PAIRpred: Partner-specific prediction of interacting residues from sequence and structure

Fayyaz ul Amir Afsar Minhas

Department of Computer Science

Colorado State University

Fort Collins, Colorado 80523, USA

fayyazafsar@gmail.com

Brian J. Geiss

Department of Microbiology, Immunology and Pathology

Colorado State University

Fort Collins, Colorado 80523, USA

brian.geiss@colostate.edu

Asa Ben-Hur (corresponding author)

Department of Computer Science

Colorado State University

Fort Collins, Colorado 80523, USA

asa@cs.colostate.edu

October 11, 2013

short title: Interface prediction using PAIRpred

keywords: interface prediction, machine learning

Abstract

This paper presents a novel partner-specific protein-protein interaction site prediction method called PAIRpred. Unlike most existing machine learning binding site prediction methods, PAIRpred uses information from *both* proteins in a protein complex to generate predict pairs of interacting residues on the two proteins. PAIRpred captures sequence and structure information about residue pairs through pairwise kernels that are used in training a support vector machine classifier. As a consequence, PAIRpred presents a more detailed model of protein binding, and offers state of the art accuracy in predicting binding sites at the protein level as well as inter-protein residue contacts at the complex level. We demonstrate PAIRpred's performance on Docking Benchmark 4.0 and recent CAPRI targets. We present a detailed performance analysis outlining the contribution of different sequence and structure features, together with a comparison to

a variety of existing interface prediction techniques. We have also studied the impact of binding-associated conformational change on prediction accuracy and found PAIRpred to be more robust to such structural changes than existing schemes. As an illustration of potential applications of PAIRpred, we provide a case study in which PAIRpred is used to analyze the nature and specificity of the interface in the interaction of human ISG15 protein with NS1 protein from influenza A virus. Python code for PAIRpred is available at: `http://combi.cs.colostate.edu/supplements/pairpred/`.

1 Introduction

Proteins form the functional backbone of all living cells. They are involved in a variety of cellular functions and processes ranging from cell signaling and transportation to structural stability and gene expression control. Most protein functions are possible only through the interaction or binding of multiple proteins, and as a consequence, the study of protein binding is important in understanding protein function and disease mechanism as well as for drug design, discovery and effectiveness studies. The regions where binding occurs can be identified by analyzing the NMR or X-ray crystallography structures of bound protein complexes, or using biological assays such as mutagenesis experiments. However, these techniques are time consuming, difficult, and expensive to perform. As a result, computational methods for predicting the binding sites in protein-protein interactions are of great importance to help guide biological or structural biology efforts. However, the computational prediction of protein binding sites from their unbound structures or sequences is a complicated task owing to the large variety of physio-chemical phenomena involved in the process, not all of

which are fully known or thoroughly understood [1]. For example, the conformation of two proteins in their bound state can be very different from their unbound configurations [2, 3]. Furthermore, protein molecules are in a constant state of motion in which both the backbone and side chains of residues exhibit significant flexibility [4, 5].

The problem of predicting binding regions in protein complexes from the *unbound* structures or sequences of the proteins involved in the complex has two flavors:

1. Partner-independent prediction: Given a protein A, find whether a residue a in the protein is involved in an interaction with *any* other protein.

2. Partner-specific prediction: Given proteins A and B, find whether a residue a in A interacts with residue b in B upon the formation of the complex $A - B$.

With this context in mind, we propose to differentiate between a *binding site* on a protein and the *interface* in a complex as follows: the region on a protein that is involved in an interaction with another protein is called its binding site, whereas, the group of interacting residues in a complex constitute the *interface* of the complex. Note that, given the interface of a complex, it is trivial to determine the binding region on each protein in the complex whereas the inverse is not. Thus, partner-independent predictors can only find protein binding sites whereas partner-specific predictors can provide information about both interfaces and the binding sites on the individual proteins.

A number of partner-independent methods have been proposed. However, in this paper, we focus on partner-specific prediction only. For reviews of partner-independent predictors, the interested reader is referred to [1, 6–8].

Partner-specific predictions provide more information about the nature of the complex as

they can tell which residues in one protein interact with which residues in the other protein. These more detailed predictions can then be used, for example, to enumerate the distinct binding modes of a protein and to find out whether a protein can bind two other proteins simultaneously or not. Furthermore, partner independent predictors ignore the fact that the binding propensity of a residue is dependent upon the nature and local environment of residues in its target protein. As a consequence, partner-specific interface predictors can be expected to be more accurate in comparison to binding site predictors, as it presents a more complete model of protein binding. This has been demonstrated by Ahmad et al. [9], and the results presented in this paper confirm these findings.

Existing techniques that can be used for the partner-specific prediction of interfaces can be divided in to three classes:

Docking methods: The objective of protein-protein docking methods is to predict the three dimensional structure of a macromolecular complex given the unbound structures of its constituent proteins. Existing methods for protein-protein docking include ZDOCK [10], HADDOCK [11] and RosettaDock [12]. For further details on different docking methods, the interested reader is referred to a recent review [13]. Docking methods typically produce a large number of putative complexes which are then ranked using a ranking criterion to identify the near-native structure. Once the predicted structure of a protein complex is available from docking, the binding interface can be easily recovered. However, docking solves a more general and complex problem than interface prediction as its primary objective is to construct the correct three dimensional structure of the protein complex. Docking methods are hampered by a lack of complete understanding

of the factors involved in complex formation such as binding associated conformational changes [1]. As a consequence, docking methods do not fare well in cases with large conformational change. For example, ZDOCK is able to find near-native structures for 33 rigid-body complexes but for only two non-rigid body complexes in its top 10 predictions on a data set of 124 rigid and 52 non-rigid complexes [10]. Docking methods can benefit from binding site predictions as the correct identification of the interface of the complex can limit the degrees of conformational freedom in docking. Some machine learning schemes such as have been used for the scoring of docking conformations to predict which one is closest to the native structure [14, 15]. Some docking methods employ partner-independent predictors to accomplish this. However, partner-specific interface predictions can be expected to play a better role [16].

Template based methods: With the growth in the number of protein complexes in PDB, both template based interface predictors and template based docking schemes have attracted attention. In these methods, a protein complex is modeled using sequence or structural similarity to a known template protein complex. Template based methods can either use sequence or structural homology or interface similarity [17]. However, these methods are applicable only when template complexes or interfaces exist for a query complex. A recent comparison between template based and docking methods shows that both types of methods have comparable performance [18].

Machine learning methods: Direct prediction of interfaces using machine learning techniques is a relatively unexplored research area and the accuracy of existing methods in this category is low. Unlike template based methods, machine learning based tech-

niques, depending upon the features used for the prediction, are also applicable in cases where template similarity of the query proteins to known interfaces cannot be established. In this work, we focus primarily on machine learning methods.

One of the first approaches to perform partner-specific binding site predictions is InSite [19]. InSite models interactions at the motif or domain level, and predicts pairs of interacting motifs that best explain a given protein-protein interaction network. InSite does not use information about known interaction sites, and is limited by the richness of the motif library and its coverage across a given protein. Finally, it is more valuable to obtain binding site information at the residue level, as it allows for a more detailed understanding of the interaction.

Recently, Ahmad and Mizuguchi investigated the impact of performing partner-specific versus partner-independent binding site prediction [9]. Their analysis of the binding propensities of residue pairs in protein-protein interfaces clearly shows that the binding propensity of a residue is strongly dependent upon its partner in other proteins. On this basis, they hypothesized that considering residue pairs on interacting proteins in binding site prediction can improve performance and found out that this is in fact the case. Their neural network ensemble predictor (PPiPP) employed position specific scoring matrix and amino acid composition features. However, PPiPP's accuracy is low with area under the Receiver Operating Characteristic curve [20] of 72.9. The sequence-based nature of PPiPP allows the method to be applied to proteins for which only the sequence is known but at the same time it is unable to utilize the wealth of information contained in protein structures.

To the best of our knowledge, no structure based machine learning scheme for partner-

specific protein interface prediction exists in the literature. In this paper, we present a novel partner-specific SVM based interface predictor called PAIRpred (*Pa*rtner-specific *i*nteracting *r*esidue *pred*ictor) that uses both sequence information and features computed from the unbound structures. We present performance analysis of PAIRpred at both the complex level, i.e., for the prediction of interfaces, and at the protein level, i.e., for the prediction of binding sites in the individual proteins in the complex using Docking Benchmark 4.0 [21] and independent test complexes from the CAPRI experiment [13]. At the protein level, we compare PAIRpred's performance to the binding site predictor PredUS [22] and the protein level results from PPiPP [9]. At the complex level, we compared against the docking method ZDOCK [10] and PPiPP [9]. We show that considering information about the binding partner of a protein enables more accurate prediction of its binding site. Furthermore, we also study the relation between PAIRpred's performance and the degree of binding associated conformational change.

2 Methods

2.1 Data and pre-processing

In the development of PAIRpred, we have used the protein-protein docking benchmark data set 3.0 (DBD 3.0) [21]. This data set has also been used in the performance analysis of PPiPP [9] and allows a direct performance comparison. DBD 3.0 contains 124 non-redundant complexes of pairs of proteins for which both the bound and unbound X-ray crystallography structures are known. The proteins structures in DBD 3.0 have resolutions better than 3.25

Å and a minimum sequence length of 30. No two complexes in DBD 3.0 share the same SCOP [23] family-family pair [24] and have sequence identity of more than 30% in both chains. Further testing was performed on version 4.0 of DBD which contains a total of 176 complexes including those already in DBD 3.0.

2.2 Interacting residue-pair definition

We define two residues belonging to two different proteins in a complex to be interacting if the distance between any two heavy atoms of those residues in the *bound* conformations of their proteins is less than or equal to 6.0 Å. All other residue pairs from the two proteins on that complex were taken as negative examples. Similar definitions have been used in previous studies (see [9] and references therein). Defining interacting residues in this way resulted in a total of about 11,500 positive examples in DBD 3.0, i.e., 93 interacting residue pairs per complex on average. The average number of residues pairs, in overall, for a complex is around 67,000.

2.3 Feature extraction

We extracted both sequence and structure features at the residue level from the *unbound* structure of each protein. When the three dimensional of proteins forming a complex is not available, PAIRpred can make predictions based on its sequence alone. We have used a number of existing programs and methods from the literature to extract features from protein sequences and structures (see Figure 1).

- Structure based features

The following features have been computed directly from the structure.

Relative Accessible Surface Area (x_a^{rASA}) The relative accessible surface area (rASA) from a given protein structure was computed using STRIDE [25].

Residue depth (x_a^{RD}) Residue depth is defined as the minimum distance of a residue from the surface of the protein and has been computed using MSMS [26]. The residue depth values produced by MSMS were normalized to have the range from 0 to 1. x_a^{RD} and x_a^{rASA} are combined to form a single surface exposure feature denoted by x_a^{exp}. We found that residue depth carries complimentary information to that in rASA for residue interaction prediction.

Half Sphere Amino Acid Composition (x_a^{HSAAC}) Hamelryck [27] found that the geometry and physiochemical characteristics of the regions in the direction of the side chain of a residue (called the 'up' direction) and in its opposite direction (called the 'down' direction) can be very different from each another. Based upon this observation, we computed a feature (called HSAAC) that captures the amino acid composition in the direction of the side chain of a residue $x_a^{HSAAC_u}$ and in the direction opposite to the side chain $x_a^{HSAAC_d}$. The amino acid composition in a direction is defined as the number of times a particular amino acid occurs in that direction within a minimum atomic distance threshold of 8.0 Å from the residue of interest. Thus, HSAAC combines surface accessibility and amino acid composition within the neighborhood of a residue. These amino acid composition vectors in the two directions are then normalized to have unit norm to get $x_a^{HSAAC_u}$ and $x_a^{HSAAC_d}$ which are then concatenated to get x_a^{HSAAC}. We utilized Biopython

(Cock et al., 2009) to compute x_a^{HSAAC}.

Protrusion Index (x_a^{CX}) The protrusion index of a non-hydrogen atom is defined as the proportion of the volume of a sphere with a radius of 10.0 Å centered at that atom that is not filled with atoms [28]. The protrusion index has been calculated using PSAIA [29]. The protrusion index for single residue is a 6 dimensional vector comprising the mean, standard deviation, maximum and minimum of the protrusion values of all atoms in the residue along with the mean and standard deviation of the protrusion values of only its side chain atoms. Each element of this vector is normalized to have the range from 0 to 1.

- Sequence based features

We ran PSI-BLAST [30] against the non-redundant 'nr' database [31] to compute the Position Specific Scoring Matrix (PSSM) and the Position Specific Frequency Matrix (PSFM) for a given protein. The following sequence based features are then computed:

Profile Features (x_a^{PSSM}, x_a^{PSFM}) In order to extract the profile features for a residue from the PSSM, we took the PSSM columns within a length 11 window centered at that residue. This 20×11 matrix is converted to a single 220 dimensional unit vector denoted by x_a^{PSSM}. x_a^{PSFM} is constructed in a similar manner from the PSFM.

Predicted Relative Accessible Surface Area (x_a^{prASA}) To determine whether predicted rASA can be used instead of the true rASA, we used SPINE X [32] to predict rASA using the PSI-BLAST data. The predicted rASA is denoted by prASA to emphasize the fact that it has been predicted from sequence.

2.4 Pairwise classification using SVMs

We model the interface prediction problem as a classification problem in which a classification example i is a pair of residues from two different proteins in a complex. Each example i is represented by $((a_i, b_i), y_i)$, where (a_i, b_i) is a pair of residues and y_i is the associated label, indicating whether the two residues interact ($y_i = +1$) or not ($y_i = -1$). Figure 2 illustrates this concept.

As a classifier, we use a support vector machine (SVM) [33] trained over a set of N labeled examples. An SVM finds the optimal separating boundary between two classes by simultaneously maximizing the margin between them and minimizing the cost of misclassification over the training data. For an overview of SVMs in computational biology, the interested reader is referred to [34]. Due to its large-margin nature, an SVM can offer good accuracy over previously unseen examples during testing.

In order to use an example in training or classification, a classifier needs the feature representation for the pair of residues in that example. However, it is easier and computationally more efficient to extract features for a single residue on a protein than for a pair of residues on two different proteins. Thus, we would like to be able to use the residue level features directly to generate predictions at the residue pair level. This is where our use of SVMs for classification offers a significant advantage in comparison to other classifiers. Unlike other classifiers, such as the neural networks employed in PPiPP, SVMs can operate without requiring the explicit feature representation of an example by using a *kernel function* [35]. A kernel function is, in essence, a dot product that measures the degree of similarity between two examples. In this work, we construct a *pairwise* kernel of the form $K((a, b), (a', b'))$

which can directly score the similarity between examples (a, b) and (a', b') by comparing the feature representations of individual residues in these examples. The pairwise kernel eliminates the need of constructing an explicit feature representation of each example because the scoring function of the SVM can be expressed only in terms of this pairwise kernel as: $f_{AB}((a, b)) = \sum_{i=1}^{N} \alpha_i y_i K((a_i, b_i), (a, b))$. In this scoring function, the values of α_i are obtained through training.

One of the interesting features of using pairwise kernels in the SVM is that these kernels can themselves be built from kernels over individual residues. Such residue level kernels, denoted by $K_r\,(a, b)$, compare the explicit feature representations of residues a and b to score the degree of similarity between them. The problem of constructing pairwise kernels from kernels over individual objects has been studied in the machine learning and bioinformatics communities [36–39]. We constructed the pairwise kernel K_{pw} for our SVM as the additive combination of one or more of the following pairwise kernels from the literature:

$$K_{tppk}((a, b), (a', b')) = K_r(a, a')K_r(b, b') + K_r(a, b')K_r(b, a')$$

$$K_{mlpk}((a, b), (a', b')) = (K_r(a, a') - K_r(a, b') - K_r(b, a') + K_r(b, b'))^2$$

$$K_{sum}((a, b), (a', b')) = K_r(a, a') + K_r(b, b') + K_r(a, b') + K_r(b, a').$$

Here, K_{tppk} is the tensor product pairwise kernel (TPPK) proposed by Ben-Hur and Noble [36]. TPPK detects high similarity between examples (a, b) and (a', b') if a, expressed in terms of its feature representation, is similar to one of the residues in (a', b') and b is also similar to the other residue in the other example. It can be shown that the feature space of TPPK consists of products of features of the underlying residue kernel K_r.

$K_{mlpk}((a, b), (a', b'))$ is the metric learning pairwise kernel (MLPK) [37]. If the feature

representation of a residue a is given by $\phi(a)$, then the MLPK kernel can be written as:

$$K_{mlpk}((a,b),(a',b')) = ((\phi(a) - \phi(b))^T(\phi(a') - \phi(b')))^2.$$

This shows that the MLPK is a homogeneous polynomial kernel of degree 2 between pairs after mapping a pair (a,b) to the vector $\Phi_{mlpk}((a,b)) = \phi(a) - \phi(b)$. Vert et al. have shown that MLPK performs slightly better than TPPK for predicting protein-protein interactions and that their additive combination performs better than either of the kernels [37].

Given the feature space representation $\phi(a)$ of a residue a, the direct sum pairwise kernel can be written as [38, 40]:

$$K_{sum}((a,b),(a',b')) = (\phi(a) + \phi(b))^T(\phi(a') + \phi(b')).$$

This shows that the sum kernel uses the underlying feature map $\Phi_{sum}((a,b)) = \phi(a) + \phi(b)$.

We found that the simple kernel K_{sum} performed better than both TPPK and MLPK for our problem. However, the additive combination of the three kernels performed better than any of the individual kernels (see the results section for more details). Finally, each pairwise kernel K_{pw} is normalized as $K((a,b),(a',b')) = \frac{(K_{pw}((a,b),(a',b')))}{\sqrt{K_{pw}((a,b),(a,b))K_{pw}((a',b'),(a',b'))}}$ for use in the SVM.

To produce a pairwise prediction for an example (a,b), PPiPP [9] concatenates the feature representation of the two residues in the example in both orders $\begin{bmatrix}\phi(a)\\ \phi(b)\end{bmatrix}$ and $\begin{bmatrix}\phi(b)\\ \phi(a)\end{bmatrix}$. In comparison to PPiPP, our pairwise kernel based approach is computationally more efficient as it requires no duplication of the data. Moreover, pairwise kernels in our formulation directly model the inter-dependencies within individual feature components.

The residue kernel K_r used in constructing the pairwise kernel in PAIRpred is itself an unweighted summation of one or more of the following kernels, which are computed using

the features described in section 2.3:

$$K_{profile}(a,b) = g(x_a^{PSSM}, x_b^{PSSM}; \gamma_{PSSM}) + g(x_a^{PSFM}, x_b^{PSFM}; \gamma_{PSFM})$$

$$K_{HSAAC}(a,b) = g(x_a^{HSAAC}, x_b^{HSAAC}; \gamma_{HSAAC})$$

$$K_{prASA}(a,b) = g(x_a^{prASA}, x_b^{prASA}; \gamma_{prASA})$$

$$K_{exp}(a,b) = g(x_a^{exp}, x_b^{exp}; \gamma_{exp})$$

$$K_{CX}(a,b) = g(x_a^{CX}, x_b^{CX}; \gamma_{CX}).$$

In the above equations, $g(a,b;\gamma) = exp(-\gamma\|a-b\|^2)$ is the Gaussian kernel. The γ parameter in the Gaussian kernel controls the decay of the exponential function. If γ is set too high or too low, the exponential function can saturate at 0 or 1 which will inhibit effective learning from the training data. We chose the values of these parameters so that, for the majority of non-identical input vectors for a kernel, the similarity score does not saturate and maintains good dynamic range. This heuristic is inspired by the literature about parameter selection in radial basis function neural networks [41]. Once chosen in this manner, these parameters were not changed to optimize accuracy. The selected values of these parameters are as follows: $\gamma_{PSSM} = \gamma_{PSFM} = \gamma_{HSAAC} = 0.5, \gamma_{CX} = 1.0$ and $\gamma_{exp} = \gamma_{prASA} = 3.0$. As discussed in the results section, these values give good performance over test data. Training and classification has been performed using the SVM implementation in the machine learning library PyML [42].

2.5 Post-processing

A binding site or interface is a collection of spatially neighboring residues whose binding propensities are correlated. Keeping this in mind, we smoothed the prediction score for a

pair of residues by averaging prediction scores within their local neighborhoods through the following post-processing step:

$$f'_{AB}((a,b)) = \frac{1}{2}\left(\frac{\sum_{b' \in N(b)} f_{AB}((a,b'))}{|N(b)|} + \frac{\sum_{a' \in N(a)} f_{AB}((a',b))}{|N(a)|}\right), \quad (1)$$

where $f_{AB}((a,b))$ is the raw PAIRpred discriminant score from the trained SVM and $N(r)$ is the set of the 10 neighboring residues of residue r on the same protein including r itself. Thus the post-processed scores is the sum of the averages of the prediction scores of a residue on one protein with a set of residues on the other protein. As discussed in the results section, this simple post-processing scheme improves the prediction performance significantly.

2.6 Performance evaluation

Performance evaluation was carried out in two stages. In the first stage we compared different kernel designs, and residue-level features using five-fold cross-validation at the complex level. In this cross-validation procedure, examples from all complexes in our data set were divided into 5 folds such that all examples from a complex are found in exactly one fold. To reduce computational time during model selection, the 5 fold cross-validation was done using a class-size balanced sample from DBD 3.0 in which the number of randomly chosen negative examples for a complex is equal to the number of positive examples in it. For each fold, the value of the parameter C that controls the cost of misclassification over training data in the SVM was selected by performing a similar nested 5-fold cross-validation. The value of C was selected from $\{0.1, 1.0, 10.0, 100.0\}$. The classification function values and the known true labels of the examples were used to compute the Receiver Operating Characteristic (ROC) curve for each complex. The average of the area (expressed as a percentage) under

the ROC curve for all complexes, labeled as AUC, has been used as the performance statistic for selecting the optimal model.

In the second stage of performance evaluation, we performed a leave-one-complex-out cross-validation analysis with the optimal kernel design selected in the first stage. In this cross-validation procedure, a classifier is trained on a balanced set of examples extracted from all but one of the complexes, and testing is performed on *all* pairs of residues from the left-out complex. This evaluation protocol is identical to the one used for PPiPP [9] and allows a direct and fair comparison between the two methods. The average area (expressed as percentage) under the ROC curves for all complexes (AUC) is used as a performance metric as it allows a quantitative comparison with other interface prediction methods. However, AUC scores are not easy to interpret in this setting. In cases with highly unbalanced data with a big difference in the number of positive and negative test examples as we have here, AUC can give a false impression of accuracy. For these reasons we propose a measure of accuracy that is specifically designed for this domain. Our measure, which we call RFPP (rank of the first positive prediction), is defined as follows: RFPP(p) = q, if p % of the complexes tested have at least one true positive interacting residue pair among the top q predictions. Thus, an ideal classifier will have RFPP(100) = 1, i.e., in every complex, the top scoring prediction from the classifier belongs to the interface. In comparison to an ROC curve, this measure is more informative for the biologist as it tells us directly how often the top ranking predictions can be expected to correspond to known interactions.

We also evaluate the performance of PAIRpred for binding site prediction at the single protein level (i.e., binding site prediction) and compare it to existing partner-independent methods. Pairwise predictions of interacting residues at the complex level (from Equation

(1)) are converted into predictions at the protein level for each protein as follows: $f_A(a_j) = \max_{b_j \in B} f'_{AB}((a_j, b_j))$ and $f_B(b_j) = \max_{a_j \in A} f'_{AB}((a_j, b_j))$. AUC scores for an individual protein can then be easily computed.

3 Results and Discussion

3.1 Comparison of residue and pairwise representations

We analyzed and compared different feature representations and pairwise kernel formulations in order to see the contribution of different features towards prediction accuracy and the impact of pairwise kernel design. Figure 3 shows the complex-wise averaged ROC curve for different feature and kernel combinations. In order to compare different feature representations, we chose to use $K_{pw} = K_{mlpk} + K_{tppk} + K_{sum}$ as the pairwise kernel. Our first step was to analyze the accuracy when our method is restricted to using sequence-based features only, which include sequence profile and relative accessible surface area predicted from sequence. As shown in figure 3a, profile features alone give an AUC of 79.4, and adding the predicted rASA (i.e., $K_r = K_{profile} + K_{prASA}$) increases the AUC to 80.4. For the profile-based features we found that the combination of PSSM and PSFM features performed slightly better than either of the two alone (results not shown).

The addition of structure-based features provides a big boost in performance: the combination of true surface accessibility features (K_{exp}) with the profile features ($K_{profile}$) gives an AUC of 86.2 compared to 79.4 for the profile-based features alone and 80.4 using the combination of profile and predicted rASA features. Such an improvement is to be expected

because most of the residues involved in the interaction have high surface accessibility. However, the use of predicted rASA did not result in such a big increase. This is because the protein-wise averaged correlation between predicted and true rASA values for binding residues is low ($r = 0.56$, against $r = 0.76$ for non-interacting residues). Thus, the use of a better sequence based predictor of surface accessibility can help improve the accuracy of the sequence based predictions in future.

Addition of HSAAC and protrusion index based features ($K_{HSAAC} + K_{CX}$) improves the accuracy of the method even further (AUC of 87.1). For the rest of the analyses in the paper we have used $K_r = K_{profile} + K_{exp} + K_{HSAAC} + K_{CX}$.

The choice of a pairwise kernel has a strong influence on accuracy. Figure 3b show the ROC curves for different pairwise kernel formulations. K_{mlpk}, K_{tppk}, and K_{sum} produce AUC scores of 82.0, 86.7, and 86.9 respectively, while adding all three provides an AUC score of 87.1. For the rest of the analyses in the paper we have used $K_{pw} = K_{mlpk} + K_{tppk} + K_{sum}$.

In order to test the variability of the results, the cross-validation procedure in model selection was repeated 5 times with change both in the randomly selected negative examples and the membership of complexes in different folds. We then evaluated the mean and standard deviation of different cross-validation runs. The maximum standard deviation in the AUC scores for any kernel combination was 0.2. This shows that these results are robust to changes in the training data.

3.2 Prediction using residue exposure alone

As discussed above, residue exposure features result in a big improvement in accuracy. To explore the contribution of the residue exposure features (rASA, reside depth, and mean protrusion), we computed the sum of the residue exposure of the two residues in each example. Using this combination as a ranking criterion we computed the AUC score for each complex. This naïve way of classification yields some interesting results. The average AUC scores for all complexes from rASA, reside depth (RD) and the mean protrusion value are 71.9, 69.4 and 71.2 respectively. These results are only marginally inferior to the leave-one-complex-out cross-validation results from PPiPP (AUC = 72.9) [9]. Since rASA and RD are both measures of the surface accessibility of a residue, the AUC values for these features clearly reflect the known fact that surface residues are more likely to participate in protein-protein interactions. The AUC score of the protrusion index shows that the residues that interact have few atoms around them. This includes surface atoms, and especially those atoms on the surface that lie in cavities or protrude out from their local neighborhood. The protrusion index captures more local shape information than rASA and the two can be complementary to one another. The fact that pairwise summation of surface exposure features provides good results explains why the pairwise sum kernel K_{sum} was able to perform better than the other two pairwise kernels.

The same ranking criterion over the relative accessible surface area predicted from sequence using SPINE X gives an AUC of only 0.56. This clearly shows that these predictions need to be more accurate to be effective in finding interfaces in protein complexes.

3.3 Results for leave-one-complex-out cross-validation

For comparison with other methods we used the optimal kernel combination found through kernel evaluation (section 3.1) and recomputed its performance using the leave-one-complex-out cross-validation protocol detailed in Section 2.6. Results of this analysis are reported in table I. The AUC scores for interface prediction, averaged across the 123 complexes in DBD 3.0, for sequence and structure kernels are 80.9 and 87.3, respectively. It is interesting to note that these scores from leave-one-complex-out cross-validation over all examples are very close to those obtained with the balanced sample. Evaluation over all the 176 complexes in DBD 4.0 gives an AUC score of 87.0 with the structure kernel.

At the protein level, the AUC scores of PAIRpred for sequence and structure kernels for DBD 3.0 are 70.8 and 77.0 (with post-processing), respectively. It can also be noted that post-processing increases the performance of the method. This is particularly true at the protein level.

3.4 Comparison with PPiPP and ZDOCK

PPiPP [9] is a recently proposed sequence based method for partner-specific predictions that uses an ensemble of neural networks trained with a more elaborate version of our profile representation with different window sizes [9]. Table I shows the results of leave-one-complex-out cross-validation for DBD 3.0 using PPiPP. Even with the sequence features alone, PAIRpred gives better AUC and RFPP scores than PPiPP. As shown in figure 4a, PAIRpred's performance at the complex level (i.e., for interface prediction) is superior to PPiPP not only in overall AUC but also in the number of true positives within the first 10%

false positives.

PPiPP offers better accuracy than other published sequence based methods for binding site prediction such as PSIVER and SPPIDER (results given in [9]). PAIRpred's performance at the protein level (i.e., for binding site prediction) is also superior to PPiPP using either sequence features alone or in conjunction with protein structure (see Table I and Figure 4b).

We also compared PAIRpred with the docking method ZDOCK [10] over the 176 complexes in DBD 4.0. For this purpose, we have used, for each complex, the top 2000 predictions in the 15-degrees sampling data available online for ZDOCK v. 3.02. For each ZDOCK prediction for a complex, we computed the pairwise minimum inter-atomic distance between all residues of the two proteins in the predicted complex. The inverse of this distance was used as a ranking criterion in the evaluation of the AUC score at the complex level. The AUC score of a ZDOCK prediction tells us how good that prediction is at identifying the known interface in the complex and is directly comparable to the AUC scores given earlier for PAIRpred and PPiPP. For a given complex, we computed the maximum AUC score in the top N ZDOCK predictions and then averaged these scores across all complexes for a given value of N to obtain the results shown in Figure 5. These results show that PAIRpred is better than the best of the top 11 ZDOCK predictions. The AUC score of the top prediction by ZDOCK is roughly comparable to that of PPiPP.

3.5 Comparison with partner-independent predictions

In order to test the hypothesis that a partner-specific predictor can perform better than partner-independent predictors, we developed an SVM based binding site predictor (referred

to as *vanilla SVM*) using the same structural features as in PAIRpred and compared its leave-one-protein-out cross validation performance to the PAIRpred results at the protein level. Figure 4b shows the ROC curve for vanilla SVM which gives an AUC score of 72.6. PAIRpred's performs much better than the vanilla SVM. This clearly shows that partner-specific predictors can offer superior performance in comparison to partner-independent ones even when the same residue level features are used. Moreover, PAIRpred's AUC score of 70.8 with the sequence features alone is only marginally inferior to vanilla SVM even though the latter employs structure based features. As a matter of fact, PAIRpred with sequence features alone gives better true positive rates than the vanilla SVM consistently for false positive rates less than 0.4.

At the protein level, PAIRpred's performance using structure based features can be roughly contrasted to PredUS [22], a recently published structure based binding site predictor. PredUS performs better than other similar predictors available in the literature and gives an AUC score of 73.9 over 188 chains in DBD 3.0. It must be noted that a direct comparison between the performance of the two methods is not possible because of differences in their evaluation data sets, interface definitions, and cross-validation protocols. PAIRpred's performance with structure features can be expected to be equal or slightly better than that of PredUS as PAIRpred gives an AUC score of 77.0 over 248 proteins in DBD 3.0.

3.6 Spatial proximity of PAIRpred predictions

In order to see whether the top predictions by PAIRpred are spatially close, we compared pairwise distances between residues in our top predictions with a random sample of residues.

More specifically, we computed the pairwise distances among the top 20 residue predictions from PAIRpred for each protein and also between the remaining pairs of residues from each protein. The average of the pairwise distances in the top predictions is 15.6 Å and 20.1 Å for the remaining pairs. These distances are significantly different (with a p-value of 4.7×10^{-25} using the Wilcoxon Rank Sum test on all complexes in DBD 4.0). This indicates the top PAIRpred predictions exhibit spatial clustering.

Furthermore, we found that the difference between the mean pairwise distances across the top predictions and the remaining residues in a protein is inversely correlated with the its AUC (correlation coefficient of -0.49, 2 tailed p-value of 1.1×10^{-21}). Thus, this difference in distances is a rough indication of the quality of prediction.

3.7 Effects of conformational change

Proteins can undergo significant conformation change upon binding as buried residues can become exposed and vice versa. In order to observe the effects of the degree of conformational change on the accuracy of PAIRpred, we plotted the AUC of a complex against the root mean square deviation (RMSD) between the bound and the unbound states over the interface residue for in the complex. A large RMSD value for a complex corresponds to a large binding-associated conformation change. Figure 6a shows that the accuracy decreases with increase in conformational change. This effect was also observed for PPiPP. However, PAIRpred performs much better than PPiPP for complexes with large conformational change. Based on the degree of conformational change, the complexes in the docking benchmark datasets have been divided into three categories: rigid body, medium difficulty and hard. Figure 6a shows

the prediction performance across complexes in these categories. As expected, PAIRpred performs better for rigid body complexes in comparison to the other two categories that involve larger conformational changes.

We investigated the effects of conformational change on PAIRpred performance at the residue level as well. As we had access to both the bound and the unbound states of each protein, we were able to calculate the absolute difference in rASA for a residue between the two states of the protein. A large difference is indicative of a large conformational change in the environment around that residue. For a pair of residues we define the degree of conformational change as the sum of the changes in the individual residues, and denote it as $\Delta rASA(a, b)$. AUC exhibits a high negative correlation (see figure 6b) with $\Delta rASA$ (correlation coefficient of -0.97, p-value of 1.5×10^{-3}). AUC vs. change in residue depth shows a similar trend. This demonstrates the inherent difficulty of predicting residue-residue interactions in protein complexes that undergo a large conformational change. This difficulty is exacerbated by the fact that there is only a small amount of training data (24 complexes in DBD 4.0) available for such cases. Furthermore, the standard deviation of AUC scores for complexes from the hard category in DBD 4.0 shown in figure 6a is much larger in comparison to other categories. This suggests that effective handling of complexes with large conformational change requires a larger number of training examples with this property.

3.8 Evaluation on CAPRI targets

In order to further analyze the performance of PAIRpred, we tested it on nine recent targets from the Critical Assessment of Protein Interactions (CAPRI) experiment [13]. We used all

heteromeric protein complexes published after 2007 for which both the bound and unbound X-ray crystallography structures are available. For this task, PAIRpred was trained using DBD 4.0, and results of this analysis are reported in Table II. This table shows that PAIRpred is able to predict the interface with good accuracy for most targets. For seven out of these nine targets, the top 15 PAIRpred predictions contain at least one true positive. It is interesting to note that even for complexes involving large conformational changes, such as 3BX1 and 2WPT, the first true positive lies within the top 10 predictions. PAIRpred does not perform well on two targets: 3FM8 and 2VDU. These targets have proven to be very challenging for docking methods as well: only 1% and 4% of the models predicted by docking methods in CAPRI have an acceptable complex structure for 3FM8 and 2VDU, respectively [43].

3.9 Application to Human ISG15-Influenza A NS1 interaction

Due to its partner-specific nature and state of the art accuracy, PAIRpred can be used to study the nature and mechanics of an interface beyond what is possible with partner-independent predictors. In this section, we demonstrate PAIRpred's capabilities beyond the simple prediction of an interface by using the interaction between ISG15 protein in human and mouse and NS1 protein from Influenza A virus as a case study.

The influenza B virus is known to infect only human and non-human primates and the cause of this specific behavior have been investigated in [44] through a study of the bound and unbound structures of NS1 protein from the virus and the ISG15 protein in humans and other species. We have used PAIRpred to study the binding between these two proteins and

compare the findings from this computational analysis to the results published in [44].

We first predicted the interface of the complex from the unbound structures of the two proteins using both PPiPP and PAIRpred and used the known interface to compare the performance of the two methods. The unbound PDB structures of NS1 and ISG15 are available as 1XEQ [45] and 1Z2M [46]. The complex structure (PDB ID: 3SDL) has two chains each of NS1 and ISG15 [47]. There is no significant conformational change in NS1 upon binding to ISG15 with only a disorder to order change in a short C-terminal polypeptide sequence. ISG15 undergoes modest conformational change upon binding NS1 with a backbone RMSD of 1.05 Å. We obtained the predictions from the unbound proteins by training PAIRpred on DBD 3.0 to allow for a comparison with PPiPP, and used structure-based features. This complex is not a part of training sets of PAIRpred or PPiPP. The AUC scores for PPiPP and PAIRpred for this complex are 67.2 and 92.4, respectively. The first true positive detected by PAIRpred is the top-most prediction, whereas the first true positive detected by PPiPP occurs at rank 174. PAIRpred is able to find more than half of the interacting residue pairs within its top 100 predictions (see Figure 7a). The predictions correspond very closely to the interactions discussed in [44]. We also compared the interface prediction performance of PAIRpred to that of ZDOCK for this complex by using the inverse of the inter-residue distance from ZDOCK predictions as a ranking criterion as described in Section 3.4. It was found that the AUC score from PAIRpred is better than the best of the top 13 ZDOCK predictions for this complex.

Next we used, PAIRpred predictions in order to identify the residues that are crucial for binding. Specifically, we conducted an *in silico* mutagenesis experiment in which we changed the NS1: L88 residue involved in our top prediction (ISG15: L10, NS1: L88) to an

alanine. We also recapitulated one mutagenesis experiments reported in (Guan et al., 2011) which involved changing NS1: F34 (which also interacts with ISG15: L10) to an alanine. The (ISG15: L10, NS1: F34) interaction is originally ranked 8th in PAIRPRed predictions for this complex. We obtained the predicted structure after the mutations using I-TASSER (Roy et al., 2010). In comparison to the wild-type predictions for (ISG15: L10, NS1: L88) and (ISG15: L10, NS1: F34), we observed a decrease of 25% and 53% in prediction scores for L88 and F34 mutations in NS1, respectively (see Figure 7b). The prediction scores for other interacting residues were essentially unchanged. These results indicate that both these residues are, as experimentally determined in [44], very important for this interaction.

As stated earlier, NS1 binds specifically to ISG15 from human and non-human primates and does not bind to mouse ISG15. Guan et al. [44] attribute this binding specificity to residues 47-52 and 76-80 in the sequence alignment of ISG15s from these three species. We obtained the unbound structure of mouse ISG15 using I-TASSER. We then compared the PAIRpred prediction scores for (human ISG15,NS1) complex to those from the (mouse ISG15, NS1) interaction. This comparison allowed us to identify the ISG15 residues that are interacting in (human ISG15, NS1) complex but undergo a large decrease in their prediction scores in the (mouse ISG15, NS1) interaction. These locations (in order of decreasing magnitude of change in predictions scores) are 76, 77, 72, 74 and 49. This strengthens the claim made in [44].

These analyses clearly demonstrate the usefullness of partner-specific predictions generated from PAIRpred as the mutagenesis studies explained above cannot be performed with conventional partner-independent predictors.

3.10 Using PAIRpred

PAIRpred has been implemented in Python and its architecture allows extension in future to include more residue-level features or pairwise kernels. Complete implementation of PAIRpred, together with the pre-trained classifier, can be downloaded at `http://combi.cs.colostate.edu/supplements/pairpred/`. PAIRpred users need to supply the FASTA sequence files or, when available, the PDB format structure files as input. PAIRpred then automatically extracts features from these files and produces predictions using a pre-trained SVM. Users also have the option of training the classifier on their own data sets. PAIRpred generates its prediction for a complex as a simple text file which contains the pairwise interaction scores for each pair of residues from the two proteins in the input. This pairwise prediction file can then be used to generate protein-level binding site predictions through scripts available as part of the PAIRpred package. PAIRpred implementation also provides PyMOL scripts for visualizing top PAIRpred predictions both at the complex and protein levels as shown in Figure 7a.

4 Conclusions

We have presented a new method for predicting the interface of a protein complex called PAIRpred that offers state-of-the-art accuracy for both interface and binding site prediction. The proposed scheme is able to make accurate predictions using either sequence information alone or in conjunction with structure-based features. There are very few machine learning based methods that perform partner-specific prediction of interactions, and PAIRpred provides a large improvement over the recently published PPiPP method. We investigated the

merit of sequence and structure-based features and found that using structure provides a big improvement in performance. Furthermore, the analysis of the accuracy of PAIRpred shows much better scaling of performance with respect to the degree of conformational change upon complex formation in comparison to PPiPP. However, there is still plenty of room for improvement, especially for complexes that exhibit a large degree of conformational change upon binding. In the future we plan on adding features to capture shape complementarity between binding interfaces, information about correlated mutations [48, 49], protein flexibility and predictors of degree of conformational change [50] in order to improve the predictions even further. Moreover, PAIRpred can potentially improve the accuracy of docking methods if used as a filter or by direct incorporation into the energy function [6].

References

[1] 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, May 2009.

[2] O. Keskin, A. Gursoy, B. Ma, and R. Nussinov, "Principles of protein-protein interactions: what are the preferred ways for proteins to interact?," *Chemical reviews*, vol. 108, pp. 1225–1244, Apr. 2008. PMID: 18355092.

[3] J.-P. Changeux and S. Edelstein, "Conformational selection or induced-fit? 50 years of debate resolved," *F1000 Biology Reports*, vol. 3, Sept. 2011.

[4] K. Gunasekaran and R. Nussinov, "How different are structurally flexible and rigid binding sites? sequence and structural features discriminating proteins that do and do not undergo conformational change upon ligand binding," *Journal of Molecular Biology*, vol. 365, pp. 257–273, Jan. 2007.

[5] M. N. Wass, A. David, and M. J. Sternberg, "Challenges for the prediction of macromolecular interactions," *Current Opinion in Structural Biology*, vol. 21, pp. 382–390, June 2011.

[6] H.-X. Zhou and S. Qin, "Interaction-site prediction for protein complexes: a critical assessment," *Bioinformatics*, vol. 23, pp. 2203–2209, Sept. 2007. PMID: 17586545.

[7] Y. Ofran, "Prediction of protein interaction sites," in *Computational Protein-Protein Interaction*, pp. 167–184, CRC Press, 2009.

[8] S. Leis, S. Schneider, and M. Zacharias, "In silico prediction of binding sites on proteins," *Current Medicinal Chemistry*, vol. 2010, no. 17, pp. 1550–1562, 2010.

[9] S. Ahmad and K. Mizuguchi, "Partner-aware prediction of interacting residues in protein-protein complexes from sequence data," *PLoS ONE*, vol. 6, Dec. 2011. PMID: 22194998 PMCID: 3237601.

[10] B. G. Pierce, Y. Hourai, and Z. Weng, "Accelerating protein docking in ZDOCK using an advanced 3D convolution library," *PLoS ONE*, vol. 6, p. e24657, Sept. 2011.

[11] S. J. de Vries, A. D. J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, and A. M. J. J. Bonvin, "HADDOCK versus HADDOCK: new features

and performance of HADDOCK2.0 on the CAPRI targets," *Proteins*, vol. 69, pp. 726–733, Dec. 2007. PMID: 17803234.

[12] S. Chaudhury, M. Berrondo, B. D. Weitzner, P. Muthu, H. Bergman, and J. J. Gray, "Benchmarking and analysis of protein docking performance in rosetta v3.2," *PLoS ONE*, vol. 6, p. e22477, Aug. 2011.

[13] J. Janin, "Docking predictions of protein-protein interactions and their assessment: The CAPRI experiment," in *Identification of Ligand Binding Site and Protein-Protein Interaction Area* (I. Roterman-Konieczna, ed.), no. 8 in Focus on Structural Biology, pp. 87–104, Springer Netherlands, Jan. 2013.

[14] A. J. Bordner and A. A. Gorin, "Protein docking using surface matching and supervised machine learning," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 2, p. 488502, 2007.

[15] O. Martin and D. Schomburg, "Efficient comprehensive scoring of docked protein complexes using probabilistic support vector machines," *Proteins: Structure, Function, and Bioinformatics*, vol. 70, no. 4, p. 13671378, 2008.

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

[17] N. Tuncbag, A. Gursoy, and O. Keskin, "Prediction of proteinprotein interactions: unifying evolution and structure at protein interfaces," *Physical Biology*, vol. 8, p. 035006, June 2011.

[18] T. Vreven, H. Hwang, B. G. Pierce, and Z. Weng, "Evaluating template-based and template-free protein-protein complex structure prediction," *Briefings in Bioinformatics*, July 2013. PMID: 23818491.

[19] H. Wang, E. Segal, A. Ben-Hur, Q.-R. Li, M. Vidal, and D. Koller, "InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale," *Genome Biology*, vol. 8, no. 9, p. R192, 2007. PMID: 17868464 PMCID: 2375030.

[20] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, pp. 24–38, Jan. 2006.

[21] H. Hwang, T. Vreven, J. Janin, and Z. Weng, "Protein-protein docking benchmark version 4.0," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 15, p. 31113114, 2010.

[22] Q. C. Zhang, L. Deng, M. Fisher, J. Guan, B. Honig, and D. Petrey, "PredUs: a web server for predicting protein interfaces using structural neighbors," *Nucleic Acids Research*, vol. 39, pp. W283–W287, May 2011.

[23] A. Andreeva, D. Howorth, S. E. Brenner, T. J. P. Hubbard, C. Chothia, and A. G. Murzin, "SCOP database in 2004: refinements integrate structure and sequence family data," *Nucleic acids research*, vol. 32, pp. D226–229, Jan. 2004. PMID: 14681400.

[24] H. Hwang, B. Pierce, J. Mintseris, J. Janin, and Z. Weng, "Protein-protein docking benchmark version 3.0," *Proteins: Structure, Function, and Bioinformatics*, vol. 73, no. 3, p. 705709, 2008.

[25] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins*, vol. 23, p. 566579, 1995.

[26] M. Sanner, A. Olson, and J.-C. Spehner, "Reduced surface: an efficient way to compute molecular surfaces," *Biopolymers*, vol. 38, no. 3, pp. 305–320, 1996.

[27] T. Hamelryck, "An amino acid has two sides: A new 2D measure provides a different view of solvent exposure," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, no. 1, p. 3848, 2005.

[28] A. Pintar, O. Carugo, and S. Pongor, "CX, an algorithm that identifies protruding atoms in proteins," *Bioinformatics (Oxford, England)*, vol. 18, pp. 980–984, July 2002. PMID: 12117796.

[29] J. Mihel, M. iki, S. Tomi, B. Jeren, and K. Vlahoviek, "PSAIA protein structure and interaction analyzer," *BMC Structural Biology*, vol. 8, p. 21, Apr. 2008.

[30] S. F. Altschul, T. L. Madden, A. A. Schffer, 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. PMID: 9254694 PMCID: 146917.

[31] S. McGinnis and T. L. Madden, "BLAST: at the core of a powerful and diverse set of sequence analysis tools," *Nucleic Acids Research*, vol. 32, pp. W20–W25, July 2004.

[32] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "SPINE x: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," *Journal of Computational Chemistry*, vol. 33, pp. 259–267, Jan. 2012.

[33] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, p. 273297, 1995.

[34] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schlkopf, and G. Rtsch, "Support vector machines and kernels for computational biology," *PLoS Comput Biol*, vol. 4, p. e1000173, Oct. 2008.

[35] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, p. 3846, Jan. 2005.

[36] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, pp. i38–i46, June 2005.

[37] J.-P. Vert, J. Qiu, and W. Noble, "A new pairwise kernel for biological network inference with support vector machines," *BMC Bioinformatics*, vol. 8, p. S8, Dec. 2007.

[38] A. Bar-hillel and D. Weinshall, "Boosting margin based distance functions for clustering," in *In Proceedings of the Twenty-First International Conference on Machine Learning*, p. 393400, 2004.

[39] C. Brunner, A. Fischer, K. Luig, and T. Thies, "Pairwise support vector machines and their application to large scale problems," *Journal of Machine Learning Research*, vol. 13, p. 22792292, Aug. 2012.

[40] C. Brunner, A. Fischer, K. Luig, and T. Thies, "Pairwise kernels, support vector machines, and the application to large scale problems," tech. rep., Technische Universitat Dresden Institut fur Numerische Mathematik, 2011.

[41] S. S. . Haykin, *Neural networks a comprehensive foundation*. Prentice Hall, 2nd ed. ed., 1999.

[42] A. Ben-Hur, "PyML: machine learning using python (http://pyml.sourceforge.net/)," Dec. 2012.

[43] M. F. Lensink and S. J. Wodak, "Docking and scoring protein interactions: CAPRI 2009," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 15, p. 30733084, 2010.

[44] R. Guan, L.-C. Ma, P. G. Leonard, B. R. Amer, H. Sridharan, C. Zhao, R. M. Krug, and G. T. Montelione, "Structural basis for the sequence-specific recognition of human ISG15 by the NS1 protein of influenza b virus," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 13468–13473, Aug. 2011.

[45] C. Yin, J. A. Khan, G. V. T. Swapna, A. Ertekin, R. M. Krug, L. Tong, and G. T. Montelione, "Conserved surface features form the double-stranded RNA binding site of non-structural protein 1 (NS1) from influenza a and b viruses," *Journal of Biological Chemistry*, vol. 282, pp. 20584–20592, May 2007.

[46] J. Narasimhan, "Crystal structure of the interferon-induced ubiquitin-like protein ISG15," *Journal of Biological Chemistry*, vol. 280, pp. 27356–27365, June 2005.

[47] R. Guan, L.-C. Ma, P. G. Leonard, B. R. Amer, H. Sridharan, C. Zhao, R. M. Krug, and G. T. Montelione, "Structural basis for the sequence-specific recognition of human ISG15 by the NS1 protein of influenza b virus," *Proceedings of the National Academy of Sciences*, vol. 108, pp. 13468–13473, Aug. 2011.

[48] S. D. Dunn, L. M. Wahl, and G. B. Gloor, "Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction," *Bioinformatics*, vol. 24, pp. 333–340, Feb. 2008.

[49] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, "Protein 3D structure computed from evolutionary sequence variation," *PLoS ONE*, vol. 6, p. e28766, Dec. 2011.

[50] J. Marsh and S. Teichmann, "Relative solvent accessible surface area predicts protein conformational changes upon binding," *Structure*, vol. 19, pp. 859–867, June 2011.

Figure 1 Residue-level feature extraction in PAIRpred. The feature representation for residue a is denoted by $\boldsymbol{x}_a$. Different components of the feature representation are denoted by the superscript (e.g., x_a^{rASA} indicates the relative accessible surface area for residue a). Each box also indicates the program used to extract a given set of features.

Figure 2 Overview of PAIRpred. (i) Extract residue-level features from sequence and unbound structures (see Figure 1 for details). (ii) Construct pairwise kernel from the residue-level kernel $K_r(a_i, a_j)$. (iii) Use the pairwise kernel to train the SVM and classify each residue pair in the query proteins.

Figure 3 Selecting the optimal sequence/structure representation by comparing ROC curves for different kernel designs. Shown are the averaged ROC curves computed using 5-fold cross-validation over complexes in DBD 3.0. The inset shows the true positive rate (TPR) vs. false positive rate (FPR) for up to first 10 % false positives. The legend shows the AUC scores for the different kernels used. (a) Results for different residue kernels K_r using the pairwise kernel $K_{pw} = K_{mlpk} + K_{tppk} + K_{sum}$. The curves illustrate the increase in performance as additional structural information is added to the sequence-based kernel. Recall that $K_{profile}$ is the PSI-BLAST profile kernel; K_{prASA} uses predicted rASA; K_{exp} is the residue exposure kernel; K_{HSAAC} is the half-sphere exposure kernel; K_{CX} uses protrusion-index features. (b) Results for different pairwise kernels K_{pw} with residue kernel $K_r = K_{profile} + K_{exp} + K_{HSAAC} + K_{CX}$.

Figure 4 Comparison between PAIRpred, PPiPP [9] and vanilla SVM at the complex and protein level predictions on DBD 3.0. PAIRpred-seq refers to PAIRpred based only on sequence features.

Figure 5 The maximum AUC from top N ZDOCK predictions in comparison to PAIRpred and PPiPP [9].

Figure 6 Effect of conformational change on PAIRpred performance. (a) AUC vs. RMSD for each complex in DBD 3.0 and the 53 new complexes in DBD 4.0 using leave-one-complex-out cross-validation. The legend shows mean AUC and standard deviation (within paranthesis) of complexes in each category for each data set. (b) Relationship between AUC and the change in rASA ($\Delta rASA$). Residue pairs were binned into groups based on their $\Delta rASA$ and the AUC score was computed for all the residues within each bin using our 5-fold cross-validation scheme on DBD 3.0 complexes.

Figure 7 (a) PAIRpred predictions for human ISG15 and influenza B NS1 mapped onto the 3D structure of the complex (PDB ID: 3SDL). The red dotted lines indicate the true positives in the top predictions with the width of the line proportional to the prediction score. The circled area is expanded in (b). (b) Results of in silico mutagenesis. The blue residues were changed to alanines. Notice the change in the prediction score (indicated by the width of the orange dotted lines) for the mutated residues between the wild-type (left) and the mutant (right).

Table I: PAIRpred and PPiPP performance. We compare the performance of PAIRpred and PPiPP [9] using Area Under the ROC Curve (AUC) and the rank of the first positive prediction (RFPP). RFPP(p) indicates that p percent of the proteins achieve that level of performance. For example, on DBD 4.0 without post processing, the second PAIRpred prediction is part of the interface for 10% of the complexes. PAIRpred results are provided for two residue kernels: the sequence-based kernel, and for the kernel that uses all the features computed from sequence and structure.

Dataset	**Method**		**RFPP (p)**					**AUC**	
			10%	**25%**	**50%**	**75%**	**90%**	**Complex**	**Protein**
DBD 3.0 (124 complexes)	**PPiPP**		9	19	78	297	760	72.9	66.1
	PAIRPred								
	$K_r = K_{profile} + K_{prASA}$	No post-processing	2	13	68	257	804	80.9	70.8
	$K_r = K_{profile} + K_{exp} + K_{HSAAC} + K_{CX}$	No post-processing	1	5	22	89	282	87.3	73.4
		With post-processing	1	3	16	103	272	88.7	77.0
DBD 4.0 (176 complexes)	$K_r = K_{profile} + K_{exp} + K_{HSAAC} + K_{CX}$	No post-processing	2	6	19	75	340	87.0	73.1
		With post-processing	1	3	18	101	282	87.8	75.4

Table II: PAIRpred evaluation results on recent heteromeric protein complex targets from CAPRI with both bound and unbound X-ray structures available. The degree of conformational change for the ligand and receptor proteins has been measured as the backbone root mean square deviation. The maximum sequence identity of a protein from the CAPRI set to any protein in DBD 4.0 has been calculated using local sequence alignment. The AUC score for each complex and the rank of the first positive prediction (RFPP) is reported.

Complex ID in PDB	Target ID in CAPRI	Ligand Backbone RMSD (Å)	Receptor Backbone RMSD (Å)	Max. Seq Id. of ligand to DBD4	Max. Seq Id. of receptor to DBD4	AUC	RFPP
4G9S	T58	0.3	0.7	28 %	27%	89.7	4
4EEF	T56	0.7	0.5	27 %	29 %	76.3	1
3R2X	T50	0.5	0.6	29 %	26 %	90.3	15
3U43	T47	0.9	1.5	60 %	55 %	88.9	2
2WPT	T41	2.0	0.7	62 %	66 %	85.8	1
3E8L	T40	0.2	0.4	100 %	28 %	92.1	9
3FM8	T39	0.0	1.6	28 %	25 %	79.6	71
3BX1	T32	2.0	0.4	30 %	56 %	89.7	10
2VDU	T29	1.1	0.4	28 %	27 %	82.9	302

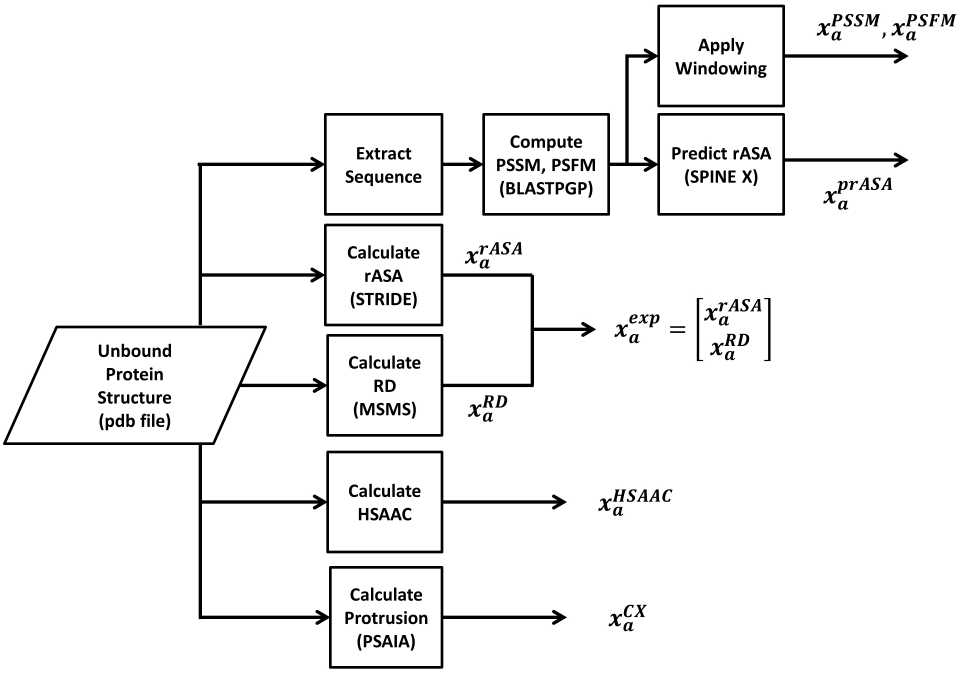


Figure 1

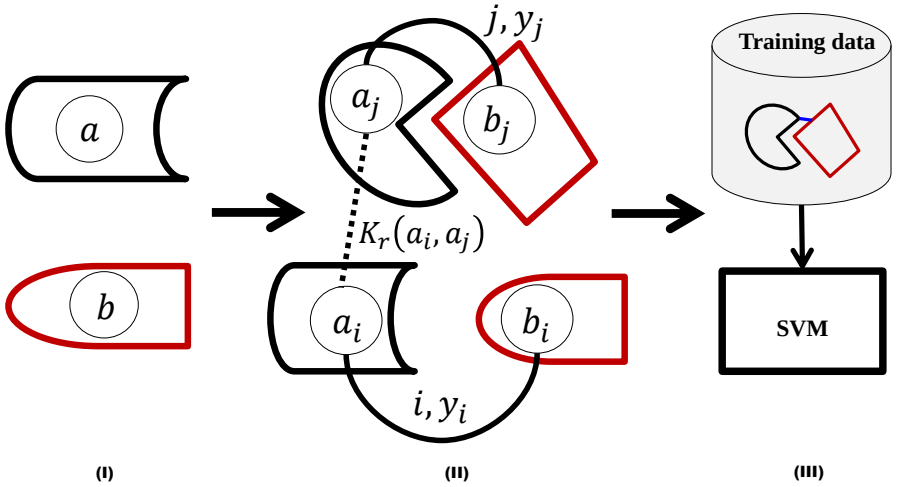
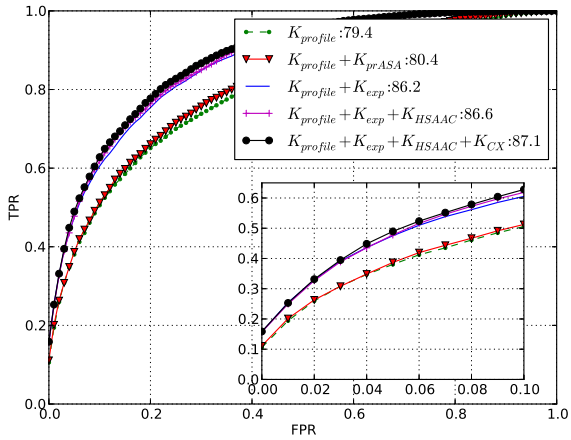
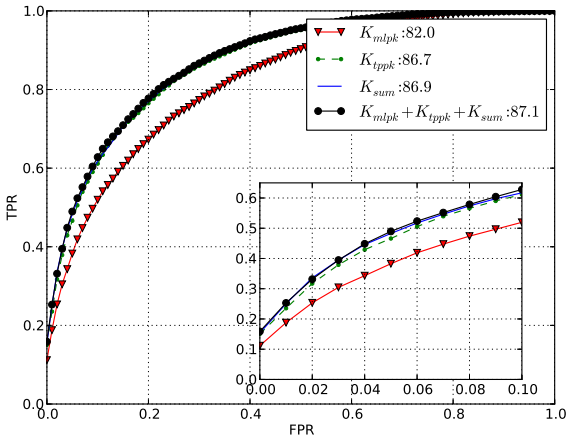


Figure 2

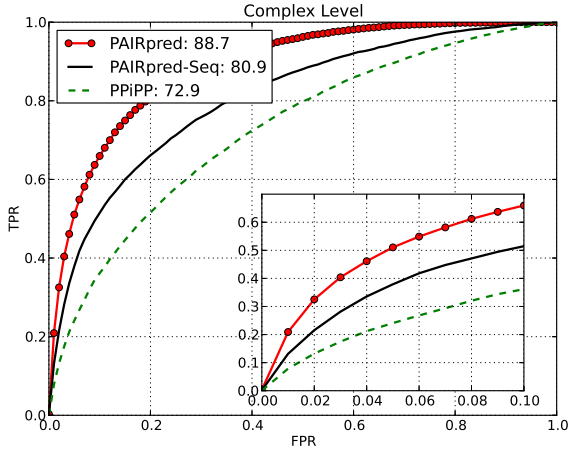


(a)

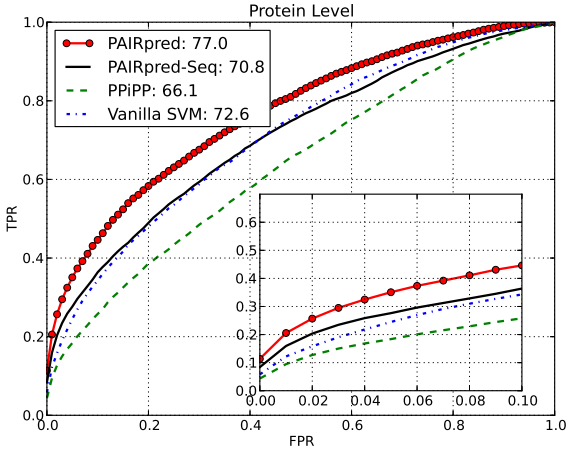


(b)

Figure 3



(a)



(b)

Figure 4

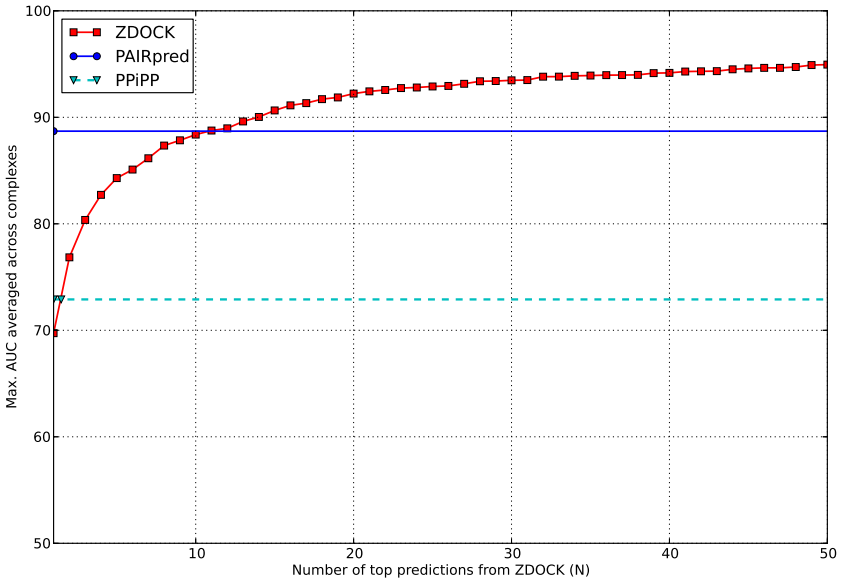
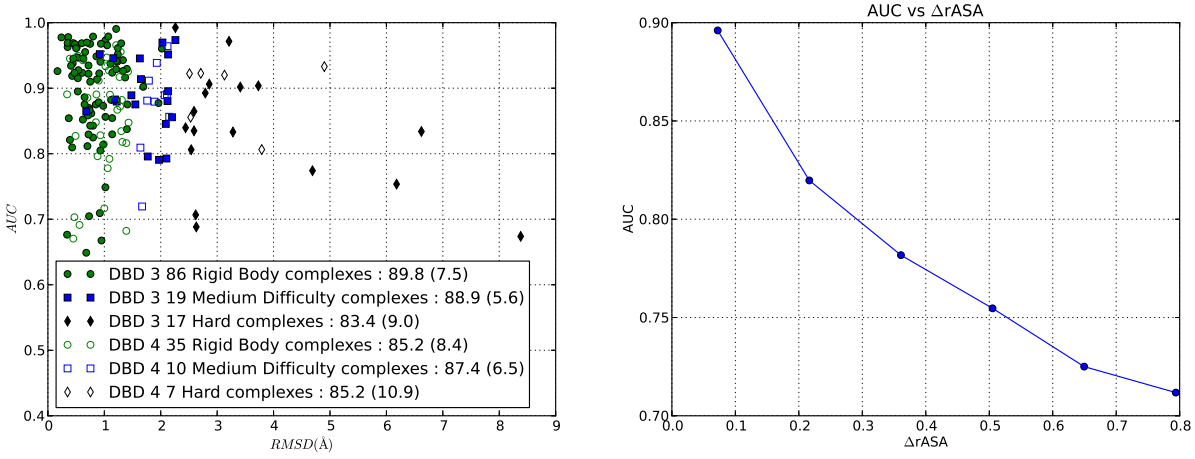
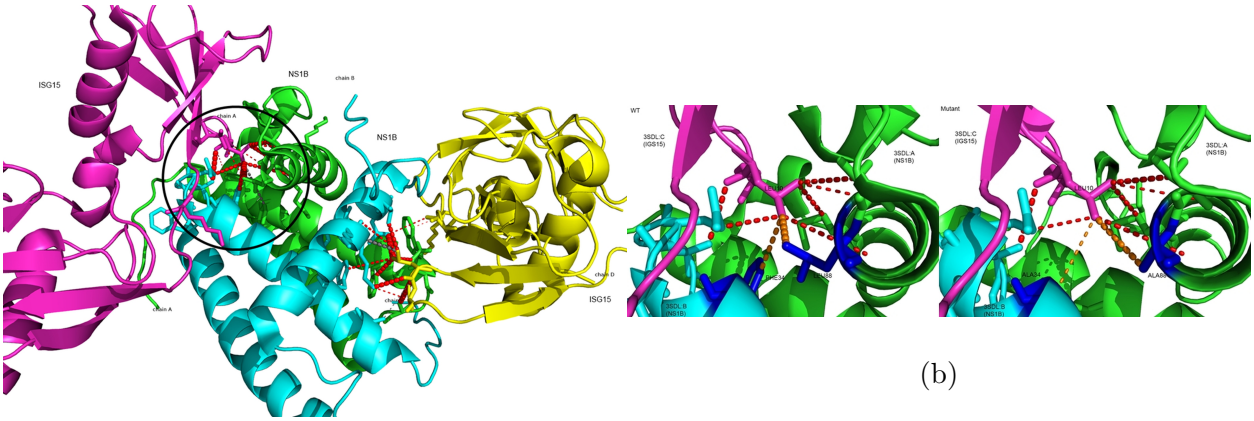


Figure 5



(a) (b)

Figure 6


ISG15
NS1B
chain B
chain A
NS1B
chain A
chain D
ISG15
WT
3SDLC
(ISG15)
3SDLA
(NS1B)
3SDLB
(NS1B)
Mutant
3SDLC
(ISG15)
3SDLA
(NS1B)
3SDLB
(NS1B)

(a)

(b)

Figure 7