2, pp. 11–21, Jan. 2002.

[143] I. N. Shindyalov and P. E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.," *Protein Engineering*, vol. 11, pp. 739–747, Sept. 1998.

[144] M. L. Connolly, "The molecular surface package.," *Journal of molecular graphics*, vol. 11, pp. 139–141, June 1993.

[145] C. Barber and D. Dobkin, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software*, 1996.

[146] D. La, J. Esquivel-Rodriguez, V. Venkatraman, B. Li, L. Sael, S. Ueng, S. Ahrendt, and D. Kihara, "3D-SURFER: software for high-throughput protein surface comparison and analysis," *Bioinformatics*, vol. 25, pp. 2843–2844, Oct. 2009.

VITA

VITA

David La's scientific career started when he pursued his Master's degree in the area of Bioinformatics (2003-2005) from the California State Polytechnic University, Pomona, where he worked with Dennis R. Livesay (presently at University of North Carolina at Charlotte) on a novel method for predicting protein functional sites with phylogenetic motifs that resulted in 6 peer-review publications. With this surge of motivation, he came to Purdue University, West Lafayette to pursue his Ph.D. starting in 2005, where he worked with Daisuke Kihara on the computational prediction of protein-protein interaction sites inspired by the well established computational methods in phylogenetics, which has lead to the development of new and interesting ways to discover mutation patterns. His doctoral study at Purdue University resulted in 9 additional peer-review publications. He plans to continue his research career in the area of protein-protein interface design with David Baker at the University of Washington, Seattle.